ISBN: 978-981-94-4198-3
2025 the 15th International Workshop on Computer Science and Engineering (WCSE 2025)
doi: 10.18178/wcse.2025.06.032

Evaluating Performance of Multiclass Chest Radiograph Classification Using Soft Voting of CNNs and Transformers

Hui-Chu Chiu [1], Deng-Yiu Chiu [2], Usama Yasir Khan [3], Chia-Ching Chang [3,4],

Cheng-Hsuan Juan [5], Yao-Hsien Lee [6], Chen-Shu Wang [7], Chun-Jung Juan [3,8+]

1. Ph.D. Program of Management, Chung-Hua University, Hsinchu, Taiwan.
2. Department of Information Management, Chung-Hua University, Hsinchu, Taiwan.
3. Department of Medical Imaging, China Medical University Hsinchu Hospital, Hsinchu, Taiwan
4. Department of Management Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan.
5. Department of Obstetrics and Gynecology, Cheng Ching Hospital Chung Kang Branch, Taichung, Taiwan.
6. Department of Finance, Chung Hua University, Hsinchu, Taiwan.
7. Department of Information and Finance Management, Taipei, Taiwan.
8. Department of Medical Imaging, Medical University Hospital, Taichung, Taiwan; Department of Radiology, School of Medicine, College of Medicine, China Medical University, Taichung, Taiwan; Department of Biomedical Engineering and Environmental Sciences, National Tsing Hua University, Hsinchu, Taiwan; Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

Abstract. Accurate and sensitive classification of multiple thoracic diseases using chest radiographs remains a significant challenge in computer-aided diagnosis. While Convolutional Neural Networks (CNNs) and Transformer-based models have shown strong performance individually is often limited. This study proposes a deep learning framework to improve the area under the curve (AUC) for multiclass detection of 14 chest diseases. We utilized the NIH ChestX-ray14 dataset, which includes 112,120 frontal-view chest radiographs from 30,805 unique patients, annotated with 14 disease labels. Predictions from five models—including a CNN model (DenseNet121), two Transformer models (Vision Transformer and Swin Transformer), and two hybrid models (Hybrid ViT–DenseNet121 and Hybrid Swin–DenseNet121)—were integrated using soft voting to enhance classification robustness. Soft voting achieved an AUC of 0.8437, significantly higher than all individual or hybrid models (all $P < 0.05$), except DenseNet121 (AUC = 0.8433; $P = 0.76$). This framework significantly improves AUC for multiclass chest disease detection and offers a robust approach for more reliable and sensitive clinical screening applications.

Keywords: Chest radiograph, convolutional neural network, transformer

1. Introduction

Interpretation of chest X-rays (CXRs) remains challenging, especially when differentiating subtle pathologies or overlapping disease patterns [1]. Deep learning, particularly convolutional neural networks (CNNs), has significantly advanced automated CXR analysis. CNN architectures including ResNets and DenseNets have demonstrated high accuracy in detecting multiple thoracic diseases simultaneously [2, 3]. Transformers, including but not limited to Vision Transformers (ViTs) and Swin Transformers, capture long-range dependencies and hierarchical representations, showing promise in medical image classification [4-6]. Hybrid architectures combining CNNs with Transformers have been proposed to leverage the strengths of both local and global feature extraction [7-9], leading to improved performance in complex classification tasks. Despite progress, several limitations persist. Deep learning models often struggle with class imbalance and insufficient sensitivity [10]. This study proposes a deep learning framework to improve the area under the curve (AUC) for multiclass detection of 14 chest diseases on radiographs.

+ Corresponding author. Tel.: + 886-3-5580558
E-mail address: peterjuancj@gmail.com.

2. Materials and Methods

2.1. Patients and Dataset:

This study employed the NIH ChestX-ray14 dataset, a widely used benchmark public dataset released by the National Institutes of Health. Institutional review board approval was not required for using the public dataset retrospectively. The dataset consists of 112,120 frontal chest radiographs from 30,805 unique patients, collected over a span of several years. Each image is annotated with one or more of 14 thoracic disease labels: atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia. Manuscript may be submitted to "Scientific Report" or other journals.

2.2. Imaging preprocessing

All chest radiographs were resized to 224 × 224 pixels to conform with input requirements of the pretrained models. Pixel intensity values were normalized using ImageNet statistics, with a mean of (0.485, 0.456, 0.406) and a standard deviation of (0.229, 0.224, 0.225) across all three RGB channels. To simulate real-world deployment conditions and allow for equitable model performance comparison, no additional image preprocessing or data augmentation techniques—such as rotation, scaling, or flipping—were applied during training or validation. This design choice was intentional to evaluate each model's robustness and generalization ability under naturally acquired, unaltered clinical images, as would be encountered in actual diagnostic workflows.

2.3. Detection AI Models

All models shared configurations during training process including a loss function of binary cross entropy with logit loss, Adam optimizer, CosineAnnealingLR as the learning rate scheduler, a batch size of 32, a total of 100 epochs with early stopping based on validation loss to prevent overfitting, an initial learning rate of $\mathbf{1\times10^{-4}}$, and a sigmoid score for each of 14 classes of chest disease. Pos_weight was computed dynamically to deal with the class imbalance. The best model was saved according to the minimum of validation loss.

2.3.1 DenseNet121:

As the baseline convolutional neural network (CNN), we adopted DenseNet121 , a densely connected architecture that facilitates feature reuse and has demonstrated strong performance in medical image classification tasks [11]. DenseNet121 was initialized with pretrained weights from the ImageNet dataset and fine-tuned for the multi-label classification task on the NIH ChestX-ray14 dataset using a binary cross-entropy (BCE) loss function.

2.3.2 Vision Transformer (ViT):

The Vision Transformer (ViT) models long-range dependencies and is beneficial for global feature extraction [4]. We used the base variant of ViT (vit_base_patch16_ 224) with approximately 86M parameters. The input image was divided into non-overlapping 16×16 patches.

2.3.3 Swin Transformer:

The Swin Transformer applies self-attention within shifted windows to capture both local and global context efficiently [5]. We used the base variant of Swin transformer (swin_base_patch4_window7_224) with approximately 88M parameters, image resolution of 224×224, and window size of 7×7. The input image was divided into non-overlapping 4×4 patches.

2.3.4 Hybrid ViT–DenseNet121:

We constructed a hybrid architecture by sequentially combining DenseNet121 with a Vision Transformer (ViT) module. The DenseNet121 served as the feature extraction backbone, generating rich hierarchical features. These

extracted features were then passed to the ViT module, which enabled long-range dependency modeling and global self-attention across spatial positions, complementing the localized convolutional representations.

2.3.5 Hybrid Swin–DenseNet121:

We developed a hybrid architecture by sequentially integrating DenseNet121 with a Swin Transformer module. DenseNet121 functions as the feature extraction backbone, producing hierarchical feature maps. These features are subsequently fed into the Swin Transformer, which applies hierarchical self-attention within shifted windows to capture long-range dependencies and spatial relationships.

2.3.6 Soft Voting Ensemble:

To increase robustness and generalization, the outputs of the aforementioned models with an area under the curve (AUC) higher than 0.8 were combined using soft voting. In this ensemble method, the predicted class probabilities from each model were averaged to generate a consensus prediction for each disease category.

2.4. Evaluation Metrics

Performance was measured using Area Under the ROC Curve (AUC) according to the eq. 1.

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \quad (1)$$

2.5. Statistical Analysis

DeLong's test was used for pairwise AUC comparison between models and ensemble approaches [12]. A p-value < 0.05 was considered statistically significant.

3. Results

3.1. Model Performance in Multiclass Classification

Among all individual models, DenseNet121 achieved the highest overall performance with an AUC of 0.8433. The Vision Transformer (ViT) and Swin Transformer followed with moderate AUCs of 0.7640 and 0.8279, respectively. Hybrid architectures demonstrated intermediate performance. The ViT–DenseNet121 hybrid attained an AUC of 0.8217, while the Swin–DenseNet121 achieved 0.8269. Across all models, emphysema and pneumothorax consistently yielded the highest AUCs, whereas hernia and fibrosis showed the lowest.

3.2. Performance of Soft Voting Ensemble

To improve robustness and exploit model complementarity, a soft voting ensemble was constructed using models with individual AUCs greater than 0.8: DenseNet121, Swin Transformer, and both hybrid models. This ensemble yielded an enhanced average AUC of 0.8437, surpassing all individual and hybrid models except for DenseNet121, with no statistically significant difference (P = 0.76, DeLong's test). However, the ensemble showed statistically significant improvements over the Swin Transformer and both hybrid models (all P < 0.05).

The soft voting ensemble also led to more consistent class-wise predictions. For example, although the standalone ViT model misclassified several cases of cardiomegaly, these were correctly identified by the ensemble. Moreover, the ensemble reduced variance in predicted probabilities across models, indicating better generalization and reduced overfitting. Notably, the ensemble also improved sensitivity for rare conditions such as hernia and pleural thickening.

3.3. Class-specific Performance of Deep Learning Models

The class-specific diagnostic performance of various models is illustrated in Fig. 1. AUROC scores ranged from 0.6624 (nodule, detected by ViT) to 0.9424 (hernia, detected by DenseNet121). Ten disease categories-atelectasis, cardiomegaly, effusion, mass, pneumothorax, consolidation, edema, emphysema,

and hernia-achieved mean AUROC values above 0.8, reflecting reliable classification performance. In contrast, infiltration, nodule, pneumonia, and pleural thickening were more challenging to detect, with mean AUROC values below 0.8.

4. Discussion and Conclusion

This study systematically evaluated the performance of convolutional and transformer-based deep learning models for multiclass classification of chest radiographs. DenseNet121 demonstrated the highest diagnostic accuracy among all individual models, reaffirming the robustness and suitability of convolutional neural networks (CNNs) for medical image analysis. In contrast, the Vision Transformer (ViT) and Swin Transformer, though promising, exhibited slightly lower AUC, suggesting that pure transformer-based models may require further domain-specific optimization to perform at par with CNNs in radiographic interpretation tasks.

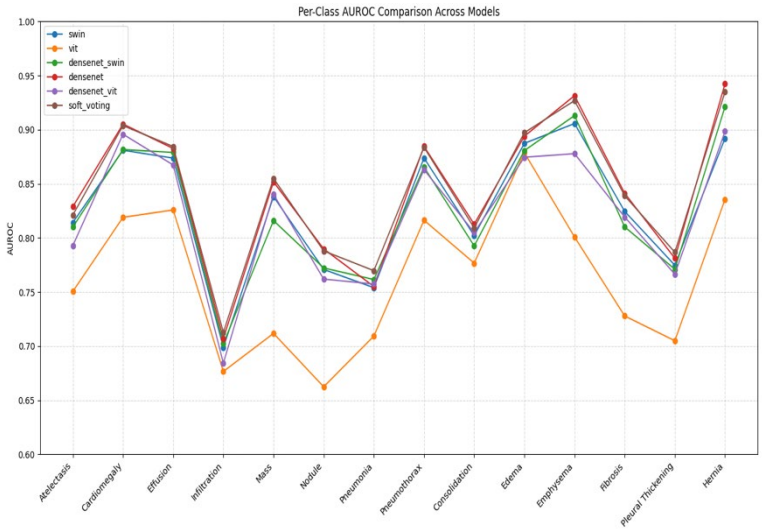

Fig. 1. Class-specific AUROCs for different deep learning models.

The hybrid architectures, ViT–DenseNet121 and Swin–DenseNet121, achieved intermediate performance, indicating that fusing local feature extraction from CNNs with the global attention mechanisms inherent to transformers can enhance classification outcomes. However, the most notable performance improvement was observed with the soft voting ensemble model, which integrated predictions from high-performing models. This ensemble not only achieved the highest average AUC (0.8437) among all tested models—excluding DenseNet121, with which the difference was not statistically significant—but also significantly outperformed the Swin Transformer and both hybrid models.

Importantly, ensemble learning contributed to improved diagnostic stability, especially for rare conditions such as hernia and pleural thickening. These findings underscore the potential of ensemble approaches to mitigate challenges associated with class imbalance and overfitting in clinical datasets. The observed reduction in inter-model variance further suggests improved generalizability—an essential requirement for models intended for clinical deployment.

Class-specific analysis revealed consistent diagnostic accuracy for diseases with distinctive radiographic features, such as emphysema and pneumothorax. In contrast, conditions like nodule, pneumonia, and infiltration presented lower AUCs across all models, reflecting the intrinsic difficulty in detecting such pathologies due to subtle imaging cues and interobserver variability in their interpretation. Despite the encouraging findings, this study has several limitations. The dataset, while labeled by expert consensus, may still include annotation noise, which could affect training outcomes. Additionally, the models were evaluated on a single public dataset, and their generalizability to external populations remains to be verified. Future work should focus on external validation, incorporation of multi-modal data (e.g., clinical and laboratory features), and the development of advanced learning paradigms such as self-supervised or federated learning.

Incorporating class-balanced loss functions and uncertainty quantification techniques may further enhance model performance and reliability.

In conclusion, while DenseNet121 remains a strong baseline for chest radiograph classification, hybrid and ensemble approaches show considerable promise in enhancing diagnostic performance and robustness. Ensemble learning, in particular, presents a viable path forward for achieving clinically reliable AI-based diagnostic support systems in radiology.

5. Acknowledgements

C.J.J. received support partly by the Taiwan Ministry of Science and Technology under grant MOST 111-2314-B-039-036 and partly by the China Medical University Hsinchu Hospital under grant CMUHCH-DMR-112-001

6. References

[1] A. Quintero-Rincon, R. Di-Pasquale, K. Quintero-Rodriguez, and H. Batatia, “Computer-based quantitative image texture analysis using multi-collinearity diagnosis in chest X-ray images,” *PLoS One,* vol. 20, no. 4, pp. e0320706, 2025.

[2] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” *arXiv:1711.05225* 2017.

[3] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman, “Weakly supervised medical diagnosis and localization from multiple resolutions,” *arXiv:1803.07703*, 2018.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv:2010.11929 [preprint]*, 2020.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” *arXiv:2103.14030* 2021.

[6] J. Ko, S. Park, and H. G. Woo, “Optimization of vision transformer-based detection of lung diseases from chest X-ray images,” *BMC Med Inform Decis Mak,* vol. 24, no. 1, pp. 191, Jul 8, 2024.

[7] N. Ashraf, A. Mamun, Hasnat, Abdullah, and G. R. Alam, “SynthEnsemble: A Fusion of CNN, Vision Transformer, and Hybrid Models for Multi-Label Chest X-Ray Classification,” *arXiv:2311.07750* 2023.

[8] X. Fu, R. Lin, W. Du, A. Tavares, and Y. Liang, “Explainable hybrid transformer for multi-classification of lung disease using chest X-rays,” *Sci Rep,* vol. 15, no. 1, pp. 6650, Feb 24, 2025.

[9] Z. R. Murphy, K. Venkatesh, J. Sulam, and P. H. Yi, “Visual Transformers and Convolutional Neural Networks for Disease Classification on Radiographs: A Comparison of Performance, Sample Efficiency, and Hidden Stratification,” *Radiol Artif Intell,* vol. 4, no. 6, pp. e220012, Nov, 2022.

[10] A. Majkowska, S. Mittal, D. F. Steiner, J. J. Reicher, S. M. McKinney, G. E. Duggan, K. Eswaran, P. H. Cameron Chen, Y. Liu, S. R. Kalidindi, A. Ding, G. S. Corrado, D. Tse, and S. Shetty, “Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation,” *Radiology,* vol. 294, no. 2, pp. 421-431, Feb, 2020.

[11] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks." pp. 4700-4708.

[12] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach,” *Biometrics,* vol. 44, no. 3, pp. 837-45, Sep, 1988.