International Journal of Robotics and Control Systems
Vol. 6, No. 1, 2026, pp. 454-474
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

A Comparative Study of Respiratory Diseases Classification Using Grad-CAM-Based DenseNet Architectures

Yuri Pamungkas [a,1,*], Shoffi Izza Sabilla [a,2], Padma Nyoman Crisnapati [b,3], Gao Yulan [c,4], Yamin Thwe [d,5]

[a] Department of Medical Technology, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
[b] Department of Mechatronics Engineering, RMUTT, Khlong Luang, Pathum Thani 12110, Thailand
[c] Department of Mechanical Engineering, Guizhou University of Engineering Science, Qixingguan 551700, China
[d] Department of Big Data Management and Analytics, RMUTT, Khlong Luang, Pathum Thani 12110, Thailand
[1] yuri@its.ac.id; [2] shoffi@its.ac.id; [3] crisnapati@rmutt.ac.th; [4] gaoyulan1987@163.com; [5] yamin_t@mail.rmutt.ac.th
* Corresponding Author

ARTICLE INFO

Article history
Received August 25, 2025
Revised September 19, 2025
Accepted February 06, 2026

Keywords
Respiratory Disease Classification;
Chest X-ray (CXR) Imaging;
DenseNet Architectures;
Deep Learning in Medical Imaging;
Grad-CAM Interpretability

ABSTRACT

Respiratory diseases such as COVID-19, tuberculosis, and pneumonia remain major global health concerns, and CXR imaging plays a crucial role in their early detection and diagnosis. However, manual interpretation of chest radiographs is time-consuming and subject to variability among clinicians. Deep learning offers a promising solution to support automated diagnosis, although challenges remain regarding optimal model selection and interpretability. The contribution of this study is a comparative evaluation of DenseNet121, DenseNet169, and DenseNet201 architectures, combined with Grad-CAM to enhance transparency in decision-making. Two openly accessible collections of chest radiograph images were employed. Dataset-1 consisted of 6,432 images of Normal, Pneumonia, and COVID-19, while Dataset-2 included 15,421 images of Normal, Pneumonia, and Tuberculosis. Each model was trained for 50 epochs under four optimizers, namely Adam, Adamax, and SGD. Performance was assessed using metrics evaluation and Grad-CAM was utilized to depict the areas that significantly shaped the model's predictions. The results demonstrated that DenseNet169 consistently achieved the most balanced performance across datasets and optimizers. On Dataset-1 with Adam optimization, it reached an accuracy of 97.46%, precision of 96.05%, recall of 96.24%, F1-score of 96.10%, and specificity of 97.68%. On Dataset-2, it achieved 97.11% accuracy, 96.17% precision, 95.67% recall, 95.69% F1-score, and 97.83% specificity. These outcomes confirm that DenseNet169 is particularly well-suited for screening applications where sensitivity is critical. Grad-CAM depictions additionally confirmed that the model concentrated on diagnostically pertinent pulmonary regions, thereby strengthening clinical trust. In conclusion, DenseNet169 proved to be the most robust and reliable architecture for respiratory disease categorization, while Grad-CAM enhanced model interpretability. These results emphasize the promise of DenseNet-driven strategies as supportive instruments in medical image analysis and indicate opportunities for continued enhancement in clinical practice.

1. Introduction

Respiratory diseases, such as pneumonia, tuberculosis, and COVID-19, continue to be major contributors to global morbidity and mortality. Ensuring timely and accurate identification of these conditions is crucial to reducing the impact of disease and improving health results [1]–[3]. Traditional diagnostic approaches, including CXR and CT scans, depend greatly on the expertise of radiologists, which can be subjective, labor-intensive, and susceptible to variability between observers [4], [5]. These limitations are even more pronounced in resource-limited settings, where a shortage of radiologists often leads to delayed or inaccurate diagnoses [6]. Hence, there is a pressing demand for CAD systems that are automated, reliable, and interpretable to assist clinical decision-making in managing respiratory diseases [7]. In the past few years, deep learning methods, especially CNNs, have shown pronounced advances across medical image-analysis tasks, including classification, detection, and segmentation [8]–[10]. Among these, DenseNet architectures have emerged as powerful models due to their capacity for feature reuse, ability to mitigate vanishing gradient problems, and high accuracy with fewer parameters compared to other CNNs [11]. Previous research has highlighted the effectiveness of DenseNet for detecting pneumonia and COVID-19, though the majority of these studies have been restricted to single-architecture models or binary classification tasks [12]. Furthermore, many of these models function as "black boxes" with limited interpretability, which poses a significant barrier to clinical acceptance and trust [13]–[15].

To address these challenges, explainable AI approaches, including Grad-CAM, have been incorporated. Grad-CAM delivers visual explanations by emphasizing critical image areas that influence the model's outputs, thus improving clarity and comprehensibility [16]. Although prior studies have applied Grad-CAM for tasks such as COVID-19 or pneumonia detection individually [17], there is still a lack of comprehensive comparative evaluations that investigate various DenseNet architectures in combination with Grad-CAM across multiple respiratory diseases. Addressing this gap is essential to determine the most effective architecture while maintaining interpretability. The unique contribution of this research resides in its comparative analysis, which evaluates DenseNet121, DenseNet169, and DenseNet201 for classifying multiple respiratory diseases while simultaneously incorporating Grad-CAM to enhance interpretability. Unlike previous work that primarily emphasizes either predictive performance or interpretability, this research integrates both aspects by not only reporting classification results but also verifying whether the Grad-CAM highlighted regions correspond to clinically meaningful features. This dual focus strengthens both model optimization and its potential for clinical reliability.

The contribution of the research is twofold. First, it provides a comprehensive comparative evaluation of DenseNet architectures in classifying respiratory diseases, offering insights into their relative strengths and weaknesses. Second, it integrates Grad-CAM-based interpretability to enhance clinical trust and adoption of AI-based CAD systems. By demonstrating how different DenseNet variants perform and how their decisions can be interpreted, this research promotes the creation of trustworthy, explainable, and clinically relevant AI systems for diagnosing respiratory conditions. Ultimately, the findings aim to support radiologists and clinicians, particularly in resource-constrained environments, by improving diagnostic efficiency and decision-making.

2. Related Works

In recent years, the use of deep learning for diagnosing respiratory diseases through chest imaging has been widely investigated. Numerous studies have confirmed the efficacy of CNNs and transfer learning methods in distinguishing normal from pathological conditions. For instance, Elshennawy et al. [18] employed multiple architectures including ResNet152V2, CNN, MobileNetV2, and LSTM-CNN for pneumonia detection from CXR, achieving a remarkable accuracy of 99.22% with high recall and precision. Similarly, Jain et al. [19] evaluated custom CNNs along with established architectures such as ResNet50, VGG16, Inception-v3, and VGG19, reporting an accuracy of 92.31% and recall of 98% in binary classification tasks.

Beyond binary classification, researchers have expanded to multi-class approaches. Alsharif et al. [20] proposed a novel 50-layer CNN model (PneumoniaNet) to classify normal, bacterial, and viral pneumonia in pediatric CXR, attaining an impressive accuracy of 99.7%. Likewise, Nneji et al. [21] combined shallow CNNs with MobileNet-V3 and Inception-V3, incorporating image preprocessing techniques such as CECED and CLAHE. Their model classified normal, bacterial, viral, and COVID-19 cases with 98.3% accuracy and an AUC of 99.0%. Sharma et al. [22] also investigated both binary and multi-class classification using VGG16 with different classifiers, where the binary dataset achieved 92.15% accuracy, and the multi-class dataset yielded 95.4% accuracy with an AUC of 0.988.

Custom CNN modifications have also been investigated. Szepesi et al. [23] introduced dropout in convolutional layers to enhance generalization in pneumonia detection, reaching an accuracy of 97.21%. Mujahid et al. [24] integrated CNN, VGG16, ResNet50, Inception-V3, and ensemble methods, achieving a state-of-the-art performance of 99.29% accuracy with recall as high as 99.73%. Similarly, Bhatt et al. [25] used CNN ensembles, though their study highlighted certain performance limitations, reporting an accuracy of 84.12% but maintaining a high recall of 99.23%, indicating sensitivity to positive cases despite lower precision.

Apart from CNN-based methods, texture-based radiomic approaches combined with machine learning classifiers have been explored. Toro et al. [26] utilized fractal dimension, radiomics, and superpixel-based histon features, integrated with traditional classifiers like KNN, SVM, and RF. Their results varied across datasets, with the Josep-NIH dataset achieving 99% accuracy and an AUC of 0.995, highlighting the potential of hybrid approaches in medical imaging tasks. MobileNet has also shown promising results in lightweight models suitable for real-world deployment. Reshan et al. [27] applied MobileNet for binary pneumonia classification and achieved 94.23% accuracy with an AUC of 0.972, underscoring the balance between performance and computational efficiency. The following is a summary of the related works presented in Table 1.

Overall, prior studies demonstrate that CNN-based methods, particularly advanced architectures such as VGG, Inception, ResNet, and MobileNet, have achieved high accuracy in respiratory disease classification. However, the majority of works focus on performance metrics alone while neglecting interpretability. Although some approaches have integrated ensembles or hybrid features, systematic comparisons of different DenseNet architectures with interpretability methods like Grad-CAM are still limited. This research seeks to address that gap by providing a comparative evaluation of DenseNet121, DenseNet169, and DenseNet201, while incorporating Grad-CAM to enhance clinical transparency and trust.

3. Methodology

3.1. Dataset

Within this study, two open-access CXR datasets were employed to examine the performance of various DenseNet models in identifying respiratory diseases. These datasets were carefully chosen to reflect diverse diagnostic conditions and to facilitate both binary and multi-class classification experiments. Dataset-1 (Fig. 1(a)) includes radiographic chest images classified into three categories (COVID-19, Pneumonia, and Normal) with a total of 6,432 samples [28]. Moreover, the dataset was additionally split into training and testing subsets to guarantee consistent assessment. Specifically, the training subset comprised 460 COVID-19 cases, 1,266 Normal cases, and 3,418 Pneumonia cases, while the testing set consisted of 116 COVID-19 cases, 317 Normal cases, and 855 Pneumonia cases. This dataset enabled a comprehensive evaluation of model capability in differentiating viral and bacterial respiratory infections from healthy cases.

Dataset-2 (Fig. 1(b)) contains chest X-ray images from three different classes, such as Tuberculosis, Pneumonia, and Normal, totaling 15,421 images [29]. To provide balanced evaluation conditions, the training set consisted of 1,788 Tuberculosis images, 9,188 Normal images, and 4,145 Pneumonia images. The testing set was equally distributed, with 100 images for each category (100 Tuberculosis, 100 Normal, and 100 Pneumonia). This dataset facilitated the exploration of model

robustness across larger and more diverse samples, particularly for differentiating Tuberculosis from other respiratory diseases. Both datasets were preprocessed prior to training, including resizing images to match the input dimensions of DenseNet architectures and applying normalization to standardize pixel values. The use of two distinct datasets allowed this study to perform comparative analysis not only across architectures (DenseNet-121, DenseNet-169, and DenseNet-201) but also across different disease combinations, enhancing the generalizability and reliability of the findings.

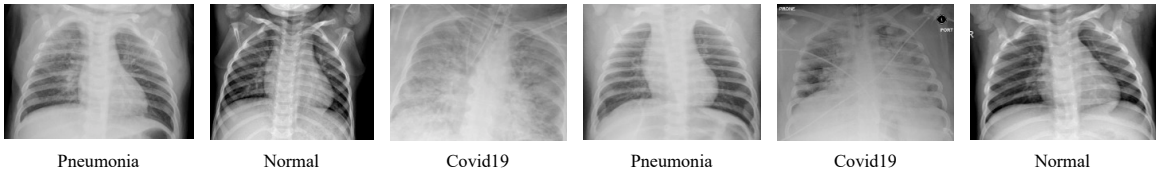
Table 1. Summary of the related works

Ref	Authors & Year	Type of Imaging	DL Techniques	Classification Task	Evaluation Results
[18]	Elshennawy, et al (2020)	CXR	ResNet152V2, MobileNetV2, CNN, LSTM-CNN	Binary (Pneumonia vs Normal)	AUC = 99.77%, Recall = 99.43%, Accuracy = 99.22%, F1-score = 99.44%, Precision = 99.44%.
[19]	Jain, et al (2020)	CXR	Custom CNN (2-layer, 3-layer), VGG16, VGG19, ResNet50, Inception-v3	Binary (Normal vs Pneumonia)	Accuracy = 92.31%, Recall = 98%, F1-Score = 94%
[20]	Alsharif, et al (2021)	CXR	PneumoniaNet (Novel 50-layer CNN)	Multi-class (Normal, Bacterial, Viral)	Specificity = 99.85%, AUC = 0.981, Recall = 99.74%, Accuracy = 99.7%, F1-score = 99.7%, Precision = 99.7%
[21]	Szepesi, et al (2022)	CXR	Custom CNN (with dropout in convolutional layers)	Binary (Pneumonia vs Normal)	Recall = 97.34%, AUC = 0.982, F1-score = 97.37%, Precision = 97.40%, Accuracy = 97.21%
[22]	Nneji, et al (2022)	CXR	Shallow CNN (LBP), MobileNet-V3 (CECED), Inception-V3 (CLAHE)	Multi-class (Normal, Bacterial, Viral, COVID-19)	AUC = 99.0%, Recall = 98.9%, Specificity = 99.2%, Accuracy = 98.3%, F1-score = 99.0%, Precision = 98.8%
[23]	Mujahid, et al (2022)	CXR	CNN, VGG-16, ResNet50, Inception-V3, Ensembles	Binary (Pneumonia vs Normal)	Accuracy = 99.29%, Precision = 98.83%, F1-score = 99.28%, Recall = 99.73%, AUC = 99.30
[24]	Toro, et al (2022)	CXR	Texture-based features (Radiomics, Fractal Dimension, Superpixel-based Histon) + ML classifiers (KNN, SVM, RF)	Binary (Pneumonia vs Normal)	Dataset 1 (GWCMCx - Paediatric): F1-Score = 0.93, Accuracy = 93%, AUC = 0.954. Dataset 2 (Josep-NIH): F1-Score = 0.98, Accuracy = 99%, AUC = 0.995
[25]	Bhatt, et al (2023)	CXR	CNN ensemble	Binary (Pneumonia vs Normal)	Recall = 99.23%, F1-score = 88.56%, Accuracy = 84.12%, Precision = 80.04%
[26]	Sharma, et al (2023)	CXR	VGG16 + Neural Networks (compared with VGG16 + SVM, KNN, RF, NB)	Binary (Pneumonia vs Normal) and Multi-class (Normal, Pneumonia, COVID-19)	Dataset 1 (Binary): F1-score = 0.937, Accuracy = 92.15%, AUC = 0.974, Recall = 0.9308, Precision = 0.9428. Dataset 2 (Multi-class): Recall = 0.954, AUC = 0.988, Precision = 0.954, Accuracy = 95.4%, F1-score = 0.954
[27]	Reshan, et al (2023)	CXR	MobileNet	Binary (Pneumonia vs Normal)	Precision = 93.75%, Accuracy = 94.23%, AUC = 0.972, F1-score = 95.96%, Recall = 98.28%

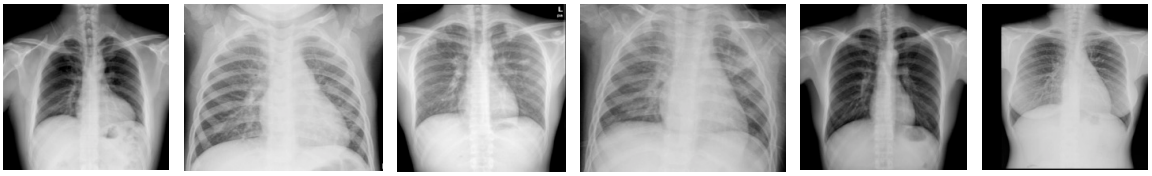
3.2. Preprocessing

Before training the DenseNet models, all CXR images underwent preprocessing to maintain consistency and enhance model generalization [30]. Each image was scaled to 224×224 pixels in RGB

format to align with DenseNet input requirements, and pixel intensities were normalized to the range [0,1] to stabilize the training phase [31]. To overcome the challenge of limited dataset size and minimize overfitting, extensive data augmentation was applied [32]. This included random rotations between -50° and +50°, as well as random zooming in or out within a scale factor of 0.5 to 1.5, enabling the models to better adapt to variations in orientation and scale [33]. Brightness adjustments were also randomly applied within the range [0.3, 1.0] to replicate diverse exposure conditions commonly encountered in radiographic imaging [34]. Additionally, horizontal and vertical flips were introduced to increase orientation variability [35], while random shifts within ±0.3 of the image dimensions along height and width were performed to help the models handle positional variations of anatomical features [36]. Altogether, these preprocessing techniques enriched the dataset with a wide range of transformations while preserving clinically meaningful features [37], thereby improving the robustness and generalizability of DenseNet architectures when used for respiratory disease classification tasks [38].



(a) CXR dataset-1



(b) CXR dataset-2

Fig. 1. CXR dataset for respiratory diseases classification

3.3. DenseNet Architectures

The fundamental principle of DenseNet is dense connectivity, in which every layer obtains information not just from the directly prior layer but also from all the preceding ones [39]. In other words, the feature maps generated at any layer are passed directly to every subsequent layer inside the same structural block [40]. Such a connectivity pattern encourages reutilization of features, enhances gradient propagation, alleviates the vanishing gradient issue, and substantially decreases the parameter count relative to traditional CNN frameworks [41]. In medical image analysis, such as CXR, this is highly advantageous because DenseNet allows the model to derive richer and more meaningful feature representations from comparatively limited datasets while maintaining efficiency in computation [42].

If the result produced by the l^{th} layer is represented as x_l, then rather than relying solely on the result from the directly prior layer, DenseNet formulates x_l as a function derived from the merging of all earlier feature maps. This relation can be formulated as:

$$x_l = H_l([x_0, x_1, x_2, \dots, x_{l-1}]) \quad (1)$$

Here, $x_0, x_1, \dots, x_{l-1}$ denote the feature maps generated by the earlier layers, and $[\cdot]$ signifies concatenation along the channel dimension. The function $H_l(\cdot)$ is the transformation applied at the l^{th} layer, which generally involves a series of processes including batch normalization (BN), ReLU activation, and convolution. This may be mathematically represented as:

$$H_l(x) = W_l * \left(\sigma\left(BN(x)\right)\right) \quad (2)$$

where BN(·) represents batch normalization, $\sigma(\cdot)$ is the non-linear activation function (ReLU), and $W_l * (\cdot)$ indicates convolution with weights W_l.

An important parameter in DenseNet is the growth rate, symbolized by k, which specifies how many additional feature maps are produced by every layer. If the input to a dense block has m_0 initial feature maps, then after l layers the overall count of feature maps becomes:

$$m_l = m_0 + k \times l \tag{3}$$

This linear growth ensures that each layer expands the feature space while still reusing information from previous layers. By design, this mitigates redundancy and promotes feature reuse, thereby improving efficiency compared to traditional CNNs that often learn overlapping features in different layers.

Since the concatenation of features causes the number of feature maps to increase rapidly, DenseNet introduces transition layers between dense blocks to regulate model complexity. A transition layer usually comprises a 1×1 convolution step followed by a 2×2 average pooling process, reducing both the spatial resolution and the channel count. To further control model size, a compression factor $\theta \in (0,1]$ is often applied, such that the number of feature maps after a transition layer is calculated as:

$$m_{\text{out}} = \lfloor \theta \cdot m_{\text{in}} \rfloor \tag{4}$$

where m_{in} denotes the number of input feature maps and θ is typically set to 0.5. This ensures that while the dense connectivity promotes expressive feature learning, the overall model remains computationally efficient.

Based on Table 2, all three models begin with the same initial configuration. The input image size is normalized to 224×224×3, followed by a stem comprising a 7×7 convolution with stride 2 and a 3×3 maximum pooling process. This stage serves as an initial feature extractor before the data is passed into the series of dense blocks. Each dense block consists of a defined set of layers, where each layer performs batch normalization, ReLU activation, a 1×1 convolution (bottleneck), and a 3×3 convolution, with the output concatenated with all previous feature maps within the block. The key differences among the three models lie in the depth of the dense blocks. DenseNet121 has dense blocks arranged with layer counts of 6, 12, 24, and 16, in that order. DenseNet169 increases the depth in later stages with distributions of 6, 12, 32, and 32 layers, while DenseNet201 additionally enlarges the third block to 48 layers, followed by 32 layers in the final block. Between each dense block, a transition layer made up of a 1×1 convolution operation combined with a 2×2 average pooling step with a compression factor (θ) of 0.5 is employed to control the growth of feature maps and reduce spatial resolution. This ensures that the increasing model depth remains computationally tractable.

After the dense blocks, all three variants utilize the same final layers, which include batch normalization (BN), ReLU nonlinearity, and a global average pooling (GAP) layer to combine the extracted features. The classification stage is realized via a dense (fully connected) layer and then a softmax function, generating probability outputs across the target categories. For training, the models were optimized using multiple optimizers, including Adam, Adamax, and SGD, while the categorical crossentropy loss function was applied to handle the multi-class classification task.

3.4. Grad-CAM

Grad-CAM is an explainable AI approach developed to provide visual interpretability for deep learning models, particularly CNNs [43]. The main goal of Grad-CAM is to emphasize the portions of the image that have the greatest influence on the model's classification result [44]. By overlaying heatmaps onto the original medical images, Grad-CAM helps clinicians and researchers identify the anatomical or pathological features that most strongly influenced the predictions [45]. This interpretability is particularly relevant within medical image processing tasks, including chest

radiograph evaluation, where trust, transparency, and clinical validation are essential for adoption [46].

Table 2. Architecture of three DenseNet models used in this study

Stage	DenseNet121	DenseNet169	DenseNet201
Input	224 × 224 × 3	224 × 224 × 3	224 × 224 × 3
Stem	7 × 7 Conv (stride 2), 3 × 3 MaxPool	7 × 7 Conv (stride 2), 3 × 3 MaxPool	7 × 7 Conv (stride 2), 3 × 3 MaxPool
Dense Block 1	6 layers	6 layers	6 layers
Transition Layer 1	Conv 1 × 1, AvgPool 2 × 2 (θ = 0.5)	Conv 1 × 1, AvgPool 2 × 2 (θ = 0.5)	Conv 1 × 1, AvgPool 2 × 2 (θ = 0.5)
Dense Block 2	12 layers	12 layers	12 layers
Transition Layer 2	Conv 1 × 1, AvgPool 2 × 2 (θ = 0.5)	Conv 1 × 1, AvgPool 2 × 2 (θ = 0.5)	Conv 1 × 1, AvgPool 2 × 2 (θ = 0.5)
Dense Block 3	24 layers	32 layers	48 layers
Transition Layer 3	Conv 1 × 1, AvgPool 2 × 2 (θ = 0.5)	Conv 1 × 1, AvgPool 2 × 2 (θ = 0.5)	Conv 1 × 1, AvgPool 2 × 2 (θ = 0.5)
Dense Block 4	16 layers	32 layers	32 layers
Final Layers	Batch Normalization, ReLU, Global AvgPool	Batch Normalization, ReLU, Global AvgPool	Batch Normalization, ReLU, Global AvgPool
Classifier	FC + Softmax	FC + Softmax	FC + Softmax
Training Configuration	Optimizers: “Adam, Adamax, SGD” Loss: “categorical crossentropy”	Optimizers: “Adam, Adamax, SGD” Loss: “categorical crossentropy”	Optimizers: “Adam, Adamax, SGD” Loss: “categorical crossentropy”

The Grad-CAM technique works by utilizing the gradient signals that propagate into the last convolutional layers of a CNN to evaluate the relevance of each feature map for a given class prediction [47]. Let y^c represent the output score for class c (prior to applying the softmax function), and let A^k denote the k^{th} feature map of a layer performing convolution operations. The importance coefficient α_k^c for the feature map A^k is subsequently derived as the overall mean of the gradients of y^c with respect to A^k.

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5}$$

where Z refers to the total pixel count in the feature map, and $\frac{\partial y^c}{\partial A_{ij}^k}$ represents the gradient of the class score relative to the activation at the spatial position (i, j). The localization map specific to a class $L_{\text{Grad-CAM}}^c$ is subsequently produced through the application of a weighted sum of the feature maps and subsequently passed through a ReLU function.

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{6}$$

The ReLU activation guarantees that solely the features that contribute in favor of the target class are shown, while uninformative or contradictory features are attenuated. Finally, the generated heatmap is upscaled to match the resolution of the original input image and superimposed to provide clearer visual interpretations.

In this study, Grad-CAM was applied to three DenseNet architectures (DenseNet121, DenseNet169, and DenseNet201) to provide interpretability in respiratory disease classification tasks. For each case in the dataset, heatmaps were generated to assess whether the model attended to clinically meaningful regions within the chest cavity, such as opacities, consolidations, or cavitations,

depending on the disease [48]. By integrating Grad-CAM into the analysis pipeline, it was possible to confirm that correct predictions were often supported by attention to relevant lung regions, whereas misclassifications were linked to attention on less informative or irrelevant areas [49]. This allowed not only validation of model decisions but also a deeper understanding of each architecture's behavior under different conditions [50]. The incorporation of Grad-CAM offers significant advantages in the context of medical imaging. It provides transparency that strengthens clinical trust by ensuring that diagnostic decisions are not based on spurious correlations [51]. It also facilitates error analysis and model improvement by revealing systematic biases or inconsistencies in the decision-making process [52]. Furthermore, Grad-CAM contributes to medical education by serving as a visual tool that illustrates how deep learning models interpret complex radiographic patterns, thereby bridging the gap between computational methods and clinical expertise [53].

4. Results and Discussion

This section presents the outcomes of the experiments from the comparative study of DenseNet models for classifying respiratory diseases. Two CXR datasets were utilized in this work. The first dataset (CXR Dataset-1) comprised 6,432 images classified into three groups: Pneumonia, COVID-19, and Normal. The second dataset (CXR Dataset-2) contained 15,421 images, also grouped into three categories, specifically Pneumonia, Tuberculosis, and Normal. To guarantee reliable training and testing, both datasets were separated into training and testing subsets [54]. Specifically, for CXR Dataset-1, 80% of the images were allocated for training and 20% for testing, while for CXR Dataset-2, 98% of the samples were employed for training and the rest 2% for testing.

The classification task was performed using three state-of-the-art DenseNet architectures (DenseNet121, DenseNet169, and DenseNet201). Each architecture was trained under identical conditions for 50 epochs, but with variations in the choice of optimizers, namely Adam, Adamax, and SGD. This variation is allowed for a comparative assessment not only of model depth and architectural complexity but also on how various optimization algorithms influence model convergence and classification accuracy [55]. The evaluation of the models was conducted by employing a broad range of performance indicators typically utilized in medical image classification [56]. The following are performance metrics from DenseNet architectures in respiratory diseases classification.

Fig. 2 shows the comparative performance of DenseNet121, DenseNet169, and DenseNet201 on CXR Dataset-1 (Pneumonia, COVID-19, and Normal) when trained with the Adam optimizer. All three models performed strongly across the evaluation metrics, though some differences were apparent. DenseNet169 achieved the highest accuracy at 97.46%, followed closely by DenseNet201 with 97.10%, while DenseNet121 reached 96.74%. Precision values were relatively stable, ranging from 95.56% for DenseNet201 to 96.05% for DenseNet169, with DenseNet121 scoring 95.74%. More variation was seen in recall: DenseNet169 again led with 96.24%, DenseNet201 followed with 95.33%, while DenseNet121 dropped to 92.16%. The F1-score reflected the same trend, with DenseNet169 on top at 96.10%, DenseNet201 at 95.38%, and DenseNet121 at 93.80%. Specificity remained consistently high, with DenseNet169 and DenseNet201 slightly outperforming DenseNet121 (97.68% and 97.31% compared with 96.34%).

Fig. 3 presents the results on CXR Dataset-2 (Pneumonia, Tuberculosis, and Normal), also optimized with Adam. Here, DenseNet169 once again delivered the best performance across almost all metrics. It achieved the highest accuracy (97.11%), with DenseNet121 following at 96.44% and DenseNet201 trailing at 94.44%. Precision results showed a similar pattern, with DenseNet169 at 96.17%, DenseNet121 at 95.40%, and DenseNet201 at 93.16%. In recall, DenseNet169 maintained its advantage at 95.67%, compared with 94.67% for DenseNet121 and 91.67% for DenseNet201. The F1-score echoed these results: 95.69% for DenseNet169, 94.69% for DenseNet121, and 91.68% for DenseNet201. Specificity was strong for all three models, with DenseNet169 slightly ahead at 97.83%, followed by DenseNet121 at 97.33%, and DenseNet201 at 95.83%.

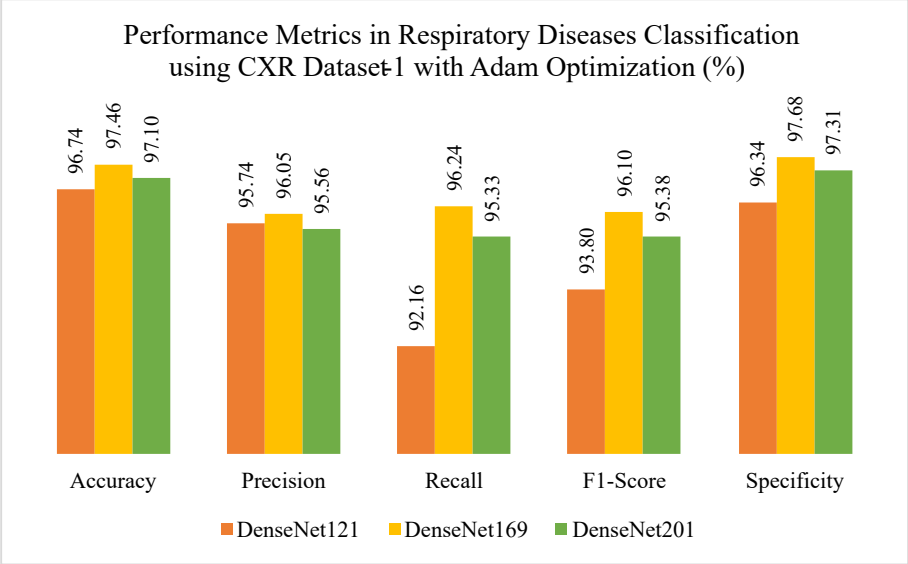


Fig. 2. Respiratory diseases classification using CXR dataset-1 with Adam optimization

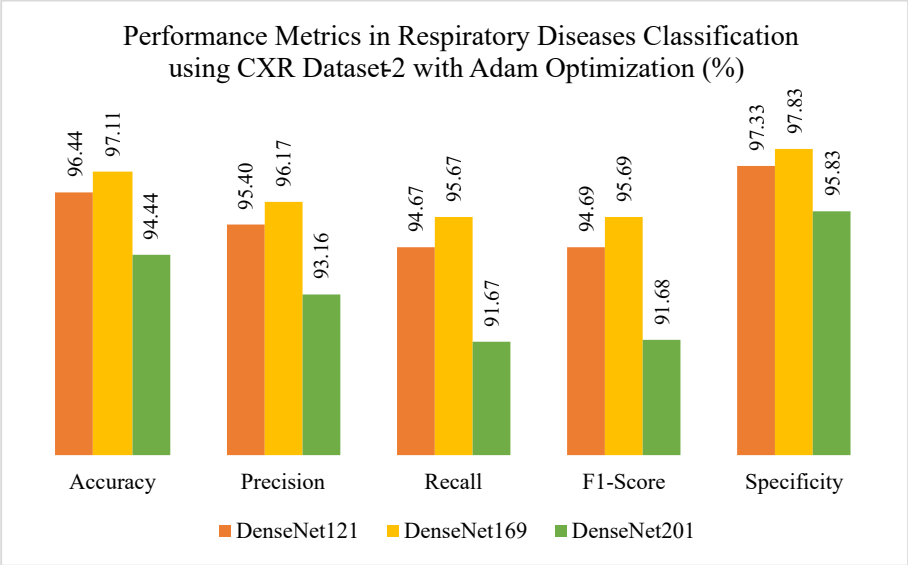


Fig. 3. Respiratory diseases classification using CXR dataset-2 with Adam optimization

When comparing the two datasets, several patterns emerge. On Dataset-1, all three models performed at a high and relatively comparable level, with DenseNet169 standing out as the most balanced architecture, particularly in recall and F1-score. DenseNet201 also achieved strong accuracy and specificity, but its recall was slightly lower, while DenseNet121, though still accurate, showed clear limitations in sensitivity. On Dataset-2, the differences became more pronounced. DenseNet169 consistently outperformed the other two models, not only in accuracy but also in recall, F1-score, and specificity. DenseNet121 remained competitive, especially in specificity, whereas DenseNet201 underperformed compared with its shallower counterparts, with notable drops in recall and F1-score. Taken together, these findings suggest that DenseNet169 provides the most balanced and reliable performance across datasets, while DenseNet121 demonstrates strength in specificity. DenseNet201, despite being the deepest model, does not necessarily translate its complexity into superior results, underscoring the importance of aligning architectural depth with dataset characteristics and optimization strategies [57].

Fig. 4 presents the classification performance of DenseNet121, DenseNet169, and DenseNet201 on CXR Dataset-1 (Pneumonia, COVID-19, and Normal) using the Adamax optimizer. Overall, all

three models achieved strong results across the evaluation metrics, though DenseNet169 and DenseNet201 showed a more balanced performance compared with DenseNet121. In terms of accuracy, DenseNet169 obtained the highest score at 97.46%, closely followed by DenseNet201 at 97.20%, while DenseNet121 reached 96.64%. Precision values were also favorable, ranging from 95.04% for DenseNet121 to 96.75% for DenseNet169, with DenseNet201 at 95.36%. A clearer distinction appeared in recall, where DenseNet201 achieved the best result at 96.29%, followed by DenseNet169 at 94.95%, while DenseNet121 lagged at 93.36%. The F1-score reflected a similar pattern: DenseNet169 scored 95.82% and DenseNet201 95.77%, both outperforming DenseNet121 at 94.07%. Specificity was high across all three models, with DenseNet201 slightly ahead at 97.55%, followed by DenseNet169 at 97.08% and DenseNet121 at 96.74%.

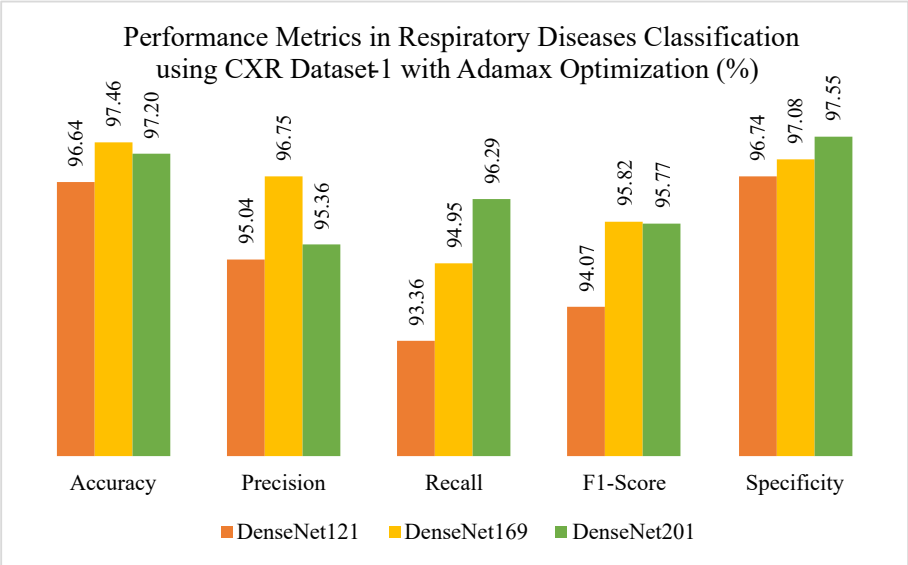


Fig. 4. Respiratory diseases classification using CXR dataset-1 with Adamax optimization

Fig. 5 shows the comparative performance on CXR Dataset-2 (Pneumonia, Tuberculosis, and Normal) under Adamax optimization. In this case, DenseNet201 emerged as the strongest model across most metrics. It achieved the highest accuracy at 96.67%, narrowly surpassing DenseNet169 at 96.22% and DenseNet121 at 95.56%. Precision followed the same trend, with DenseNet201 leading at 95.65%, compared to 95.16% for DenseNet169 and 94.33% for DenseNet121. For recall, DenseNet201 again ranked first with 95.00%, while DenseNet169 reached 94.33% and DenseNet121 trailed at 93.33%. The F1-score confirmed DenseNet201's balanced performance, achieving 95.04%, followed by DenseNet169 at 94.39% and DenseNet121 at 93.36%. Specificity remained consistently high, with DenseNet201 at 97.50%, DenseNet169 at 97.17%, and DenseNet121 at 96.67%.

When comparing the two datasets, some interesting patterns emerge. On Dataset-1, DenseNet169 stood out with the highest accuracy (97.46%) and precision (96.75%), while DenseNet201 excelled in recall (96.29%) and specificity (97.55%). DenseNet121, although competitive, consistently showed lower recall (93.36%) and F1-score (94.07%), indicating a greater risk of missing positive cases. On Dataset-2, however, DenseNet201 clearly established itself as the best overall performer, achieving the highest accuracy, precision, recall, F1-score, and specificity. DenseNet169 remained strong and stable, while DenseNet121 again trailed behind, particularly in recall (93.33%) and F1-score (93.36%). Taken together, these findings suggest that Adamax optimization tends to favor deeper architectures when applied to larger and more diverse datasets such as Dataset-2, allowing DenseNet201 to fully leverage its representational capacity. On smaller datasets such as Dataset-1, however, DenseNet169 provided a more stable and balanced performance, striking an effective compromise between depth and generalization. This comparison highlights the importance of matching model complexity with dataset size and variability, and underscores that while Adamax

improves robustness across DenseNet architectures, the relative advantage of each variant depends on the nature of the dataset used [58].

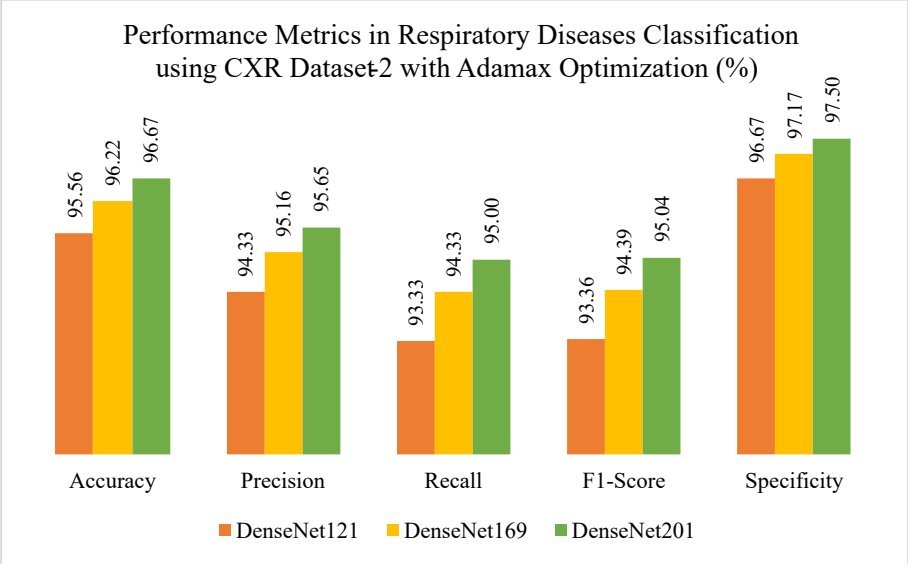


Fig. 5. Respiratory diseases classification using CXR dataset-2 with Adamax optimization

Fig. 6 presents the performance of DenseNet121, DenseNet169, and DenseNet201 on CXR Dataset-1 (Pneumonia, COVID-19, and Normal) trained with the Stochastic Gradient Descent (SGD) optimizer. All three models performed at a high level, though DenseNet169 demonstrated the most consistent balance across metrics. It achieved the highest accuracy at 97.20%, with DenseNet201 close behind at 97.00% and DenseNet121 at 96.64%. Precision remained stable across the models, ranging from 94.06% for DenseNet121 to 95.99% for DenseNet169 and 95.82% for DenseNet201. A clearer gap was seen in recall, where DenseNet169 led with 95.26%, followed by DenseNet201 at 93.78% and DenseNet121 at 93.03%. The F1-score reflected the same trend, with DenseNet169 achieving 95.61%, compared with 94.76% for DenseNet201 and 93.90% for DenseNet121. Specificity was high across all three models, exceeding 96%, with DenseNet169 again slightly ahead at 97.12%.

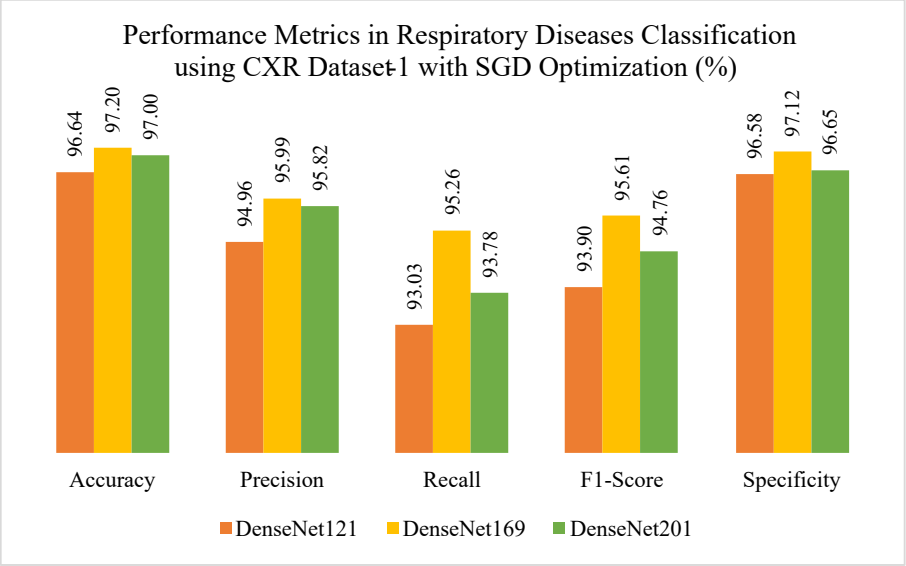


Fig. 6. Respiratory diseases classification using CXR dataset-1 with SGD optimization

Fig. 7 shows the results on CXR Dataset-2 (Pneumonia, Tuberculosis, and Normal) using the same SGD optimization. Here, DenseNet169 once again emerged as the strongest performer. It

reached the highest accuracy at 96.67%, followed by DenseNet121 at 96.22% and DenseNet201 at 96.00%. Precision was also best for DenseNet169 at 95.65%, compared with 94.89% for DenseNet121 and 94.57% for DenseNet201. The recall scores followed the same order, with DenseNet169 at 95.00%, DenseNet121 at 94.33%, and DenseNet201 at 94.00%. The F1-scores mirrored this distribution, as DenseNet169 achieved 95.04%, outperforming DenseNet121 at 94.33% and DenseNet201 at 94.04%. Specificity values were uniformly strong, with DenseNet201 at the top with 97.50%, while both DenseNet169 and DenseNet121 recorded 97.17%.

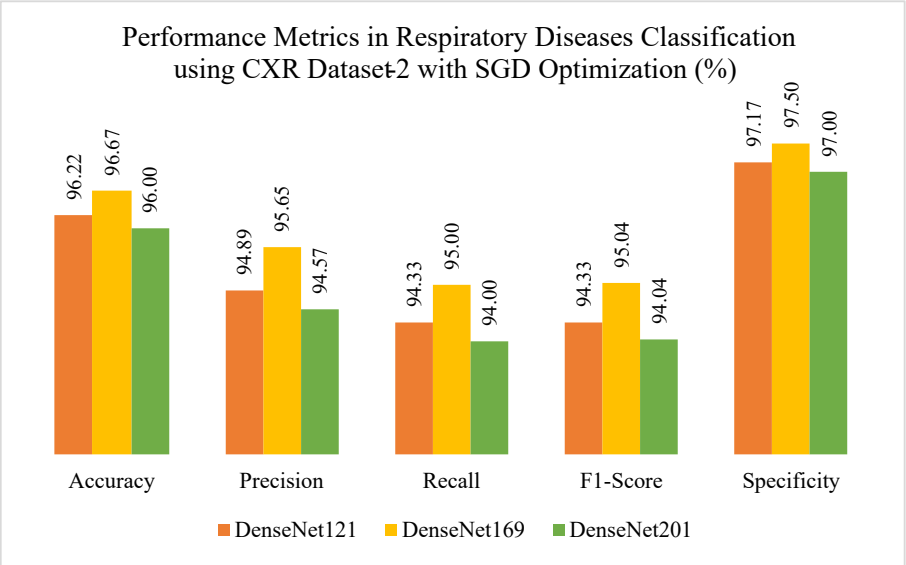


Fig. 7. Respiratory diseases classification using CXR dataset-2 with SGD optimization

When the results of Fig. 6 and Fig. 7 are compared, it becomes clear that SGD allowed all three DenseNet variants to perform robustly across both datasets, though the influence of dataset characteristics shaped the relative outcomes. On Dataset-1, DenseNet169 offered the best balance, delivering the highest accuracy, recall, and F1-score. DenseNet201 was competitive in accuracy but fell short in recall, while DenseNet121, though accurate, had the weakest sensitivity, suggesting a higher chance of missing positive cases. On Dataset-2, the performance across the models was more evenly distributed, but DenseNet169 again led most metrics, particularly accuracy, precision, and recall. DenseNet121 also proved reliable, showing strong accuracy and the highest specificity, while DenseNet201, despite matching this specificity, recorded slightly lower recall and F1-scores. Taken together, these findings suggest that DenseNet169 consistently provides the most stable and balanced outcomes under SGD optimization, excelling especially in recall and F1-score (two metrics that are highly relevant in clinical screening), where minimizing false negatives is critical. DenseNet121 demonstrated strength in specificity, making it useful in confirmatory contexts where reducing false positives is a priority [59]. DenseNet201, while strong in specificity, showed relatively weaker sensitivity, underscoring the point that deeper networks do not always guarantee superior performance without careful tuning [60]. Overall, SGD proved effective in supporting generalization across DenseNet variants, with DenseNet169 standing out as the most reliable architecture for harmonizing sensitivity, specificity, and overall classification performance.

The comparative analysis of DenseNet121, DenseNet169, and DenseNet201 (Fig. 2 to Fig. 7) reveals important insights into how model depth, optimizer choice, and dataset characteristics interact to influence classification performance. A consistent trend observed throughout the experiments is that DenseNet169 often provided the most balanced performance across accuracy, precision, recall, F1-score, and specificity. This suggests that its intermediate depth offers an optimal trade-off between representational capacity and generalization ability, allowing it to capture complex patterns without overfitting to the training data [61]. In contrast, DenseNet201, despite its greater depth and theoretical representational power, did not consistently outperform the shallower variants [62]. Its weaker recall

and F1-scores, particularly on Dataset-2 with Adam and SGD, indicate a tendency toward reduced sensitivity, possibly due to overfitting or difficulty in optimization when working with finite and imbalanced medical datasets. DenseNet121, while delivering competitive accuracy and particularly strong specificity, repeatedly underperformed in recall, highlighting its vulnerability to missing positive cases. Another critical factor emerging from the results is the influence of dataset size and variability. On Dataset-1 (smaller, 6,432 images), DenseNet169 consistently excelled, while DenseNet201 performed competitively but not decisively better. This pattern shifted on Dataset-2 (larger, 15,421 images), where Adamax optimization allowed DenseNet201 to fully leverage its depth and achieve superior results across nearly all metrics. This finding underscores the importance of aligning architectural complexity with dataset scale: deeper models may require larger and more diverse datasets to demonstrate their true potential, whereas mid-depth models such as DenseNet169 remain more robust across smaller datasets. The contrast between datasets highlights a practical challenge in medical imaging research, where data availability often varies widely depending on disease type and patient population [63].

The role of optimization methods also deserves attention. Adam generally produced stable and strong results across both datasets, favoring DenseNet169 in particular. Adamax, on the other hand, seemed to enhance the performance of deeper networks, allowing DenseNet201 to outperform in Dataset-2, where the model could take advantage of the optimizer's stability with sparse gradients. In comparison, SGD delivered solid generalization across all three models but showed a consistent advantage for DenseNet169, reinforcing its reliability across datasets. These findings demonstrate that optimizer choice is not merely a technical detail but a critical determinant of how well a given architecture can adapt to specific dataset conditions. From a clinical perspective, the trade-offs observed between recall and specificity have direct implications for deployment [64]. Models such as DenseNet169, which consistently achieved high recall, are particularly valuable in screening scenarios where minimizing false negatives is essential, for example in detecting early-stage pneumonia or tuberculosis. DenseNet121, with its high specificity, could serve better in confirmatory diagnostic workflows, reducing the burden of false positives on clinicians. DenseNet201, while showing promise in larger datasets, requires careful tuning to avoid underperforming in sensitivity, a crucial limitation if applied in real-world clinical environments where missed cases could have serious consequences.

To complement the quantitative performance metrics, several representative examples of classification results are presented from both datasets (Fig. 8 and Fig. 9). These visual outputs provide an intuitive understanding of how the DenseNet models operate in practice, highlighting their ability to correctly identify different respiratory diseases across varied clinical scenarios. Each chest X-ray image is displayed with its actual label, the predicted class, and the probability score that reflects the model's level of confidence in its decision.

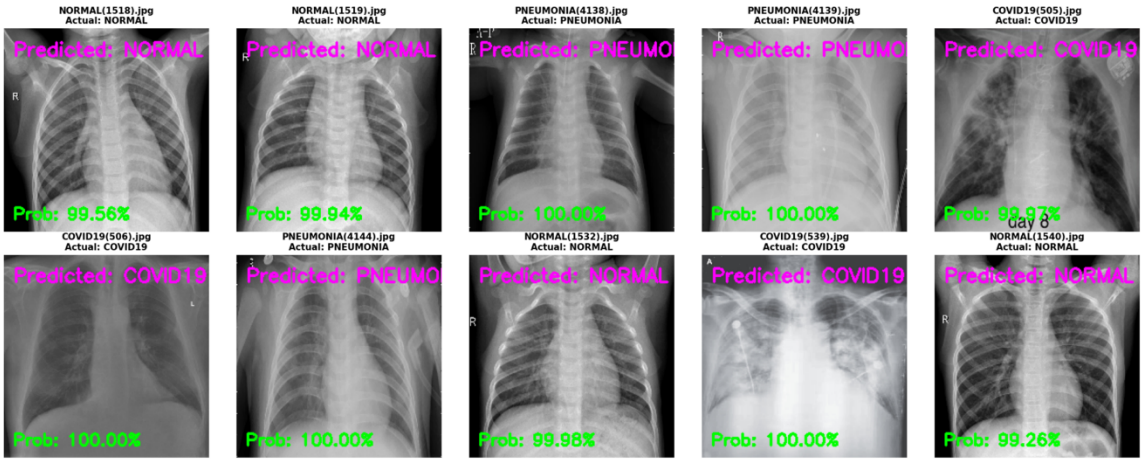


Fig. 8. Example of classification results on dataset-1

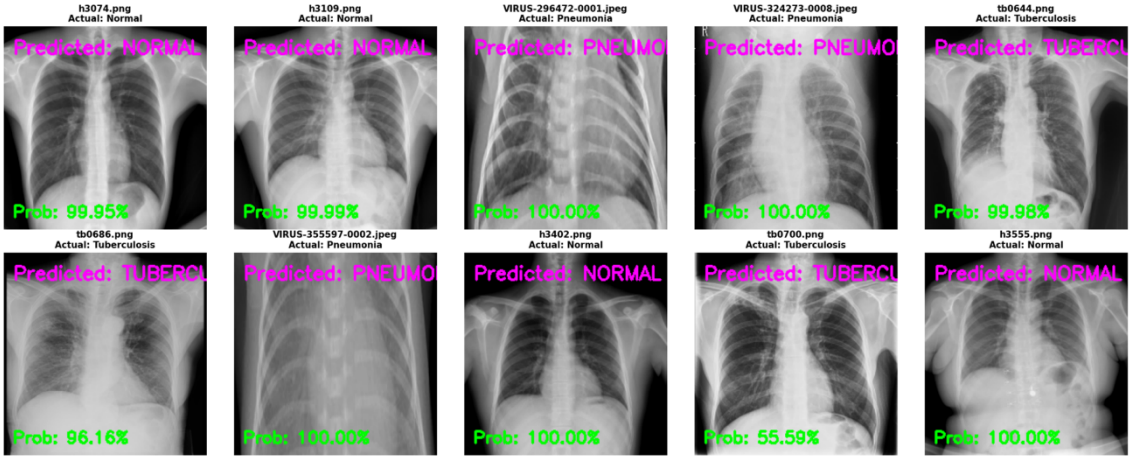


Fig. 9. Example of classification results on dataset-2

From Dataset-1 (COVID-19, Pneumonia, and Normal), the models consistently produced accurate predictions with very high confidence levels, often exceeding 99%, as illustrated in Fig. 8. These examples demonstrate the robustness of the models in differentiating between healthy lungs and pathological findings associated with pneumonia or COVID-19. Similarly, for Dataset-2 (Tuberculosis, Pneumonia, and Normal), the models also performed strongly, with most predictions reaching near-perfect confidence (Fig. 9). However, some tuberculosis cases showed lower confidence scores compared to pneumonia or normal cases, reflecting the greater variability and complexity of tuberculosis manifestations on chest X-rays.

To move beyond numerical metrics and predicted probabilities, Grad-CAM was utilized to deliver visual explanations of the DenseNet models (Fig. 10, Fig. 11, and Fig. 12). This approach emphasizes the areas within chest radiographs that exerted the greatest impact on the model's classification outcome, providing critical understanding of whether the predictions rely on patterns of true clinical significance [65]. By comparing the original image, the Grad-CAM heatmap, and the Grad-CAM overlay, it is possible to assess both the dependability of the model and its applicability for integration into clinical practice [66].

The Grad-CAM outputs clearly demonstrate that the models focus on pathologically relevant regions of the lungs when predicting disease categories. In pneumonia and COVID-19 cases, the heatmaps concentrated on areas of increased opacity and patchy infiltrates, which are consistent with known radiological features (Fig. 10 and Fig. 11). Similarly, in tuberculosis cases, the highlighted regions corresponded to localized consolidations or cavitary changes (Fig. 12). Importantly, in normal cases, the activation was minimal and diffuse, suggesting that the model accurately recognized the lack of pathological features without focusing on non-essential anatomical structures such as the ribs or diaphragm. These visualizations not only confirm the reasoning pathway of the model but also increase confidence by demonstrating consistency with clinical standards [67].

From a critical perspective, the Grad-CAM visualizations highlight both strengths and limitations. On one hand, the ability of the models to localize abnormalities aligns well with clinical reasoning, reinforcing their potential role as decision-support tools [68]. This is particularly valuable in screening workflows, where interpretability helps clinicians verify automated outputs [69]. On the other hand, some heatmaps showed relatively broad or diffuse activations that extended beyond the lung fields [70]. Such cases raise concerns about whether the model occasionally relies on spurious features or image artifacts [71]. This underscores the importance of coupling high performance with interpretability to prevent overreliance on "black-box" predictions [72].

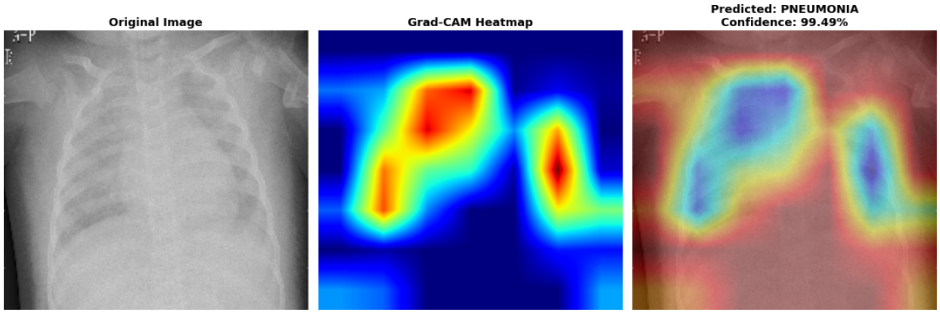


Fig. 10. Example of Grad-CAM heatmap and overlay of Pneumonia CXR

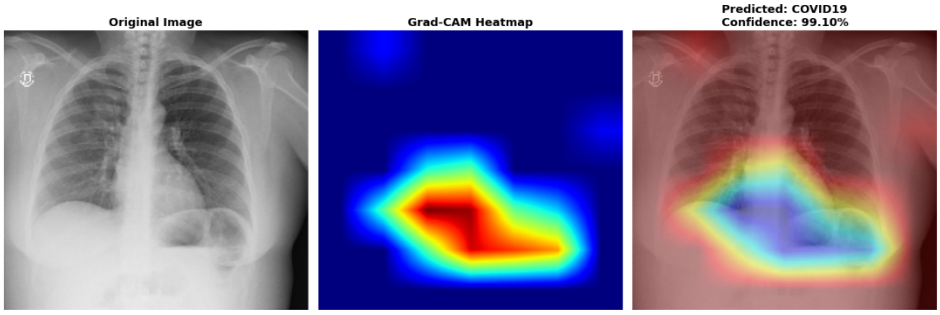


Fig. 11. Example of Grad-CAM heatmap and overlay of Covid19 CXR

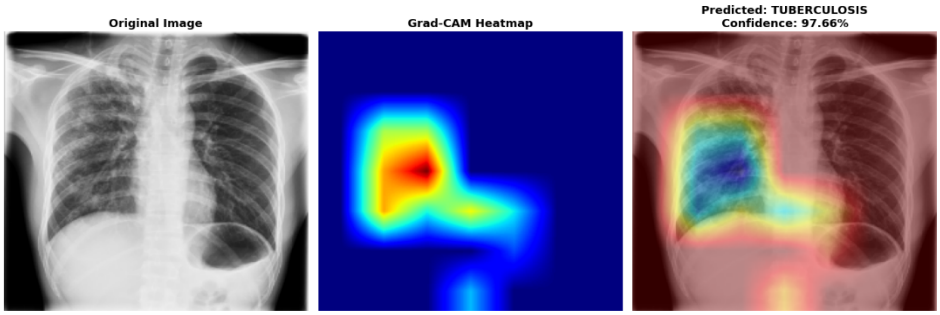


Fig. 12. Example of Grad-CAM heatmap and overlay of Tuberculosis CXR

5. Conclusion

This study conducted a comparative evaluation of three DenseNet architectures, namely DenseNet121, DenseNet169, and DenseNet201, for the purpose of classifying respiratory diseases through chest radiographic images. Two datasets were analyzed in order to reflect different diagnostic contexts. The initial dataset was composed of Normal, Pneumonia, and COVID-19 cases, while the subsequent dataset included Normal, Pneumonia, and Tuberculosis cases. The models were trained and tested under several optimizers, including Adam, Adamax, and SGD. Their performance was assessed using conventional evaluation measures, including accuracy, precision, recall, F1-score, and specificity. To complement these quantitative outcomes, Grad-CAM visualization was employed to enhance interpretability and to reveal the image regions that most significantly influenced the model's predictions. Across the experiments, DenseNet169 consistently demonstrated the most balanced and reliable performance. This model achieved superior recall and F1-score, which makes it a strong candidate for screening applications where sensitivity is of primary importance. DenseNet121 achieved high and stable specificity, which suggests its usefulness in confirmatory diagnostic settings where reducing false positives is critical. DenseNet201, although deeper in architecture, did not consistently outperform the other two models. Its advantage became more evident when applied to the

larger second dataset under Adamax optimization, which indicates that deeper networks require greater data diversity and larger sample sizes to fully realize their potential.

The application of Grad-CAM offered meaningful understanding of how the models carried out their decision-making process. The visual outputs indicated that the networks predominantly concentrated on lung regions of clinical significance when producing predictions. This enhanced confidence in the models' reliability and highlighted the value of interpretability tools as a bridge between algorithmic performance and clinical trust. However, there were also instances where the heatmaps appeared diffuse or extended beyond the lung fields. Such cases suggest that further refinement is still required in order to achieve consistently precise localization across all conditions. In conclusion, DenseNet169 emerged as the most robust and consistent architecture across both datasets and optimization strategies. DenseNet121 and DenseNet201 demonstrated distinct strengths in specificity and deeper feature representation, which highlights the importance of aligning model selection with both dataset characteristics and clinical priorities. The combination of strong classification performance and explainability through Grad-CAM underscores the potential of DenseNet-based models as decision-support tools in radiology. Future work should focus on expanding dataset diversity, improving model generalization for diseases with heterogeneous presentations such as tuberculosis, and refining interpretability techniques to ensure reliability and acceptance in clinical practice.

Author Contribution: All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding: This research was funded by the Research Department of the Institut Teknologi Sepuluh Nopember (ITS) 2025, grant number 2366/PKS/ITS/2025.

Acknowledgment: The authors would like to acknowledge the Department of Medical Technology, Institut Teknologi Sepuluh Nopember, for the facilities and support in this research. The authors also gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2025.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] R. Cioboata, V. Biciusca, M. Olteanu, and C. M. Vasile, "COVID-19 and tuberculosis: Unveiling the dual threat and shared solutions perspective," *Journal of Clinical Medicine*, vol. 12, no. 14, p. 4784, Jan. 2023, https://doi.org/10.3390/jcm12144784.

[2] N. A. Téllez-Navarrete, J. Romero-Tendilla, A. Morales, E. Becerril, N. Alvarado-Peña, M. A. Salazar-Lezama, P. Garciadiego-Fossas, E. Cadena-Torres, L. Chavez-Galan, and L. A. Ramón-Luing, "Assessment of the impact of COVID-19 on tuberculosis care at a tertiary hospital: Integrating lessons from COVID-19 learned," *Frontiers in Public Health*, vol. 13, p. 1505914, 2025, https://doi.org/10.3389/fpubh.2025.1505914.

[3] L. Parolina, N. Pshenichnaya, I. Vasilyeva, I. Lizinfed, N. Urushadze, V. Guseva, O. Otpushchennikova, O. Dyachenko, and P. Kharitonov, "Clinical characteristics of COVID-19 in patients with tuberculosis and factors associated with disease severity," *International Journal of Infectious Diseases*, vol. 124, pp. S82–S89, Apr. 2022, https://doi.org/10.1016/j.ijid.2022.04.041.

[4] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest X-ray analysis: A survey," *Medical Image Analysis*, vol. 72, p. 102125, Aug. 2021, https://doi.org/10.1016/j.media.2021.102125.

[5] R. Najjar, "Redefining radiology: A review of artificial intelligence integration in medical imaging," *Diagnostics*, vol. 13, no. 17, p. 2760, Aug. 2023, https://doi.org/10.3390/diagnostics13172760.

[6] Y. Liao, H. Liu, and I. Spasić, "Deep learning approaches to automatic radiology report generation: A systematic review," *Informatics in Medicine Unlocked*, vol. 39, p. 101273, Jan. 2023, https://doi.org/10.1016/j.imu.2023.101273.

[7] Q.-M. Liao, W. Hussain, Z.-X. Liao, S. Hussain, Z.-L. Jiang, Y.-H. Zhu, H.-Y. Luo, X.-Y. Ji, H.-W. Wen, and D.-D. Wu, "Computer-aided application in medicine and biomedicine," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, p. 221, Aug. 2025, https://doi.org/10.1007/s44196-025-00936-y.

[8] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Frontiers in Public Health*, vol. 11, p. 1273253, 2023, https://doi.org/10.3389/fpubh.2023.1273253.

[9] P. K. Mall, P. K. Singh, S. Srivastav, V. Narayan, M. Paprzycki, T. Jaworska, and M. Ganzha, "A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities," *Healthcare Analytics*, vol. 4, p. 100216, Dec. 2023, https://doi.org/10.1016/j.health.2023.100216.

[10] M. E. Rayed, S. M. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. F. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, vol. 47, p. 101504, Apr. 2024, https://doi.org/10.1016/j.imu.2024.101504.

[11] A. W. Saleh, G. Gupta, S. B. Khan, N. A. Alkhaldi, and A. Verma, "An Alzheimer's disease classification model using transfer learning DenseNet with embedded healthcare decision support system," *Decision Analytics Journal*, vol. 9, p. 100348, Dec. 2023, https://doi.org/10.1016/j.dajour.2023.100348.

[12] H. A. Sanghvi, R. H. Patel, A. Agarwal, S. Gupta, V. Sawhney, and A. S. Pandya, "A deep learning approach for classification of COVID and pneumonia using DenseNet-201," *International Journal of Imaging Systems and Technology*, vol. 33, no. 1, pp. 18–38, 2023, https://doi.org/10.1002/ima.22812.

[13] C. Vieira, L. Rocha, M. Guimarães, and D. Dias, "Exploring transparency: A comparative analysis of explainable artificial intelligence techniques in retinography images to support the diagnosis of glaucoma," *Computers in Biology and Medicine*, vol. 185, p. 109556, Dec. 2024, https://doi.org/10.1016/j.compbiomed.2024.109556.

[14] S. U. Hassan, S. J. Abdulkadir, M. S. M. Zahid, and S. M. Al-Selwi, "Local interpretable model-agnostic explanation approach for medical imaging analysis: A systematic literature review," *Computers in Biology and Medicine*, vol. 185, p. 109569, Dec. 2024, https://doi.org/10.1016/j.compbiomed.2024.109569.

[15] Q. Xu, W. Xie, B. Liao, C. Hu, L. Qin, Z. Yang, H. Xiong, Y. Lyu, Y. Zhou, and A. Luo, "Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review," *Journal of Healthcare Engineering*, vol. 2023, no. 1, p. 9919269, Feb. 2023, https://doi.org/10.1155/2023/9919269.

[16] M. Ennab and H. Mcheick, "Advancing AI interpretability in medical imaging: A comparative analysis of pixel-level interpretability and Grad-CAM models," *Machine Learning and Knowledge Extraction*, vol. 7, no. 1, p. 12, Feb. 2025, https://doi.org/10.3390/make7010012.

[17] A. Shah and M. Shah, "Advancement of deep learning in pneumonia/COVID-19 classification and localization: A systematic review with qualitative and quantitative analysis," *Chronic Diseases and Translational Medicine*, vol. 8, no. 3, pp. 154–171, 2022, https://doi.org/10.1002/cdt3.17.

[18] N. M. Elshennawy and D. M. Ibrahim, "Deep-pneumonia framework using deep learning models based on chest X-ray images," *Diagnostics*, vol. 10, no. 9, p. 649, Aug. 2020, https://doi.org/10.3390/diagnostics10090649.

[19] R. Jain, P. Nagrath, G. Kataria, V. S. Kaushik, and D. J. Hemanth, "Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning," *Measurement*, vol. 165, p. 108046, Dec. 2020, https://doi.org/10.1016/j.measurement.2020.108046.

[20] R. Alsharif, Y. Al-Issa, A. M. Alqudah, I. A. Qasmieh, W. A. Mustafa, and H. Alquran, "PneumoniaNet: Automated detection and classification of pediatric pneumonia using chest X-ray images and CNN approach," *Electronics*, vol. 10, no. 23, p. 2949, Nov. 2021, https://doi.org/10.3390/electronics10232949.

[21] P. Szepesi and L. Szilágyi, "Detection of pneumonia using convolutional neural networks and deep learning," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 3, pp. 1012–1022, 2022, https://doi.org/10.1016/j.bbe.2022.08.001.

[22] G. U. Nneji, J. Cai, J. Deng, H. N. Monday, E. C. James, and C. C. Ukwuoma, "Multi-channel based image processing scheme for pneumonia identification," *Diagnostics*, vol. 12, no. 2, p. 325, Jan. 2022, https://doi.org/10.3390/diagnostics12020325.

[23] M. Mujahid, F. Rustam, R. Álvarez, J. L. V. Mazón, I. de la T. Díez, and I. Ashraf, "Pneumonia classification from X-ray images with Inception-V3 and convolutional neural network," *Diagnostics*, vol. 12, no. 5, p. 1280, May 2022, https://doi.org/10.3390/diagnostics12051280.

[24] C. O.-Toro, A. G.-Pedrero, M. L.-Saavedra, and C. G.-Martín, "Automatic detection of pneumonia in chest X-ray images using textural features," *Computers in Biology and Medicine*, vol. 145, p. 105466, 2022, https://doi.org/10.1016/j.compbiomed.2022.105466.

[25] H. Bhatt and M. Shah, "A convolutional neural network ensemble model for pneumonia detection using chest X-ray images," *Healthcare Analytics*, vol. 3, p. 100176, Nov. 2023, https://doi.org/10.1016/j.health.2023.100176.

[26] S. Sharma and K. Guleria, "A deep learning-based model for the detection of pneumonia from chest X-ray images using VGG-16 and neural networks," *Procedia Computer Science*, vol. 218, pp. 357–366, 2023, https://doi.org/10.1016/j.procs.2023.01.018.

[27] M. S. A. Reshan, K. S. Gill, V. Anand, S. Gupta, H. Alshahrani, A. Sulaiman, and A. Shaikh, "Detection of pneumonia from chest X-ray images utilizing MobileNet model," *Healthcare*, vol. 11, no. 11, p. 1561, May 2023, https://doi.org/10.3390/healthcare11111561.

[28] P. Patel, "Chest X-ray (COVID-19 and pneumonia)," *Kaggle*, 2019, https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia.

[29] Roshan, "Imbalanced tuberculosis and pneumonia dataset," *Kaggle*, 2022. https://www.kaggle.com/datasets/roshanmaur/imbalanced-tuberculosis-and-pnuemonia-dataset.

[30] J. Colin and N. Surantha, "Interpretable deep learning for pneumonia detection using chest X-ray images," *Information*, vol. 16, no. 1, p. 53, Jan. 2025, https://doi.org/10.3390/info16010053.

[31] L. A. Abraham, G. Palanisamy, and G. Veerapu, "Transparent brain tumor detection using DenseNet169 and LIME," *Scientific Reports*, vol. 15, no. 1, p. 28185, Aug. 2025, https://doi.org/10.1038/s41598-025-13233-7.

[32] M. A. Widneh, A. A. Alemu, and D. D. Getie, "Exploring batch normalization's impact on dense layers of multiclass and multilabel classifiers," *International Journal of Intelligent Systems*, vol. 2025, no. 1, p. 1466655, 2025, https://doi.org/10.1155/int/1466655.

[33] Y. Zakaria, S. A. Mokhtar, H. Baraka, and M. Hadhoud, "Improving small and cluttered object detection by incorporating instance-level denoising into single-shot alignment network for remote sensing imagery," *IEEE Access*, vol. 10, pp. 51176–51190, 2022, https://doi.org/10.1109/ACCESS.2022.3174087.

[34] W. Ching, J. Robinson, and M. McEntee, "Patient-based radiographic exposure factor selection: A systematic review," *Journal of Medical Radiation Sciences*, vol. 61, no. 3, pp. 176–190, Aug. 2014, https://doi.org/10.1002/jmrs.66.

[35] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, and future directions," *Journal of Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021, https://doi.org/10.1186/s40537-021-00444-8.

[36] M. V. Sanida, T. Sanida, A. Sideris, and M. Dasygenis, "An advanced deep learning framework for multi-class diagnosis from chest X-ray images," *J–Multidisciplinary Scientific Journal*, vol. 7, no. 1, pp. 48–71, 2024, https://doi.org/10.3390/j7010003.

[37] A. Mumuni and F. Mumuni, "Automated data processing and feature engineering for deep learning and big data applications: A survey," *Journal of Information and Intelligence*, vol. 3, no. 2, pp. 113–153, 2025, https://doi.org/10.1016/j.jiixd.2024.01.002.

[38] A. T. Tran, T. Zeevi, and S. Payabvash, "Strategies to improve the robustness and generalizability of deep learning segmentation and classification in neuroimaging," *BioMedInformatics*, vol. 5, no. 2, p. 20, Apr. 2025, https://doi.org/10.3390/biomedinformatics5020020.

[39] Y.-D. Zhang, S. C. Satapathy, X. Zhang, and S.-H. Wang, "COVID-19 diagnosis via DenseNet and optimization of transfer learning setting," *Cognitive Computation*, vol. 16, no. 4, pp. 1649–1665, Jul. 2024, https://doi.org/10.1007/s12559-020-09776-8.

[40] K. Liu, Y. Yang, Y. Tian, and H. Mao, "Image dehazing technique based on DenseNet and the denoising self-encoder," *Processes*, vol. 12, no. 11, p. 2568, Nov. 2024, https://doi.org/10.3390/pr12112568.

[41] J. Sturekova, P. Kamencay, P. Sykora, and R. Hlavata, "A comparison of convolutional neural network transfer learning regression models for remote photoplethysmography signal estimation," *AI*, vol. 6, no. 2, p. 24, Feb. 2025, https://doi.org/10.3390/ai6020024.

[42] S. K. Betha, D. R. Dev, K. Sunkara, P. V. Kodavanti, and A. Putta, "An efficient attention DenseNet with LSTM for lung disease detection and classification using X-ray images supported by adaptive R2-UNet-based image segmentation," *Archives of Physiology and Biochemistry*, vol. 131, no. 6, pp. 977–1007, 2025, https://doi.org/10.1080/13813455.2025.2524182.

[43] D. Muhammad and M. Bendechache, "Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis," *Computational and Structural Biotechnology Journal*, vol. 24, pp. 542–560, Dec. 2024, https://doi.org/10.1016/j.csbj.2024.08.005.

[44] C. V. Aravinda, K. B. Sudeepa, S. Pradeep, P. Suraksha, and M. Lin, "Leveraging compact convolutional transformers for enhanced COVID-19 detection in chest X-rays: A Grad-CAM visualization approach," *Frontiers in Big Data*, vol. 7, p. 1489020, 2024, https://doi.org/10.3389/fdata.2024.1489020.

[45] P. Dusza, T. Banzato, S. Burti, M. Bendazzoli, H. Müller, and M. Wodzinski, "Comparative evaluation of CAM methods for enhancing explainability in veterinary radiography," *Scientific Reports*, vol. 15, no. 1, p. 29690, 2025, https://doi.org/10.1038/s41598-025-14060-6.

[46] E. H. Houssein, A. M. Gamal, E. M. G. Younis, and E. Mohamed, "Explainable artificial intelligence for medical imaging systems using deep learning: A comprehensive review," *Cluster Computing*, vol. 28, no. 7, p. 469, 2025, https://doi.org/10.1007/s10586-025-05281-5.

[47] J. Manoharan and Y. Sivagnanam, "A novel human action recognition model by Grad-CAM visualization with multi-level feature extraction using global average pooling with sequence modeling by bidirectional gated recurrent units," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, p. 118, 2025, https://doi.org/10.1007/s44196-025-00848-x.

[48] C. C. Ukwuoma, Z. Qin, M. B. Heyat, F. Akhtar, A. Smahi, J. K. Jackson, S. F. Qadri, A. Y. Muaad, H. N. Monday, and G. U. Nneji, "Automated lung-related pneumonia and COVID-19 detection based on novel feature extraction framework and vision transformer approaches using chest X-ray images," *Bioengineering*, vol. 9, no. 11, p. 709, Nov. 2022, https://doi.org/10.3390/bioengineering9110709.

[49] P. Kaushik, E. Jain, V. Kukreja, S. Hariharan, M. Krishnamoorthy, V. Ahuja, A. Bhattacherjee, R. K. Kaushal, and S.-Y. Chen, "Modelling radiological feature fusion and explainable AI in pneumonia detection: A graph-based deep learning and transformer approach," *Results in Engineering*, vol. 26, p. 105225, May 2025, https://doi.org/10.1016/j.rineng.2025.105225.

[50] C. M. Tsai and J.-D. Lee, "Dynamic ensemble learning with gradient-weighted class activation mapping for enhanced gastrointestinal disease classification," *Electronics*, vol. 14, no. 2, p. 305, Jan. 2025, https://doi.org/10.3390/electronics14020305.

[51] C. Y. T. Wong, F. Antaki, P. W. Court, A. Y. Ong, and P. A. Keane, "The role of saliency maps in enhancing ophthalmologists' trust in artificial intelligence models," *Asia-Pacific Journal of Ophthalmology*, vol. 13, no. 4, p. 100087, Jul. 2024, https://doi.org/10.1016/j.apjo.2024.100087.

[52] M. M. Musthafa, T. R. Mahesh, V. Vinoth Kumar, and S. Guluwadi, "Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with ResNet-50," *BMC Medical Imaging*, vol. 24, no. 1, p. 107, 2024, https://doi.org/10.1186/s12880-024-01292-7.

[53] D. Tang, J. Chen, L. Ren, X. Wang, D. Li, and H. Zhang, "Reviewing CAM-based deep explainable methods in healthcare," *Applied Sciences*, vol. 14, no. 10, p. 4124, May 2024, https://doi.org/10.3390/app14104124.

[54] S. Shahrabadi, T. Adão, E. Peres, R. Morais, L. G. Magalhães, and V. Alves, "Automatic optimization of deep learning training through feature-aware-based dataset splitting," *Algorithms*, vol. 17, no. 3, p. 106, Mar. 2024, https://doi.org/10.3390/a17030106.

[55] E. Hassan, M. Y. Shams, N. A. Hikal, and S. Elmougy, "The effect of choosing optimizer algorithms to improve computer vision tasks: A comparative study," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 16591–16633, 2023, https://doi.org/10.1007/s11042-022-13820-0.

[56] D. R. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: A survey," *Evolutionary Intelligence*, vol. 15, no. 1, pp. 1–22, 2022, https://doi.org/10.1007/s12065-020-00540-3.

[57] M. Bundea and G. M. Danciu, "Pneumonia image classification using DenseNet architecture," *Information*, vol. 15, no. 10, p. 611, Oct. 2024, https://doi.org/10.3390/info15100611.

[58] T. Zhou, X. Ye, H. Lu, X. Zheng, S. Qiu, and Y. Liu, "Dense convolutional network and its application in medical image analysis," *BioMed Research International*, vol. 2022, pp. 1–22, Apr. 2022, https://doi.org/10.1155/2022/2384830.

[59] S. Anari, S. Sadeghi, G. Sheikhi, R. Ranjbarzadeh, and M. Bendechache, "Explainable attention based breast tumor segmentation using a combination of UNet, ResNet, DenseNet, and EfficientNet models," *Scientific Reports*, vol. 15, no. 1, p. 1027, 2025, https://doi.org/10.1038/s41598-024-84504-y.

[60] M. Rahman, P. Roy, S. S. Frizell, and L. Qian, "Evaluating pretrained deep learning models for image classification against individual and ensemble adversarial attacks," *IEEE Access*, vol. 13, pp. 35230–35242, 2025, https://doi.org/10.1109/ACCESS.2025.3544107.

[61] F. van der Sluis and E. L. van den Broek, "Model interpretability enhances domain generalization in the case of textual complexity modeling," *Patterns*, vol. 6, no. 2, p. 101177, Feb. 2025, https://doi.org/10.1016/j.patter.2025.101177.

[62] J.-R. Paredes-Núñez, C. Rodríguez, L.-J. Varela-Serpa, and C. Navarro, "The challenge of deep learning for the prevention and automatic diagnosis of breast cancer: A systematic review," *Diagnostics*, vol. 14, no. 24, p. 2896, Dec. 2024, https://doi.org/10.3390/diagnostics14242896.

[63] A. Alabduljabbar, S. U. Khan, A. Alsuhaibani, F. Almarshad, and Y. N. Altherwy, "Medical imaging datasets, preparation, and availability for artificial intelligence in medical imaging," *Journal of Alzheimer's Disease Reports*, vol. 8, no. 1, pp. 1471–1483, Jul. 2024, https://doi.org/10.3233/ADR-240129.

[64] G. Inbasakaran and J. Anitha Ruth, "Clinical-ready CNN framework for lung cancer classification: Systematic optimization for healthcare deployment with enhanced computational efficiency," *Intelligence-Based Medicine*, vol. 12, p. 100292, 2025, https://doi.org/10.1016/j.ibmed.2025.100292.

[65] E. Mahamud, N. Fahad, M. Assaduzzaman, S. M. Zain, K. O. M. Goh, and M. K. Morol, "An explainable artificial intelligence model for multiple lung diseases classification from chest X-ray images using fine-tuned transfer learning," *Decision Analytics Journal*, vol. 12, p. 100499, 2024, https://doi.org/10.1016/j.dajour.2024.100499.

[66] G. P. Usha and J. S. R. Alex, "Advanced Grad-CAM extensions for interpretable aphasia speech keyword classification: Bridging the gap in impaired speech with XAI," *Results in Engineering*, vol. 24, p. 103414, Dec. 2024, https://doi.org/10.1016/j.rineng.2024.103414.

[67] F. M. Talaat, S. A. Gamel, R. M. El-Balka, M. Shehata, and H. ZainEldin, "Grad-CAM-enabled breast cancer classification with a 3D Inception-ResNet V2: Empowering radiologists with explainable insights," *Cancers*, vol. 16, no. 21, p. 3668, Oct. 2024, https://doi.org/10.3390/cancers16213668.

[68] N. Shifa, M. Saleh, Y. Akbari, and S. Al-Maadeed, "A review of explainable AI techniques and their evaluation in mammography for breast cancer screening," *Clinical Imaging*, vol. 123, p. 110492, May 2025, https://doi.org/10.1016/j.clinimag.2025.110492.

[69] S. C. Nouis, V. Uren, and S. Jariwala, "Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: A qualitative study of healthcare professionals' perspectives in the UK," *BMC Medical Ethics*, vol. 26, p. 89, Jul. 2025, https://doi.org/10.1186/s12910-025-01243-z.

[70] K. Preechakul, S. Sriswasdi, B. Kijsirikul, and E. Chuangsuwanich, "Improved image classification explainability with high-accuracy heatmaps," *iScience*, vol. 25, no. 3, p. 103933, Mar. 2022, https://doi.org/10.1016/j.isci.2022.103933.

[71] M. Z. Naser, "From failure to fusion: A survey on learning from bad machine learning models," *Information Fusion*, vol. 120, p. 103122, Mar. 2025, https://doi.org/10.1016/j.inffus.2025.103122.

[72] S. Asadi, P. Jimeno-Sáez, A. López-Ballesteros, and J. Senent-Aparicio, "Comparison and integration of physical and interpretable AI-driven models for rainfall–runoff simulation," *Results in Engineering*, vol. 24, p. 103048, 2024, https://doi.org/10.1016/j.rineng.2024.103048.