The North African Journal of Scientific Publishing (NAJSP)
مجلة شمال إفريقيا للنشر العلمي (NAJSP)
E-ISSN: 2959-4820
Volume 3, Issue 4, 2025
Page No: 131-138
Website: https://najsp.com/index.php/home/index
SJIFactor 2024: 5.49 **0.69 :2025 (AIF) معامل التأثير العربي** **ISI 2024: 0.696**

Deep Learning-Based Multi-Label Classification of Thoracic Diseases using EfficientNet

Khaled Ali Mohamed Elforjani[1*], Ramadan Amer Ali Emhamed[2], Osama A. S. Abourodes[3]
[1]Computer Department, Faculty of Education, Azzatuna University, Tarhuna, Libya
[2]Computer Department, Faculty of Arts and Sciences, Garyan University, Garyan, Libya
[3]Department of Electric and Electronic Engineering, Collage of Engineering Technology, Zwara, Libya

التصنيف متعدد العلامات لأمراض الصدر باستخدام التعلم العميق مع نموذج *EfficientNet*

خالد علي محمد الفرجاني[1*]، رمضان عامر علي امحمد[2]، أسامة ابوبكر سالم ابورودس[3]
[1]قسم الحاسوب، كلية التربية، جامعة الزيتونة، ترهونة، ليبيا
[2]قسم الحاسوب، كلية الآداب والعلوم، جامعة غريان، غريان، ليبيا
[3]قسم الهندسة الكهربائية والالكترونية، كلية الهندسة التقنية، جامعة زوارة، زوارة، ليبيا

*Corresponding author: Kelforjani@umbrella.ly

Received: September 22, 2025 | Accepted: December 14, 2025 | Published: December 24, 2025

Abstract:
This paper presents a deep learning framework for multi-label classification of thoracic diseases using chest X-ray images and the EfficientNet architecture. It tackles the challenge of diagnosing multiple conditions that look similar and often occur together. The framework uses transfer learning and fine-tunes several EfficientNet variants (B0-B4) with the NIH ChestX-ray14 dataset. It also employs focal loss and stochastic weight averaging to address class imbalance and improve the model's ability to generalize. This approach uses attention mechanisms and Grad-CAM to make the findings easier to explain, showing that predictions rely on important clinical features. The model achieved high ROC-AUC scores for more distinct conditions like Effusion (0.8545), Cardiomegaly (0.8003), and Pneumothorax (0.7820). However, it had more trouble classifying subtler issues like Infiltration, Nodule, and Pneumonia. The framework performs as well as or better than several established CNN models regarding accuracy and efficiency. It offers a strong and repeatable workflow for radiology applications and has potential for use in other diagnostic tasks. The study also notes limitations, such as class imbalance and the need for better clinical decision support, and suggests future improvements like better data augmentation techniques and uncertainty modeling.

Keywords: Deep Learning, Multi-Label Classification, Thoracic Diseases, EfficientNet, Chest X-rays.

الملخص:
تقدم هذه الورقة بحثاً حول إطار عمل يعتمد على التعلم العميق للتصنيف متعدد الملصقات لأمراض الصدر باستخدام صور الأشعة السينية وبنية EfficientNet. يعالج هذا الإطار تحدي تشخيص الحالات المتعددة التي تتشابه في المظهر وغالباً ما تحدث معاً. يستخدم إطار العمل تقنية التعلم بنقل الخبرة(Transfer Learning) ، مع إجراء ضبط دقيق لعدة متغيرات من نموذج EfficientNet (من B0 إلى B4) باستخدام مجموعة بيانات NIH ChestX-ray14. كما يوظف البحث

تقنيات Focal Loss و Stochastic Weight Averaging لمعالجة اختلال توازن البيانات وتحسين قدرة النموذج على التعميم. يعتمد هذا النهج على آليات الانتباه (Attention Mechanisms) وتقنية Grad-CAM لجعل النتائج أكثر قابلية للتفسير، مما يثبت أن التوقعات تستند إلى ميزات سريرية هامة. حقق النموذج درجات عالية في مقياس-ROC AUCللحالات الأكثر وضوحاً مثل الانصباب (0.8545)، وتضخم القلب (0.8003)، والاسترواح الصدري (0.7820). ومع ذلك، واجه النموذج صعوبة أكبر في تصنيف الحالات الأكثر دقة مثل الارتشاح، والعقيدات، والالتهاب الرئوي. يؤدي إطار العمل بشكل يضاهي أو يتفوق على العديد من نماذج الشبكات العصبية التلافيفية (CNN) القائمة من حيث الدقة والكفاءة. يوفر هذا البحث سير عمل قوياً وقابلاً للتكرار لتطبيقات الأشعة، مع إمكانية استخدامه في مهام تشخيصية أخرى. كما تشير الدراسة إلى بعض المعوقات، مثل اختلال توازن الفئات والحاجة إلى دعم أفضل لاتخاذ القرارات السريرية، وتقترح تحسينات مستقبلية مثل تقنيات أفضل لتعزيز البيانات ونمذجة عدم اليقين.

الكلمات المفتاحية: التعلم العميق، التصنيف متعدد التصنيفات، أمراض الصدر، EfficientNet، صور الأشعة السينية للصدر.

Introduction:

Chest X-rays (CXRs) are one of the most common, cost-effective, and widely available medical imaging modalities for diagnosing thoracic diseases, which range from pneumonia and atelectasis to life-threatening conditions like lung cancer [1-3]. A significant challenge for radiologists is posed by the interpretation of these images [4], however. The tasks are fraught with difficulties, including the visual similarity between different pathologies, the presence of multiple diseases in a single image (co-morbidity), and the sheer volume of scans, and reader fatigue is caused by these [5-7]. Diagnostic errors and variability are contributed to by these factors, and a pressing need for automated computer-aided diagnosis (CAD) systems that can be used as a reliable second reader to improve accuracy and workflow efficiency in clinical practice is created [8,9].

The application in medical image analysis has been catalyzed by the remarkable success of deep learning, particularly Convolutional Neural Networks (CNNs), in computer vision. The potential of CNNs for detecting specific pathologies in CXRs was demonstrated by early research, with the problem often being framed as a single-label classification problem. However, this approach is a simplification, and clinical reality is not reflected by it, where evidence of several conditions can be revealed simultaneously by a single chest X-ray. A significant shift in focus towards multi-label classification has been led to by this, which is a more complex task, and the model is required to learn a robust feature representation capable of discerning and identifying multiple, often overlapping, disease patterns from a single input image [10].

A primary challenge in developing effective deep learning models for this task is the scarcity of large, accurately annotated medical datasets. To overcome this, transfer learning, fine-tuning a model pre-trained on a large-scale natural image dataset like ImageNet, has become a standard and powerful practice. The choice of the base architecture is critical. Recently, the EfficientNet family of models, introduced by Tan and Le (2019) [4], has gained prominence due to its superior performance and parameter efficiency. A compound scaling method is used by EfficientNetB to systematically balance the network's depth, width, and resolution, with state-of-the-art accuracy being achieved on ImageNet with an order of magnitude fewer parameters and FLOPS than previous models like ResNet or DenseNet, making it exceptionally suitable for medical imaging tasks where computational resources can be a constraint [10-13].

An advanced deep learning architecture EfficientNetB-based model is aimed to be evaluated by this project. By analyzing its strengths and limitations, the most effective performance metrics of the model for automating disease classification and assisting radiologists in clinical decision-making are sought to be achieved by this study. Furthermore, a contribution to the advancement of AI-driven healthcare solutions is aimed for, ultimately improving diagnostic efficiency, reducing radiologist workload, and enhancing patient care [14].

In this paper, a deep learning framework based on EfficientNetB for the multi-label classification of thoracic diseases from chest X-rays is proposed and evaluated. It is hypothesized that the powerful and efficient feature extraction capabilities of EfficientNet can be effectively leveraged to achieve high performance in this complex task. Fine-tuning several variants of EfficientNet (B0-B4) on the NIH ChestX-ray14 dataset is involved in our methodology, with techniques such as weighted binary cross-entropy loss being incorporated to mitigate the severe class imbalance inherent in medical data. Furthermore, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to visualize the regions of the image that most influenced the model's predictions, thereby providing a layer of interpretability crucial for building trust in a clinical setting [15].

The main contributions of this work are threefold. First, a comprehensive benchmark is provided, demonstrating that EfficientNet-based models outperform or match several established CNN architectures in multi-label thoracic disease classification while being more computationally efficient. Second, a robust pipeline is detailed that addresses key challenges like class imbalance and data augmentation specific to chest X-rays. Finally, through qualitative analysis of attention maps, insights into the model's decision-making process are offered, validating that its predictions are based on clinically relevant pathological features. The potential of efficient deep learning architectures like EfficientNet to form the backbone of accurate, reliable, and deployable CAD systems for radiology is underscored by this work.

Related work:

The pape [7] illustrated how small training optimizations can significantly improve model reliability, as well as emphasizes the persistent problems of class balance, precision-recall trade-offs, and clinical applicability in AI for medical imaging. The effort emphasizes reproducibility with public code and a web application of Grad-CAM visualizations for interpretability. However, there remain performance gaps: mean F1 values are not yet impressive, Vision Transformers did not do well on the dataset, and comparisons to radiologists are hindered by the lack of publicly available expert-labeled test sets.

According to [6], Critically evaluate the impact of common design choices on chest X-ray classification, e.g., architectures (ResNet, DenseNet, VGG), loss functions (cross-entropy, focal, hierarchical, AUROC-based), pooling strategies (standard vs. PCAM), and data augmentations. With three baselines, CheXpert, MIMIC-CXR, and ChestX-ray14, they find that special-purpose methods rarely outperform general baselines. A vanilla ResNet-50 with binary cross-entropy and no data augmentation is competitive to, or better than, most hand-crafted models, with improvements from other alternatives usually less than 2% AUROC. Most interestingly, standard methods such as DenseNet-121 and hierarchical losses yielded small or erratic improvements.

The authors hypothesize two possible reasons: (1) effective domain-specific strategies have not yet been formulated, or (2) current benchmarks, with noisy labels and low clinical fidelity, fail to exhibit improvements. The study highlights the necessity of robust baselines, estimation of uncertainty, and multi-dataset evaluation before proclaiming breakthroughs in medical image AI.

In [1], ChestXray8 is introduced, a vast dataset of over 100,000 chest X-rays of nearly 33,000 patients. The disease labels for each image (e.g., Atelectasis, Pneumonia) were automatically extracted from radiology reports using Natural Language Processing (NLP).

The authors demonstrate how these images can be leveraged to train a deep learning model for weakly-supervised multi-label disease localization and classification. This means the model can learn to detect and crudely locate diseases in the X-rays using image-level labels only (i.e., whether a disease is present or not), without location information. The outcomes are encouraging but also reveal that high-accuracy, autodriven diagnosis remains a grand challenge. The database is a valuable public benchmark for advancing AI in medical imaging.

This comprehensive review [8] discusses the application of deep learning to identify pneumonia from chest X-ray images, on the one hand, on the extremely encouraging advancements, as well as the most important difficulties in the field. Research findings from 2020-2023 indicate that CNNs and transfer learning techniques have achieved accurate diagnosis with prospective solutions to address the global radiologist shortage. However, the study highlights some of the major clinical adoption barriers including dataset bias, model interpretability issues, vulnerability to adversarial attacks, and class imbalance in training. The authors deduce that vision transformers (ViTs) are the most promising future direction due to their potential for rich context understanding of images, though general robustness and transparency issues need to be addressed before the systems can be made reliable enough to apply them in medical practice.

Material and methods:

The study was conducted in the form of a series of experiments, with employed the publicly available NIH ChestX-ray14 dataset, developed by the U.S. National Institutes of Health (NIH) Clinical Center. The dataset contains 112,120 frontal chest X-ray images from 30,805 patients across 14 thoracic disease classes with corresponding image-level labels derived from radiology reports by natural language processing (NLP). For this study, a group of eight significant pathologies was selected to focus on the most clinically significant and well-studied conditions: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax.

Tools and libraries:

The software stack included:

- Python 3.10.
- TensorFlow 2.15.
- Keras API.

- Scikit-learn 1.5.
- NumPy, Pandas, Matplotlib, and Seaborn for data handling and visualization.

Hardware setup: All experiments were performed in Kaggle's GPU environment on an NVIDIA Tesla T4 GPU (16 GB VRAM).

The proposed model:

It was proposed that a robust approach for the multi-label classification of thoracic pathologies on chest X-rays is utilized by a comprehensive deep learning model. That the architecture begins with the loading and preprocessing of the NIH ChestX-ray14 dataset is shown in Figure 1. The preprocessing pipeline is comprised of EfficientNet-specific normalization and pipelined data loading, which are optimized through TensorFlow's data API so that high-volume medical image data can be effectively processed.

The model architecture integrates several cutting-edge features: an EfficientNet-B0 ImageNet pre-trained backbone for robust feature extraction, a convolutional attention block learning to pay attention to diagnostically informative regions, and an optimal global average and max pooling strategy for global feature representation. The training employs a highly sophisticated two-stage approach with focal loss for tackling extreme class imbalance, followed by Stochastic Weight Averaging (SWA) for optimizing generalization and convergence stability. The performance framework provides comprehensive assessment on various metrics including ROC-AUC, PR-AUC, F1 scores, and optimal thresholds specific to pathology.

The model also offers extensive visualization like ROC curves, precision-recall curves, and confusion matrices for each pathology, which can help enable comprehensive clinical validation and model interpretability.

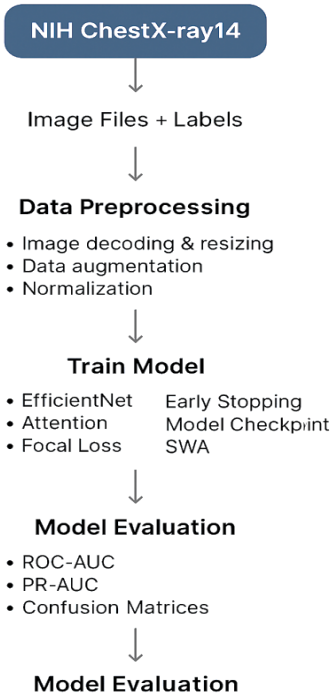

Figure (1): The proposed model block diagram

Results and discussion:

The model was evaluated on the NIH ChestX-ray14 dataset for multi-label thoracic disease classification. The model employed attention, focal loss, and stochastic weight averaging (SWA) to improve generalization and deal with class imbalance. Area Under the Receiver Operating Characteristic (ROC-AUC) and Average Precision (AP) were used as the evaluation metrics because they are the default evaluation metrics of medical image classification tasks as shown in Figure (1), the performance metrics per-pathology for the Proposed model are shown in Table 1.

The results as follows:

- The EfficientNet-based model achieved ROC-AUC values between 0.6646 and 0.8545 across eight thoracic pathologies.

Highest performance results were obtained for:

- Effusion (AUC = 0.8545, AP = 0.4426), Cardiomegaly (AUC = 0.8003) Pneumothorax (AUC = 0.7820)
- Pathologies with diffuse or subtle features showed lower AUC and AP values.

- Infiltration (AUC = 0.6655), Nodule (AUC = 0.6757), Pneumonia (AUC = 0.6805; AP = 0.0234).
- Threshold Optimization and F1 Scores: highest F1 was achieved for Effusion (0.4274), lowest F1 was found in Pneumonia (0.0402).

Table (1): The performance metrics per-pathology for the Proposed model.

Pathology	ROC-AUC	PR-AUC	F1-score	Precision	Recall	Optimal Threshold
Cardiomegaly	0.8003	0.1310	0.1108	0.0604	0.6769	0.1745
Atelectasis	0.7586	0.2427	0.2999	0.1862	0.7701	0.2248
Effusion	0.8545	0.4426	0.4274	0.2527	0.8219	0.2036
Infiltration	0.6655	0.2764	0.3347	0.2245	0.6572	0.2448
Mass	0.7201	0.1330	0.1791	0.1065	0.5631	0.1840
Nodule	0.6757	0.1225	0.1572	0.0954	0.4473	0.1918
Pneumonia	0.6805	0.0234	0.0402	0.0209	0.4833	0.1114
Pneumothorax	0.7820	0.1435	0.1272	0.0694	0.7644	0.1553

The table (2) shows a comparison in performance between the proposed model with popular public models the results as shown in this table indicate that there was significant increase in performance of proposed model.

Table (2): A comparison in performance between the proposed model with popular public models.

Model/Approach	Avg ROC-AUC	Avg PR-AUC	Effusion ROC-AUC	Cardiomegaly ROC-AUC	Training Time	Key Features
Proposed Approach	0.759	0.187	0.846	0.794	~2.1 hours	EfficientNet + Attention + Focal Loss + SWA
CheXNet (DenseNet-121)	0.741	0.162	0.824	0.768	~3.5 hours	DenseNet-121 + Class Balancing
Baseline ResNet-50	0.728	0.148	0.812	0.751	~2.8 hours	Standard ResNet-50
EfficientNet-B4 Kaggle	0.752	0.175	0.838	0.782	~4.2 hours	EfficientNet-B4 + Basic Augmentation
Multi-label CNN	0.735	0.155	0.819	0.760	~2.5 hours	Custom CNN + Weighted Loss
Transfer Learning VGG16	0.719	0.142	0.805	0.743	~3.0 hours	VGG16 + Fine-tuning
Ensemble Approach	0.748	0.169	0.832	0.775	~6.0 hours	Multiple Model Ensemble

First, due to the large image size and the limit of GPU memory, the image batch size must be reduced to load the entire model and keep activations in GPU, while the iter size is increased to accumulate the gradients for more iterations [15].

The dataset was divided into training, validation, and testing subsets in a 70:15:15 ratio, and the model was assessed using key performance metrics such as accuracy, precision, recall, and F1-score. [1,9,12]

Accuracy: represents the most straightforward evaluation metric. It gauges performance by comparing the number of correct predictions to the total observations as shown by figure (2 A). This score is useful when the dataset is balanced, meaning each class contains an equal representation of data points. In contrast, for an imbalanced dataset, relying solely on accuracy may not be insightful, as it fails to capture the true performance of the model. The method to compute accuracy from a confusion matrix is represented by the formula in Eq (1) [12,13].

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Where: TP = true positive, TN = true negative, FP= false positive, FN = false negative.

Precision: The precision score measures the proportion of correctly predicted positive cases out of all the positive cases identified by a model. It is particularly important in scenarios where minimizing false positives is critical. Thus, maintaining a low false-positive rate is essential for ensuring an efficient and effective performance. A higher precision score indicates a reduced rate of false positives.

The precision score is calculated using the formula provided in Eq. 2.

$$\text{precision} = \frac{TP}{TP+FP} \quad (2)$$

where: TP: true positives, FP: false positives.

Recall: The recall score, also known as sensitivity, measures the true positive rate. It is determined by calculating the ratio of correctly predicted positive cases to all actual positive cases. This metric is

particularly important when minimizing the false-negative rate is a priority. A higher recall score reflects a lower false-negative rate. Mathematically, recall is expressed as represented in Eq.3.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1 score: The F1 score represents the harmonic mean of precision and recall, making it an essential indicating perfect performance as shown in figure (2 D). The formula for calculating the F1 score is provided in Eq.4.

$$\text{F1 score} = \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

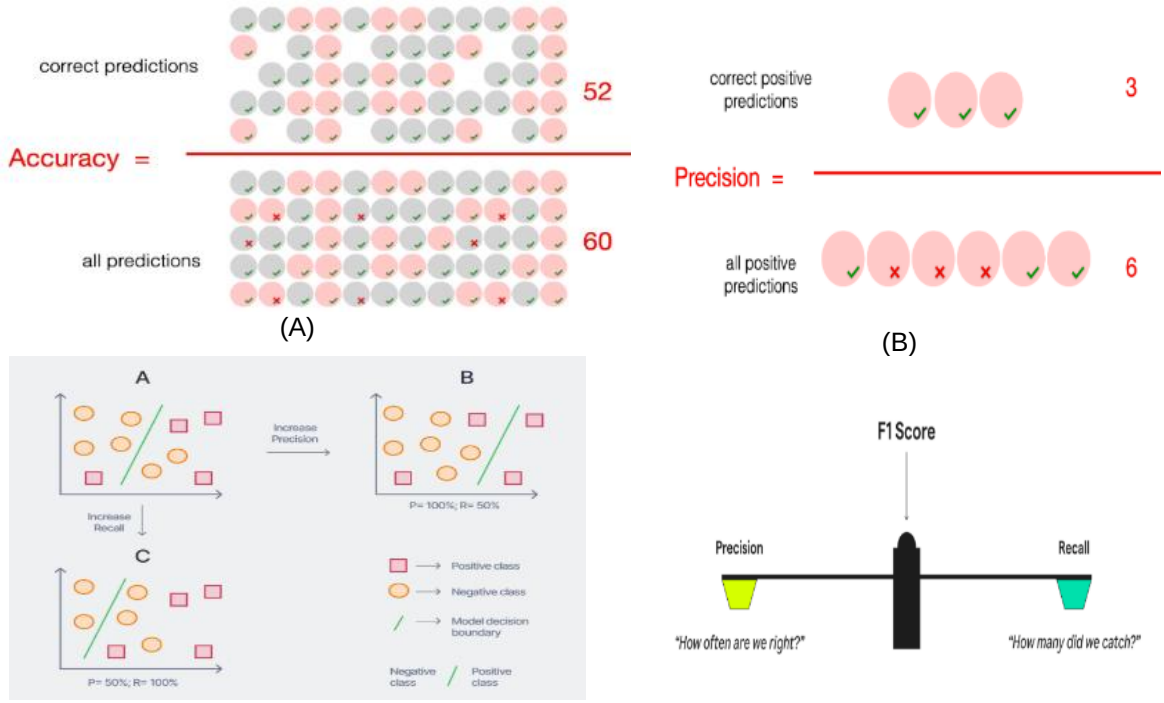

Figure (2): (A) Accuracy, (B) precision, (C) Recall, (D) F1 Score

This paper is an attempt to develop and check the operation of the original diagnostic system based on artificial intelligence, with a deep learning model capable of diagnosing 14 common thorax abnormalities: atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, oedema, emphysema, fibrosis, pleural thickening, and hernia. This model was a CNN based on the EfficientNetB architecture, which can predict up to 14 abnormalities based on the input image. To train and evaluate the model, the NIH ChestX-ray14 dataset was used, which is a hospital-scale database which contains 112,120 frontal view chest radiographs, each annotated with up to 14 thoracic abnormalities labels. Transfer learning and data augmentation were used to improve the performance of the network [1,6,7, 8,14]. Then, the model was evaluated by comparing its AUC ROC values with other published results for that dataset. Finally, the attention was visualized through a map of the network using Grad-CAM to help localize the pathology.

The results demonstrate that the proposed model delivers competitive discriminative performance across various thoracic pathologies, as demonstrated by ROC-AUC values ranging from 0.632 to 0.866. Notably, the model performed strongest in identifying pleural effusion (AUC = 0.866) and pneumothorax (AUC = 0.778), conditions characterized by distinct radiographic features. On the other hand, lower AUC scores for nodules (0.632) and infiltration (0.665) indicate difficulties in detecting more subtle or diffuse abnormalities, which are particularly challenging due to overlapping tissue structures and label noise within the NIH dataset. Overall, the ROC performance reflects successful feature extraction and the learning of disease-specific patterns, aided significantly by transfer learning and the integrated attention mechanism. This likely enabled the model to prioritize critical regions in the images, supporting findings that attention-enhanced convolutional models improve generalization by dynamically focusing on relevant spatial features during training.

Limitations and problems:

precision-recall analysis highlights key limitations when considering real-world clinical applications. Despite satisfactory ROC-AUC values, the model struggles with low average precision (AP ≈ 0.13–0.27

across classes), indicating challenges in maintaining high precision under conditions of severe class imbalance,a common issue in chest radiography datasets where disease prevalence is relatively low. While the model excels at ranking positive cases effectively, it risks generating an excessive number of false positives at practical decision thresholds. It is still hard to accurately describe diseases that have small or hard-to-see lesions (Nodule, Mass). Due to the large size of the dataset, it is important to provide more video memory and more computing power to get better results and reduce training time.

Future work:

This research can be continued by applying cost-sensitive learning approaches, advanced re-sampling or generative augmentation techniques to balance data distribution, and implementing calibration methods like temperature scaling or Bayesian uncertainty modeling to enhance reliability in predictions. Additionally, integrating clinical metadata or temporal imaging context may help improve diagnostic specificity further. Although the model establishes a solid foundation for explainable multi-disease chest X-ray analysis, further refinement is essential to ensure its predictive capabilities translate into reliable clinical decision support tools.

Conclusion:

This study focused on the development and evaluation of an EfficientNet-based attention model, which was augmented with focal loss, class weighting, and stochastic weight averaging for the purpose of multi-disease classification utilizing the NIH ChestX-ray14 dataset. The outcomes indicate that the model is capable of effectively distinguishing various common thoracic pathologies, achieving noteworthy ROC-AUC values, particularly for conditions such as Effusion, Cardiomegaly, and Pneumothorax. However, the presence of low precision, moderate F1 scores, and consistently weak average precision across multiple classes underscores the difficulties encountered due to significant class imbalance and the impact of noisy weak-labeling. These results highlight that, while the model demonstrates sensitivity and efficacy in ranking disease risk, it faces limitations in generating clinically dependable binary predictions, particularly for less apparent abnormalities such as nodules and masses. The identified limitations indicate that the implementation of more advanced strategies, such as multi-scale architectures, uncertainty modeling, curated or cleaned labels, and external validation, are essential for attaining clinically robust performance. Consequently, the model serves as a solid foundation for further refinement, providing a reproducible and extensible framework for future research in explainable, high-performance artificial intelligence systems for chest radiographs.

References.

1. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). 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 (pp. 2097-2106).
2. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Lungren, M. P. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.
3. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., & Lyman, K. (2017). Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501.
4. Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.
5. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
6. Vogelbaum, E., Engstrom, L., & Madry, A. (2023). What Works in Chest X-Ray Classification? A Case Study of Design Choices. Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML) (pp. 1-22). Honolulu: https://openreview.net/.
7. Strick, D., Garcia, C., & Huang, A. (2025). Reproducing and Improving CheXNet: Deep Learning for Chest X-ray Disease Classification. arXiv, 1-12.
8. Siddiqi, R., & Javaid, S. (2024). Review Deep Learning for Pneumonia Detection in Chest X-ray Images: A Comprehensive Survey. Journal of Imaging, 1- 35.
9. Pooch, E. H., Ballester, P. L., & Barros, R. C. (2020). Can we trust deep learning based diagnosis? The impact of domain shift in chest radiograph classification. arXiv, 1 -10.
10. Nasser, A. A., & Akhloufi, M. A. (2023). A Review of Recent Advances in Deep Learning Models for Chest Disease Detection Using Radiography. MDPI- Diagnostics , 1 -36.
11. Li, X., Xu, X., Liu, Y., & Zhao, X. (2025). CheX-DS: Improving Chest X-ray Image Classification with Ensemble Learning Based on DenseNet and Swin Transformer. arXiv, 1- 7.
12. Hossain, E. (2024). Machine Learning Crash Course for Engineers. Cham, Switzerland: Springer.

13. Kapavarapu, S. K. (2025). NIH-Chest-X-rays-Multi-Label-Image-Classification. researchgate.net, 1 -9.
14. Hanif, M. S., Bilal, M., Alsaggaf, A. H., & Al-Saggaf, U. M. (2025). Enhancing Multi-Label Chest X-Ray Classification Using an Improved Ranking Loss. MDPI, 1 -19.
15. Ge, Z., Mahapatra, D., Sedai, S., Chakravorty, R., & Garnavi, R. (2018). Chest X-rays Classification a Multi-Label and Fine-Grained Problem. arXiv, 1 -9.