ISSN:2790-1688 Volume-14-(2025)

Deep learning in Chest X-Ray Pneumonia Diagnosis: A Review of Research Advances

Xujia Liu

School of Software Engineering, Wuhan University of Technology, Wuhan, 430070, China

gemxjjxl@163.com

Abstract. Pneumonia, an acute respiratory infection affecting over 450 million people annually, remains a leading cause of global mortality. Chest X-ray (CXR) imaging serves as the primary diagnostic tool, yet its reliance on subjective radiologist interpretation often leads to delays and inconsistencies, particularly in resource-limited settings. Recent advancements in deep learning (DL) have revolutionized pneumonia diagnosis by enabling automated, high-accuracy analysis of CXR images. This review systematically examines the evolution of DL architectures for pneumonia detection, including convolutional neural networks (CNNs), vision transformers (ViTs), hybrid models, and emerging vision foundation models (VFMs). The author highlights their respective strengths, such as CNNs' localized feature extraction and ViTs' global context modeling, while addressing critical challenges like data scarcity, model interpretability, and computational barriers. Furthermore, the author discusses future directions, emphasizing hybrid model designs, federated learning for data diversity, and interpretability enhancements to bridge the gap between AI and clinical practice. By overcoming these challenges, DL-based diagnostic systems can achieve broader adoption, ultimately improving early detection and patient outcomes in both developed and developing regions.

Keywords: Deep learning; Chest X-ray; Pneumonia diagnosis; Vision foundation models.

1. Introduction

Pneumonia, an acute lung infection responsible for around 4% of global annual deaths[1], poses a significant global health threat. Chest X-ray (CXR) imaging is the primary diagnostic tool, but its reliance on radiologist interpretation often leads to subjective assessments, causing potential delays and inconsistent diagnoses, particularly in resource-constrained areas[2][3]. The advent of deep learning (DL) has transformed medical imaging by enabling automated analysis of complex patterns in radiographic data [4].This review employs a systematic literature research methodology, retrieving and analyzing relevant literature on Google Scholar over the past decade to synthesize the evolutionary trajectory of DL applications in pneumonia diagnosis, including the evolution from conventional convolutional neural networks (CNNs) to modern vision transformers (ViTs) and foundation models. The analysis focuses on architectural innovations, performance benchmarks, and clinical applicability, providing a structured comparison of their strengths and limitations, aiming to guide future research toward more robust, accessible, and clinically trustworthy systems. This review underscores the imperative to align AI innovations with clinical practice. By systematically addressing key barriers such as model interpretability and dataset heterogeneity, DL systems can enhance clinician confidence and expand access to underserved populations. The translational significance of this research lies in its potential to transform AI-driven diagnostics into practical healthcare solutions: optimized DL-based pneumonia diagnosis systems demonstrate measurable improvements in diagnostic accuracy while reducing diagnostic errors. Such advancements not only contribute to the evolution of precision medicine, but more importantly, support global health equity by making advanced diagnostic capabilities accessible across diverse healthcare settings, particularly in regions with limited medical resources.

2. Deep Learning Architectures for Pneumonia Diagnosis

2.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) serve as the foundational architecture for deep learning in medical imaging, particularly for pneumonia detection from chest X-ray (CXR) images. For example, AlexNet, a pioneering CNN, achieved early success in image classification by leveraging five convolutional layers to detect edges and textures, followed by max-pooling to reduce dimensionality [5]. While originally trained on the ImageNet dataset, AlexNet's architecture laid the groundwork for medical applications [6]. Another important model, DenseNet201, makes CNNs work better by connecting layers closely, which cuts down on unnecessary parameters and helps reuse features more effectively. [7]. In pneumonia detection, DenseNet201 achieved state-of-the-art accuracy by integrating multi-scale contextual information from CXR images. Its dense blocks allow for better gradient flow during training, making it particularly effective for distinguishing subtle differences between bacterial and viral pneumonia.

However, CNNs face limitations in capturing global anatomical context due to their local receptive field, motivating the need for architectures capable of modeling long-range dependencies, such as vision transformers (ViTs) discussed below.

2.2 Vision Transformers (ViTs)

Vision Transformers (ViTs) have revolutionized pneumonia detection by addressing the limitations of CNNs in global context modeling. Unlike CNNs, which rely on local receptive fields, ViTs leverage self-attention mechanisms to capture long-range spatial dependencies, making them particularly effective for analyzing complex medical images like chest X-rays (CXRs).

The foundational ViT architecture decomposes input images into patch embeddings, processes them through transformer blocks with multi-head self-attention, and integrates positional encoding to retain spatial information [8]. This design allows ViTs to handle images of arbitrary resolutions, a critical advantage for high-resolution medical imaging where down sampling might obscure subtle pathological features. For instance, in the proposed framework, ViT achieves 97.61% accuracy in CXR-based pneumonia detection, outperforming CNNs by effectively capturing global patterns like diffuse lung opacities and consolidation [9].Among its variants, DeiT (Data-Efficient Image Transformer) stands out for its ability to reduce reliance on large, labeled datasets through knowledge distillation [10]. By pre-training on unlabeled datasets and fine-tuning on CXRs, DeiT achieves competitive accuracy while minimizing computational costs, making it suitable for scenarios with limited medical data.

In summary, ViTs represent a paradigm shift in medical image analysis, offering superior global context modeling and adaptability to diverse CXR datasets. Their ability to combine fine-grained feature extraction with holistic image understanding positions them as indispensable tools for advancing pneumonia diagnosis. While ViTs currently require more computational resources than CNNs, ongoing optimizations in architecture design such as hybrid CNN-ViT models and pre-training strategies are bridging this gap.

2.3 Hybrid Architectures

Hybrid architecture, which integrates the capabilities of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), has emerged as a powerful solution in medical image analysis. By combining CNNs' proficiency in extracting fine-grained local features, such as lung textures in chest X-ray (CXR) images, with ViTs' strength in modeling long-range dependencies through self-attention mechanisms, these hybrid models overcome the limitations of their single-architecture counterparts.

Swin UNETR uses a U-Net backbone, which is a type of CNN architecture, to initially capture hierarchical local features like alveolar patterns in pneumonia cases. These features are then flattened into patch embeddings and fed into a Swin Transformer [11]. This sequential approach enables the model to first focus on local structures before conducting global reasoning. In brain tumor

segmentation tasks, Swin UNETR has achieved an impressive Dice score of 0.92, highlighting its effectiveness in integrating multi-scale features. UNETR, on the other hand, reformulates 3D segmentation tasks as sequence-to-sequence prediction problems. It uses a Transformer encoder to capture long-distance dependencies and global context and then connects this encoder directly to a CNN-based decoder through skip connections [12]. This design allows the decoder to utilize the global features learned by the Transformer. In multi-organ segmentation tasks on the BTCV dataset, UNETR has demonstrated state-of-the-art performance, achieving a Dice score of 0.89, and outperforming existing methods in brain tumor and spleen segmentation on the MSD dataset.

In summary, hybrid architecture offers significant advantages in medical image analysis. Models like Swin UNETR and UNETR have shown outstanding performance in various tasks, outperforming single-architecture models. However, challenges such as high computational costs and reliance on large, annotated datasets still exist. Future research will likely focus on developing more lightweight designs and self-supervised pre-training techniques to make these models more clinically applicable.

2.4 Vision Foundation Models (VFMs)

Vision Foundation Models (VFMs) have emerged as a revolutionary force in the landscape of pneumonia diagnosis within the realm of deep learning. These models, pre-trained on vast and diverse datasets, possess the ability to capture broad visual patterns and semantic information, which can be effectively transferred and adapted to the specific task of pneumonia detection in chest X-ray (CXR) images.

CLIP (Contrastive Language-Image Pretraining), a foundational VFM pre-trained on large-scale multimodal datasets, learns generalizable visual representations by aligning images with text descriptions [13]. While not explicitly designed for medical diagnosis, it has demonstrated potential in medical image analysis tasks by bridging visual concepts with language. Building upon CLIP, MedCLIP [14]is tailored for the medical domain. By decoupling images and texts for multimodal contrastive learning, it scales the usable training data and replaces the InfoNCE loss with semantic matching loss based on medical knowledge. This approach eliminates false negatives in contrastive learning, enabling MedCLIP to outperform state-of-the-art methods in zero-shot prediction, supervised classification, and image-text retrieval for medical images. Another model, MedSAM is an adaptation for medical imaging. It fine-tunes the Segment Anything Model (SAM) on a large medical image dataset with over a million image-mask pairs. MedSAM enables prompt-driven segmentation of various medical structures, offering accurate and efficient segmentation results [15].

In summary, these VFMs ranging from general multimodal models like CLIP [13]to medical-specific frameworks like MedCLIP [14], MedSAM [15], BiomedGPT [16]and RETFound [17]exemplify the transformative potential of foundation models in medical imaging. By leveraging pre-trained knowledge across diverse modalities and domains, CXR images enable more accurate and efficient pneumonia diagnosis.

3. Challenges

3.1 Data Limitations and Generalization

Most CXR datasets remain small or imbalanced, hindering the development of robust deep learning models. For example, the Pediatric CXR dataset contains only 1,583 normal images vs. 4,273 pneumonia cases [2], while the ChestX-ray14 dataset contains over 100,000 frontal-view X-ray images, its pneumonia annotations are sparse relative to other pathologies, with fewer than 1,500 positive cases explicitly labeled for pneumonia detection [18]. Such imbalances can bias model training, leading to overfitting on dominant classes.

Cross-institutional validation reveals even more fundamental limitations, with models experiencing catastrophic performance degradation (accuracies declining from 0.732 to 0.238) when tested on external datasets [19]. These generalization failures stem primarily from variations in

imaging protocols and scanner manufacturer differences. While federated learning approaches like NVIDIA Clara offer privacy-preserving solutions, they require standardized annotation protocols and homogeneous preprocessing pipelines to mitigate the 8-10% performance variations introduced by DICOM conversion differences.

3.2 Interpretability and Computational Barriers

The clinical adoption of DL-based pneumonia diagnosis faces two fundamental technological barriers. First, the opaque decision-making of convolutional neural networks (CNNs) and vision foundation models (VFMs) raises clinician skepticism, with studies showing radiologists frequently override AI recommendations due to lack of trust in unexplained predictions[20]. Second, the high computational cost of DL models hinders deployment, especially in low-resource settings. Training state-of-the-art models often requires expensive GPU infrastructure. Although efficient architectures like MobileNetV3 reduce computational costs through hardware-aware design and optimization, balancing accuracy and efficiency remains a key challenge [21]. This trade-off is critical, as over 70% of healthcare facilities in developing regions lack the infrastructure for standard DL deployment, limiting the global scalability of AI-assisted pneumonia diagnosis.

4. Future Directions

4.1 Model Architecture Innovation and Performance Optimization

Future advancements in model architecture should prioritize hybrid designs that merge the local feature extraction strengths of convolutional neural networks (CNNs) with the global contextual understanding of vision transformers (ViTs) [22]. Additionally, lightweight architectures like MobileViT, which optimize parameter efficiency through hybrid CNN-Transformer designs, should be further optimized for edge deployment in resource-constrained environments [23]. Vision foundation models (VFMs) like MedCLIP hold promise for zero-shot and few-shot learning, but their computational costs must be mitigated through techniques like knowledge distillation or parameter pruning to enable broader clinical adoption[24].

4.2 Data Quality Enhancement and Generalization

To address data limitations and improve generalization, future studies should emphasize cross-institutional data collaboration and standardized annotation protocols. Federated learning frameworks can help aggregate diverse datasets while preserving patient privacy, though challenges in DICOM conversion consistency must be resolved[25]. Data augmentation strategies should be combined with domain randomization to simulate variability in imaging protocols and scanner types[26].

4.3 Interpretability and Clinical Integration

Enhancing interpretability is critical for clinical trust and regulatory approval[27]. Techniques like Layer-wise Relevance Propagation (LRP) and Class Activation Maps (CAMs) should be integrated into models to provide pixel-level explanations of diagnostic decisions[28], highlighting regions like ground-glass opacities that align with radiologist expertise. Clinician-in-the-loop systems, which allow real-time validation of AI predictions, can bridge the gap between opaque models and clinical workflow. For computational scalability, edge-friendly architectures must be paired with cloud-based inference services to support low-resource settings, where over 70% of healthcare facilities lack advanced GPU infrastructure.Creating balanced datasets with clear pneumonia labels, instead of using limited labels in mixed disease datasets like ChestX-ray14, will also make the model stronger.

5. Conclusion

Deep learning has demonstrated significant potential in revolutionizing pneumonia diagnosis through chest X-ray analysis. From traditional CNNs to advanced vision foundation models, each

architecture offers unique advantages tailored to specific diagnostic needs: CNNs excel at localized feature extraction for fine-grained pathologies but face limitations in global context due to restricted receptive fields; vision transformers (ViTs) enhance global pattern recognition via self-attention for long-range spatial dependencies in CXRs; hybrid architectures integrate CNNs' local and ViTs'global modeling for superior segmentation despite higher computational costs; vision foundation models (VFMs) enable zero/few-shot learning through multi-modal pre-training for generalized diagnosis but confront significant computational barriers. While facing challenges such as data scarcity, interpretability issues, and computational demands, future research should focus on developing hybrid CNN-ViT models, improving data quality via federated learning, and enhancing interpretability to foster clinical adoption. By addressing these challenges, DL-based diagnostic tools can achieve broader accessibility and reliability, ultimately improving patient outcomes worldwide.

References

[1] G. A. Roth et al., "Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017," The lancet, vol. 392, no. 10159, pp. 1736–1788, 2018.

[2] D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," cell, vol. 172, no. 5, pp. 1122–1131, 2018.

[3] C. L. F. Walker et al., "Global burden of childhood pneumonia and diarrhoea," The Lancet, vol. 381, no. 9875, pp. 1405–1416, 2013.

[4] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60–88, 2017.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012, Accessed: Jun. 05, 2025.

[6] T. Rahman et al., "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray," Applied Sciences, vol. 10, no. 9, p. 3233, 2020.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708. Accessed: Jun. 05, 2025.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020, Accessed: May 17, 2025.

[9] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using Vision Transformers on chest X-rays," Scientific reports, vol. 14, no. 1, p. 2487, 2024.

[10] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in International conference on machine learning, PMLR, 2021, pp. 10347–10357. Accessed: Jun. 05, 2025.

[11] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, vol. 12962, A. Crimi and S. Bakas, Eds., in Lecture Notes in Computer Science, vol. 12962. , Cham: Springer International Publishing, 2022, pp. 272–284. doi: 10.1007/978-3-031-08999-2_22.

[12] A. Hatamizadeh et al., "Unetr: Transformers for 3d medical image segmentation," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 574–584. Accessed: Jun. 11, 2025.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.

[14] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, 2022, p. 3876. Accessed: Jun. 03, 2025.

[15] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," Nature Communications, vol. 15, no. 1, p. 654, 2024.

[16] K. Zhang et al., "Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks," arXiv e-prints, p. arXiv-2305, 2023.

[17] Y. Zhou et al., "A foundation model for generalizable disease detection from retinal images," Nature, vol. 622, no. 7981, pp. 156–163, 2023.

[18] P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," Dec. 25, 2017, arXiv: arXiv:1711.05225. doi: 10.48550/arXiv.1711.05225.

[19] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," PLoS medicine, vol. 15, no. 11, p. e1002683, 2018.

[20] F. Cabitza, R. Rasoini, and G. F. Gensini, "Unintended consequences of machine learning in medicine," Jama, vol. 318, no. 6, pp. 517–518, 2017.

[21] A. Howard et al., "Searching for mobilenetv3," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1314–1324. Accessed: May 17, 2025.

[22] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022. Accessed: May 17, 2025.

[23] H.-I. Liu et al., "Lightweight Deep Learning for Resource-Constrained Environments: A Survey," ACM Comput. Surv., vol. 56, no. 10, pp. 1–42, Oct. 2024, doi: 10.1145/3657282.

[24] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017, Accessed: May 17, 2025.

[25] M. F. Sohan and A. Basalamah, "A systematic review on federated learning in medical image analysis," IEEE Access, vol. 11, pp. 28628–28644, 2023.

[26] Z. Liu, Q. Lv, Y. Li, Z. Yang, and L. Shen, "MedAugment: Universal Automatic Data Augmentation Plug-in for Medical Image Analysis," Aug. 14, 2024, arXiv: arXiv:2306.17466. doi: 10.48550/arXiv.2306.17466.

[27] R. Thakur, "Explainable AI: developing interpretable deep learning models for medical diagnosis," Int J Multidisciplinary Res, vol. 6, 2024.

[28] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification," Frontiers in aging neuroscience, vol. 11, p. 194, 2019.