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

DOI:10.22144/ctujoisd.2025.050

BagViT: Bagged vision transformers for classifying chest X-ray images

Thi-Diem Truong[1,2,3], and Thanh-Nghi Do[1,4*]

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

[2]Faculty of Information Technology, An Giang University, Viet Nam

[3]Vietnam National University Ho Chi Minh City, Viet Nam

[4]UMI UMMISCO 209, IRD/UPMC, France

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

Article info.

Received 13 Jul 2025
Revised 15 Aug 2025
Accepted 2 Oct 2025

Keywords

Bagging, deep learning, lung disease classification, vision transformer (ViT), X-ray images

ABSTRACT

In this paper, we propose a novel ensemble method, termed Bagged Vision Transformers (BagViT), to enhance the classification accuracy for Chest X-ray (CXR) images. BagViT constructs an ensemble of independent Vision Transformer (ViT) models, each of which is trained on a bootstrap sample (sampling with replacement) drawn from the original training dataset. To enhance model diversity, we use MixUp to generate synthetic training examples and introduce training randomness by varying the number of training epochs and selectively fine-tuning the top layers of each model. Final predictions are obtained through majority voting. Experimental results on a real-world dataset collected from Chau Doc Hospital (An Giang, Vietnam) demonstrate that BagViT significantly outperforms fine-tuned baselines such as VGG16, ResNet, DenseNet, ViT. Our BagViT achieves a classification accuracy of 72.25%, highlighting the effectiveness of ensemble learning with transformer architectures in scenarios with complex CXR images.

1. INTRODUCTION

Lung diseases represent a critical and enduring challenge to global public health. Prevalent conditions such as asthma, pneumonia, tuberculosis, lung cancer, and chronic obstructive pulmonary disease (COPD) contribute substantially to global morbidity and remain among the leading causes of mortality (Yadav et al., 2023). According to the World Health Organization (WHO), pneumonia was the foremost cause of death among children under five years of age in 2019, accounting for approximately 740,180 fatalities, representing 14% of all deaths in this age group (World Health Organization, 2022). Similarly, the Global Asthma Report 2022 highlights asthma as a pervasive chronic illness, affecting an estimated 262 million individuals and resulting in over 1,000 deaths each day (Global Asthma Network, 2022).

Quick and accurate diagnosis is essential for guiding effective clinical decision-making and improving patient prognosis in cases of pulmonary disease. Chest X-ray (CXR) imaging remains a foundation in the diagnostic workflow for respiratory conditions, owing to its ability to capture high-resolution structural information of the thoracic cavity. CXR imaging facilitates the detection of various pulmonary pathologies, including pneumonia, tuberculosis, and malignancies such as lung cancer. Compared to more advanced imaging modalities such as computed tomography (CT) or magnetic resonance imaging (MRI), CXR offers a cost-effective and widely accessible solution, particularly in resource-limited settings.

Despite its advantages, the diagnostic accuracy of CXR imaging is inherently dependent on the interpretive skill of the clinician. The nuanced visual patterns associated with different pulmonary diseases demand considerable expertise, and even experienced radiologists may be prone to inter-observer variability or diagnostic oversight. Therefore, enhancing the reliability and consistency of CXR interpretation remains a pressing objective in modern medical imaging and computer-aided diagnosis.

In this paper, we propose Bagged Vision Transformers (BagViT), an ensemble learning framework based on Vision Transformers (ViT) (Dosovitskiy et al., 2020). BagViT constructs a collection of ViT models, each trained independently on a bootstrap sample drawn from the original training set. To introduce model diversity and enhance generalization, each ViT is trained for a randomly selected number of epochs (ranging from 8 to 12), and selectively fine-tuned on top layers sampled from 36, 38, 40, 42, 44. Furthermore, the MixUp technique (Zhang et al., 2018) is used to generate synthetic training examples. Final predictions are obtained through majority voting over the ensemble members. This ensemble strategy mitigates overfitting and leverages the representational power of transformer architectures, making it well-suited for high-variance classification tasks such as CXR image recognition.

Empirical evaluation on a real-world dataset collected from Chau Doc General Hospital (An Giang province, Viet Nam) demonstrates that BagViT achieves superior performance compared to fine-tuned baselines. Specifically, BagViT outperforms VGG16 (Simonyan & Zisserman, 2014), ResNet (He et al., 2016), DenseNet (Huang et al., 2017), standalone ViT (Dosovitskiy et al., 2020), achieving a top-1 classification accuracy of 72.25%.

The rest of this paper is structured as follows. Section 2 reviews related work on CXR image classification. Section 3 presents our proposed BagViT algorithm for effective CXR image classification. Section 4 shows the experimental results, followed by conclusions and future work in Section 5.

2. RELATED WORK

Deep learning has emerged as a foundation in chest X-ray (CXR)-based lung disease diagnosis, with numerous research demonstrating its effectiveness across diverse architectures and learning paradigms (Callı et al., 2021; Hage Chehade et al., 2024; Koyyada & Singh, 2024). Do et al. (2022) proposed a hybrid approach that combines fine-tuned pre-trained deep networks with a support vector machine (SVM) (Vapnik, 1995) classifier for detecting COVID-19 from CXR images. By leveraging the nonlinear decision boundaries of SVMs on top of deep networks, the method achieved superior performance, reaching a classification accuracy of 96.16%, outperforming all individual deep models. To address the pervasive issue of data imbalance in medical imaging, Galán-Cuenca et al. (2024) used Siamese networks (Chicco, 2021), achieving a 5.6% improvement in F1 score by leveraging similarity-based learning. Vo and Do (2024) integrated contrastive learning with nonlinear classifiers, resulting in an accuracy of 87.9%, while Shelke et al. (2021) applied deep convolutional architectures such as VGG-16 (Simonyan & Zisserman, 2014) and DenseNet-161 (Huang et al., 2017), reaching a detection accuracy of 98.9% for COVID-19. Chen and Lin (2024) proposed a multi-task contrastive learning framework that jointly addressed pneumonia and COVID-19 detection, demonstrating the utility of task-aware representations in CXR analysis.

Recent advances have also focused on privacy-preserving and optimization-driven strategies. Adjei-Mensah et al. (2024) introduced Cov-Fed, a federated learning model incorporating attention mechanisms, achieving an accuracy of 87.65% while preserving data privacy across institutions. Poloju and Rajaram (Poloju & Rajaram, 2024) proposed a hybrid model combining ensemble learning, Emperor Penguin Optimization, and SVM, reporting a high diagnostic accuracy of 97.5%. Additionally, Verma et al. (2024) conducted a comparative study of classical machine learning classifiers (Hastie et al., 2009) using handcrafted CXR features.

Despite these advances, CXR image classification remains inherently challenging. The domain suffers from a lack of large-scale annotated datasets, as CXR is often region-specific and requires domain expertise for accurate labeling. Moreover, the subtle visual differences between some pathological classes, like the overlapping radiographic features of pneumonia and other lung diseases, result in low inter-class variance, which makes the discrimination task difficult. In parallel, the intrinsic complexity of CXR images, which often include anatomical overlaps, imaging artifacts, and comorbid

conditions, introduces significant noise into the data. These factors collectively limit the effectiveness of conventional learning algorithms and emphasize the necessity for robust, data-efficient models capable of extracting discriminative features under noisy and ambiguous conditions.

3. PROPOSED APPROACH

Our investigation aims to create a robust machine learning method for accurately classifying CXR images. The task presents several challenges, including the limited size and heterogeneity of annotated CXR datasets, high visual similarity among certain radiographic features of pneumonia and other lung diseases, and substantial variability in image quality arising from inconsistent conditions.

Vision Transformer (ViT) (Dosovitskiy et al., 2020) has demonstrated state-of-the-art performance across various vision tasks due to its ability to:

Self-attention mechanisms model long-range dependencies;

– Learn rich semantic representations from images without requiring convolutional inductive biases;

– Be scalable and pre-trainable, enabling fine-tuning for downstream tasks like CXR image classification.

However, ViTs generally require extensive training data to achieve robust generalization. In the context of CXR images, where datasets are often small and diverse, directly training a ViT can result in significant overfitting and increased output variance.

In machine learning, model generalization performance is analyzed by the bias–variance decomposition (Breiman, 1996): bias measures errors from incorrect model assumptions, while variance reflects sensitivity to variations in the training data.

ViTs, being highly expressive, tend to have low bias but high variance, especially when trained on small or heterogeneous annotated datasets, such as CXR images. This leads to unstable predictions and poor generalization.

Therefore, our proposed Bagged Vision Transformers (BagViT) framework mitigates the high variance typically observed in small-scale and diversity of CXR image datasets by integrating three complementary strategies: data diversity via bootstrap sampling, model diversity through randomized training configurations, and robust prediction aggregation using majority voting.

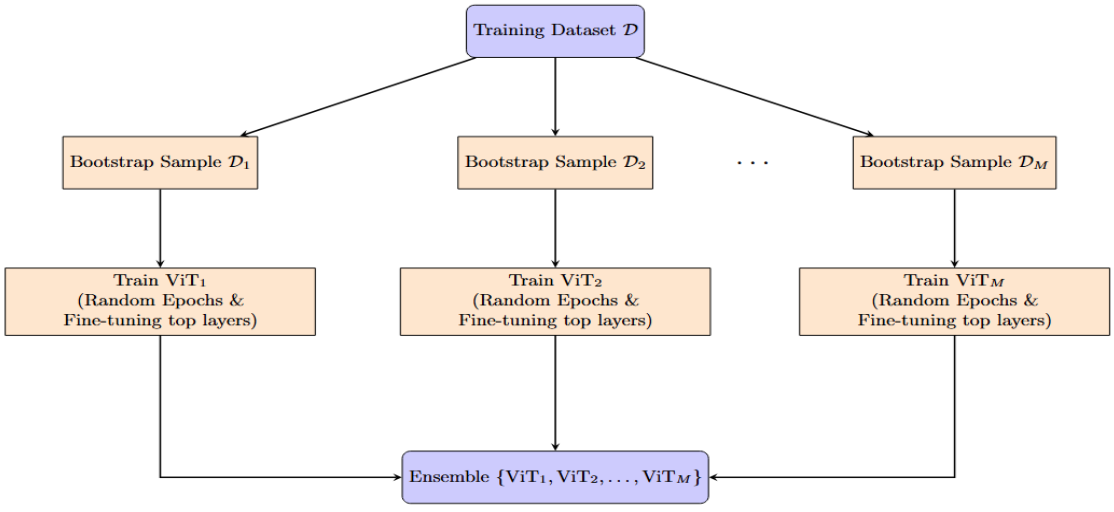

Figure 1. BagViT algorithm for training an ensemble of ViT models on bootstrap samples with random training settings

The BagViT described in Algorithm 1 and Figure 1, learns an ensemble of ViT models to improve classification performance on limited image datasets. Each ViT model is trained independently on a bootstrap sample drawn with replacement from the original training CXR image dataset,

introducing data diversity. For every model, a random number of training epochs (between 8 and 12), selectively fine-tuning top layers sampled from 36, 38, 40, 42, 44, and the data augmentation MixUp (Zhang et al., 2018) are used to further increase variance reduction. This randomness in both data and training configuration helps mitigate overfitting and stabilizes generalization, especially under small-scale and diversity of CXR image data regimes.

Algorithm 1: BagViT for training an ensemble of ViT models

Input: Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, number of models M
Output: Ensemble of Vision Transformer (ViT) models $\{\text{ViT}_1, \ldots, \text{ViT}_M\}$

for $m \leftarrow 1$ **to** M **do**
  Sample bootstrap dataset $\mathcal{D}_m$ from $\mathcal{D}$ with replacement;
  Initialize ViT model ViT_m;
  Randomly choose number of training epochs $E_m \sim \mathcal{U}\{8, 9, 10, 11, 12\}$;
  Randomly choose fine-tuning top layers $L_m \sim \mathcal{V}\{36, 38, 40, 42, 44\}$;
  Configure ViT model ViT_m with the MLP head having num_classes and L_m fine-tuning top layers;
  Train ViT_m on $\mathcal{D}_m$ using data augmentation MixUp for E_m epochs;
return $\{\text{ViT}_1, \text{ViT}_2, \ldots, \text{ViT}_M\}$;

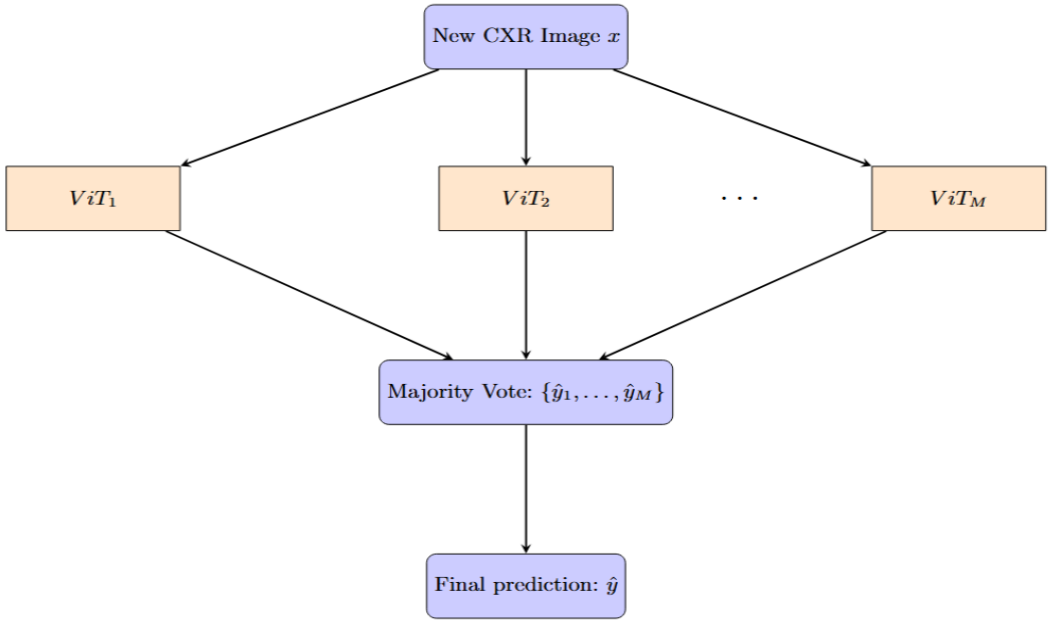

Figure 2. BagViT prediction of a new CXR image via majority voting

The class prediction of a new CXR image is described in Algorithm 2 and Figure 2 is then determined by majority voting over predictions produced by all trained ViT models.

The BagViT framework offers several key advantages for CXR image classification:

1. Reduced variance: By aggregating predictions from multiple diverse models, BagViT stabilizes ViT outputs, mitigating sensitivity to minor data variations.

2. Enhanced generalization: The ensemble improves performance on unseen CXR images, particularly in cases of domain shifts or limited training data.

3. Increased robustness to overfitting: Bootstrap sampling prevents any single model from overfitting the entire data distribution, minimizing the impact of rare or biased samples.

Algorithm 2: BagViT prediction via majority voting

Input: Ensemble of ViT models $\{\text{ViT}_1, \text{ViT}_2, \ldots, \text{ViT}_M\}$, input image x
Output: Predicted class label $\hat{y}$

```
for m ← 1 to M do
    ŷ_m ← ViT_m(x) ;                    // Predict with m-th ViT model
ŷ ← MajorityVote({ŷ_1, ŷ_2, ..., ŷ_M});
return ŷ;
```

4. EXPERIMENTAL RESULTS

4.1. Technical implementation

To assess the effectiveness of the proposed method for classifying CXR images, we developed and implemented BagViT using Python. The model architecture and training procedures were built with PyTorch (Paszke et al., 2019) and the Keras deep learning library (Chollet et al., 2015), using TensorFlow (Abadi et al., 2015) as the computational backend for efficient GPU-accelerated operations. We employed Scikit-learn (Pedregosa et al., 2011) for machine learning utilities and OpenCV (Itseez, 2015) for image processing tasks.

We also aim to evaluate the performance of the BagViT algorithm against various fine-tuning strategies applied to deep neural networks, such as VGG16 (Simonyan & Zisserman, 2014), ResNet (He et al., 2016), DenseNet (Huang et al., 2017), and ViT (Dosovitskiy et al., 2020).

4.2. Hardware setup

All experiments were performed on a high-performance Linux workstation running Ubuntu 22.04. The system was equipped with an Intel(R) Core i7-14700K processor (3.4 GHz, 20 cores), 64 GB of RAM, and a ROG Strix GeForce RTX 4090 GPU with 24 GB of GDDR6X VRAM and 16,384 CUDA cores, delivering significant acceleration for matrix computations and deep learning tasks.

4.3. Dataset description

We used a chest X-ray (CXR) dataset collected from the General Hospital of An Giang Province, developed as part of our prior work (Truong et al., 2024). The dataset comprises 17,973 CXR images extracted from the hospital's electronic medical records (EMRs), with all images undergoing standardized preprocessing prior to model training and evaluation. Ground-truth labels were derived from discharge diagnoses documented in the EMRs, including 12 pulmonary classes: Normal, Chronic Obstructive Pulmonary Disease (COPD), COVID-19, Asthma, Tuberculosis, Pulmonary Edema, Respiratory Failure, Pleural Effusion, Pneumothorax, Malignant Neoplasm, Pneumonia, and Pulmonary Collapse. The dataset was randomly partitioned into a training set containing 14,371 images and a test set with 3,602 images, as detailed in Table 1.

4.4. Tuning parameters

To optimize the performance of the deep neural networks, we used selective fine-tuning strategies tailored to the architectural characteristics of each model. Specifically, we fine-tuned the top 15 layers of the VGG16 network, targeting the deeper convolutional blocks responsible for high-level feature abstraction. For ResNet, we updated the top 100 layers, leveraging the residual structure to adapt mid-to-high-level features without destabilizing the earlier representations. In the case of DenseNet, we fine-tuned the top 50 layers to recalibrate densely connected feature pathways while preserving the lower-layer feature reuse. For the Vision Transformer (ViT), we fine-tuned the top 40 transformer layers, focusing on the deeper self-attention blocks to adapt high-level token interactions to the target domain. This layer-specific tuning approach balances model plasticity and stability, facilitating efficient domain adaptation while mitigating overfitting.

Table 1. Description of the CXR image dataset

No	Label	Train set	Test set
1	Normal	2,469	618
2	COPD	490	123
3	Covid-19	2,000	501
4	Asthma	153	39
5	Tuberculosis	657	165
6	Pulmonary Oedema	149	38
7	Respiratory Failure	1,105	277
8	Pleural Effusion	452	114
9	Pneumothorax	213	54
10	Malignant Neoplasm	128	33
11	Pneumonia	6,472	1,618
12	Pulmonary Collapse	83	22
	Total	14,371	3,602

During the training of VGG16, ResNet, DenseNet, and ViT networks, we used the Adam optimizer with a learning rate of 0.0001 and trained the models for 50 epochs.

BagViT trains 50 ViT models, each set up with a randomly chosen number of epochs ranging from 8 to 12 and a number of top layers for fine-tuning randomly selected from 36, 38, 40, 42, 44.

4.5. Classification results for CXR images

Fine-tuned models VGG16, ResNet, DenseNet, and ViT are referred to as FT-VGG16, FT-ResNet, FT-DenseNet, and FT-ViT, respectively. Our Bagged Vision Transformers model is denoted as BagViT.

We report the overall classification accuracy for CXR images in Table 2 and Figure 3. The highest accuracy is highlighted in bold, and the second-highest is shown in italics.

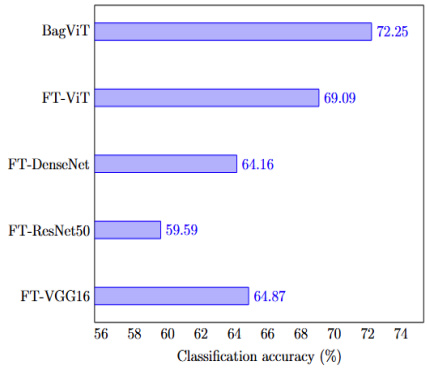

Figure 3. Overall classification accuracy for CXR images

FT-VGG16 and FT-DenseNet, among the baseline fine-tuned convolutional networks, show similar performance, with accuracies of 64.87% and 64.16%, respectively. FT-ResNet50 underperforms slightly, achieving an accuracy of 59.59%, which may be attributed to its deeper architecture requiring more data to fine-tune effectively or potential overfitting due to the dataset scale.

Table 2. Overall classification accuracy for CXR images

No	Visual approach	Accuracy (%)
1	FT-VGG16	64.87
2	FT-ResNet50	59.59
3	FT-DenseNet	64.16
4	FT-ViT	*69.09*
5	BagViT	**72.25**

Notably, FT-ViT outperforms all CNN-based models, attaining an accuracy of 69.09%. This result illustrates the strength of transformer-based architectures in capturing global contextual information, which is particularly advantageous in medical imaging tasks where long-range dependencies and subtle inter-regional features are important.

The proposed BagViT model, which incorporates ensemble learning with ViTs, achieves the highest accuracy at 72.25%, demonstrating a significant improvement over all individual models. This performance gain shows the effectiveness of ensemble-based models in enhancing generalization and robustness, especially in complex CXR image classification tasks involving noisy and heterogeneous medical data. Overall, the results indicate that ViT-based architectures, particularly when combined via bagging, randomized training configurations, and majority voting, offer a promising direction for advancing automated CXR interpretation.

We further evaluated the performance of BagViT by varying the number of ensemble members (denoted as *nbag*), with the corresponding results presented in Figure 4.

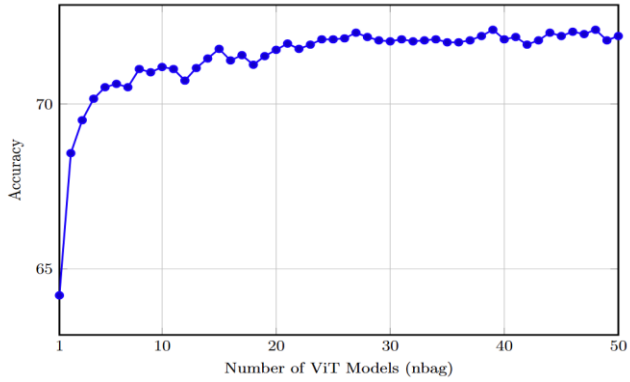

Figure 4. Classification accuracy according to the number of ViT models in the BagViT ensemble

The highest performance was attained with an accuracy of 72.25% (with 39 or 48 ViT models). This enhancement indicates that BagViT effectively boosts generalization by integrating multiple ViT-based learners through a bagging approach. The ensemble structure of BagViT likely reduces variance and enhances robustness, which is particularly advantageous in cases with subtle inter-class differences, such as CXR image classification. These results clearly show that BagViT offers a superior solution for the classification of CXR images, outperforming individual models.

5. CONCLUSIONS AND FUTURE WORK

We proposed Bagged Vision Transformers (BagViT), a robust ensemble learning framework for the classification of CXR images. BagViT constructs an ensemble of independently trained ViT models, each trained on a bootstrap sample of the original training set. To enhance model diversity, we used MixUp augmentation to generate synthetic training instances and introduced training variability by randomizing the number of training epochs and selectively fine-tuning only the top layers of each model. Final predictions are aggregated via majority voting, resulting in enhanced robustness and generalization performance.

Empirical evaluations show that BagViT substantially outperforms traditional fine-tuning approaches based on VGG16, ResNet, DenseNet, and standalone ViT models. Achieving an overall classification accuracy of 72.25%, BagViT establishes a new state-of-the-art baseline for the CXR image dataset considered in this work.

In future work, we plan to enhance ensemble diversity by incorporating additional transformer variants (e.g., Swin, DeiT) and hybrid CNN-transformer models. We also aim to explore adaptive voting strategies, such as weighted or confidence-based aggregation, to better exploit the strengths of individual models.

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 very much the College of Information Technology, Can Tho University.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous distributed systems*. https://www.tensorflow.org/

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

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Callı, 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: Cartwright, H. (eds) Artificial Neural Networks. Methods in Molecular Biology, vol 2190. Humana, New York, NY (pp. 73-94). https://doi.org/10.1007/978-1-0716-0826-5_3

Chollet, F. (2015). *Keras*. https://keras.io/

Do, T.-N., Le, V.-T., & Doan, T.-H. (2022). SVM on top of deep networks for Covid-19 detection from chest X-ray images. *Korea Institute of Information and Communication Engineering*, *20*(3), 219–225. https://doi.org/10.56977/jicce.2022.20.3.219

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). *An image is worth 16x16 words: Transformers for image recognition at scale*. https://doi.org/10.48550/arXiv.2010.11929

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

Global Asthma Network. (2022). *GAR 2022*. http://globalasthmareport.org/gar2022.html

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction, second edition*. Springer Series in Statistics.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2261–2269). https://doi.org/10.1109/CVPR.2017.243

Itseez. (2015). *Open source computer vision library*. https://github.com/itseez/opencv

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.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). *PyTorch: An imperative style, high-performance deep learning library* (Vol. 32). Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning with Python. *Journal of Machine Learning Research*, *12*, 2825--2830.

Poloju, N., & Rajaram, A. (2024). 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.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, *abs/1409.1556*. https://doi.org/10.48550/arXiv.1409.1556

Truong, T.-D., Huynh, P.-H., Nguyen, V. H., & Do, T.-N. (2024). Enhancing the efficiency of lung disease classification based on multi-modal fusion model. *Intelligent Systems and Data Science*, 55–70. https://doi.org/10.1007/978-981-97-9616-8_5

Vapnik, V. (1995). *The Nature of statistical learning theory*. New York, NY: Springer-Verlag.

Verma, S., Devarajan, G. G., & Sharma, P. K. (2024). Comparative evaluation of feature extraction techniques in chest Xray image with different classification model. *International Advanced Computing Conference*, 197–209. https://doi.org/10.1007/978-3-031-56703-2_17

Vo, T.-T., & Do, T.-N. (2024). Improving chest X-ray image classification via integration of self-supervised learning and machine learning algorithms. *Journal of Information and Communication Convergence Engineering*, *22*(2), 165–171. https://doi.org/10.56977/jicce.2024.22.2.165

World Health Organization. (2022). *Pneumonia in children*. https://www.who.int/news-room/fact-sheets/detail/pneumonia

Yadav, P., Menon, N., Ravi, V., & Vishvanathan, S. (2023). Lung-GANs: Unsupervised representation learning for lung disease classification using chest CT and X-ray images. *IEEE Transactions on Engineering Management*, *70*(8), 2774–2786. https://doi.org/10.1109/TEM.2021.3103334

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). *Mixup: Beyond empirical risk minimization*. In *6th International Conference on Learning Representations (ICLR 2018), Vancouver Convention Center, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. https://openreview.net/