Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 14 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i14.58193

PAPER

Deep Learning-Based Real-Time Classification of Thoracic Pathologies in Chest Radiographs

Hanan Sabbar[1] (✉), Hassan Silkan[1], Khalid Abbad[2]

[1]Chouaib Doukkali University, El Jadida, Morocco

[2]Sidi Mohamed Ben Abdellah University, Fes, Morocco

sabbar.h@ucd.ac.ma

ABSTRACT

Diagnosing thoracic diseases from chest radiographs remains challenging, especially in resource-limited environments. This study presents YOLOv8n-cls, a lightweight deep learning model for real-time classification of five pathologies: 1) COVID-19, 2) fibrosis, 3) normal, 4) pneumonia, and 5) tuberculosis. The model was trained on a dataset of 11,019 chest X-ray images, combining public data from NIH ChestX-ray14 and a private clinical dataset, and achieving a Top-1 accuracy of 92.23%. Preprocessing included format conversion and text removal, while data augmentation techniques such as flipping, rotation, brightness/contrast adjustment, and affine translation were applied to improve model generalization. Performance evaluation relied on confusion matrices, precision, recall, F1-score, specificity, and ROC-AUC curves. Moreover, Grad-CAM visualizations were employed to enhance interpretability and analyze misclassification patterns. YOLOv8n-cls provides a strong balance between accuracy and computational efficiency, making it suitable for real-time clinical deployment.

KEYWORDS

thoracic disease classification, chest X-ray imaging, deep learning, YOLOv8n-cls, medical image analysis, real-time diagnosis

1 INTRODUCTION

Over the past decade, deep learning has profoundly transformed the field of medical imaging, opening new opportunities for the automated detection of thoracic diseases from chest X-ray scans. Chest radiography remains among the most widely used imaging techniques worldwide due to its low cost, rapid acquisition, and broad accessibility. It plays a crucial role in detecting conditions such as pneumonia, tuberculosis, lung cancer, and pleural effusion. However, interpreting CXR images manually requires considerable time and specialized expertise. Moreover, diagnostic conclusions can differ between radiologists, which may reduce consistency and delay clinical decision-making [1].

Sabbar, H., Silkan, H., Abbad, K. (2025). Deep Learning-Based Real-Time Classification of Thoracic Pathologies in Chest Radiographs. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(14), pp. 20–37. https://doi.org/10.3991/ijoe.v21i14.58193

Article submitted 2025-08-10. Revision uploaded 2025-09-24. Final acceptance 2025-09-24.

These factors highlight the pressing need for reliable, rapid, and accurate computer-assisted diagnostic tools, particularly in high-volume hospitals and resource-limited healthcare settings. Convolutional neural networks (CNNs) have demonstrated their ability to classify thoracic diseases with accuracy levels approaching those of expert radiologists [2]. Architectures, such as ResNet, have achieved strong results, especially when transfer learning is applied, allowing efficient training even with relatively small datasets [3]. Nonetheless, conventional CNN-based systems often depend on large, annotated datasets and demand significant computational resources, which restricts their adoption in real-world clinical practice [4].

To address these limitations, this research investigates YOLOv8n-cls, a compact yet robust variant of YOLOv8n designed for chest pathology classification. The model was trained on an enhanced dataset of 11,019 chest radiographs, combining private clinical data with additional samples from the public NIH ChestX-ray14 database. The dataset was distributed across five diagnostic categories: 1) normal, 2) tuberculosis, 3) pneumonia, 4) COVID-19, and 5) fibrosis, using a stratified split of 70% for training, 15% for validation, and 15% for testing.

The training process incorporated a structured preprocessing pipeline that included DICOM to PNG conversion, image resizing to 224×224 pixels, and removal of non-clinical text using Keras OCR and interpolation masking to ensure consistent image quality. Data augmentation techniques such as horizontal flipping, small rotations (±10°), brightness and contrast adjustment, and affine translation were applied to improve robustness and mitigate overfitting.

Although YOLOv8n-cls has not been extensively studied for thoracic disease classification, its compact architecture and real-time inference capabilities make it an attractive choice for deployment in clinical settings. Compared to established CNN models such as ResNet or DenseNet, YOLOv8n-cls requires significantly fewer parameters and achieves faster inference without compromising accuracy. The objective of this work is not to propose a novel architecture but to show that a lightweight model such as YOLOv8n-cls can deliver comparable or superior results with lower computational cost, particularly when supported by a robust preprocessing strategy.

The study makes four main contributions:

1. It introduces an extended and diversified dataset of 11,019 CXR images, including by underrepresented pathologies such as fibrosis, and improved through stratified sampling techniques.
2. It proposes a clinically relevant preprocessing pipeline that preserves the diagnostic integrity of the images.
3. It compares YOLOv8n-cls with state-of-the-art architectures such as EfficientNet B4, B2, B6, and B0, as well as EfficientNetV2 XL, and EfficientNetV2 L, demonstrating superior performance with a Top-1 accuracy of 92.23% while maintaining computational efficiency.
4. It introduces YOLOv8n-cls as an efficient and adaptable solution for real-time integration into clinical workflows, supporting radiologists in delivering fast and reliable diagnoses of thoracic diseases.

2 RELATED WORK

Over the past decade, deep learning has advanced considerably in thoracic disease classification, driven by advances in CNNs and the availability of large-scale

annotated chest X-ray datasets. Early studies, such as CheXNeXt, demonstrated that CNNs could perform on par with radiologists in detecting common diseases such as pneumonia and atelectasis [5]. However, their performance often declined for more complex or less frequent conditions (for example, emphysema and cardiomegaly), highlighting limitations in replicating nuanced clinical judgment.

Researchers have explored various CNN architectures, including ResNet, VGG, DenseNet, and EfficientNet, for multi-disease classification tasks [6, 7]. For instance, ResNet-50 achieved an AUC of 0.911 when trained on 14 thoracic pathologies in a multi-label setting [8]. Yet, comparative evaluations show that no single model consistently outperforms others across different datasets. VGG16 and XceptionNet excelled on the Kaggle dataset, while ResNet-50 and DenseNet performed better on NIH ChestX-ray [9]. Transfer learning has been widely adopted to improve generalization, enabling diagnostic accuracy comparable to expert radiologists [10]. Nevertheless, the high computational cost of these deep models continues to hinder their integration into real-time clinical workflows.

Dataset quality and availability remain critical challenges. Public datasets such as NIH ChestX-ray have enabled large-scale training, but rare classes such as fibrosis and pneumothorax are underrepresented, and label inconsistencies persist [11]. To mitigate these issues, preprocessing techniques such as contrast enhancement, segmentation, and data augmentation have become standard. Ait Nasser and Akhloufi reported that carefully designed augmentation strategies can improve rare disease detection rates by up to 15% [12]. However, the lack of interpretability in deep learning models still undermines clinical trust and adoption [13].

To overcome these limitations, hybrid and transformer-based models have gained traction. Vision Transformers (ViTs), which use self-attention to capture long-range dependencies, have been applied successfully to complex thoracic pathologies such as cardiomegaly [14]. Singh (2024) enhanced DenseNet121 with transformer blocks and achieved superior results across multiple datasets [15], while Tiwari et al. (2022) reported an AUC of 0.839 using DenseNet-121 on six thoracic classes [16]. CNN-RNN hybrids such as VGG19-RNN have also shown promise, reaching 97.8% accuracy for COVID-19 detection [17]. Yet, these complex architectures require substantial computational resources, which limits their deployment in low-resource settings.

In response to this challenge, lightweight models have emerged as a practical alternative. The YOLO (You Only Look Once) framework, originally developed for object detection, has been adapted to medical image analysis due to its speed and reduced hardware requirements [18]. YOLO-based models maintain high classification accuracy while being suitable for real-time diagnosis, particularly in resource-constrained settings [19].

Recent research also explores the integration of multiple imaging modalities such as combining X-rays with CT or ultrasound to increase diagnostic reliability [20, 21]. At the same time, explainable AI techniques are becoming more important for interpreting model outputs and improving transparency in AI-assisted decision-making [22].

Despite the wide use of CNNs and transformers, the YOLOv8n-cls architecture recently introduced and optimized for classification remains underexplored in thoracic disease detection, especially for multi-class classification that includes rare pathologies. Unlike prior work focused on heavier architectures, this study leverages YOLOv8n-cls to balance diagnostic performance and computational efficiency. We evaluate it on a heterogeneous dataset and apply an optimized preprocessing pipeline to assess its practical potential as a clinically deployable solution.

3 MATERIALS AND METHODS

This work follows a systematic methodology to design, train, and evaluate the YOLOv8n-cls model for thoracic disease classification using chest radiographs. The overall process includes dataset preparation, image preprocessing, data augmentation, model optimization, and performance evaluation, aiming to ensure high diagnostic accuracy and practical clinical relevance.

The dataset consists of 11,019 chest X-ray images, compiled from both public sources (including NIH ChestX-ray14) and private clinical repositories. This hybrid composition provides diversity in imaging protocols, patient demographics, and diagnostic categories. Prior to training, each image undergoes a standardized preprocessing pipeline involving format conversion, resizing, and removal of non-clinical textual artifacts. To enhance model generalization and robustness, various data augmentation techniques are applied during training. These include random horizontal flipping, rotation, affine translations, and brightness/contrast adjustments. Model training is performed on a high-performance computing environment optimized for deep learning tasks. The YOLOv8n-cls architecture, configured for multi-class classification, achieves a balance between computational efficiency and diagnostic accuracy. Performance evaluation was conducted using standard metrics such as Top-1 accuracy, precision, recall, F1-score, and specificity. In addition, ROC-AUC curves were generated for each class to assess discriminative performance, and Grad-CAM visualizations were employed to enhance interpretability by highlighting key image regions that contributed to model predictions across the five thoracic pathologies.

Figure 1 illustrates an overview of the methodological workflow, which is further detailed in the subsequent subsections.

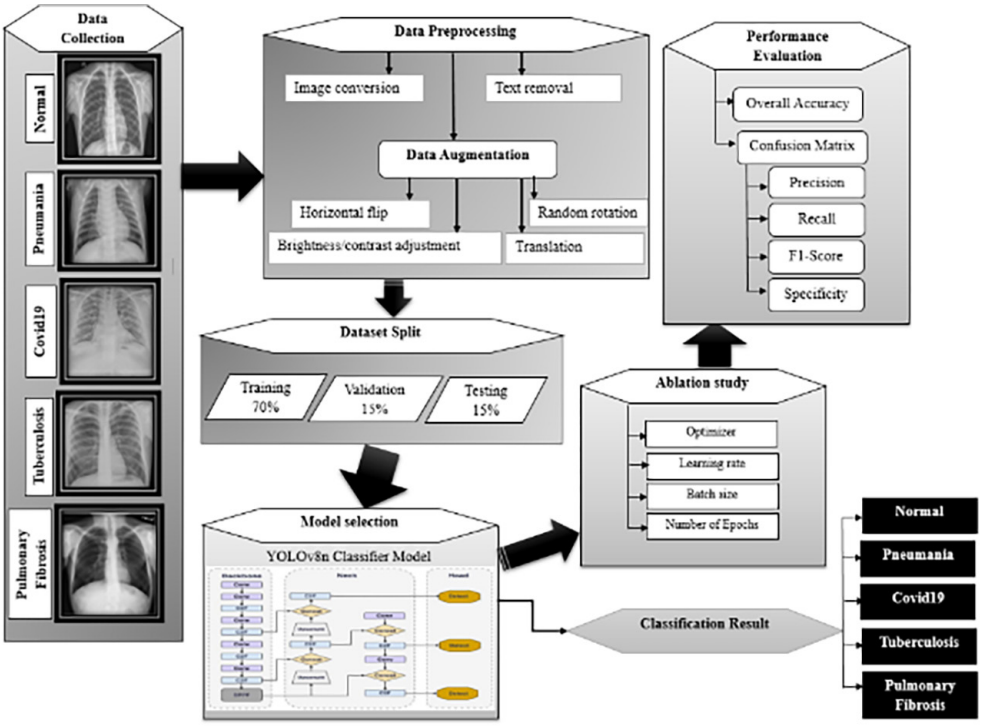


Fig. 1. Flow chart of the proposed methodology for thoracic disease classification

3.1 Dataset description

In this study, we assembled a rich and diverse dataset to enable effective classification of thoracic pathologies such as pneumonia, tuberculosis, fibrosis, and COVID-19. The collection process drew on two primary sources: publicly available chest X-ray repositories and private healthcare institutions. This combination provided a broad range of imaging conditions, patient demographics, and acquisition protocols. Incorporating private clinical data added variability and realism, thereby supporting better model generalization across different clinical environments and imaging standards.

The final dataset contains **11019** chest X-ray images, grouped into five diagnostic categories: 1) **normal** (**4011** images), 2) **tuberculosis** (**1474** images), 3) **pneumonia** (**1976** images), 4) **COVID-19** (**1706** images), and 5) **fibrosis** (**1852** images). While the dataset is generally well balanced, tuberculosis cases are somewhat fewer than other categories such as fibrosis. This imbalance motivates the use of targeted data augmentation to strengthen the model's ability to learn minority classes and reduce bias toward more frequent conditions.

For robust and unbiased evaluation, the dataset was split into training, validation, and testing subsets using stratified sampling. This ensured that class proportions remained consistent across all splits: 70% for training, 15% for validation, and 15% for testing. Table 1 presents the detailed distribution of samples per category.

Table 1. Example dataset size for thoracic disease classification

Classes	Number of Images
Normal	4011
Tuberculosis	1474
Pneumonia	1976
COVID-19	1706
Fibrosis	1852
TOTAL	**11019**

Figure 2 showcases representative X-ray images from each pathology class, highlighting the dataset's diversity and its suitability for training a robust and generalizable YOLOv8n-cls model.

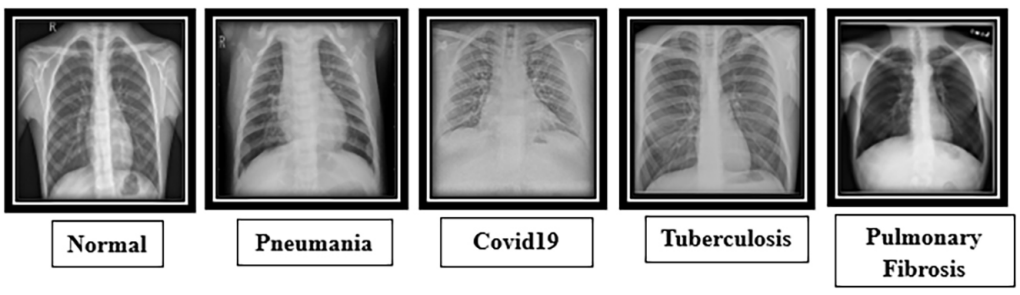


Fig. 2. Chest X-ray images with their respective labels, representing the different classes of pathologies analyzed

3.2 Radiographic data preprocessing

The preprocessing of chest X-ray images represented a fundamental step in preparing the dataset for model training. This phase was designed to ensure standardized image quality, remove irrelevant or non-clinical information, and preserve the diagnostic integrity of each scan.

Most images were originally stored in high-resolution DICOM format. To facilitate efficient processing and reduce memory usage, raw pixel data were extracted with Pydicom and converted into PNG format using the Python Imaging Library (PIL). This conversion enabled faster training workflows without compromising diagnostic detail.

To maintain consistency across the dataset, all images were resized to 224×224 pixels, matching the input requirements of the classification model. In addition, identifying text embedded in the radiographs was eliminated through a multi-step cleaning process. This included automated detection using Keras OCR, masking with interpolation, and manual verification to ensure completeness and accuracy. As illustrated in Figure 3, this cleaning process pipeline preserved all clinically significant features while enhancing overall image clarity.

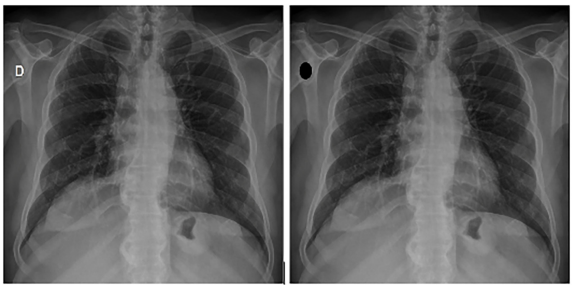


Fig. 3. Text removal applied to a chest X-ray image

By conducting these standardization steps, the preprocessing pipeline helped reduce potential training bias, mitigate overfitting, and support the model's ability to generalize across varied clinical imaging environments.

3.3 Image augmentation

To enhance the YOLOv8n-cls model's ability to generalize thoracic disease classification across varied clinical settings, a set of systematic data augmentation techniques was applied to each training image. These augmentations mimic real-world variations such as patient posture changes, positioning shifts, and variations in exposure or contrast, thereby strengthening model robustness.

The following transformations were applied using the **PyTorch** *torchvision.transforms* **module**:

- **Random Horizontal Flip (p = 0.5):** Mirrors the image left-to-right to ensure symmetry invariance.
- **Random Rotation (±10 degrees):** Simulates slight changes in patient positioning during image capture.
- **Color Jitter (brightness = 0.3, contrast = 0.3):** Varies exposure and contrast to emulate acquisition inconsistencies.
- **Random Affine (translate up to 10%):** Applies small spatial shifts along x and y axes to simulate minor misalignments.

These transformations were applied consistently and probabilistically during training to every image, generating a rich set of variations while preserving anatomical plausibility.

Table 2 presents the data augmentation techniques along with their corresponding parameters and application probabilities

Table 2. Summary of data augmentation techniques

Technique	Parameter
Random Horizontal Flip	p = 0.5
Random Rotation	degrees = ±10
Color Jitter	brightness = 0.3, contrast = 0.3
Random Affine	translate = (0.1, 0.1), degrees = 0

Figure 4 illustrates examples of the augmentation strategies implemented on chest X-ray images. Each transformation increases the variability of the dataset, helping the YOLOv8n-cls model generalize more effectively and maintain consistent performance in real clinical scenarios.

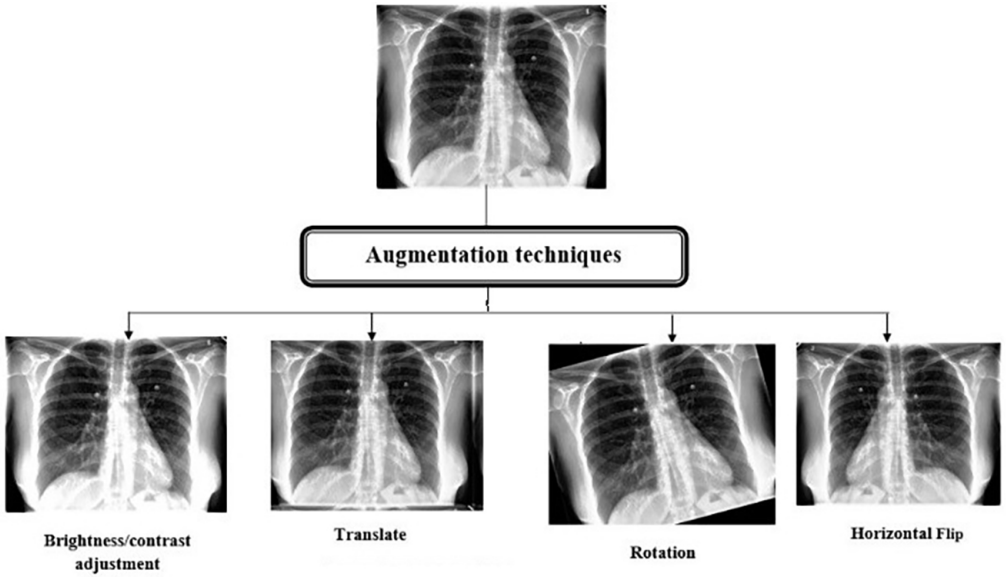


Fig. 4. Examples of augmentation techniques applied to chest X-ray images

3.4 Implementation and hardware architecture

The YOLOv8n-cls model for thoracic disease classification was trained following a carefully optimized configuration to achieve high accuracy and robust generalization. The AdamW optimizer was chosen for its efficiency in gradient computation and its stable convergence behavior. Training began with a learning rate of 1.0×10^{-3}, gradually reduced to 5.0×10^{-4} to promote steady learning and avoid abrupt parameter shifts. The configured weight decay value reduced overfitting by constraining the magnitude of the weights. The training process spanned 100 epochs, a duration sufficient to capture the complexity of thoracic radiographic patterns without incurring unnecessary computational costs. The chosen batch size value offered an effective

balance between GPU memory usage and training stability. All images were standardized to 224×224 pixels to preserve essential anatomical features while maintaining compatibility with the network's input dimensions. A momentum coefficient of 0.9 was used to smooth parameter updates and accelerate convergence.

- **Image size:** 224×224
- **Epochs:** 100
- **Optimizer:** AdamW
- **Batch size:** 32
- **Initial lr:** 1.0×10^{-3}
- **Final lr:** 5.0×10^{-4}
- **Momentum:** 0.9
- **Weight decay:** 0.0005

A preliminary sensitivity analysis was conducted to verify the robustness of the selected hyperparameters. Variations in learning rate, batch size, and weight decay were explored in early-stage experiments. The final configuration was chosen for its optimal trade-off between convergence stability, generalization performance, and computational efficiency.

All experiments were carried out in Python using specialized libraries. The Ultralytics framework handled YOLOv8 implementation, pydicom was used for DICOM file handling, Pillow managed image processing, and Keras-OCR removed non-clinical text from radiographs. Training was performed on a workstation equipped with:

- **CPU:** Intel Core i7-12700 @ 4.9 GHz (12 cores)
- **RAM:** 64 GB DDR5
- **GPU:** NVIDIA RTX A2000 with 12 GB of VRAM
- **Storage:** 1 TB NVMe SSD

CUDA acceleration was enabled throughout to ensure efficient handling of high-resolution medical images, stable batch processing, and optimized GPU memory utilization. This robust computational setup provided fast, stable training and supported the development of a high-accuracy, computationally efficient thoracic disease classification model efficient.

3.5 YOLOv8 architecture

The original YOLO was introduced by Joseph Redmon and colleagues as a unified and real-time object detection framework. Unlike traditional multi-stage pipelines, YOLO processes the entire image in a single pass using one convolutional neural network. This network predicts bounding boxes and class probabilities simultaneously, providing an efficient and cohesive solution for object detection tasks [23].

For the purpose of thoracic image classification, YOLOv8 was adapted to focus exclusively on predicting image classes, leaving aside the bounding box detection component. The adapted design integrates an optimized structure composed of five convolutional layers, four C2f modules, and a single classification head. One key modification involves replacing the earlier C3 modules with C2f blocks, which improve information flow while maintaining computational efficiency. Additionally, 3×3 convolutional filters are used to enhance the extraction of discriminative features.

As illustrated in Figure 5, the YOLOv8 architecture is organized into three main functional parts. The backbone is responsible for extracting low- and high-level visual features from the input image. The neck aggregates and refines these features across multiple scales, ensuring that both fine details and broader contextual cues are preserved. Finally, the head applies the classification layers that generate the final category predictions. This streamlined yet powerful configuration makes YOLOv8 particularly effective for medical image classification tasks, where both precision and computational efficiency are crucial.

In this study, we specifically employed the YOLOv8n-cls variant, which is the lightweight classification version of YOLOv8. This choice was motivated by its balance between speed and accuracy, making it well-suited for real-time medical imaging applications. Its reduced computational complexity allowed us to train and deploy the model efficiently while maintaining a high level of diagnostic performance across diverse thoracic disease categories.

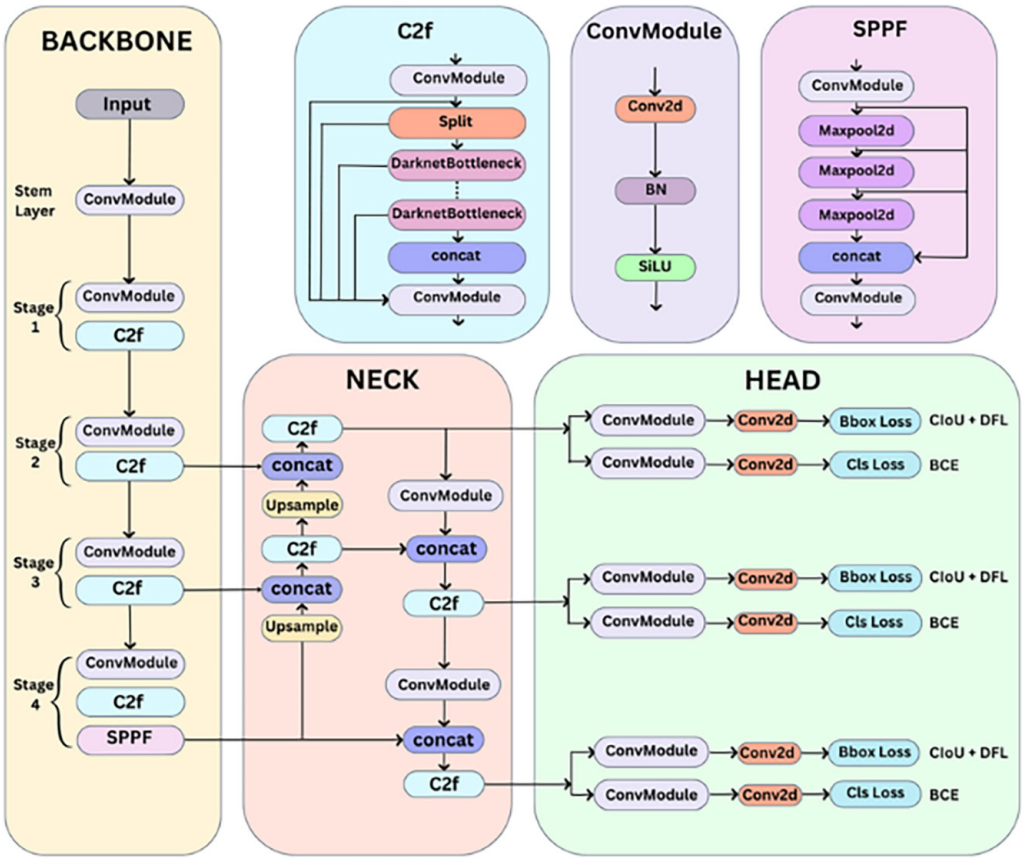


Fig. 5. YOLOv8 architecture, illustrating the structure of the backbone, neck, and head

3.6 Performance metrics

In this study, a confusion matrix is utilized to evaluate and compute the performance metrics of the YOLOv8n-cls model for thoracic disease classification. This matrix provides a detailed breakdown of the predictions made by the model during both the training and testing phases, enabling a comprehensive assessment of its performance.

To quantify the model's effectiveness, we employed several standard classification metrics, including accuracy, precision, sensitivity, F1-score and specifity. These metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \times 100 \tag{3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

$$Specifity = \frac{TN}{TN + FP} \tag{5}$$

4 RESULTS AND DISCUSSION

This section presents an updated evaluation of the YOLOv8n-cls model for thoracic disease classification from chest X-ray images. The model was assessed using **Top-1 accuracy, loss curves, confusion matrices, ROC-AUC scores, and Grad-CAM visualizations** to provide a comprehensive analysis of its performance.

4.1 Classification accuracy and training dynamics

The evaluation of the YOLOv8n-cls model demonstrates strong robustness in classifying five thoracic pathologies: 1) COVID-19, 2) pneumonia, 3) tuberculosis, 4) fibrosis, and 5) normal lungs. As illustrated in Figure 6, the Top-1 accuracy increased steadily during training and reached 92.23%, indicating a high level of discriminative power across clinically relevant classes. This performance reflects the effectiveness of the training setup, which included the AdamW optimizer, stratified sampling, and a carefully designed data augmentation strategy to enhanced diversity and reduce overfitting.

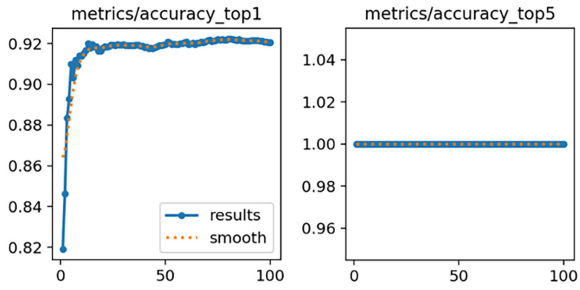


Fig. 6. Accuracy curve of YOLOv8n-cls during training

The Top-5 accuracy, illustrated in Figure 6 (right), remained constant at 100%, confirming that the correct class was consistently among the model's top five predictions. This provides an additional indicator of classification robustness in a multiclass context.

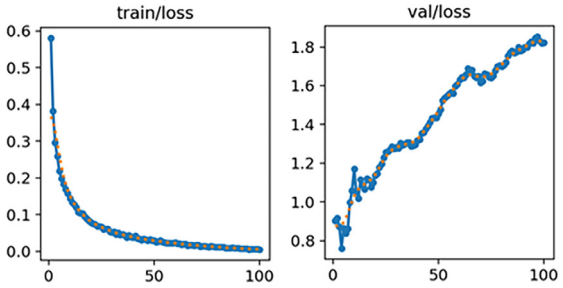


Fig. 7. Evolution of training and validation loss

Figure 7 illustrates the evolution of training and validation loss over 100 epochs. The training loss decreased progressively and converged to minimal values, indicating effective and stable learning. While the validation loss displayed slight fluctuations after epoch 60, it remained stable, reflecting the model's strong ability to maintain performance on unseen data.

These results confirm the robustness of the training process, further strengthened by the use of data augmentation techniques, which contributed significantly to the model's stability and generalization.

4.2 Confusion matrices and classification metrics

The confusion matrix in Figure 8 and the classification metrics presented in Table 3 provide a detailed and comprehensive evaluation of the YOLOv8n-cls model's performance across the five thoracic disease categories. These two elements jointly illustrate the model's strengths and highlight areas requiring refinement for enhanced clinical applicability.

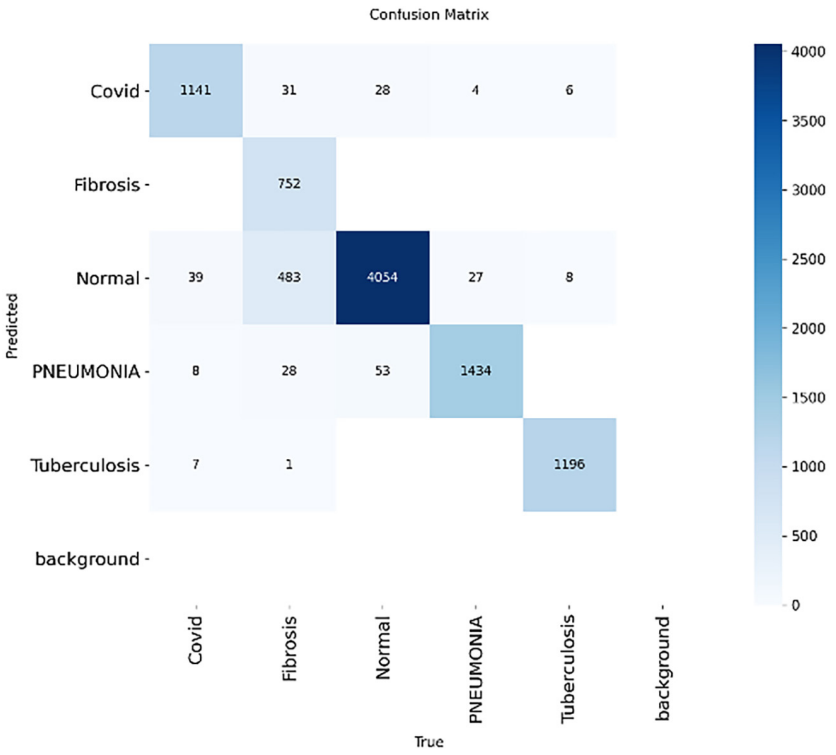


Fig. 8. Confusion matrix for the proposed model in thoracic disease classification

Table 3. Classification performance metrics for our model per class (in %)

Class	Precision	Recall	Specificity	F1-Score
Normal	98.62	98.98	99.05	98.8
Tuberculosis	98.8	99.52	99.21	99.16
Pneumonia	93.28	100	97.12	96.9
COVID-19	94.48	96.28	98.62	95.61
Fibrosis	93.55	57.91	99.34	71.72

The Normal and Tuberculosis classes achieved the most consistent and accurate results, with precision and recall rates exceeding 98%, and F1-scores reaching 98.80% and 99.16% respectively. This indicates that the model can effectively identify both healthy lungs and tuberculosis cases with high confidence, reflecting robust feature extraction and stable generalization in these categories.

Similarly, Pneumonia and COVID-19 were classified effectively, with F1-scores of 96.90% and 95.61% respectively. These results confirm the model's robustness in detecting infectious diseases with distinct radiographic patterns. COVID-19 in particular achieved a high specificity of 98.62%, underscoring the model's ability to minimize false positives in clinical diagnosis.

By contrast, Fibrosis presented a more complex challenge. Although the model achieved a high specificity of 99.34%, indicating that non-fibrotic cases were rarely misclassified as fibrosis, the recall for this class was substantially lower (57.91%), and the F1-score dropped to 71.72%. This gap suggests that a significant number of true fibrosis cases were missed, often misclassified as Normal or COVID-19, likely due to the subtle and overlapping radiological features that fibrosis shares with other thoracic conditions. These results may also reflect an underrepresentation of fibrosis images in the training set or a broader intra-class variability.

The classification metrics in Table 3 underscore the model's strong diagnostic performance in most categories, while also emphasizing the difficulty of accurately identifying fibrosis. Increasing class representation and adopting targeted augmentation strategies could help address this gap in future work.

These findings confirm that YOLOv8n-cls provides a solid and interpretable framework for thoracic disease classification, offering high diagnostic reliability for most conditions. While further refinement is required for underrepresented or visually ambiguous categories, the model demonstrates a promising balance between accuracy and efficiency, making it well suited for integration into clinical decision-support systems.

4.3 ROC curve analysis and AUC scores

The receiver operating characteristic (ROC) curve is a widely used diagnostic tool for assessing a model's discriminative power. It plots the true positive rate (TPR) against the false positive rate (FPR) across different threshold levels. The area under the curve (AUC) summarizes this information into a single metric, where higher values indicate stronger class separability, as illustrated in Figure 9.

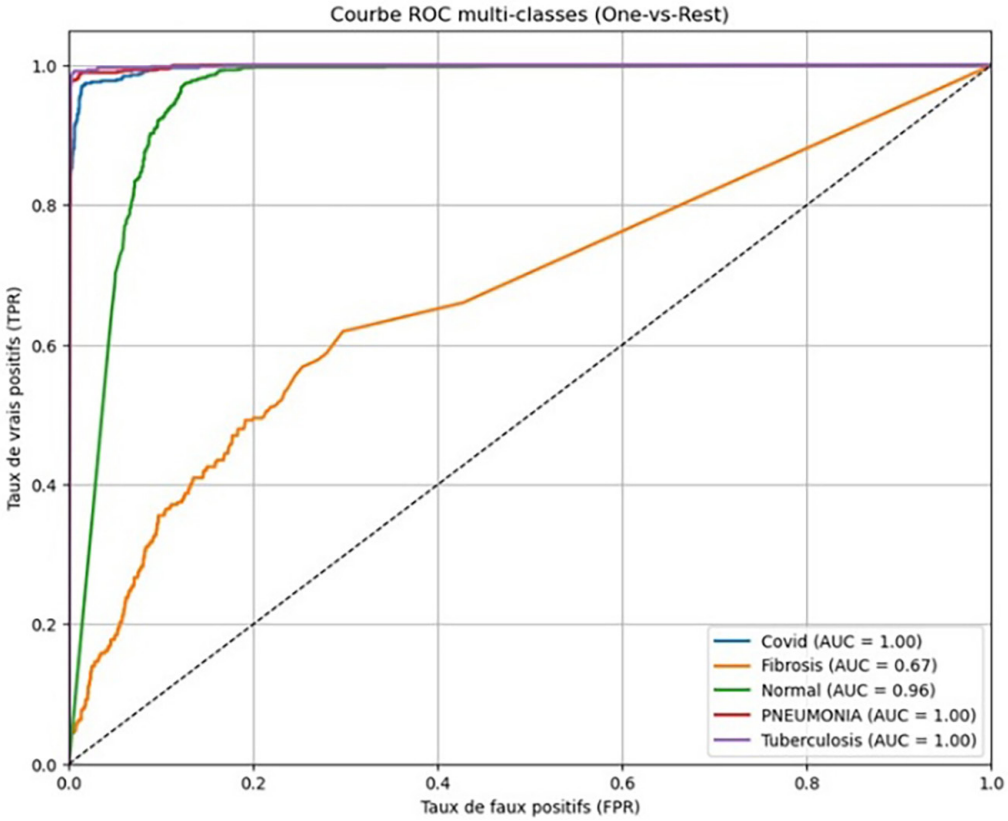


Fig. 9. Multi-class ROC curves for thoracic disease classification using one-vs.-rest strategy

In this study, the YOLOv8n-cls model demonstrated excellent discriminative performance for most thoracic disease categories. As shown in Figure 9, the ROC curves were generated using a one-vs.-rest approach for each class. The model achieved perfect AUC scores of 1.00 for COVID-19, pneumonia, and tuberculosis, confirming its robustness in identifying these pathologies with high confidence. The Normal class also produced a high AUC of 0.96, reflecting the model's ability to reliably distinguish healthy lungs from pathological cases.

However, the fibrosis class obtained a relatively low AUC of 0.67, indicating limited class separability. This outcome aligns with previous confusion matrix observations and may be attributed to the overlapping radiographic features of fibrosis with other thoracic conditions, as well as underrepresentation of fibrosis cases in the training dataset.

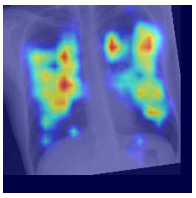
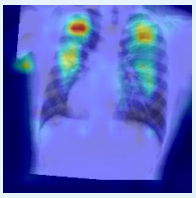
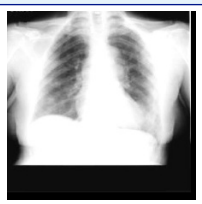
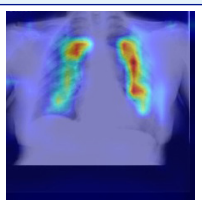
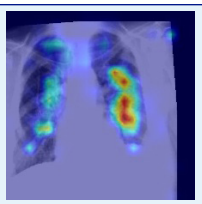
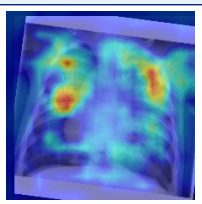
Overall, the ROC and AUC analysis reinforces the model's strength in classifying prevalent and visually distinct thoracic diseases, while underlining the need for further improvements in detecting subtle abnormalities such as those associated with fibrosis.

4.4 Grad-CAM visualizations

To provide visual insight into the inner workings of the model, we employed gradient-weighted class activation mapping (Grad-CAM) to generate class-specific attention maps for representative test images. Table 4 presents a comparative overview of the original chest radiographs alongside their corresponding Grad-CAM visualizations for each studied class.

This visual inspection helps determine whether the model's decisions are based on clinically relevant image regions, thereby enhancing the trustworthiness and interpretability of its predictions.

Table 4. Comparative visualization of original X-ray images and Grad-CAM heatmaps by class

Class	Original Image	Grad-CAM Visualization
Normal		
Tuberculosis		
COVID-19		
Pneumonia		
Fibrosis		

In the Grad-CAM visualization for normal cases, no focal activation is observed, which is consistent with the absence of pathological features. The heatmap is evenly distributed, indicating that the model does not detect any suspicious regions an expected and desirable behavior.

For Tuberculosis, the Grad-CAM map highlights strong activations in the upper lung zones, which aligns with the typical radiological manifestation of pulmonary tuberculosis. This suggests that the model successfully identifies disease-relevant patterns.

For COVID-19, activations are bilaterally distributed in the lower lung fields, particularly along the peripheral regions. This distribution reflects the characteristic pattern of COVID-19 involvement, showing the model's ability to localize diffuse ground-glass opacities.

For pneumonia, the heatmaps are more asymmetrical, with high-intensity regions in one lung field. This localization corresponds to alveolar opacities caused by bacterial pneumonia, indicating that the model correctly focuses on lobar consolidation zones.

For fibrosis, diffuse activation patterns are observed across both lungs, especially at the bases, consistent with interstitial lung involvement in fibrotic disease. These patterns demonstrate the model's ability to capture widespread structural alterations.

The Grad-CAM visualizations confirm the model's capacity to focus on clinically meaningful regions that align with radiographic findings typical of each disease, thereby supporting its potential utility in clinical decision-support settings.

5 COMPARATIVE ANALYSIS

Alongside the experiments performed with the proposed YOLOv8n-cls architecture for thoracic disease classification in chest X-ray images, several state-of-the-art deep learning models were also evaluated. The set included EfficientNet B4, B2, B6, and B0, as well as EfficientNetV2 XL, and EfficientNetV2 L. he aim was to compare their predictive performance and to assess how well each architecture could meet clinical requirements. The results of this comparative study are summarized in Table 5.

Table 5. Comparative performance of various models in thoracic disease classification

Model	Top-1 Accuracy (%)
EfficientNet B4	95.49
EfficientNetV2 XL	85.50
EfficientNet B6	92.0
EfficientNetV2 L	91.1
EfficientNet B2	77.73
EfficientNet B0	93.42

Although EfficientNet B0 and EfficientNet B4 achieved relatively high accuracy during validation, their performance dropped when evaluated on the independent test set. This variation underlines the importance of distinguishing between validation and test results when assessing the applicability of a model in real-world medical scenarios. In several cases, the observed decline in performance suggests overfitting, where the model learns specific patterns from the validation data but struggles to generalize to new, unseen cases. In contrast, the YOLOv8n-cls model maintained consistent performance across both validation and test datasets, demonstrating stronger generalization capabilities. This stability reinforces its potential for clinical deployment, where models must handle diverse and previously unseen imaging data without a significant drop in accuracy.

These findings emphasize that validation accuracy alone is not a sufficient indicator of real-world reliability in medical imaging. True robustness is demonstrated when a model sustains high performance on independent test sets, and in this respect, YOLOv8n-cls outperformed all other evaluated architectures. Figure 10 illustrates the Top-1 accuracy curves of the trained models, offering a visual comparison of their learning and generalization behaviors.

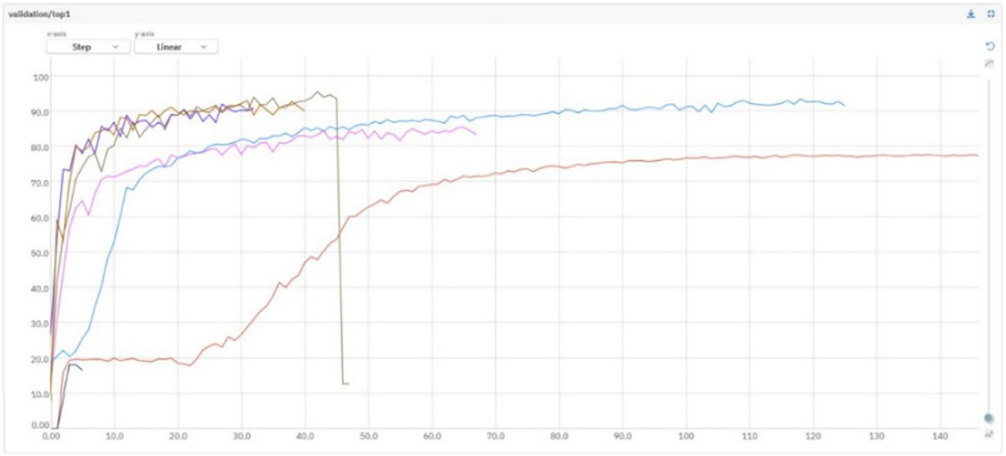


Fig. 10. Top-1 accuracy curves for trained models

6 CONCLUSION AND FUTURE WORK

This study presented a lightweight yet effective YOLOv8n-cls based framework for classifying five thoracic disease categories from chest X-rays. Trained on a large and diverse dataset of 11,019 images, the model achieved a Top-1 accuracy of 92.23%, outperforming benchmark architectures such as EfficientNet-B0 and B4. Strong results were observed for tuberculosis and pneumonia, supported by Grad-CAM visualizations and robust performance metrics.

The model's generalization was enhanced through rigorous preprocessing and evaluation protocols. However, limitations remain in handling underrepresented classes such as fibrosis. While conventional augmentation helped address class imbalance, future work will investigate advanced techniques such as GAN-based synthesis, SMOTE, and class-aware strategies to improve minority class sensitivity. Further improvements may include integrating attention mechanisms and hierarchical learning to better capture complex radiographic features. In addition, a critical next step is to conduct external validation on independent datasets from different institutions to more rigorously assess robustness and strengthen confidence in the model's clinical applicability.

This study confirms the potential of YOLOv8n-cls for scalable, interpretable, and reliable clinical deployment.

7 REFERENCES

[1] F. M. J. M. Shamrat *et al.*, "LungNet22: A fine-tuned model for multiclass classification and prediction of lung disease using X-ray images," *Journal of Personalized Medicine*, vol. 12, no. 5, p. 680, 2022. https://doi.org/10.3390/jpm12050680

[2] T. Liu, E. Siegel, and D. Shen, "Deep learning and medical image analysis for COVID-19 diagnosis and prediction," *Annual Review of Biomedical Engineering*, vol. 24, no. 1, pp. 179–201, 2022. https://doi.org/10.1146/annurev-bioeng-110220-012203

[3] M. A. A. Al-qaness *et al.*, "Chest X-ray images for lung disease detection using deep learning techniques: A comprehensive survey," *Archives of Computational Methods in Engineering*, vol. 31, no. 6, 2024. https://doi.org/10.1007/s11831-024-10081-y

[4] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nature Medicine*, vol. 28, no. 1, pp. 31–38, 2022. https://doi.org/10.1038/s41591-021-01614-0

[5] P. Rajpurkar *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Medicine*, vol. 15, no. 11, p. e1002686, 2018. https://doi.org/10.1371/journal.pmed.1002686

[6] S. Anis *et al.*, "An overview of deep learning approaches in chest radiograph," *IEEE Access*, vol. 8, pp. 182347–182354, 2020. https://doi.org/10.1109/ACCESS.2020.3028390

[7] R. Shetty and P. N. Sarappadi, "Deep learning methods on chest x-ray radiography for detection and classification of thoracic disease: A survey," *AIP Conference Proceedings*, vol. 2742, p. 020054, 2024. https://doi.org/10.1063/5.0184528

[8] A. El-Fiky, M. A. Shouman, S. Hamada, A. El-Sayed, and M. E. Karar, "Multi-label transfer learning for identifying lung diseases using chest X-rays," in *2021 International Conference on Electronic Engineering (ICEEM)*, IEEE, 2021, pp. 1–6. https://doi.org/10.1109/ICEEM52022.2021.9480622

[9] D. Pantola, D. Vatsa, and M. Gupta, "Exploring transfer learning approaches for thorax disease diagnosis," in *2023 Second International Conference on Smart Technologies for Smart Nation (SmartTechCon)*, IEEE, 2023, pp. 126–131. https://doi.org/10.1109/SmartTechCon57526.2023.10391323

[10] M. Rathi, A. Mittal, D. Agarwal, and Jahnavi, "Prediction of thorax diseases using deep and transfer learning," in *2020 Research, Innovation, Knowledge Management and Technology Application for Business Sustainability (INBUSH)*, IEEE, 2020, pp. 236–240. https://doi.org/10.1109/INBUSH46973.2020.9392161

[11] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-scale Chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2097–2106. https://doi.org/10.1109/CVPR.2017.369

[12] A. A. Nasser and M. A. Akhloufi, "A review of recent advances in deep learning models for chest disease detection using radiography," *Diagnostics*, vol. 13, no. 1, p. 159, 2023. https://doi.org/10.3390/diagnostics13010159

[13] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. https://doi.org/10.1016/j.media.2017.07.005

[14] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. https://doi.org/10.48550/arXiv.2010.11929

[15] S. Singh, "Computer-aided diagnosis of thoracic diseases in chest X-rays using hybrid cnn-transformer architecture," *arXiv preprint arXiv:2404.11843*, 2024. https://doi.org/10.48550/arXiv.2404.11843

[16] R. Tiwari, M. Verma, and S. K. Sar, "Detecting different thoracic disease using CNN-Model," in *2022 International Conference for Advancement in Technology (ICONAT)*, 2022, pp. 1–11. https://doi.org/10.1109/ICONAT53423.2022.9725940

[17] B. J. Khadhim, Q. K. Kadhim, W. K. Shams, S. T. Ahmed, and W. A. Wahab Alsiadi, "Diagnose COVID-19 by using hybrid CNN-RNN for chest X-ray," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 2, pp. 852–860, 2023. https://doi.org/10.11591/ijeecs.v29.i2

[18] G. Jocher, A. Chaurasia, and J. Qiu, "YOLOv8: A real-time object detection framework for medical imaging," Ultralytics GitHub Repository, 2023. https://github.com/ultralytics/ultralytics

[19] D. Palaniappan *et al.*, "Yolo in healthcare: A comprehensive review of detection architectures, domain applications, and future innovations," *IEEE Access*, 2025. https://doi.org/10.1109/ACCESS.2025.3599358

[20] H. Malik and T. Anees, "Multi-modal deep learning methods for classification of chest diseases using different medical imaging and cough sounds," *PLoS ONE*, vol. 19, no. 3, p. e0296352, 2024. https://doi.org/10.1371/journal.pone.0296352

[21] H. Ding, L. Fan, J. Zhang, and G. Gao, "Deep learning-based system combining chest X-ray and computerized tomography images for COVID-19 diagnosis," *British Journal of Hospital Medicine*, vol. 85, no. 8, pp. 1–15, 2024. https://doi.org/10.12968/hmed.2024.0244

[22] N. L. Rane, S. P. Choudhary, and J. Rane, "Explainable artificial intelligence (XAI) in healthcare: Interpretable models for clinical decision support," *SSRN*, 2023. https://doi.org/10.2139/ssrn.4637897

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2016, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91

8 AUTHORS

Hanan Sabbar is currently a student researcher at the LAROSERI Laboratory, specializing in medical imaging and computer vision at the Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco. Her research focuses on artificial intelligence and medical imaging, with particular interests in computer vision, deep learning, machine learning, and medical image analysis (E-mail: sabbar.h@ucd.ac.ma).

Hassan Silkan is a Research Professor at the LAROSERI Laboratory, Chouaib Doukkali University, El Jadida, Morocco. His research focuses on artificial intelligence, advanced signal processing, and image analysis (E-mail: silkan.h@ucd.ac.ma).

Khalid Abbad is a Research Professor at the Intelligent Systems and Applications Laboratory, Sidi Mohamed Ben Abdellah University, Fès, Morocco. His work focuses on artificial intelligence, advanced signal processing, and medical image analysis (E-mail: khalid.abbad@usmba.ac.ma).