Society of Medical Education & Research

2026, Volume 6, Issue 1, Page No: 53-70
ISSN: 3108-4826

Journal of Medical Sciences and Interdisciplinary Research

Deep Learning–Based Multi-Disease Classification of Chest X-Ray Images Using an Optimized VGG16 Architecture

Daniel Kim[1]*, Jihoon Park[1], Min-seo Kang[2], Hyun-woo Jung[3]

[1] Department of Radiology and Deep Learning, College of Medicine, Seoul National University, Seoul, South Korea.
[2] Department of Thoracic Imaging and AI Classification, College of Medicine, Chung-Ang University, Seoul, South Korea.
[3] Department of VGG16 Optimization for Medical Imaging, College of Medicine, Kangwon National University, Chuncheon, South Korea.

***E-mail** ✉ jaesung.kim@snu.ac.kr

Abstract

Interpreting chest radiographs is a challenging and resource-intensive clinical task, largely because of the inherent complexity of identifying a broad spectrum of pulmonary pathologies. As a result, there is a pressing need for innovative techniques that can accurately classify multiple abnormalities in chest X-ray images. This research presents an enhanced deep learning framework tailored for multi-label classification of chest X-ray images, addressing a comprehensive set of conditions, including lung opacity, normal pulmonary findings, COVID-19, bacterial pneumonia, viral pneumonia, and tuberculosis. The proposed approach utilizes a refined version of the VGG16 architecture augmented with squeeze-and-excitation (SE) blocks. This model was trained on an extensive collection of chest X-ray images and rigorously compared with leading contemporary methods using key performance indicators: accuracy, F1-score, precision, recall, and area under the curve (AUC). The modified VGG16-SE model outperformed existing techniques across every evaluation metric. It recorded an accuracy of 98.49%, an F1-score of 98.23%, a precision of 98.41%, a recall of 98.07%, and an AUC of 98.86%. The current study offers a highly effective deep learning strategy for classifying chest radiographs. Given its strong performance across diverse lung conditions, the model shows considerable promise for seamless incorporation into routine clinical practice, supporting faster and more precise diagnosis of pulmonary diseases.

Keywords: Deep learning, Chest X-ray imaging, lung diseases, Convolutional neural network, Optimizing, Multi-label categorization

Introduction

Rapid progress in medical imaging technologies has substantially improved diagnostic precision in modern healthcare, particularly in the assessment of chest X-rays—one of the most commonly performed imaging examinations worldwide. Although widely used, interpreting these images demands considerable expertise and remains challenging because of subtle radiographic signs and the specialized knowledge required for accurate diagnosis [1-3]. Chest X-rays play a central role in identifying numerous respiratory and cardiac disorders, such as lung opacities, pneumonias of different origins, tuberculosis, and various other anomalies. Each pathology presents distinct detection challenges because overlapping radiographic appearances may mimic one another. Conventional diagnostic approaches depend primarily on radiologists' subjective judgments, which are prone to considerable inter-observer variability. In contrast, automated intelligent systems present a valuable opportunity to alleviate radiologists' heavy workloads, enhance

Access this article online https://smerpub.com/

Received: 29 October 2025; **Accepted:** 07 February 2026

How to cite this article: Kim D, Park J, Kang M, Jung H. Deep Learning–Based Multi-Disease Classification of Chest X-Ray Images Using an Optimized VGG16 Architecture. J Med Sci Interdiscip Res. 2026;6(1):53-70. https://doi.org/10.51847/K7zKmQ6EV7

diagnostic reliability, and accelerate time-to-diagnosis, ultimately elevating standards of patient care [4, 5].

Recent breakthroughs in deep learning (DL) have shown strong potential in medical image interpretation, particularly through convolutional neural networks (CNNs), which have achieved outstanding results in visual recognition and classification [6, 7]. These sophisticated architectures excel at extracting complex patterns from imaging data and frequently outperform conventional image analysis methods in both speed and diagnostic precision. The capacity of CNNs to automatically derive meaningful representations from large volumes of unlabeled image data offers a transformative prospect for medical imaging, promising quicker, more accurate, and more widely accessible diagnostic processes [8, 9]. Nevertheless, despite notable improvements and refinement efforts, CNNs frequently deliver less-than-optimal results when deployed in practical medical imaging scenarios. This shortfall arises from multiple factors that limit effective generalization in real clinical environments [10-12]. A key obstacle is the substantial variation in image quality caused by differences in acquisition devices, patient positioning, and procedural variations. These inconsistencies introduce noise that models struggle to accommodate, compromising accuracy and reliability.

Furthermore, infrequent or unusual disease manifestations represent another critical barrier. When such cases appear rarely in training datasets, the model's capacity to correctly detect and categorize them declines, increasing misclassification rates for less common conditions. In addition, the inherent complexity of multi-label classification in chest X-rays intensifies these difficulties. Pathological features often display only minor distinctions, requiring models that can effectively discern subtle, overlapping characteristics. Standard CNN designs tend to underperform in such settings, especially when class boundaries are not sharply defined. A further significant concern is the class imbalance commonly observed in medical datasets. Conditions with lower prevalence receive insufficient representation, which distorts the learning process [13, 14]. Consequently, models become biased toward dominant classes, reducing generalization and reliability when applied to underrepresented pathologies in clinical practice.

This study proposes an optimized deep learning framework engineered for multi-label classification of chest X-ray images. This task is particularly demanding because of the intricate and frequently overlapping radiographic patterns associated with respiratory disorders. The framework targets several prevalent and clinically significant conditions, namely lung opacity, COVID-19, normal pulmonary appearance, viral pneumonia, bacterial pneumonia, and tuberculosis. Each condition possesses characteristic yet sometimes similar imaging features that can hinder accurate differentiation. Extensive experiments were performed on a sizable repository of labeled chest X-ray images to thoroughly evaluate the effectiveness of the suggested approach. The breadth and diversity of this dataset ensure that the model is tested under realistic conditions and can accommodate the variability encountered in everyday clinical settings. Such dataset characteristics are vital for enabling the model to generalize effectively to new cases and for ensuring that laboratory performance reliably extends to practical healthcare applications.

The primary contributions of this work are summarized as follows:

- We introduce a novel adaptation of the established VGG16 architecture to markedly improve its effectiveness in medical image diagnosis. Renowned for its depth and strong performance in visual tasks, the original VGG16 serves as an excellent base. We enhance this foundation by incorporating squeeze-and-excitation (SE) blocks, which dynamically improve the network's focus on the most relevant features. These blocks explicitly capture channel-wise interdependencies within convolutional feature maps and recalibrate feature responses according to their importance. This mechanism enables the model to emphasize salient information while suppressing less useful signals, thereby boosting predictive precision. The resulting architecture retains the core advantages of VGG16 while adding advanced feature refinement capabilities. This integration is especially beneficial in medical diagnostics, where detecting and interpreting fine-grained patterns is essential. Overall, the modified VGG16 with SE blocks exemplifies the value of merging classical deep learning designs with contemporary attention mechanisms to develop more effective and reliable analytical tools for healthcare.

- We tackle the prevalent problem of class imbalance in medical imaging datasets. The study employs chest X-ray images distributed across six categories: lung opacity, normal pulmonary states, COVID-19, viral pneumonia, bacterial pneumonia, and tuberculosis. The original dataset exhibited notable imbalance across these categories, which can adversely affect model training. To mitigate this issue and strengthen model robustness, we applied a comprehensive data augmentation strategy. This technique generates additional training samples by applying geometric transformations to existing images, including scaling, horizontal flipping, zooming, width shifting, and rotation. These augmentation methods effectively balanced the dataset, achieving more equitable representation across all classes. This balancing not only reduces bias but also enriches the training data, establishing a stronger basis for developing a dependable diagnostic model.

The remainder of this paper is structured as follows: Section 2 reviews relevant literature and provides contextual background. Section 3 details the proposed methodology, encompassing the lung disease dataset, the data augmentation approach, and the implementation of the modified VGG16 architecture with SE blocks. Section 4 reports the experimental setup and presents the obtained results. Section 5 discusses the optimization techniques utilized in this research. Finally, Section 6 concludes the study and outlines directions for future work.

Related work

The application of deep learning (DL) techniques for identifying abnormalities in chest X-ray images has gained considerable momentum in recent years. This growing interest stems from the remarkable capacity of these algorithms to detect intricate patterns and subtle anomalies that often escape traditional analytical methods. Within medical research, artificial intelligence (AI) has emerged as a prominent tool, particularly for supporting the diagnosis of various health conditions. Numerous studies employing AI in medical diagnostics have reported encouraging results, highlighting the accuracy and reliability of these sophisticated technologies [15-23]. The current section reviews the methodologies adopted by earlier investigators in this field.

In Ref. [24], the authors developed a 2D-CNN model to classify chest X-ray images into six categories: viral pneumonia, tuberculosis, fibrosis, normal lung conditions, bacterial pneumonia, and COVID-19. Their investigation compared the performance of standard VGG16 and VGG19 architectures with the proposed 2D-CNN. The 2D-CNN achieved 96.75% accuracy, while the VGG16 and VGG19 models attained 90.43% and 89.51%, respectively. In Ref. [25], a deep learning framework was introduced to enhance the identification of multiple pulmonary diseases, including opacity, fibrosis, normal lungs, viral pneumonia, tuberculosis, and COVID-19-related pneumonia. The study evaluated the standard VGG19 architecture alongside an improved variant using a dataset comprising 48,582 chest X-ray images. The conventional VGG19 model reached 96.57% accuracy, whereas the enhanced version achieved 98.88%.

Ref. [26] proposed a multi-scale CNN for a six-class classification task to detect bacterial pneumonia, tuberculosis, fibrosis, normal lung conditions, viral pneumonia, and COVID-19. The model was trained on 5,700 chest X-ray images. The researchers assessed the original VGG19 and VGG16 architectures, as well as a VGG16 variant incorporating multi-scale feature mapping. The VGG19 model achieved 95.61% accuracy, and the baseline VGG16 reached 95.79%. With the multi-scale enhancements, the VGG16 model's accuracy improved to 97.47%. In [27], the authors presented a ResNet50 model leveraging deep feature extraction to classify chest X-ray images into five categories: tuberculosis, bacterial pneumonia, viral pneumonia, normal lung conditions, and COVID-19. Using a dataset of 2,186 images, the ResNet50 architecture achieved 91.60% accuracy.

Ref. [28] examined the effectiveness of various ResNet architectures, including ResNet-50, ResNet-101, and ResNet-152, for diagnosing different lung pathologies. The analysis utilized a dataset of 21,885 chest X-ray images categorized into four classes: lung opacity, pneumonia, COVID-19, and normal cases. The study evaluated the diagnostic proficiency of these models in distinguishing key respiratory conditions. Results indicated that the ResNet-50 model attained 91% accuracy, the ResNet-101 model reached 93%, and the ResNet-152 model achieved the highest accuracy of 94%. Ref. [29] introduced a multi-class model for differentiating COVID-19, normal lung conditions, lung opacity, and viral pneumonia from chest X-ray images.

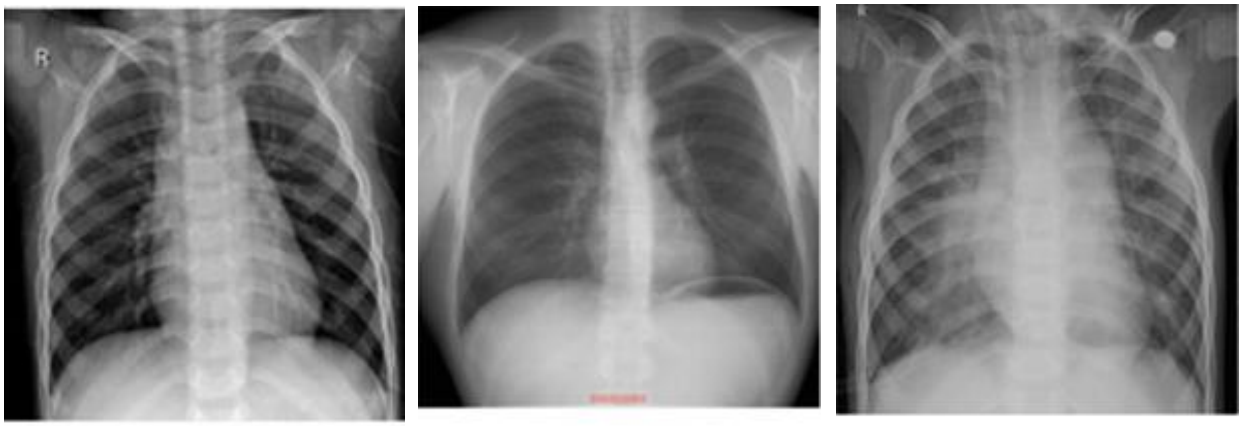
This work assessed both the standard MobileNetV2 network and a customized adaptation. The customized MobileNetV2 variant outperformed the original, achieving 95.80% accuracy compared to 90.47% for the baseline model.

Several studies in medical imaging have reported strong performance in binary or small-scale classification tasks. However, accuracy tends to decline substantially as the number of classes increases. This reduction is primarily attributable to the difficulty of differentiating among multiple overlapping conditions, especially when radiographic features show only subtle distinctions. These challenges become more pronounced in multi-class scenarios, where precise discrimination among lung infections requires recognizing fine-grained details. These limitations hinder the practical application of these models in real-world clinical environments, where patients frequently present with complex, co-existing pulmonary conditions. In such settings, the capability to accurately identify multiple lung pathologies is not merely advantageous but essential. Consequently, there is a clear need for advanced CNN architectures that can reliably manage multi-class classification of lung diseases with greater precision and robustness. Creating such frameworks would offer substantial benefits in clinical practice by enabling more accurate diagnoses and facilitating targeted treatment approaches for patients with complex pulmonary disorders. This pressing requirement underscores the importance of ongoing research and the advancement of state-of-the-art diagnostic tools capable of meeting the evolving demands of contemporary healthcare. Continued progress in DL methodologies is vital to elevate the quality of medical diagnostics and improve patient outcomes amid increasingly diverse and complex medical presentations.

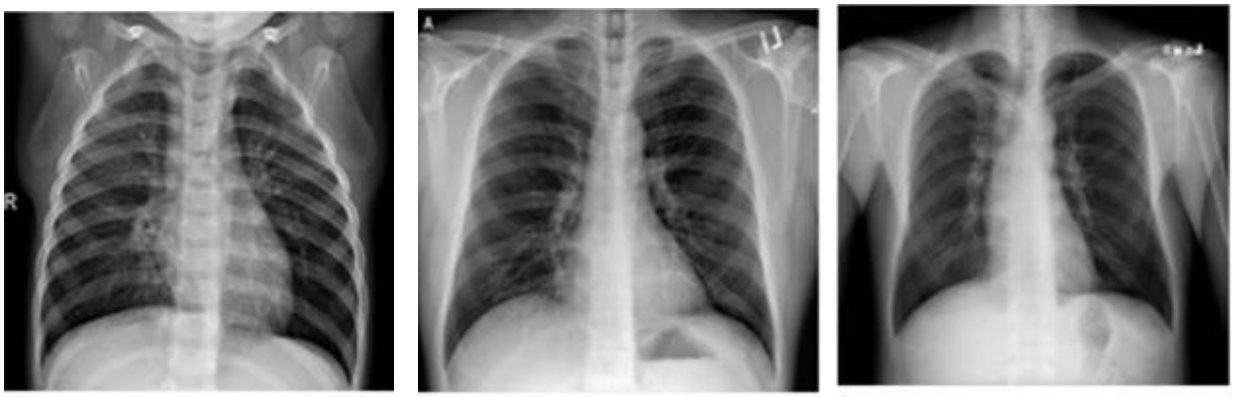
Materials and Methods

Lung disease dataset and partitioning

This study focused on multi-class classification using three separate medical imaging datasets, each consisting exclusively of chest X-ray images. The combined dataset contained 27,445 images encompassing normal cases and five distinct categories of lung pathologies. The first dataset, obtained from Ref. [30], contributed a varied collection of 3,616 COVID-19 positive cases, 10,192 normal images, 6,012 images showing lung opacity, and 1,345 images of viral pneumonia. The second dataset, sourced from Ref. [31], originally included 3,500 tuberculosis images and 3,500 normal chest X-rays; only the tuberculosis images were incorporated into the present analysis. The third dataset, acquired from Ref. [32], provided 2,780 images of bacterial pneumonia and 1,493 images of viral pneumonia; only the bacterial pneumonia cases were selected. **Figure 1** displays representative chest X-ray examples from the compiled dataset across the different diagnostic categories. In addition, **Figure 2** depicts the distribution of lung conditions within the assembled collection.



(a) (b) (c)



(d) (e) (f)

Figure 1. Representative chest X-ray samples from the compiled dataset.

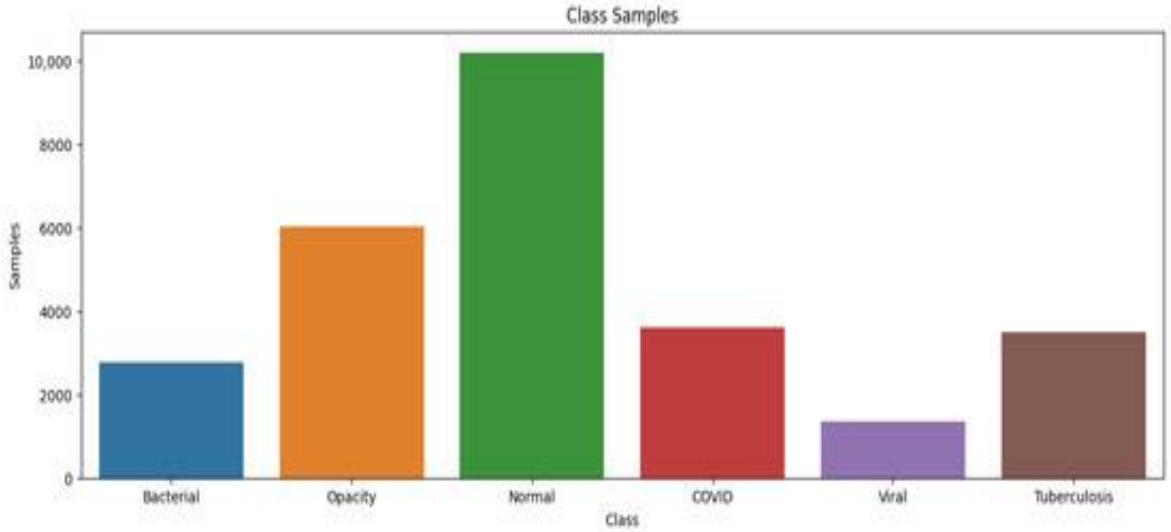


Figure 2. Distribution of lung disease categories within the dataset.

In this investigation, 80% of the overall dataset was utilized for the training phase. Specifically, 60% of the data, corresponding to 17,566 samples, served for model training, while an additional 20%, amounting to 4,390 samples, was reserved for validation. The final 20% of the dataset, comprising 5,489 samples, was allocated for testing [33, 34]. This partitioning strategy facilitated a comprehensive evaluation of model performance across the targeted lung conditions: bacterial pneumonia, COVID-19, normal cases, lung opacity, tuberculosis, and viral pneumonia. **Table 1** summarizes the data distribution for each category across the training, validation, and testing sets. For example, bacterial pneumonia was divided into 1,766 training, 441 validation, and 573 testing instances. Similarly, COVID-19 cases were allocated as 2,305 for training, 576 for validation, and 735 for testing.

Table 1. Dataset partitioning across training, validation, and testing sets.

Lung condition	Training collection	Validation collection	Testing collection
Bacterial pneumonia	1766	441	573

COVID-19	2305	576	735
Normal	6570	1643	1979
Opacity	3829	957	1226
Tuberculosis	2233	558	709
Viral pneumonia	863	215	267
Total	**17,566**	**4390**	**5489**

Image augmentation strategies

In this study, image augmentation techniques [35, 36] were employed to mitigate the constraints posed by the relatively modest size of the training dataset and to strengthen the overall training process. These techniques [37, 38] apply a range of transformations that expand the training data volume while introducing diversity in scale, orientation, and appearance. This increased variability helps reduce overfitting to the specific features of the original training samples and improves the model's ability to generalize to previously unseen data.

We applied a full range of augmentation methods to all lung condition categories except the normal class, as detailed in **Table 2**, to balance the training dataset. As a result, each disease category in the training set was expanded to 6,570 images. These augmentation procedures were critical for enhancing the diversity and resilience of the training collection. The applied transformations included random rotations, height and width shifts, brightness adjustments, zooming, shearing, horizontal flipping, and fill mode modifications. Collectively, these operations generated a more varied and representative training dataset, thereby supporting the development of an accurate and reliable model capable of performing well under different imaging conditions.

Table 2. Details of the data augmentation techniques applied to the training set.

Parameter	Value
Brightness range	[0.55, 1.35]
Horizontal flip	True
Rotation	[+28, −28]
Width shift	[0.75, 1.35]
Height shift	0.30
Zoom	[0.45, 0.95]
Fill mode	Nearest
Sheare	0.40

The parameters in **Table 2** were selected based on empirical evaluation and established best practices in deep learning for medical imaging applications. The objective was to design effective augmentation pipelines that realistically replicate variability observed in clinical environments and promote robust generalization across heterogeneous imaging conditions. Ranges for brightness, rotation, width shift, height shift, and zoom were optimized through iterative testing to provide meaningful diversity while preserving the clinical integrity of the radiographic images. Fixed fractional values for height shift and shear parameters were determined following preliminary trials that yielded optimal model accuracy and stability. No additional distribution spread was implemented, as the selected parameters sufficiently emulated the natural variations and distortions typical of clinical chest X-rays without introducing implausible artifacts.

These augmentation examples demonstrate how each transformation alters the original image, encouraging the model to learn invariant and discriminative features of underlying anatomical structures despite such variations. This strategy proved essential for maintaining strong model performance across diverse imaging scenarios. By exposing the network to augmented data, its ability to generalize from the training set to new instances improved significantly, increasing its diagnostic reliability in real-world clinical applications (**Figure 3**).

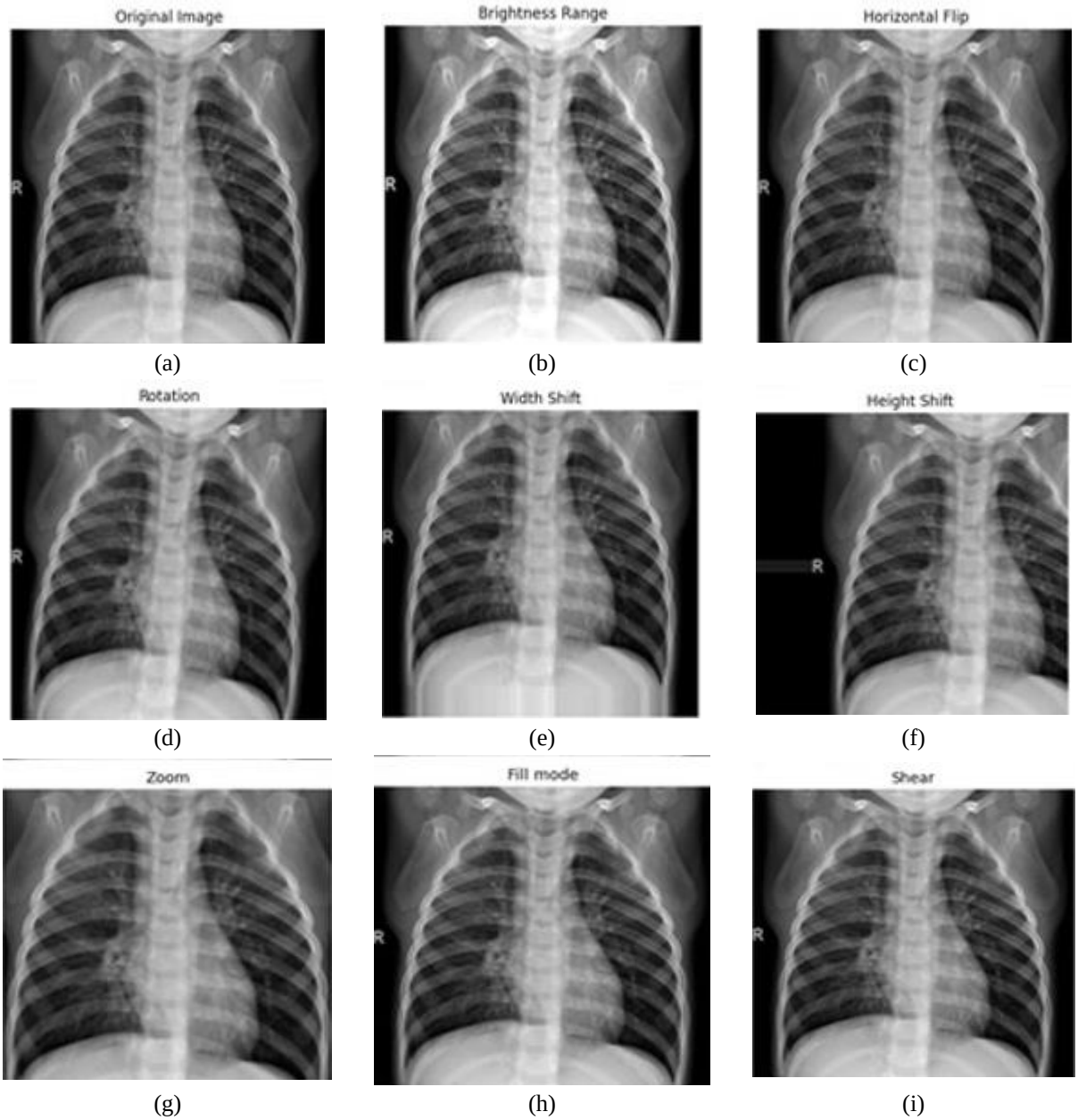


Figure 3. Examples of data augmentation techniques applied to a viral pneumonia image.

VGG16 architecture

The VGG16 architecture [39] is characterized by its simplicity and effectiveness, consisting of 16 layers in total—13 convolutional layers and 3 fully connected layers. A defining feature of VGG16, distinguishing it from prior deep learning models, is its greater network depth achieved through the consistent use of small (3 × 3) convolution filters, which contributed to superior performance. VGG16's design strategy relies on progressively stacking multiple convolutional layers with small receptive fields. This approach increases network depth while keeping the parameter count manageable, thereby improving computational efficiency and scalability. Using uniform 3 × 3 filters across all convolutional layers simplifies hyperparameter selection by removing the need to optimize varying filter sizes across layers—a common complication in network design. This uniformity also supports straightforward scaling to accommodate different levels of complexity and facilitates adaptation to a wide range of image recognition applications.

Every convolutional layer in VGG16 incorporates ReLU (Rectified Linear Unit) activation functions, which introduce non-linearity essential for learning intricate patterns from the input images. Following the activation, max-pooling operations are applied at regular intervals to reduce the spatial dimensions of the feature maps. This downsampling step decreases computational demands as the network grows deeper, helps control overfitting, and produces more abstract feature representations with fewer parameters. At the conclusion of the convolutional stack, VGG16 employs three fully connected layers, with the final layer utilizing a softmax activation function to produce class probability distributions.

SE block

The squeeze-and-excitation (SE) block [40] constitutes an innovative component in convolutional neural network design that enhances representational power by explicitly modeling channel-wise interdependencies. This lightweight module has shown notable improvements in CNN performance across diverse tasks and datasets, with only marginal increases in computational cost and no major structural changes. The SE block consists of two primary operations: squeeze and excitation. In the squeeze phase, global average pooling compresses spatial information from each feature map into a single-channel descriptor. This operation reduces each channel to a scalar value, effectively capturing a global summary of the feature responses across spatial dimensions and providing a compact representation for subsequent processing.

During the excitation phase, a self-gating mechanism leverages these channel-wise statistics to adaptively recalibrate the feature maps. A gating operation with sigmoid activation learns inter-channel dependencies and generates modulation weights. These weights are then applied to the original feature maps, selectively amplifying informative channels while attenuating less relevant ones. This dynamic recalibration enables the network to emphasize the most task-relevant features. Consequently, integrating SE blocks substantially strengthens the network's representational capacity. For example, in image classification applications [41], SE blocks allow models to prioritize more discriminative visual cues, improving accuracy. **Figure 4** illustrates the structural flowchart of the SE block.

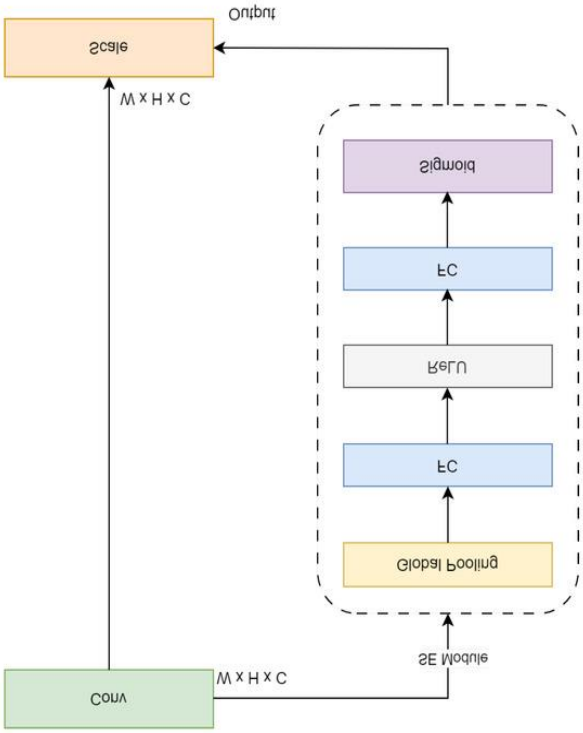


Figure 4. Flowchart of the SE block architecture. The SE block improves model performance through channel-wise feature recalibration.

The process begins with global average pooling, which condenses spatial information into a channel descriptor. This descriptor is subsequently passed through fully connected layers utilizing ReLU and sigmoid activations to produce scaling weights for each channel. The resulting weights are multiplied with the original convolutional feature maps, thereby emphasizing the most relevant features.

Proposed VGG16-SE architecture

Figure 5 presents the overall structure of the modified VGG16 model enhanced with SE blocks, specifically designed for the multi-label classification of chest X-ray images. The network accepts input images of standard dimensions (224 × 224 pixels with three color channels). It then proceeds through a series of convolutional blocks in which the number of filters progressively increases with network depth. The initial two convolutional layers each utilize 64 filters to extract low-level features from the input. Immediately after these layers, an SE block recalibrates the feature maps by dynamically adjusting channel-wise importance weights.

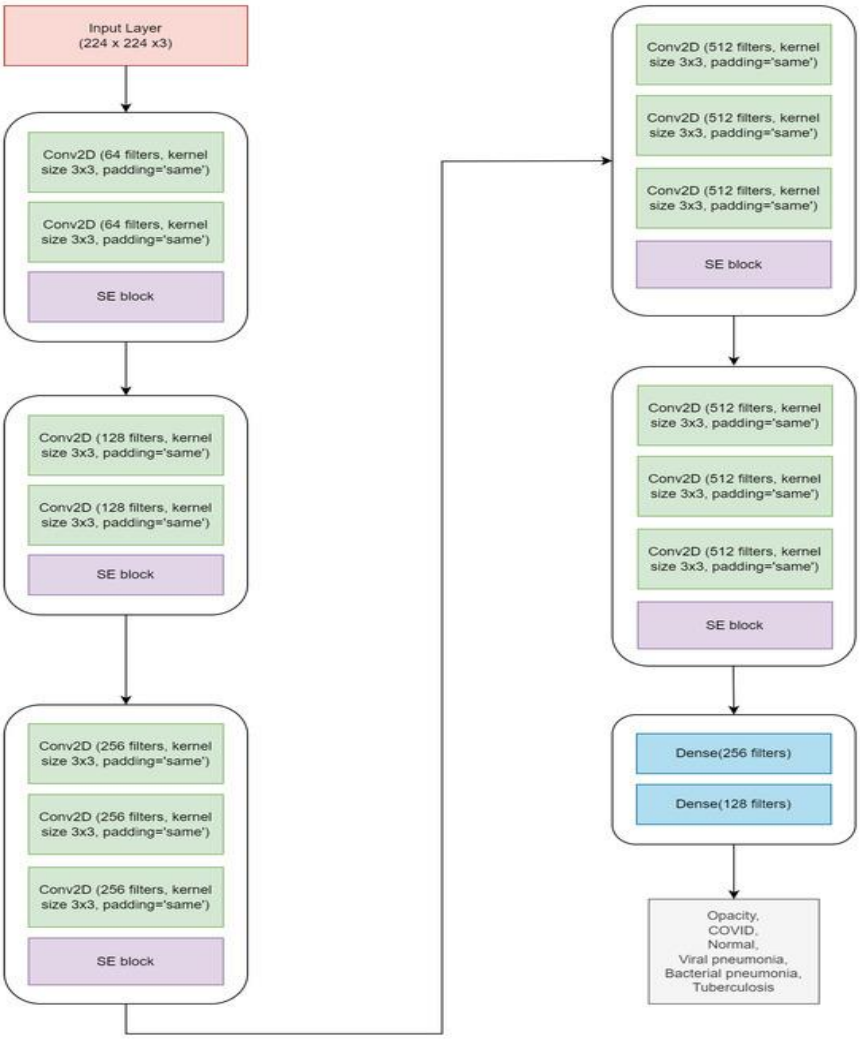


Figure 5. Overview of the proposed network architecture.

As the architecture deepens, the number of filters in subsequent convolutional blocks doubles sequentially—to 128, then 256, and finally 512—each block followed by an SE module. These progressive increases enable the extraction of increasingly complex and abstract feature representations. The interleaved SE blocks further ensure that the most task-relevant features receive greater emphasis throughout the network. In the deepest layers, where 512 filters are employed, the SE blocks play a particularly important role in handling the high-dimensional feature space. By selectively amplifying informative channels and attenuating less relevant ones, they help maintain focus on diagnostically significant information and prevent the dilution of critical signals. After the convolutional backbone, the model transitions to a classification head comprising three dense layers. The final dense layer applies a softmax activation function to generate probability distributions across the six target classes: lung opacity, COVID-19, normal, viral pneumonia, bacterial pneumonia, and tuberculosis.

Integrating SE blocks into the VGG16 framework introduces an adaptive channel recalibration mechanism through squeeze-and-excitation operations. This addition allows the network to dynamically modulate the contribution of each feature channel within convolutional layers, thereby enhancing its discriminative capability. In medical image analysis, such as identifying lung diseases from chest X-rays, SE blocks enable the model to capture subtle pathological distinctions that the standard VGG16 architecture may overlook. Such improvements are especially valuable in clinical practice, where diagnostic precision directly impacts clinical decision-making and patient prognosis. **Table 3** summarizes the modified VGG16-SE architecture.

Table 3. Detailed architecture of the proposed VGG16 model with SE blocks.

Layer (Type)	Output shape	Number of params
conv2d (Conv2D)	(None, 224, 224, 64)	1792
conv2d_1 (Conv2D)	(None, 224, 224, 64)	36,928
se_block (SEBlock)	(None, 224, 224, 64)	580
conv2d_2 (Conv2D)	(None, 224, 224, 128)	73,856
conv2d_3 (Conv2D)	(None, 224, 224, 128)	147,584
se_block_1 (SEBlock)	(None, 224, 224, 128)	2184
conv2d_4 (Conv2D)	(None, 224, 224, 256)	295,168
conv2d_5 (Conv2D)	(None, 224, 224, 256)	590,080
conv2d_6 (Conv2D)	(None, 224, 224, 256)	590,080
se_block_2 (SEBlock)	(None, 224, 224, 256)	8464
conv2d_7 (Conv2D)	(None, 224, 224, 512)	1,180,160
conv2d_8 (Conv2D)	(None, 224, 224, 512)	2,359,808
conv2d_9 (Conv2D)	(None, 224, 224, 512)	2,359,808
se_block_3 (SEBlock)	(None, 224, 224, 512)	33,312
conv2d_10 (Conv2D)	(None, 224, 224, 512)	2,359,808
conv2d_11 (Conv2D)	(None, 224, 224, 512)	2,359,808
conv2d_12 (Conv2D)	(None, 224, 224, 512)	2,359,808
se_block_4 (SEBlock)	(None, 224, 224, 512)	33,312
dense (Dense)	(None, 256)	131,328
dense_1 (Dense)	(None, 128)	32,896
dense_2 (Dense)	(None, 6)	774
Total params:		14,957,538
Trainable params:		14,957,538
Non-trainable params:		0

Investigation outcomes

Experimental setup

The experimental environment for this study was set up on a Windows 11 Pro system with 32 GB of RAM and an NVIDIA RTX 3060 GPU with 12 GB of memory. The implementation was carried out in Python, utilizing Keras as the primary deep learning library with TensorFlow serving as the backend. Hyperparameters were selected through empirical optimization to balance training efficiency and model performance. All models were trained for 30 epochs, at which point performance metrics such as accuracy and loss demonstrated stable convergence, providing an optimal balance between computational time and effectiveness. The Adam optimizer was employed for its adaptive learning rate properties, which are particularly suitable for managing the fluctuating gradients common in neural network optimization. A learning rate of 0.0001 was adopted to ensure stable and gradual convergence, facilitating consistent loss minimization without excessively rapid or sluggish updates. A mini-batch size of 32 was chosen to balance computational efficiency and training stability while making effective use of GPU memory. Finally, categorical cross-entropy was used as the loss function, given its suitability for multi-class classification problems. **Table 4** summarizes the hardware configuration and selected hyperparameters utilized in the experiments.

Table 4. Experimental setup and hyperparameter configuration.

Parameter	Specification / Value
Operating System	Windows 11 Pro

System Memory (RAM)	32 GB
Programming Language	Python
Deep Learning Framework	Keras with TensorFlow backend
Graphics Processing Unit (GPU)	NVIDIA RTX 3050 (12 GB)
Training Epochs	30
Optimization Algorithm	Adam
Learning Rate	0.0001
Mini-Batch Size	32
Loss Criterion	Cross-entropy

Performance metrics

Evaluation metrics play a fundamental role in determining the proficiency and reliability of classification models. These measures are essential not only for assessing performance during training but also for evaluating the model's ability to generalize effectively to unseen data. Key indicators such as accuracy, precision, recall, and F1-score provide comprehensive insights into model effectiveness and robustness [42-44].

$$\text{Accuracy} = \frac{(\text{TPos} + \text{TNeg})}{(\text{TPos} + \text{FNeg} + \text{FPos} + \text{TNeg})} \quad (1)$$

$$\text{Precision} = \frac{\text{TPos}}{(\text{TPos} + \text{FPos})} \quad (2)$$

$$\text{Recall} = \frac{\text{TPos}}{(\text{TPos} + \text{FNeg})} \quad (3)$$

$$\text{F}_1 - \text{score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

Where TNeg denotes the number of true negative instances correctly identified by the model. FPos represents false positives, where negative instances are incorrectly classified as positive. TPos refers to true positive instances that the model accurately recognizes. In contrast, FNeg indicates false negative cases where positive instances are mistakenly classified as negative.

Table 5 presents a comparative analysis of the overall performance between the baseline VGG16 model and the proposed VGG16-SE variant. In terms of accuracy, which reflects the proportion of correct predictions across all classes, the modified VGG16-SE model achieved 98.49%, substantially outperforming the standard VGG16 model's 96.78%. This gain demonstrates improved classification reliability and reduced false positives and false negatives. Precision increased from 95.91% in the original VGG16 to 98.41% in the enhanced model, indicating more reliable positive predictions and fewer erroneous disease identifications. Recall (sensitivity) rose from 96.41% to 98.07%, signifying a stronger ability to detect actual positive cases—an essential attribute in medical diagnostics where missed detections can have significant clinical consequences. Finally, the F1-score, representing the harmonic mean of precision and recall, improved from 96.12% to 98.23%. This balanced metric confirms consistent enhancements in both precision and recall attributable to the incorporation of SE blocks.

Table 5. Overall performance metrics comparison.

Metric	Classic VGG16	Modified VGG16-SE
Accuracy	0.9678	0.9849
Precision	0.9591	0.9841
Recall	0.9641	0.9807
F1-score	0.9612	0.9823

The comparative evaluation in **Table 6** shows notable improvements in precision, recall, and F1-score for various lung pathologies when comparing the standard VGG16 architecture with its SE-enhanced counterpart. For bacterial pneumonia, the baseline VGG16 achieved 97.65% precision and a 95.91% F1-score. In contrast, the modified VGG16-SE model attained a recall of 98.78%, resulting in an improved F1-score of 98.09%. This enhancement reflects the model's superior capacity to detect true cases while minimizing missed diagnoses. Regarding COVID-19, the VGG16-SE variant demonstrated exceptional performance with a precision of 99.46%, a recall of 99.73%, and an F1-score of 99.59%, markedly surpassing the baseline VGG16's precision of 98.50% and F1-score of 98.37%.

For the normal category, integrating SE blocks increased precision (from 96.41% to 97.91%) and F1-score (from 97.01% to 98.54%). These gains are particularly valuable for reducing false positives, thereby limiting unnecessary patient concern and follow-up procedures. For lung

opacity—a condition characterized by often subtle imaging findings—the modified model improved precision from 96.75% to 98.92% and F1-score from 95.67% to 97.81%, enabling more reliable differentiation from other pathologies and supporting appropriate clinical management. Tuberculosis detection remained highly accurate in both architectures, with the SE-enhanced model delivering a modest yet positive refinement in F1-score.

The viral pneumonia category, which presented the greatest difficulty for the baseline VGG16 (precision of 86.55% and F1-score of 90.13%), showed the largest gains. The VGG16-SE model increased precision to 96.92% and F1-score to 95.64%, demonstrating the effectiveness of SE blocks in addressing conditions with overlapping radiographic features.

Table 6. Classification performance of the baseline VGG16 and the proposed VGG16-SE model across different lung conditions.

	Classic VGG16			Modified VGG16-SE		
Lung Condition	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Bacterial	0.9765	0.9424	0.9591	0.9742	0.9878	0.9809
COVID-19	0.9850	0.9823	0.9837	0.9946	0.9973	0.9959
Normal	0.9641	0.9763	0.9701	0.9791	0.9919	0.9854
Opacity	0.9675	0.9462	0.9567	0.9892	0.9674	0.9781
Tuberculosis	0.9958	0.9972	0.9965	0.9986	0.9958	0.9972
Viral	0.8655	0.9401	0.9013	0.9692	0.9438	0.9564

Statistical analysis

Table 7 summarizes the outcomes of the statistical evaluation performed using p-values and t-tests to determine the significance of the observed performance differences. The p-value for the baseline VGG16 model was 0.0506, slightly exceeding the standard significance threshold of 0.05 and indicating that the performance gains were not statistically significant. Conversely, the p-value for the modified VGG16-SE model was 0.0368, falling below the 0.05 threshold and confirming that the improvements associated with SE block integration are statistically significant.

Table 7. Statistical analysis results.

Model	**p-Value**	**t-Test**
Classic VGG16	0.0506	1.3450
Modified VGG16-SE	0.0368	1.8404

The t-test statistic for the standard VGG16 model was 1.3450, suggesting a moderate performance differential relative to the enhanced version. Notably, the t-test value for the modified VGG16-SE model reached 1.8404, reflecting a more pronounced magnitude of improvement. These results collectively affirm the efficacy of the SE block, demonstrating that the performance gains are both statistically significant and practically meaningful.

Confusion matrices and ROC curves

The confusion matrix for the baseline VGG16 model, presented in **Figure 6**, demonstrates generally strong classification performance across several lung conditions while also revealing specific areas of misclassification, particularly among pathologies with similar radiographic appearances. The model reliably identifies COVID-19 and normal cases; however, it frequently confuses bacterial and viral pneumonia, as shown by 33 bacterial pneumonia cases incorrectly assigned to the viral pneumonia category. Distinction between lung opacity and tuberculosis remains relatively robust, although minor overlaps with other classes persist. Overall, the model handles conditions with more distinctive features effectively but shows limitations when differentiating subtle or overlapping patterns.

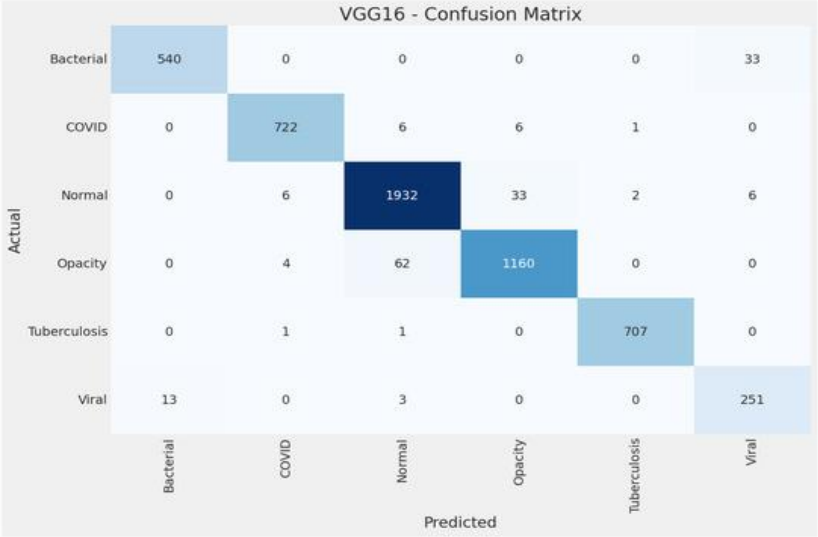


Figure 6. Confusion matrix for lung conditions using the baseline VGG16 model.

In comparison, the confusion matrix for the modified VGG16 model incorporating SE blocks, shown in **Figure 7**, indicates substantial improvements in classification accuracy and a marked reduction in misclassification rates across nearly all categories. The enhanced model demonstrates superior differentiation between bacterial and viral pneumonia, with considerably fewer erroneous assignments than the standard VGG16. The inclusion of SE blocks enables more effective focus on discriminative image features, thereby minimizing confusion arising from overlapping clinical and radiographic characteristics. Misclassifications involving normal cases and other pathologies are also notably reduced, reflecting the model's improved generalization and discriminative capabilities across diverse presentations.

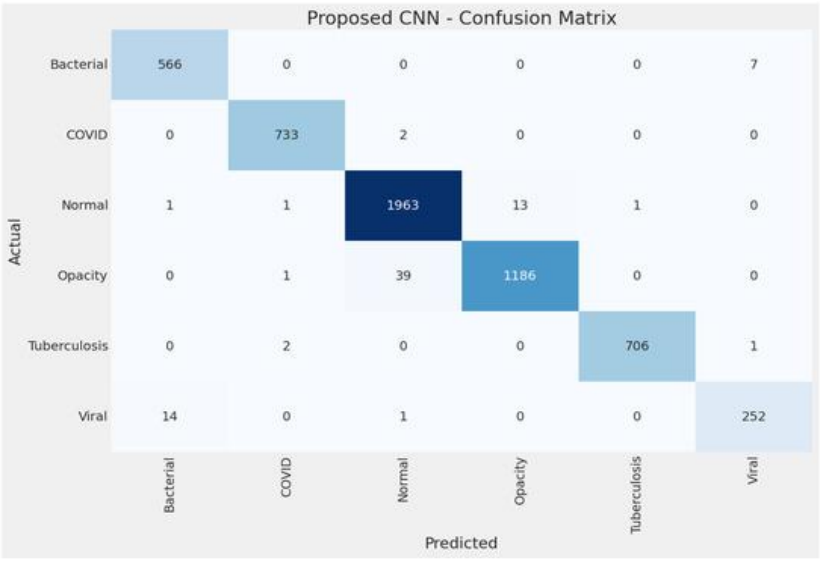


Figure 7. Confusion matrix for lung conditions using the VGG16 model with SE blocks.

The receiver operating characteristic (ROC) curves for the baseline VGG16 architecture (**Figure 8**) and the enhanced VGG16-SE model (**Figure 9**) offer a detailed assessment of discriminative performance across the six diagnostic categories: bacterial pneumonia, COVID-19, normal, lung opacity, tuberculosis, and viral pneumonia.

These curves evaluate each model's ability to distinguish between the presence and absence of specific conditions. For the standard VGG16 model in **Figure 8**, the ROC curves indicate strong overall performance, with high area under the curve (AUC) values across categories. Particularly noteworthy are the excellent results for COVID-19 (AUC of 0.9900) and tuberculosis (AUC of 0.9983), reflecting high sensitivity and specificity. Performance remains solid but comparatively lower for bacterial and viral pneumonia, suggesting opportunities for further architectural refinement to improve discriminative power in these areas.

The receiver operating characteristic (ROC) curves for the baseline VGG16 architecture (**Figure 8**) and the enhanced VGG16-SE model (**Figure 9**) offer a detailed assessment of discriminative performance across the six diagnostic categories: bacterial pneumonia, COVID-19, normal, lung opacity, tuberculosis, and viral pneumonia. These curves evaluate each model's ability to distinguish between the presence and absence of specific conditions. For the standard VGG16 model in **Figure 8**, the ROC curves indicate strong overall performance, with high area under the curve (AUC) values across categories. Particularly noteworthy are the excellent results for COVID-19 (AUC of 0.9900) and tuberculosis (AUC of 0.9983), reflecting high sensitivity and specificity. Performance remains solid but comparatively lower for bacterial and viral pneumonia, suggesting opportunities for further architectural refinement to improve discriminative power in these areas.

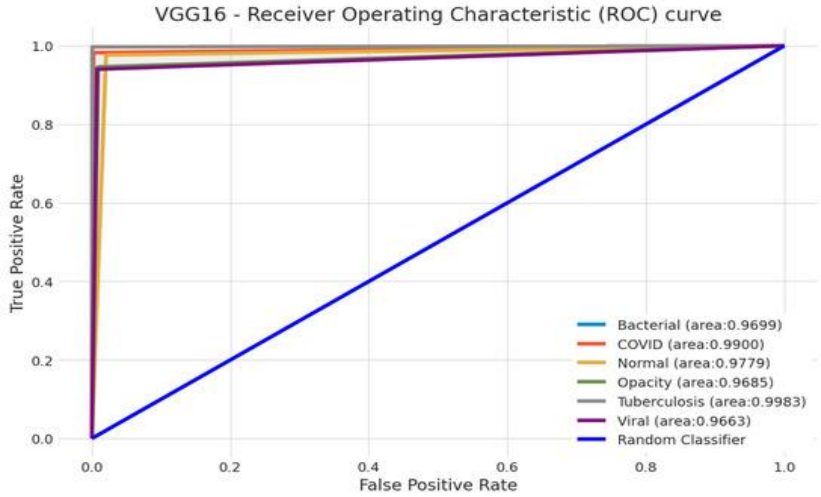


Figure 8. ROC curves for lung conditions using the baseline VGG16 model.

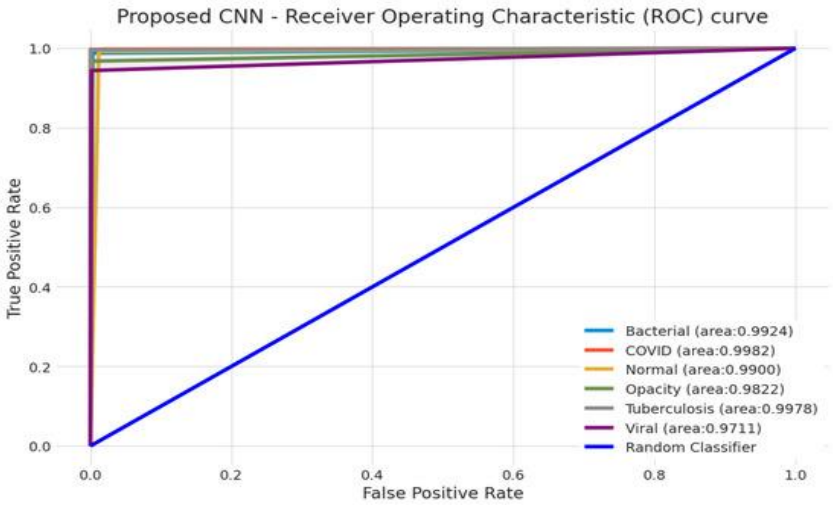


Figure 9. ROC curves for lung conditions using the modified VGG16 model with SE blocks.

By contrast, the ROC curves for the VGG16-SE model in **Figure 9** demonstrate considerable enhancement across all categories. The incorporation of SE blocks strengthens the model's ability to emphasize diagnostically relevant features, yielding AUC values approaching 1.00 for several conditions, including COVID-19 and tuberculosis, indicating near-perfect classification. For instance, the AUC for bacterial pneumonia increased from 0.9699 in the baseline model to 0.9924 in the SE-enhanced version. These advancements highlight the substantial contribution of SE blocks in boosting sensitivity and minimizing false positives and negatives. Such improvements hold significant clinical value, as accurate multi-class categorization of chest X-ray findings is essential for effective treatment planning. The superior sensitivity and specificity of the VGG16-SE model can help reduce diagnostic errors and provide clinicians with a dependable decision-support tool for pulmonary disease assessment.

Comparative analysis with other relevant studies

Table 8 presents a comparative overview of recent research on multi-class lung disease classification using various deep learning approaches, highlighting differences in methodologies and achieved accuracies. Each study investigated different convolutional neural network architectures applied to chest X-ray images to identify multiple pulmonary conditions.

Ref. [24] employed a standard 2D-CNN architecture on a dataset of 14,948 images spanning six disease categories, attaining an accuracy of 96.75%. This result establishes a respectable benchmark for relatively straightforward CNN models in complex image classification tasks. In comparison, Ref. [25] investigated a modified VGG19 architecture on a substantially larger dataset of 48,582 images across six categories, achieving a higher accuracy of 98.88%. These findings show that targeted modifications to established architectures such as VGG19 can yield considerable gains in diagnostic performance.

Table 8. Comparison with other relevant studies on lung disease identification.

Reference	No. of classes	Dataset size (Images)	Classification model	Accuracy (%)
[24]	6	14,948	2D Convolutional Neural Network (2D-CNN)	96.75
[25]	6	48,582	Enhanced VGG19	98.88
[26]	6	5700	VGG16 with Multi-Scale Features	97.47
[27]	5	2186	ResNet50	91.60
[28]	4	21,885	ResNet-152	94.00
[29]	4	21,165	Modified MobileNetV2	95.80
**Proposed method	6	49,299	VGG16 Incorporating SE Blocks	98.49

Another study [26] utilized a VGG16 model augmented with multi-scale feature processing on a dataset of 5,700 images, reaching an accuracy of 97.47%. This approach demonstrates how multi-scale techniques can enhance sensitivity to features of varying scales and improve overall classification effectiveness. Ref. [27] applied ResNet50 to a smaller dataset of 2,186 images across five categories, achieving 91.60% accuracy. Meanwhile, Ref. [28] employed the deeper ResNet-152 architecture on 21,885 images across four categories, achieving 94.00% accuracy. These results emphasize the influence of both dataset size and network depth on model performance.

Ref. [29] introduced modifications to the MobileNetV2 architecture—a lightweight model suitable for mobile applications—using 21,165 images across four disease categories and achieving 95.80% accuracy. This work highlights the viability of efficient models in medical imaging applications, achieving a favorable balance between computational demands and accuracy. The present study contributes a VGG16 architecture enhanced with SE blocks, evaluated on a large dataset of 49,299 images across six categories. The proposed model attained an accuracy of 98.49%, demonstrating that integrating SE blocks can markedly improve the performance of established architectures. By adaptively recalibrating channel-wise feature responses, SE blocks enable more effective emphasis on informative features, thereby enhancing overall diagnostic precision.

Results and Discussion

This study examined the relative performance of the conventional VGG16 architecture and its enhanced counterpart, VGG16-SE, which incorporates squeeze-

and-excitation blocks, for the multi-class classification of lung pathologies from chest X-ray images. Experimental results, analyzed through confusion matrices and ROC curves, showed consistent improvements in key evaluation metrics—including precision, recall, and F1-score—when using the VGG16-SE model. These gains can be attributed to the SE blocks' capacity to dynamically recalibrate channel-wise feature responses, enabling the model to prioritize the most relevant characteristics for each diagnostic class. Such refinement is particularly important in medical imaging, where accurate classification directly influences clinical decision-making and therapeutic strategies.

The superior accuracy and robustness of the modified VGG16-SE model carry important implications for clinical applications and patient care. By reducing false positive and false negative rates, the model offers clinicians a more trustworthy diagnostic aid, which may support earlier identification of pulmonary diseases. This capability is especially valuable for time-sensitive conditions such as tuberculosis and COVID-19, where prompt and precise diagnosis is closely linked to treatment success and improved patient prognosis. Furthermore, the enhanced precision and recall support integrating this model into routine radiological workflows, serving as an effective second reader to assist radiologists, minimize interpretive errors, and ultimately improve the quality of patient management.

Integrating advanced components such as SE blocks into convolutional neural networks represents a meaningful step forward in applying deep learning to medical image analysis. This approach introduces an attention-like mechanism that adaptively modulates feature channel contributions, amplifying informative signals while attenuating less relevant ones. Consequently, the model achieves greater sensitivity and specificity in detecting pathological patterns. By automating routine diagnostic tasks, providing reliable decision support, and enabling sophisticated medical image analysis, systems like the VGG16-SE model can alleviate workload pressures on healthcare systems and improve both operational efficiency and clinical outcomes.

Successful adoption of such advanced AI tools in clinical environments depends significantly on comprehensive training and education of healthcare professionals. Equipping medical staff with the knowledge and skills to use these technologies effectively is essential for their seamless incorporation into daily practice. Targeted training initiatives focused on the practical use of AI in diagnostics, supported by continuous education and system updates, will be instrumental in developing a workforce capable of fully harnessing the benefits of artificial intelligence in healthcare.

Conclusion

This investigation demonstrates a meaningful advancement in the multi-label classification of lung diseases from chest X-ray images through the integration of SE blocks into the standard VGG16 architecture. Incorporating these blocks substantially improved precision, recall, and F1-score across a range of pulmonary conditions, increasing the reliability and accuracy of automated diagnosis in clinical settings. Such enhancements are vital for enabling timely and appropriate patient care. The findings affirm that strategic architectural modifications, including adding SE blocks to conventional CNNs, can significantly improve model performance. The proposed VGG16-SE architecture shows a stronger ability to discern intricate patterns and subtle radiographic distinctions that unmodified networks often miss. This capability is especially advantageous when addressing complex and overlapping lung pathologies, where early and accurate detection can profoundly affect clinical outcomes.

Future research directions may include expanding the training dataset to encompass a wider variety of pathological presentations and greater demographic diversity. Incorporating data from more heterogeneous populations would further improve the model's generalizability, ensuring robust performance across different patient groups and accounting for variations in disease manifestation and imaging protocols influenced by demographic and regional factors.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Mondal MRH, Bharati S, Podder P, Kamruzzaman J. Deep Learning and Federated Learning for

Screening COVID-19: A Review. Bio Med Inform. 2023;3(3):691-7.
2. Ibrokhimov B, Kang JY. Deep learning model for COVID-19-infected pneumonia diagnosis using chest radiography images. Bio Med Inform. 2022;2(4):654-70.
3. Andersen AM, Rasmussen BS, Graumann O, Overgaard S, Lundemann M, Haubro MH, see. Minimal hip joint space width measured on X-rays by an artificial intelligence algorithm—a study of reliability and agreement. Bio Med Inform. 2023;3(3):714-23.
4. Intxausti E, Skočaj D, Cernuda C, Zugasti E. A Methodology for Advanced Manufacturing Defect Detection through Self-Supervised Learning on X-ray Images. Appl Sci. 2024;14(7):2785.
5. Singh AP, Singh A, Kumar A, Agarwal H, Yadav S, Gupta M. Development of an artificial neural network-based image retrieval system for lung disease classification and identification. Eng Proc. 2024;62(1):2.
6. An Q, Chen W, Shao W. A Deep Convolutional Neural Network for Pneumonia Detection in X-ray Images with Attention Ensemble. Diagnostics. 2024;14(4):390.
7. Rana S, Hosen MJ, Tonni TJ, Rony MAH, Fatema K, Hasan MZ, et al. DeepChestGNN: A Comprehensive Framework for Enhanced Lung Disease Identification through Advanced Graphical Deep Features. Sensors. 2024;24(9):2830.
8. Hussein HI, Mohammed AO, Hassan MM, Mstafa RJ. Lightweight deep CNN-based models for early detection of COVID-19 patients from chest X-ray images. Expert Syst Appl. 2023;223:119900.
9. Bista R, Timilsina A, Manandhar A, Paudel A, Bajracharya A, Wagle S, et al. Advancing tuberculosis detection in chest X-rays: A YOLOv7-based approach. Information. 2023;14(12):655.
10. Kvak D, Chromcová A, Biroš M, Hrubỳ R, Kvaková K, Pajdaković M, et al. Chest X-ray abnormality detection by using artificial intelligence: A single-site retrospective study of deep learning model performance. Bio Med Inform. 2023;3(1):82-101.
11. Althenayan AS, AlSalamah SA, Aly S, Nouh T, Mahboub B, Salameh L, et al. COVID-19 Hierarchical Classification Using a Deep Learning Multi-Modal. Sensors. 2024;24(8):2641.
12. Louati H, Louati A, Lahyani R, Kariri E, Albanyan A. Advancing Sustainable COVID-19 Diagnosis: Integrating Artificial Intelligence with Bioinformatics in Chest X-ray Analysis. Information. 2024;15(4):189.
13. Pan CT, Kumar R, Wen ZH, Wang CH, Chang CY, Shiue YL. Improving Respiratory Infection Diagnosis with Deep Learning and Combinatorial Fusion: A Two-Stage Approach Using Chest X-ray Imaging. Diagnostics. 2024;14(5):500.
14. Kim S, Rim B, Choi S, Lee A, Min S, Hong M. Deep learning in multi-class lung diseases' classification on chest X-ray images. Diagnostics. 2022;12(4):915.
15. Sanida T, Dasygenis M. A novel lightweight CNN for chest X-ray-based lung disease identification on heterogeneous embedded system. Appl Intell. 2024;54(6):4756-80.
16. Bhosale RD, Yadav D. Customized convolutional neural network for pulmonary multi-disease classification using chest X-ray images. Multimed Tools Appl. 2024;83(6):18537-71.
17. Pradeep Dalvi P, Reddy Edla D, Purushothama B, Dharavath R. COVID-19 detection from Chest X-ray images using a novel lightweight hybrid CNN architecture. Multimed Tools Appl. 2024:1-23.
18. Rao GE, Rajitha B, Srinivasu PN, Ijaz MF, Woźniak M. Hybrid framework for respiratory lung diseases detection based on classical CNN and quantum classifiers from chest X-rays. Biomed Signal Process Control. 2024;88:105567.
19. Chauhan S, Edla DR, Boddu V, Rao MJ, Cheruku R, Nayak SR, et al. Detection of COVID-19 using edge devices by a light-weight convolutional neural network from chest X-ray images. BMC Med Imaging. 2024;24(1):1.
20. Ye RZ, Lipatov K, Diedrich D, Bhattacharyya A, Erickson BJ, Pickering BW, et al. Automatic ARDS surveillance with chest X-ray recognition using convolutional neural networks. J Crit Care. 2024;82:154794.
21. Taleb A, Kirchler M, Monti R, Lippert C. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA; 2022. p. 20908-21.
22. Liu C, Cheng S, Shi M, Shah A, Bai W, Arcucci R. Imitate: Clinical prior guided hierarchical vision-language pre-training. arXiv. 2023;arXiv:2310.07355.

23. Liu C, Wan Z, Cheng S, Zhang M, Arcucci R. Etp: Learning transferable ECG representations via ECG-text pre-training. In: Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Republic of Korea; 2024. p. 8230-4.
24. Sultana A, Nahiduzzaman M, Bakchy SC, Shahriar SM, Peyal HI, Chowdhury ME, et al. A real time method for distinguishing COVID-19 utilizing 2D-CNN and transfer learning. Sensors. 2023;23(9):4458.
25. Sanida MV, Sanida T, Sideris A, Dasygenis M. An advanced deep learning framework for multi-class diagnosis from chest X-ray images. J. 2024;7(1):48-71.
26. Sarkar O, Islam MR, Syfullah MK, Islam MT, Ahamed MF, Ahsan M, et al. Multi-Scale CNN: An Explainable AI-Integrated Unique Deep Learning Framework for Lung-Affected Disease Classification. Technologies. 2023;11(5):134.
27. Al-Timemy AH, Khushaba RN, Mosa ZM, Escudero J. An efficient mixture of deep and machine learning models for COVID-19 and tuberculosis detection using X-ray images in resource limited settings. In: Artificial Intelligence for COVID-19. Cham: Springer; 2021. p. 77-100.
28. Hasanah SA, Pravitasari AA, Abdullah AS, Yulita IN, Asnawi MH. A deep learning review of resnet architecture for lung disease Identification in CXR Image. Appl Sci. 2023;13(23):13111.
29. Sanida T, Sideris A, Tsiktsiris D, Dasygenis M. Lightweight neural network for COVID-19 detection from chest X-ray images implemented on an embedded system. Technologies. 2022;10(2):37.
30. Kaggle. COVID-19 Radiography Dataset [Internet]. Available from: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database/activity (accessed on 20 May 2024).
31. Rahman T, Khandakar A, Kadir MA, Islam KR, Islam KF, Mazhar R, et al. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. IEEE Access. 2020;8:191586-601.
32. Kermany D, Zhang K, Goldbaum M. Labeled optical coherence tomography (OCT) and chest X-ray images for classification. Mendeley Data. 2018;2:651.
33. Rácz A, Bajusz D, Héberger K. Effect of dataset size and train/test split ratios in QSAR/QSPR multi-class classification. Molecules. 2021;26(4):1111.
34. Nguyen QH, Ly HB, Ho LS, Al-Ansari N, Le HV, Tran VQ, et al. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Math Probl Eng. 2021;2021:4832864.
35. Alomar K, Aysel HI, Cai X. Data augmentation in classification and segmentation: A survey and new strategies. J Imaging. 2023;9(2):46.
36. Xu M, Yoon S, Fuentes A, Park DS. A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognit. 2023;137:109347.
37. Hao X, Liu L, Yang R, Yin L, Zhang L, Li X. A review of data augmentation methods of remote sensing image target recognition. Remote Sens. 2023;15(3):827.
38. Makhlouf A, Maayah M, Abughanam N, Catal C. The use of generative adversarial networks in medical image augmentation. Neural Comput Appl. 2023;35(33):24055-68.
39. Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev. 2020;53(8):5455-516.
40. Jin X, Xie Y, Wei XS, Zhao BR, Chen ZM, Tan X. Delving deep into spatial pooling for squeeze-and-excitation networks. Pattern Recognit. 2022;121:108159.
41. Wang L, Qi Y, Mathiopoulos PT, Zhao C, Mazhar S. An Improved SAR Ship Classification Method Using Text-to-Image Generation-Based Data Augmentation and Squeeze and Excitation. Remote Sens. 2024;16(7):1299.
42. Orozco-Arias S, Piña JS, Tabares-Soto R, Castillo-Ossa LF, Guyot R, Isaza G. Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements. Processes. 2020;8(6):638.
43. Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics. 2021;10(5):593.
44. Pham BT, Jaafari A, Avand M, Al-Ansari N, Dinh Du T, Yen HPH, et al. Performance evaluation of machine learning methods for forest fire modeling and prediction. Symmetry. 2020;12(6):1022.