CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2025.049

Incorporating self-attention into DenseNet for multi-label chest X-ray image classification

Tri-Thuc Vo[1], and Thanh-Nghi Do[1,2*]

[1]College of Information Technology, Can Tho University, Viet Nam

[2]UMI UMMISCO 209 (IRD/UPMC), Sorbonne University, Pierre and Marie Curie University - Paris 6, France

*Corresponding author (dtnghi@cit.ctu.edu.vn)

Article info.

Received 15 Jul 2025
Revised 17 Aug 2025
Accepted 8 Oct 2025

Keywords

DenseNet, chest X-ray image, multi-label classification, self-attention

ABSTRACT

This paper presents DNet-nSA, a novel deep learning architecture designed to enhance multi-label classification of chest X-ray (CXR) images by integrating n self-attention blocks into the DenseNet framework. While convolutional neural networks (CNNs) are effective at identifying local patterns, they frequently face challenges in capturing long-range dependencies and global context, which are essential for detecting spatially distributed abnormalities in CXR images. By embedding self-attention mechanisms, DNet-nSA allows the network to better capture non-local interactions and highlight diagnostically relevant regions. We propose and evaluate two variants: DNet-1SA and DNet-2SA, corresponding to the number of self-attention modules used. Experiments conducted on the ChestX-ray14 dataset demonstrate that the proposed models outperform the baseline DenseNet, the contrastive learning approach MoCoR101, and the self-supervised learning model MoBYSwinT, achieving a notable AUC of 0.822, confirming the effectiveness of self-attention in improving multi-label CXR image classification.

1. INTRODUCTION

Chest X-ray imaging plays an important role in diagnosing lung diseases due to its ability to provide detailed images of the lung structures. According to the World Health Organization, lung diseases such as chronic obstructive pulmonary disease (COPD) and lung cancer are leading causes of high mortality rates globally. COPD accounts for over 3 million deaths annually, while lung cancer was estimated to cause 1.8 million deaths in 2020. CXR images help detect and diagnose lung diseases such as pneumonia, tuberculosis as well as lung cancer. Moreover, the cost of X-ray imaging is typically lower than that of many other imaging methods, such as CT scans or MRIs, making it a popular option at various healthcare facilities. In addition to its cost-effectiveness, X-rays have the advantage of easy accessibility in most healthcare settings. Diagnosis through imaging methods like CXR images relies heavily on the experience and expertise of the doctor. While CXR images provide detailed images of the lung structure, the interpretation relies on the physician's observational and analytical skills, which can pose a potential risk of misdiagnosis.

In this paper, we introduce DNet-nSA, a novel deep learning architecture designed to improve CXR image classification by incorporating *n* self-attention blocks (Vaswani et al., 2017) into the DenseNet (Huang et al., 2017). The addition of self-

attention significantly improves the models' ability to capture long-range dependencies and global context, which traditional convolutional layers often fail to represent adequately due to their inherently local patterns in CXR images. Therefore, our proposed DNet-nSA enhances the network's ability to capture non-local interactions and emphasize diagnostically significant regions. We propose and assess two variants, DNet-1SA and DNet-2SA, based on the number of self-attention modules used. Evaluations on the ChestX-ray14 dataset show that our models outperform the baseline DenseNet (Huang et al., 2017), the contrastive learning approach MoCoR101 (Sowrirajan et al., 2021), and the self-supervised learning model MoBYSwinT (Vo & Do, 2024a), achieving an AUC of 0.822. The empirical results demonstrate the efficacy of self-attention in advancing multi-label CXR image classification.

The remainder of this paper is as follows: Section 2 reviews related work on lung disease classification using CXR images. Section 3 presents the proposed method. Section **Error! Reference source not found.** shows the experimental results and analysis. Finally, the paper concludes with a summary and directions for future work.

2. RELATED WORK

Deep learning has been extensively used for CXR-based lung disease diagnosis (Çallı et al., 2021; Hage Chehade et al., 2024; Koyyada & Singh, 2024). Galán-Cuenca et al. (2024) applied Siamese networks (Chicco, 2021) to handle data imbalance, improving the F1 score by 5.6%. Vo and Do (2024b) used contrastive learning with nonlinear classifiers, achieving 87.9% accuracy. Shelke et al. (2021) used VGG-16 (Simonyan & Zisserman, 2015) and DenseNet-161 (Huang et al., 2017) for Covid-19 detection, reaching 98.9%. Chen and Lin (2024) proposed a multi-task contrastive learning model for pneumonia and COVID-19. Adjei-Mensah et al. (2024) introduced Cov-Fed, a federated model with attention, achieving 87.65%. Poloju et al. (Poloju & Rajaram, 2025) combined ensemble methods with Emperor Penguin Optimization and SVM (Vapnik, 2000), reaching 97.5%. Verma et al. (2024) compared seven classifiers (Hastie et al., 2009) on CXR features like LBP, HOG, and pixel descriptors.

However, CXR images often exhibit multiple pathologies simultaneously, such as both pneumonia and lung cancer. Diagnoses may vary across doctors due to differing expertise, causing inconsistencies. Datasets like ChestX-ray14 (X. Wang et al., 2017), CheXpert (Irvin et al., 2019), VinDr-CXR (Nguyen et al., 2022), and Padchest (Bustos et al., 2020) reflect this reality by providing multi-label annotations. Therefore, multi-label classification is essential for detecting multiple conditions in a single image and improving diagnostic accuracy.

Multi-label CXR classification is challenging and has attracted much research interest (Hasanah et al., 2025). HydraViT (Öztürk et al., 2025) combines a transformer and multi-branch module to learn disease co-occurrence, improving AUC by up to 2.1%. Wang et al. (2024) used local and global graphs to model pathology correlations. Hasanah et al. (2024) fused CheXNet with the Feature Pyramid Network to extract multi-scale features. Vo and Do (2024a) applied self-supervised contrastive learning with SwinT-compact, reaching 0.809 AUC. Zhao and Wang (2025) used large-kernel CNNs and GCNs for long-range dependency and disease relation modeling. Lu et al. (2024) proposed CvTGNet, combining Vision Transformer (Dosovitskiy et al., 2020) and GCN to enhance CXR diagnosis.

3. INTEGRATING SELF-ATTENTION INTO DENSENET FOR MULTI-LABEL CHEST X-RAY IMAGE CLASSIFICATION

Our investigation aims to develop an advanced deep learning architecture for accurately classifying multi-label CXR images, as illustrated in Figure 1.

In past years, deep convolutional neural networks (CNNs) such as VGG-16 (Simonyan & Zisserman, 2015), DenseNet (Huang et al., 2017), ResNet (He et al., 2015), Inception (Szegedy et al., 2016), and EfficientNet (Tan & Le, 2021) have achieved remarkable success in image classification. However, these architectures are fundamentally limited by the locality of convolutional operations, which constrain their receptive regions and hinder the modeling of long-range dependencies. This limitation becomes critical in domains like CXR image analysis, where diagnostically relevant patterns are distributed across spatially distant regions of the image. Capturing such global context is essential for accurate multi-label classification in medical imaging.

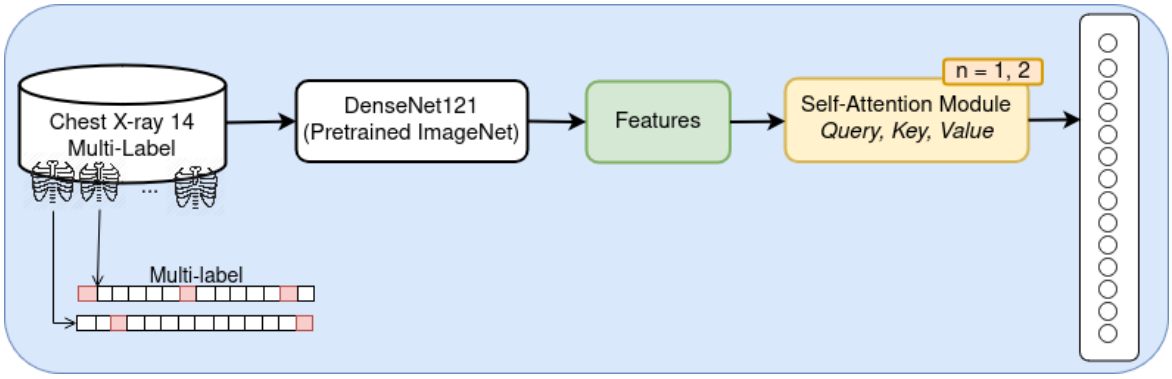

Figure 1. DenseNet with n Self-Attention (DNet-nSA) for multi-label chest X-ray images

To overcome the limitations of conventional CNNs in capturing global context, we propose enhancing deep neural networks with self-attention mechanisms. The main idea is to integrate self-attention into a CNN-based architecture to enable the model to explore and aggregate information dynamically across all spatial locations within the feature map. Unlike standard convolutional layers, which operate on fixed, local receptive regions, self-attention performs the network to model long-range dependencies by allowing each spatial position to attend to every other position. This global interaction is particularly beneficial for medical imaging tasks, such as CXR image analysis. By embedding self-attention into convolutional backbones, the model gains the ability to reason holistically about the image content, leading to improved performance in complex multi-label classification scenarios.

The underlying mechanics of self-attention are detailed as follows. Each self-attention block is based on the scaled dot-product attention mechanism, a core component of Transformer architectures. Given an input feature map $F \in R^{H\times W\times C}$, the block first projects it into three distinct representations: query Q, key K, and value V, via learned linear transformations implemented as dense layers (see Equation 1). These tensors are reshaped to a 2D format to facilitate efficient attention computation over spatial dimensions.

$$Q = FW_Q, \quad K = FW_K, \quad V = FW_V \qquad (1)$$

The attention weights are obtained by computing pairwise similarities between the query and key vectors, scaled by the dimensionality of the query space to ensure numerical stability. A softmax operation is applied to yield a normalized attention map, which is then used to aggregate the value representations. This process enables the model to incorporate information from all spatial locations, effectively capturing long-range dependencies and global context (see Equation 2).

$$A = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right), \quad O = AV \qquad (2)$$

The output of the attention module is then reshaped back to the original spatial dimensions and fused with the input feature map via a residual connection. Finally, a layer normalization step is applied to stabilize training dynamics and promote better gradient flow across layers (see Equation 3).

$$\hat{F} = \text{Reshape}(O) + F, \text{Output} = \text{LayerNorm}\left(\hat{F}\right) (3)$$

This architecture enhances the representational power of convolutional backbones like DenseNet, particularly in tasks where spatially distant features are semantically correlated, such as in CXR image analysis.

Stacking multiple self-attention layers allows the model to iteratively refine its understanding of spatial dependencies, enhancing its capacity progressively to model complex, spatially distributed patterns.

Algorithm 1 illustrates how to build the model for DenseNet enhanced with stacked self-attention blocks. A base model (i.e. DenseNet121) extracts intermediate features, which are refined through self-attention to capture global context, for improving multi-label CXR image classification.

Algorithm 1: DenseNet with stacked Self-attention blocks (DNet-nSA)

Input:
Image shape (H, W, C), training data (X_{train}, y_{train}), num_labels
Base model $DenseNet$,
Number of blocks n

Output:
Model

Function `BuildModel`($input\_shape, num\_labels$):
  $DNet$ = DenseNet121(weights="imagenet", include_top=False, input_shape=input_shape);
  $F \leftarrow$ output feature map from base model;
  $F' \leftarrow$ `StackSelfAttention`(F, num_blocks=n) ; // Stack nSA
  $x \leftarrow$ GlobalAveragePooling2D(F');
  $output \leftarrow$ Dense(num_labels, activation="sigmoid")(x);
  Create $Model$ with $input$ and $output$;
  return $Model$
Function `StackSelfAttention`(F, num_blocks):
  for $i \leftarrow 1$ **to** num_blocks **do**
    $F \leftarrow$ `SelfAttentionBlock`(F)
  return F
Function `SelfAttentionBlock`(F):
  $C \leftarrow$ number of channels in F;
  Compute $Q = \text{Dense}(C/8)(F)$;
  Compute $K = \text{Dense}(C/8)(F)$;
  Compute $V = \text{Dense}(C)(F)$;
  Reshape Q, K, V to 2D shape $(H \times W, C)$;
  $A = \text{Softmax}\left(\frac{QK^\top}{\sqrt{C/8}}\right)$; // Attention map
  $O = AV$; // Attention output
  Reshape O to shape of F;
  $\hat{F} \leftarrow F + O$; // Residual connection
  Normalize $\hat{F}$ with LayerNorm;
  return $\hat{F}$

4. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed method for multi-label CXR classification, we assess model performance using the Area Under the ROC Curve (AUC). As a standard metric for multi-label tasks, AUC provides a robust measure of the model's discriminative ability across all classes, offering comprehensive insight into its predictive performance.

We first developed a fine-tuned DenseNet (Huang et al., 2017) baseline, optimized for the multi-label classification of CXR images. Building upon this foundation, we created DNet-nSA, an enhanced architecture that incorporates stacked self-attention blocks atop the original DenseNet backbone to capture global contextual dependencies. Moreover, we proposed the DNet-MHSA architecture for the integration of a DNet121 backbone with a multi-head self-attention mechanism. The training DNet-nSA, DNet-MHSA were implemented in Python using the Keras API (Chollet, 2015) with TensorFlow (Abadi et al., 2016) as the backend, leveraging GPU acceleration for efficient computation.

We are particularly interested in comparing our method with the contrastive learning approach MoCoR101 (Sowrirajan et al., 2021), which utilizes MoCo with a ResNet101 backbone (He et al., 2015), and with the self-supervised learning model MoBYSwinT (Vo & Do, 2024a), based on MoBY (Xie et al., 2021) and the Swin Transformer architecture (Liu et al., 2021).

All experiments were conducted on a high-performance Ubuntu 22.04 system with an Intel Core i7-14700K CPU (20 cores, 64 GB RAM) for preprocessing. Model training leveraged a ROG Strix RTX 4090 GPU (24 GB VRAM, 16,384 CUDA cores) to accelerate deep learning computations.

4.1. Chest X-Ray image dataset

We evaluated the proposed method on the ChestX-Ray14 dataset (Wang et al., 2017), a large-scale NIH-released collection of 112,120 frontal CXR images from 30,805 patients (from 1992 to 2015), annotated with 14 thoracic pathologies. The dataset is split into training (70%), validation (15%), and test (15%) sets, with 78,484, 16,818, and 16,818 images, respectively (see Table 1). Each image is multi-labelled based on the presence of one or more conditions (e.g., Atelectasis, Effusion, Pneumonia), facilitating the training of multi-label classification models (see examples in Figure 2). Labels are binary per class (0: absence, 1: presence).

Table 1. Multi-label Chest X-Ray 14 dataset

Class	Train (0)	Train (1)	Valid (0)	Valid (1)	Test (0)	Test (1)
Class 0	70,387	8,097	15,080	1,738	15,094	1,724
Class 1	76,528	1,956	16,403	415	16,413	405
Class 2	69,116	9,368	14,812	2,006	14,875	1,943
Class 3	64,569	13,915	13,844	2,974	13,813	3,005
Class 4	74,456	4,028	15,934	884	15,948	870
Class 5	74,054	4,430	15,873	945	15,862	956
Class 6	77,488	996	16,604	214	16,597	221
Class 7	74,789	3,695	15,997	821	16,032	786
Class 8	75,235	3,249	16,129	689	16,089	729
Class 9	76,863	1,621	16,485	333	16,469	349
Class 10	76,724	1,760	16,446	372	16,434	384
Class 11	77,305	1,179	16,557	261	16,572	246
Class 12	76,096	2,388	16,277	541	16,362	456
Class 13	78,327	157	16,789	29	16,777	41

(a) Mass, Pleural Thickening (b) Effusion, Consolidation, Edema (c) Atelectasis, Pleural Thickening, Effusion, Pneumothorax

Figure 2. Samples of Chest X-Ray images with multi-label annotations

4.2. Tuning parameters

Fine-tuning was performed by retraining the top layers of each architecture, with the best performance achieved by updating the top 10 layers of DenseNet121. All models, including the baseline DNN, the proposed DNN with self-attention blocks (DNet-nSA, DNet-MHSA), MoCoR101, and MoBYSwinT, were trained using a batch size of 32 for 20 epochs with early stopping, optimized with Adam (learning rate 0.0001). The AUC metric was employed as the optimization objective to guide learning in the multi-label classification setting.

4.3. Classification results

Table 2 presents the Area Under the Curve (AUC) performance of six different models for multi-label classification of CXR images across 14 classes. The compared models include MoCoR101 (A ResNet-101 backbone pre-trained using MoCo), MoBYSwinT (A Swin Transformer pre-trained with MoBY), DNet121 (DenseNet-121 baseline), DNet-1SA and DNet-2SA (DenseNet with one and two self-attention blocks), and DNet-MHSA (DenseNet with Multi-Head Self-Attention). The highest average AUC is highlighted in bold, and the second-highest is shown in italics. Figure 3 visualizes the average AUC per model.

Table 2. Multi-label CXR image classification results of different models

Class	MoCoR101	MoBYSwinT	DNet121	DNet-1SA	DNet-2SA	DNet-MHSA
Class 0	0.775	0.779	0.803	0.796	0.780	0.806
Class 1	0.868	0.906	0.899	0.911	0.891	0.859
Class 2	0.862	0.858	0.879	0.889	0.884	0.880
Class 3	0.696	0.699	0.716	0.709	0.720	0.714
Class 4	0.780	0.799	0.839	0.849	0.851	0.837
Class 5	0.695	0.708	0.761	0.751	0.776	0.752
Class 6	0.714	0.746	0.736	0.737	0.756	0.713
Class 7	0.845	0.850	0.873	0.877	0.873	0.866
Class 8	0.776	0.791	0.802	0.796	0.795	0.773
Class 9	0.874	0.887	0.875	0.882	0.890	0.864
Class 10	0.843	0.833	0.876	0.865	0.882	0.875
Class 11	0.764	0.785	0.756	0.776	0.789	0.781
Class 12	0.749	0.769	0.758	0.763	0.783	0.775
Class 13	0.864	0.923	0.734	0.822	0.838	0.867
Average	0.793	0.809	0.807	*0.816*	**0.822**	0.811

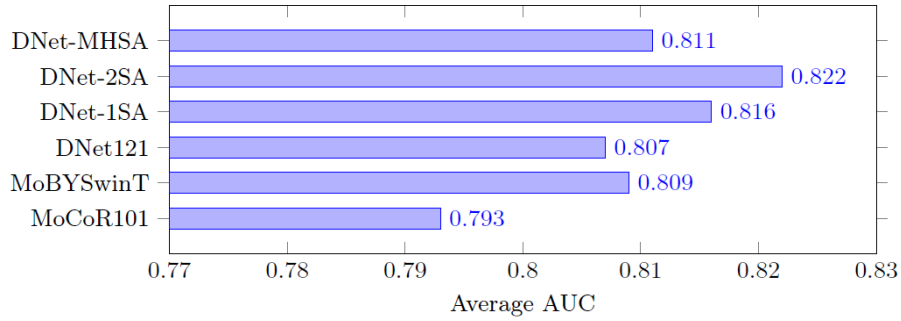

Figure 3. Comparison of average AUC across different models

Among all models, DNet-2SA achieves the highest overall performance with an average AUC of 0.822. This demonstrates the effectiveness of incorporating two self-attention blocks into the DenseNet architecture. The improvement is consistent across most classes, suggesting that deeper attention can help the model capture more complex relationships in medical images. DNet-1SA follows with an average AUC of 0.816, indicating that even a single self-attention block provides significant gains over the baseline DNet121, which records an average AUC of 0.807. Interestingly, DNet-MHSA achieves 0.811, slightly lower than DNet-2SA, suggesting that multi-head self-attention does not necessarily outperform simpler attention schemes in this setting.

The MoBYSwinT model, which uses a Swin Transformer pre-trained with self-supervised learning, performs well with an average AUC of 0.809. It even outperforms DNet121, highlighting the potential of transformer-based models in medical image analysis. However, MoBYSwinT still falls short of the attention-enhanced DenseNet variants. Notably, MoBYSwinT is the best in Class 13, achieving an AUC of 0.923, significantly outperforming all other models for that class. This may reflect the capacity of transformer models to better capture long-range dependencies or underrepresented patterns when provided with sufficient pretraining. On the other hand, MoCoR101 based on ResNet-101 and MoCo

pretraining, shows the lowest overall performance (average AUC of 0.793), suggesting that its self-supervised features are less transferable to the medical imaging domain compared to other methods evaluated.

From a class-wise perspective, Class 1 and Class 9 consistently achieve high AUCs across all models, often exceeding 0.87. In contrast, Class 3, Class 5, and Class 6 are among the most challenging, with AUCs generally below 0.78. These disparities may arise from class imbalance, ambiguous visual features, or high inter-class similarity in those disease types. With a range from 0.734 using DNet121 to 0.923 using MoBYSwinT, Class 13 demonstrates a significant difference in performance between models, suggesting the models are sensitive to feature extraction methods within this class.

Overall, the results suggest that attention mechanisms integrated into CNN backbones, particularly DNet-2SA, offer the most reliable performance for multi-label CXR classification. Although transformer-based models like MoBYSwinT show promising results, especially on certain difficult classes, they may require larger datasets or more specialized fine-tuning to fully exploit their advantages. DenseNet-121 remains a strong baseline, and the consistent performance gains observed with added attention validate the use of attention-based enhancements in medical image analysis. Future work may focus on hybrid architectures that combine CNN and transformer components, or ensemble methods that leverage the strengths of both model families to further boost classification accuracy across all classes.

5. CONCLUSION AND FUTURE WORKS

We presented DNet-nSA, a novel architecture that integrates self-attention mechanisms into DenseNet to improve multi-label classification of CXR images. Our approach addresses a key limitation of traditional CNNs, by embedding n self-attention blocks to enhance spatial feature representation, their inability to effectively model long-range dependencies and global context. We introduced two variants, DNet-1SA and DNet-2SA, and demonstrated their effectiveness on the ChestX-ray14 dataset. The proposed models outperformed several strong baselines, including the original DenseNet, the contrastive learning approach MoCoR101, and the self-supervised learning model MoBYSwinT. Notably, DNet-2SA achieved an AUC of 0.822, demonstrating the benefit of incorporating self-attention for multi-label CXR image classification. These results confirm that self-attention provides valuable enhancements to convolutional architectures in the context of medical image classification.

In future work, we plan to extend DNet-nSA by exploring adaptive attention placement and dynamic routing strategies for better computational efficiency and performance trade-offs. We also aim to evaluate the model on additional multi-label CXR datasets such as CheXpert and VinDr-CXR to further validate its generalizability across diverse clinical settings. Finally, integrating explainability techniques to visualize attention maps may enhance interpretability and support clinical decision-making.

ACKNOWLEDGMENT

This research has received support from the Vietnamese Ministry of Education and Training's scientific research project, code B2025-TCT-01. We would like to thank the College of Information Technology, Can Tho University, very much.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). *TensorFlow: A system for large-scale machine learning* (No. arXiv:1605.08695). arXiv. https://doi.org/10.48550/arXiv.1605.08695

Adjei-Mensah, I., Zhang, X., Agyemang, I. O., Yussif, S. B., Baffour, A. A., Cobbinah, B. M., Sey, C., Fiasam, L. D., Chikwendu, I. A., & Arhin, J. R. (2024). Cov-Fed: Federated learning-based framework for COVID-19 diagnosis using chest X-ray scans. *Engineering Applications of Artificial Intelligence*, *128*, 107448. https://doi.org/10.1016/j.engappai.2023.107448

Bustos, A., Pertusa, A., Salinas, J.-M., & de la Iglesia-Vayá, M. (2020). PadChest: A large chest X-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, *66*, 101797. https://doi.org/10.1016/j.media.2020.101797

Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K. G., & Murphy, K. (2021). Deep learning for chest X-ray analysis: A survey. *Medical*

Image Analysis, *72*, 102125. https://doi.org/10.1016/j.media.2021.102125

Chen, G.-Y., & Lin, C.-T. (2024). Multi-task supervised contrastive learning for chest X-ray diagnosis: A two-stage hierarchical classification framework for COVID-19 diagnosis. *Applied Soft Computing*, *155*, 111478. https://doi.org/10.1016/j.asoc.2024.111478

Chicco, D. (2021). Siamese Neural Networks: An Overview. In H. Cartwright (Ed.), *Artificial Neural Networks* (pp. 73–94). Springer US. https://doi.org/10.1007/978-1-0716-0826-5_3

Chollet, F. (2015). *Keras*. Seattle, WA, USA.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020, October 22). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv.Org. https://arxiv.org/abs/2010.11929v2

Galán-Cuenca, A., Gallego, A. J., Saval-Calvo, M., & Pertusa, A. (2024). Few-shot learning for COVID-19 chest X-ray classification with imbalanced data: An inter vs. intra domain study. *Pattern Analysis and Applications*, *27*(3), 69. https://doi.org/10.1007/s10044-024-01285-w

Hage Chehade, A., Abdallah, N., Marion, J.-M., Hatt, M., Oueidat, M., & Chauvet, P. (2024). A systematic review: Classification of lung diseases from chest X-ray images using deep learning algorithms. *SN Computer Science*, *5*(4), 405. https://doi.org/10.1007/s42979-024-02751-2

Hasanah, U., Avian, C., Darmawan, J. T., Bachroin, N., Faisal, M., Prakosa, S. W., Leu, J.-S., & Tsai, C.-T. (2024). CheXNet and feature pyramid network: A fusion deep learning architecture for multilabel chest X-Ray clinical diagnoses classification. *The International Journal of Cardiovascular Imaging*, *40*(4), 709–722. https://doi.org/10.1007/s10554-023-03039-x

Hasanah, U., Leu, J.-S., Avian, C., Azmi, I., & Prakosa, S. W. (2025). A systematic review of multilabel chest X-ray classification using deep learning. *Multimedia Tools and Applications*, *84*(23), 26719–26753. https://doi.org/10.1007/s11042-024-20172-4

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (No. arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). *Densely connected convolutional networks*. 4700–4708. https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison* (No. arXiv:1901.07031). arXiv. https://doi.org/10.48550/arXiv.1901.07031

Koyyada, S. P., & Singh, T. P. (2024). A Systematic Survey of Automatic Detection of Lung Diseases from Chest X-Ray Images: COVID-19, Pneumonia, and Tuberculosis. *SN Computer Science*, *5*(2), 229. https://doi.org/10.1007/s42979-023-02573-8

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows* (No. arXiv:2103.14030). arXiv. https://doi.org/10.48550/arXiv.2103.14030

Lu, Y., Hu, Y., Li, L., Xu, Z., Liu, H., Liang, H., & Fu, X. (2024). *CvTGNet: A novel framework for chest X-ray multi-label classification*. https://doi.org/10.1145/3649153.3649216

Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T. T., Dinh, D. H., Do, C. D., Doan, L. T., Nguyen, C. N., Nguyen, B. T., Nguyen, Q. V., Hoang, A. D., Phan, H. N., Nguyen, A. T., Ho, P. H., … Vu, V. (2022). VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, *9*(1), 429. https://doi.org/10.1038/s41597-022-01498-w

Öztürk, Ş., Turalı, M. Y., & Çukur, T. (2025). HydraViT: Adaptive multi-branch transformer for multi-label disease classification from chest X-ray images. *Biomedical Signal Processing and Control*, *100*, 106959. https://doi.org/10.1016/j.bspc.2024.106959

Poloju, N., & Rajaram, A. (2025). Hybrid technique for lung disease classification based on machine learning and optimization using X-ray images. *Multimedia Tools and Applications*, *84*(21), 23531–23553. https://doi.org/10.1007/s11042-024-19959-2

Shelke, A., Inamdar, M., Shah, V., Tiwari, A., Hussain, A., Chafekar, T., & Mehendale, N. (2021). Chest X-ray classification using deep learning for automated COVID-19 screening. *SN Computer Science*, *2*(4), 300. https://doi.org/10.1007/s42979-021-00695-5

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition* (No. arXiv:1409.1556). arXiv. https://doi.org/10.48550/arXiv.1409.1556

Sowrirajan, H., Yang, J., Ng, A. Y., & Rajpurkar, P. (2021). MoCo Pretraining Improves Representation and transferability of chest X-ray models. *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, 728–744.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the inception architecture for computer vision*. 2818–2826.

Tan, M., & Le, Q. (2021). EfficientNetV2: Smaller models and faster training. *Proceedings of the 38th International Conference on Machine Learning*, 10096–10106. https://proceedings.mlr.press/v139/tan21a.html

Vapnik, V. (2000). *The nature of statistical learning theory* (2nd ed.).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Verma, S., Devarajan, G. G., & Sharma, P. K. (2024). Comparative evaluation of feature extraction techniques in chest X ray image with different classification model. In D. Garg, J. J. P. C. Rodrigues, S. K. Gupta, X. Cheng, P. Sarao, & G. S. Patel (Eds.), *Advanced Computing* (pp. 197–209). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56703-2_17

Vo, T.-T., & Do, T.-N. (2024a). Enhancing efficiency of multi-label X-ray image classification with self-supervised learning based on compact swin transformers. In T. K. Dang, J. Küng, & T. M. Chung (Eds.), *Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications* (pp. 153–167). Springer Nature. https://doi.org/10.1007/978-981-96-0434-0_11

Vo, T.-T., & Do, T.-N. (2024b). Improving chest X-ray image classification via integration of self-supervised learning and machine learning algorithms. *Journal of Information and Communication Convergence Engineering*, *22*, 165–171. https://doi.org/10.56977/jicce.2024.22.2.165

Wang, G., Wang, P., & Wei, B. (2024). Multi-label local awareness and global co-occurrence priori learning improve chest X-ray classification. *Multimedia Systems*, *30*(3), 132. https://doi.org/10.1007/s00530-024-01321-z

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017, May 5). *ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases*. arXiv.Org. https://doi.org/10.1109/CVPR.2017.369

Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., & Hu, H. (2021). *Self-supervised learning with swin transformers* (No. arXiv:2105.04553). arXiv. https://doi.org/10.48550/arXiv.2105.04553

Zhao, X., & Wang, X. (2025). Multi-label chest X-ray image classification based on long-range dependencies capture and label relationships learning. *Biomedical Signal Processing and Control*, *100*, 107018. https://doi.org/10.1016/j.bspc.2024.107018