XAI-Enabled Equivariant Vision Transformer for Pediatric Pneumonia Detection

Hiba Tasnim., Dr. Thasni T

Department of Computer Science and Engineering College of Engineering Trivandrum, India

DOI: https://doi.org/10.51244/IJRSI.2026.1303000211

Received: 30 March 2026; Accepted: 05 April 2026; Published: 17 April 2026

ABSTRACT

Pneumonia remains the leading infectious cause of death in children under five, claiming over one million lives annually, with diagnostic delays in low-resource settings often exacerbated by rotated, flipped, or poorly aligned chest X-rays acquired from restless infants. Although convolutional neural networks (CNNs) and standard Vision Transformers (ViTs) have driven automated detection accuracies beyond 95% on benchmark datasets, their performance degrades significantly in real-world pediatric imaging due to limited geometric invariance and lack of clinical interpretability. This survey reviews advances reported recent years in deep learning approaches for pediatric pneumonia detection from chest X-rays, covering CNN-based hi- erarchies, global-context Vision Transformers, multimodal fusion with clinical markers, emerging equivariant transformer designs, and explainable AI techniques. Despite substantial progress in accuracy and sensitivity, persistent challenges include orienta- tion sensitivity, reliance on heavy data augmentation or non- imaging inputs, post-hoc interpretability, and limited standalone deployment in resource-constrained environments. The analy- sis highlights the need for geometrically robust, intrinsically interpretable, and clinically deployable models to bridge the gap between benchmark performance and reliable real-world pediatric screening.

Index Terms: Deep learning, Convolutional Neural Networks, Pneumonia, Pediatric Pneumonia, Chest X-ray, Explainable AI.

INTRODUCTION

Pneumonia remains the leading cause of death among children under five, claiming over 1.4 million lives annually because of diagnostic delays in primary-care settings [1]. Early antibiotic treatment can cut mortality by up to 40%, yet nearly half of all pediatric cases are still misdiagnosed in low- resource areas, mainly because chest X-rays of restless infants often come out rotated, flipped, or badly centred [2]. High disagreement between radiologists and the severe shortage of pediatric specialists only make things worse.

The first attempts at automated chest X-ray analysis used convolutional neural networks (CNNs). Basic CNNs reached 93.75% accuracy in simple normal vs pneumonia classification by learning hierarchical features layer by layer [3]. Today's CNNs add multi-scale processing and channel attention, push- ing sensitivity past 96% even on highly imbalanced pediatric datasets [4]. They are fast, small, and easy to run on phones, but they treat any rotation or flip as a completely new image accuracy can drop 15–20% with just a 90° turn unless heavy data augmentation is used [5].

Vision Transformers (ViTs) have significantly advanced medical image analysis by leveraging global self-attention mechanisms over image patches, achieving 95.48% accuracy and 96.18% F1-score through effective modeling of long- range anatomical dependencies [6]. Enhanced variants such as DeiT and Cross-ViT further improve performance through knowledge distillation and multi-scale feature fusion, reaching 96.61% accuracy and 99.42% recall [7], [8]. Moreover, large- scale pre-training has enabled ViTs to outperform conventional CNN-based approaches across various medical imaging tasks [5]. However, a key limitation lies in the use of standard posi- tional encodings, which disrupt the model's ability to preserve geometric symmetry. Consequently, rotated or misaligned pe- diatric chest X-rays are often interpreted as entirely different inputs, leading to substantial performance degradation. This issue is particularly critical in pediatric imaging, where

data acquisition is frequently affected by motion artifacts due to patient non-cooperation.

Only one recent work directly tackles this. Xu et al. [9] proposed GE-ViT, which replaces standard positional encoding with an E(2)-equivariant version that preserves both translation and rotation symmetry. By tying attention weights to group neighbourhoods, their model produces stable, steerable fea- tures without ever having to relearn what "up" means exactly what misaligned infant scans need.

Practical CNN-based systems have also matured fast. Menaka et al. [10] built T-ResNet for binary classification and turned it into real software used in clinics. Saikrishna et al. [11] wrapped MobileNet into a Django web app for bacterial/viral/COVID-19 typing. Annepaka et al. [12] showed a simple custom CNN can hit 99.77% accuracy on Kaggle pediatric sets. Behera et al. [13] added Squeeze-and-Excitation blocks plus Grad-CAM to EfficientNetB7 (98.31% accuracy with visible heatmaps), and Manaf et al. [14] used GAN- generated images plus Flask deployment to fight data shortage. Multimodal models that mix X-rays with blood results or clinical notes now exceed 94 percentage accuracy even when some data are missing [15], and two-stage designs can beat radiologists at bacterial-vs-viral classification [2]. Interpretability is still only post-hoc (Grad-CAM, etc.) [16], and no published transformer backbone has native equivari- ance except the single attempt above [9]. Other interesting directions include CNN-LSTM on lung sounds [17] and self- supervised equivariant ResNets [18], but none solve the full real-world pediatric puzzle.

This review examines the rapid progress made from 2023 to 2025 across these seventeen key studies, focusing on architectural robustness to real-world orientation problems, clinical interpretability, and actual deployment feasibility in low-resource settings.

LITERATURE REVIEW

A. CNN-Based Architectures for Pneumonia Detection

Convolutional Neural Networks (CNNs) pioneered auto- mated CXR analysis using localized hierarchical filters to extract edges, textures, and consolidation patterns. Singh et al.[4] developed a lightweight custom CNN with ReLU-activated convolutional layers, batch normalization, and dropout. Inputs are resized to 224×224, normalized, and augmented (rota- tion, flip) to learn spatial hierarchies low layers detect lung boundaries, deep layers identify opacities enabling binary classification in pediatric pneumonia. The design supports edge deployment in low-resource settings.

Hasan et al. [1] discussed how CNNs continue to lead this area, especially when using transfer learning with models like VGG16 and ResNet50. Their review explains that keeping the early layers fixed helps retain general image features, while fine-tuning the later layers allows the model to better adapt to pediatric chest X-rays. Singh [7] explored DenseNet- 121, where each layer builds directly on features learned earlier, making it effective for picking up small, multi-scale pneumonia signs.

Yao et al. [5] introduced a pathology-focused CNN as the first part of the AMPNet pipeline, using global-local attention to emphasize true lung consolidations instead of irrelevant image details. More recent work has continued to improve localization by adding stronger attention mechanisms and combining multiple models.Hasan et al. [1] integrated channel- wise attention in residual blocks to amplify opacity channels while suppressing ribs/heart—mimicking radiologist focus. Singh et al. [4] applied knowledge distillation to compress large teacher reasoning into compact students. Li et al. [2] used a CNN encoder in RMT for local CXR features before multimodal fusion.

Chen et al. [6] leveraged EfficientNet with compound scal ing to adapt to variable scan quality. Matsoukas et al. [9] showed pretrained CNNs rely on translational invariant kernels but need heavy augmentation. Ranaweera and Pathirana [3] noted CNNs' local receptive fields limit long-range context in diffuse viral pneumonia. No CNN embeds native equivari ance or clinician aligned interpretability, leaving orientation brittleness and black-box decisions unaddressed.

People take big pre-trained models like VGG, ResNet, DenseNet, EfficientNet, MobileNet or Xception and fine-tune them on the small public paediatric datasets. For example, one group tried seven different CNNs together as an ensemble and got around 90–91% accuracy when splitting normal, bacterial and viral cases [18]. Another team checked four transfer- learning models on kids aged 1–5 years and found DenseNet-

121 worked best at 86.8%, with Xception almost the same [20]. Someone else mixed features from EfficientNetB0 with LightGBM and hit 96.16% on the usual 5,856-image set [23]. A different paper built six simple custom CNNs from scratch and still reached 93.59% [22]. A few teams also put MobileNet inside web apps using Flask or Django so nurses can actually use them right now. The common trick is heavy data flipping and rotating during training because CNN performance de- grades significantly under geometric transformations such as rotation and inversion. Without those fake copies, accuracy drops 15–20% on real baby scans that are always crooked. No CNN yet knows on its own that a lung looks the same however you turn the picture.

B. Vision Transformer Models in Medical Imaging

Vision Transformers (ViTs) replace localized convolutions with global self-attention over image patches to capture long range dependencies. Ranaweera and Pathirana [3] split the chest X-rays into uniform patches, converted them into linear embeddings, and then applied multi-head attention to capture how these patches relate to one another, allowing the model to represent widespread infiltrates more effectively. Preprocess- ing techniques like gamma correction and data augmentation further improve the model's reliability by helping it handle variations in exposure common in pediatric scans.

Singh et al. [8] trained DeiT using a distillation token during its ImageNet stage and then adapted it to pediatric chest X- rays so the model could better capture global lung patterns. In another study, Singh [7] presented Cross-ViT, which analyzes both small and large image patches at the same time and then merges their attention outputs, improving the model's ability to detect early signs of disease.

Chen et al. [6] adapted a base ViT with layered self-attention to focus on vascular markings and ground-glass opacities, modeling non-local patterns without inductive biases. Mat- soukas et al. [9] demonstrated self-supervised masked image modeling enables transfer to data-scarce pediatric CXR tasks. Multimodal extensions include Li et al. [2] using a ViT branch in RMT with BERT-like text encoding and mask attention to ignore missing inputs, and Yao et al. [5] integrating global local ViT-style attention in AMPNet to fuse pathology with blood embeddings.

Xu et al. [10] proposed GE-ViT, overcoming positional encoding's disruption of equivariance through a novel E(2) equivariant operator that preserves translation and rotation symmetry. Patches are embedded with group-equivariant posi- tional encodings, ensuring self-attention respects affine trans- formations while using tied weights for parameter efficiency. This guarantees steerable, generalizable representations, highly relevant for pediatric CXR misalignment without relearning orientations.

Vision Transformers are slowly coming into paediatric X- rays, but not many papers use them yet because the datasets are tiny compared to ImageNet. The biggest proper paediatric dataset so far is PediCXR from Vietnam with over 9,000 scans and actual bounding boxes drawn by a senior kid radiologist for 36 different findings [19].

Despite strengths, standard ViTs lack geometric priors, treating rotation/flip as learned patterns. Matsoukas et al.[9] noted relearning per orientation. Ranaweera and Pathi- rana [3] relied on heavy augmentation, not true invariance. Interpretability is post-hoc: Grad-CAM in Chen et al. [6] shows where, not why. Hasan et al. [1] and Singh et al. [4] applied ViTs in ensembles or distillation, but compute limits deployment. No ViT enforces native equivariance or clinician aligned heatmaps.

C. Multimodal Fusion Systems for Pediatric Pneumonia

Multimodal approaches combine chest X-rays with clinical notes or laboratory markers to strengthen

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

pneumonia diagnosis, especially when pediatric data are limited. Li et al.[2] proposed RMT, which uses two parallel branches: a ViT- style encoder for CXR patches and a BERT-like module for clinical text. During fusion, masked attention down-weights any missing inputs, and the model is trained to predict both diagnosis and severity at the same time, allowing it to handle incomplete records more reliably.

Yao et al. [5] presented AMPNet, a two-stage system. The first stage uses a pathology-focused CNN with global–local attention to highlight true areas of consolidation. The second stage processes blood indicators such as WBC and CRP using a 1D CNN, and then merges both streams with joint attention to distinguish bacterial from viral infections. Other studies make use of simpler metadata fusion. Hasan et al.[1] attached demographic information to CNN feature vectors before applying weighted fusion. Singh et al. [4] compressed multimodal features into smaller CNNs, and Chen et al. [6] incorporated patient history into EfficientNet through direct concatenation.

Transformer based fusion extends attention across different data types. Ranaweera and Pathirana [3] guided ViT patch attention using text prompts. Singh et al. [8] and Singh[7] inserted blood-related tokens into DeiT and Cross-ViT, merging them with scale-aware attention. Matsoukas et al. [9] used masked-modality modeling to estimate missing clinical inputs.

Even with these advances, multimodal systems have major drawbacks: they depend on non-image data that are often unavailable in most primary care settings, they are com- plex and slow, and they risk mismatching information across modalities. More importantly, the CXR branch still lacks equivariance, making models sensitive to rotations or flips. Their explanations only operate at the modality level, and none are optimized for mobile deployment. As a result, improving CXR only robustness and spatial interpretability remains an open challenge.

D. Explainable AI (XAI) and Pediatric Specific Challenges

Explainable AI (XAI) plays an essential role in building clinician confidence and meeting regulatory expectations in pediatric imaging. Chen et al. [6] used Grad-CAM with ViT models to produce heatmaps that point out regions such as ground-glass opacities and lung consolidation, helping clini- cians review model decisions after the prediction. Hasan et al. [1] applied SHAP to CNN ensembles to measure how specific pixels and metadata—such as density in the lower lung zones—contribute to a decision. Still, both techniques are post-hoc, require heavy computation, and mainly show where the model focuses rather than the reasoning behind it.

Pediatric cases bring added complications, including differ- ences in anatomy, frequent motion blur, and limited datasets. Ranaweera and Pathirana [3] attempted to reduce variabil- ity through augmentation, but no existing approach directly uses age-specific priors. Singh et al. [8] and Singh [7] used attention-rollout methods in DeiT and Cross-ViT to trace long- distance feature relationships, but the results remain more suitable for research than everyday clinical use. Li et al. [2] and Yao et al. [5] expanded XAI to multimodal models through modality-level scores and localized attention maps, which help with severity assessment, but these systems break down when some inputs are missing and still lack geometric stability.

Xu et al. [10] moved the field further by introducing E(2)- equivariant positional encoding in GE-ViT. By structuring attention around group neighborhoods, their method keeps feature responses consistent under rotation and translation, leading to more reliable and efficient heatmaps—especially important when pediatric scans include motion or alignment issues. Godbin & Jasmine [23] switch to LightGBM after pulling features from EfficientNet or DenseNet because Light- GBM instantly tells you which features mattered most and runs much faster than SHAP. Kumar et al. [24] took only blood tests and vital signs (no picture at all), fed them to Random Forest, got 96% accuracy for survival prediction, and then checked the same data with different XAI tools – SHAP, LIME, ELI5 and QLattice all agreed the top five clues are ICU transfer, potassium level, creatinine, cyanosis and sodium. Real kid problems almost never get fixed: big thymus that looks like a tumour, skinny chests in malnourished children, blurry shots because the baby won’t stop crying, or explaining the result to worried parents in simple words.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

Despite these efforts, current XAI approaches remain post- hoc, non-equivariant, and not ready for real-time deployment. No existing system provides stable, rotation-robust explana- tions that align with clinical expectations or meet pediatric- specific needs such as handling motion artifacts and supporting clear communication with families. Achieving dependable, pediatric-appropriate XAI is still a major unresolved challenge.

E. Equivariant Learning Approaches for Geometric Robust- ness

Equivariant learning tries to handle changes in image orien- tation in a more direct way. In many CNN-based approaches, rotations and flips are usually managed through data augmen- tation and pooling layers. However, this approach is not always reliable, especially when the test images have orientations that were not well represented during training. Equivariant models attempt to overcome this by incorporating symmetry into the network design itself, so that the learned features respond consistently when the input is rotated, flipped, or shifted. This becomes important in pediatric chest X-rays, where images are often slightly misaligned because of patient movement during acquisition.

Earlier studies have applied these ideas mainly in convolution-based models by changing how filters work, so that they behave more consistently under transformations. This allows the model to focus more on meaningful lung features instead of being affected by the orientation of the image. Compared to standard CNNs, these methods can reduce the dependence on large amounts of augmented data and may perform better when the dataset is small or contains variability. Recently, similar concepts have been introduced into transformer-based models. For instance, the E(2)-equivariant Vision Transformer modifies positional encoding in a way that helps the attention mechanism remain stable even when the image is rotated [10]. There are also self-supervised methods where the model is trained using different transformed versions of the same image and learns to produce similar outputs [16]. Although these approaches show promise, they are still not widely applied in pediatric pneumonia detection. Most current methods continue to rely on augmentation rather than explicitly handling geometric variations at the model level, suggesting that this area still needs further exploration.

Comparative Analysis

A clear methodological progression is observed from CNN- based hierarchical feature extraction to Vision Transformer- driven global context modeling, and finally to multimodal fusion with auxiliary clinical inputs. Attention mechanisms en- hance localization and interpretability across paradigms, while equivariant designs ([10]) address geometric brittleness. De- spite high accuracy on controlled benchmarks, no framework combines native rotation invariance, intrinsic explainability, and pediatric specific deployment highlighting the need for unified, robust, and clinician trusted systems in real-world diagnostic pipelines.

Table 1 gathers the papers published from 2023 to 2025 and puts them side by side so the differences jump out immediately: what backbone each team used, the highest accuracy or sensitivity they claimed, whether they did anything serious about rotated or upside-down films, how they tried to explain their predictions, whether the model needs blood results or notes to work, and whether anyone actually ran it on a cheap phone.

Looking at the table from top to bottom, three clear stages appear. The first half is all convolutional: small custom CNNs, MobileNet, EfficientNet, T-ResNet, and similar [4], [11]–[14], [17]. These are the fastest, the lightest, and the only ones that have already been turned into real apps nurses can open on a phone or tablet today. Some hit 99.77 percentage accuracy on public datasets. But scroll to the column about rotation and every single one says the same thing: without stuffing the training set full of artificially flipped and spun images, the model falls flat the moment a real pediatric X-ray arrives sideways.

The middle rows switch to Vision Transformers — DeiT, Cross-ViT, ordinary ViTs with masked pre-training [3], [6]– [8]. They regularly beat the CNNs on sensitivity because they finally look at the whole lung at once instead of tile by tile. Yet the rotation column still looks grim: a turned- around film is treated as something the network has never seen before. Only one paper, Xu et al. [10], actually builds rotation- handling into the architecture itself, and even there the tests are done on clean, computer-generated rotations rather than

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

the wobbly scans of real babies. The very best numbers sit in the multimodal corner (RMT and AMPNet [2], [5]). They sneak in blood counts or typed clinical notes and jump past 97 percentage accuracy even when some data are missing. The catch is written in plain sight: none of them can run without those extra inputs, so they are useless the moment a child walks in and the lab results are still hours away.

Challenges

Among the recent studies studies examined between 2023 and 2025, none has managed to deliver a truly standalone pe- diatric pneumonia detector that works solely from the chest X- ray and still satisfies the basic demands of real-world clinical practice. Every published model either collapses in accuracy when the radiograph is rotated, flipped, poorly centerd, or blurred by a child's movement, or it only stays accurate because the training process was flooded with thousands of artificially transformed copies of each image. At the same time, no architecture has shown it can run in real time comfortably under 200 milliseconds per image on the cheap, low-memory Android phones and tablets that are the only computing devices found in most rural health centers and district-hospital casualty wards in low and middle income countries.

On top of that, none of the reviewed works produces visual explanations that remain anchored to the correct part of the lung no matter how the X-ray is oriented, and none properly copes with the everyday pediatric confounders a large thymus that looks like a mediastinal mass, a narrow chest in a malnourished child, or a big overlapping heart shadow without leaning on extra laboratory results or clinical notes. The complete absence of any single published system that simultaneously solves these four problems (true geometric robustness, genuine lightweight speed on low-end phones, orientation stable heatmaps, and pediatric specific anatomical resilience) remains the biggest roadblock preventing safe, large-scale rollout of automated screening exactly where child- hood pneumonia continues to kill the most children.

DISCUSSION

Papers published from 2020 to 2025 regularly celebrate accuracies above 95–99 percentage on clean, carefully aligned test sets, yet these numbers crumble the instant a chest X- ray from a squirming toddler arrives upside-down or off- centre — a routine event in any pediatric ward. Both CNNs and Vision Transformers lose 15–50 percentage points the moment the image is rotated or mirrored [9]. The standard fix is to flood training with thousands of artificially rotated and flipped copies, but the models simply learn to recognise those synthetic versions rather than acquiring any real grasp of rotational symmetry.

Table I Comparative Summary of Reviewed Studies on Automated Pneumonia Detection from Chest X-rays

Study	**Primary Modality / Cohort Focus**	**Model Architecture / Core Innovation**	**Dataset / Cohort Details**	**Reported Performance (Metric)**
[1] Hasan et al. (2024)	Chest X-ray (CXR) / **Mixed (pediatric + adult)**	CNN Ensemble with transfer learning and channel-wise attention gating	Mixed pediatric and adult CXR cohorts from public datasets	Accuracy 98.7%, Sensitivity 96.2%, F1-score 0.94 (ensemble)
[2] Li et al. (2024)	CXR + clinical text / **Pedi- atric**	Robust Multimodal Transformer (RMT) with dual-branch encoding and mask at- tention	Custom pediatric dataset with CXR and structured clinical notes	Accuracy >94% (modality-agnostic), F1-score 0.92 (severity grading)
[3] Ranaweera and Pathirana	CXR / **Pediatric**	Standard Vision Transformer (ViT) with gamma correction and	Guangzhou Women and Chil- dren's Medical Center pedi- atric CXR	Accuracy 95.48%, F1-score 96.18%, Specificity 94.2%

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

(2024)		augmentation	dataset	
[4] Singh et al. (2023)	CXR / **Mixed (mostly pedi-atric)**	Lightweight custom CNN with hierar- chical convolution and knowledge dis- tillation	Kaggle Pneumonia CXR dataset (mixed age)	Accuracy93.75% (binary), Inference <50 ms on edge
[5] Yao et al. (2024)	CXR + blood biomarkers / **Pediatric**	Two-stage AMPNet with pathology- aware CNN and global-local attention fusion	Custom pediatric cohort with CXR and WBC/CRP values	F1 (Bacterial): 0.7781, F1(Viral): 0.8123 vs. radiologist F1: 0.5970
[6] Chen et al. (2023)	CXR / **Mixed (adult-focused)**	Base Vision Transformer (ViT-patch16) with layered self-attention	COVID-19CXR dataset (mixed age)	Accuracy 99.57% (3- class), AUC 0.998
[7] Singh (2024)	CXR / **Pediatric**	Cross-ViT with dual-scale patch pro- cessing and attention fusion	Kaggle pediatric pneumonia CXR dataset	Accuracy 88.25%, Recall 99.42%, High sensitivity to subtle infiltrates
[8] Singh et al. (2024)	CXR / **Pediatric**	DeiT (Data-efficient Image Transformer) with distillation token	Kaggle pediatric CXR dataset	Accuracy 97.61%, Sensitivity 95%, Efficient fine-tuning
[9] Matsoukas et al. (2023)	CXR / **Adult + general med-ical**	Pretrained ViTs with self-supervised masked image modeling	General medical imaging benchmarks	Matches CNN base- lines; >20% drop under rotation
[10] Xu et al. (2023)	General images (CXR appli-cable) / **No clinical cohort**	GE-ViT with E(2)-equivariant positional encoding and group convolution	ImageNet + synthetic symme-try benchmarks	8–12% gain over standard ViT under rotation; parameter-efficient steerability

To better organize the existing research, the reviewed meth- ods can be grouped into four main categories based on their design and data usage. CNN-based approaches focus on ex- tracting local spatial features and are widely used due to their efficiency and suitability for deployment. Transformer-based models extend this by capturing global relationships across the image, although they often require large-scale training and lack robustness to geometric variations. Multimodal methods combine chest X-rays with additional clinical information to improve performance, but their dependence on non-imaging data limits practical use in many settings

Explanation tools fare no better. Whether researchers use Grad-CAM, SHAP, or attention-rollout maps, the coloured overlays are generated after the fact. Feed the same film in a different orientation and the highlighted “suspicious” region often leaps to the opposite lung or vanishes completely. For a doctor in a remote clinic who needs to point at a phone screen and convince a frightened parent that the machine really sees pneumonia, an explanation that dances around with every tilt of the image is useless.

The highest-scoring systems almost always sneak in blood counts, CRP values, or typed clinical notes to boost their numbers [2], [5]. In the rural health posts and overcrowded emergency rooms where most children

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue III March 2026

actually die of pneu- monia, those laboratory results are either unavailable or still handwritten on a scrap of paper. Pure chest-X-ray diagnosis therefore remains the only feasible route for mass screening, but the current generation of models is nowhere near ready for that job.

Future Scope

The journey from high benchmark numbers to actual lives saved in rural clinics still faces substantial hurdles. A pressing need exists for diagnostic systems that do not falter when a restless child's X-ray is taken upside-down or at an odd angle something radiologists handle effortlessly but current networks treat as an entirely new image. Equally important are explanation tools that point to the same lung region whether the scan is rotated or not, so a village health worker using a cheap phone can still show a parent exactly why the model suspects pneumonia. Making such capabilities run smoothly on low-cost Android devices without internet access remains one of the biggest practical challenges ahead.

Data scarcity and the unique appearance of infant chests continue to limit progress. Future work should explore ways to teach models the normal developmental changes a large thymus in a six-month old, a narrow chest in a malnourished toddler, so they are less easily confused by anatomy rather than disease. Learning effectively from thousands of unlabelled adult films before fine-tuning on the few hundred labelled pediatric cases, or borrowing knowledge from related tasks such as adult pneumonia or tuberculosis detection, could dramatically reduce the need for expensive annotations. At the same time, building mechanisms that flag uncertain or poor- quality images for human review would prevent over-confident mistakes in critical situations.

Ultimately, laboratory-level performance means little until it is tested where pneumonia kills most. Large-scale prospec- tive studies in district hospitals and primary-health centres of low-income regions are overdue, along with deliberate checks that the same accuracy holds for severely malnourished children, those with HIV, or those from minority communi- ties. Simple additions a thirty-second recording of breathing sounds through a smartphone microphone or a pulse-oximeter trace might boost confidence without requiring blood tests. Only when models prove themselves robust, understandable, lightweight, and fair in these real environments will artificial intelligence truly help cut childhood pneumonia deaths in the places that need it most.

Despite the progress made in automated pediatric pneumo- nia detection, a few key issues still remain. One major problem is that many models are not robust to changes in image orientation, so their performance drops when chest X-rays are rotated, misaligned, or affected by motion. Another limitation is the heavy reliance on data augmentation, where models learn to handle transformations indirectly instead of actually understanding them. Finally, there is still no single approach that brings together robustness, interpretability, and ease of deployment, which makes it difficult to use these models effectively in real-world and low-resource environments.

CONCLUSION

The rapid evolution of deep learning has significantly ad- vanced automated pneumonia detection from chest X-rays, with convolutional neural networks establishing foundational hierarchical feature extraction, Vision Transformers introduc- ing powerful global context modeling, multimodal systems enhancing diagnostic depth through auxiliary data integra- tion, and explainable AI techniques beginning to address clinical trust. These paradigms have collectively improved detection accuracy, sensitivity to subtle pediatric anomalies, and resilience in data scarce environments. However, critical challenges persist: geometric brittleness under rotation or flip common in restless infant imaging, post-hoc and computa- tionally heavy interpretability, over reliance on non-imaging inputs in multimodal frameworks, and absence of real-time, clinician-aligned deployment pipelines.

Despite methodological maturity, no existing approach na- tively embeds orientation invariance, delivers intrinsic, radi- ologist readable heatmaps, or ensures standalone robustness across primary care settings. Pediatric specific constraints such as anatomical variability, motion artifacts, limited labeled data, and the need

for transparent parent-clinician communication remain largely unaddressed. These gaps underscore a clear need for a unified framework that combines equivariant pro- cessing, attention driven explainability, and mobile integration without sacrificing performance or accessibility.

Future systems must move beyond research benchmarks to deliver trustworthy, deployable, and pediatric optimized diag- nostics transforming AI from a supportive tool into a reliable partner in global child health. By closing these foundational gaps, automated pneumonia detection can achieve bedside impact, reducing diagnostic delays, guiding precise treatment, and ultimately saving young lives in both emergency and low resource contexts.

REFERENCES

1. M. R. Hasan, S. M. A. Ullah, and S. M. R. Islam, "Recent advancement of deep learning techniques for pneumonia prediction from chest X-ray image," Medical Reports, vol. 7, p. 100106, 2024.
2. J. Li, Z. Nan, G. Qi et al., "Assessing severity of pediatric pneumonia using multimodal transformers with multi-task learning," Digital Health, vol. 10, pp. 1–19, 2024.
3. K. Ranaweera and P. N. Pathirana, "Leveraging Vision Transformers for Enhanced Accuracy in Pneumonia Detection from Medical Imaging Data," in Proc. 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2024, pp. 1–4.
4. G. Singh, K. Guleria, and S. Sharma, "An Efficient Convolutional Neural Network Model for Pneumonia Detection in Chest X-ray Images," in Proc. 2024 12th International Conference on Internet of Everything, Microwave, Embedded, Communication and Networks (IEMECON), 2024, pp. 1–5.
5. D. Yao, Z. Xu, Y. Lin, and Y. Zhan, "Accurate and intelligent diagnosis of pediatric pneumonia using X-ray images and blood testing data," Frontiers in Bioengineering and Biotechnology, vol. 11, p. 1058888, 2023.
6. T. Chen, I. Philippi, Q. B. Phan et al., "A vision transformer machine learning model for COVID-19 diagnosis using chest X-ray images," Healthcare Analytics, vol. 5, p. 100332, 2024.
7. G. Singh, "Comparative Analysis of Vision Transformers and Traditional Deep Learning Approaches for Automated Pneumonia Detection in Chest X-Rays," arXiv preprint arXiv:2507.10589, 2025.
8. S. Singh, M. Kumar, B. K. Verma et al., "Efficient pneumonia detection using Vision Transformers on chest X-rays," Scientific Reports, vol. 14, p. 2487, 2024.
9. J. Ko, S. Park, and H. G. Woo, "Optimization of vision transformer- based detection of lung diseases from chest X-ray images," BMC Medical Informatics and Decision Making, vol. 24, no. 1, p. 191, 2024.
10. R. Xu, K. Yang, K. Liu, and F. He, "E(2)-Equivariant Vision Trans- former," in Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence, 2023, pp. 2356–2366.
11. K. Menaka, S. R. G, B. P. K, R. U, and S. S. Prabhu, "Detection and Recognition of Pneumonia in Chest X-ray Images using Deep Learning Methods," in 2025 International Conference on Emerging Technologies in Computing and Communication (ETCC), 2025, pp. 1–6, doi: 10.1109/ETCC65847.2025.11108354.
12. G. V. Saikrishna and C. Lakshmi, "Pneumonia Detection using Convolu- tional Neural Network," in Proceedings of the 6th International Confer- ence on Inventive Research in Computing Applications (ICIRCA-2025), 2025, pp. 1426–1431, doi: 10.1109/ICIRCA65293.2025.11089799.
13. N. Annepaka, N. Garg, D. Saxena, P. Tanwar, M. R. Chinta, K. Choud- hary, E. P. Kumar, and R. Kumar, "A Smart Approach to Pneumonia Detection Using Deep Learning," in 2024 2nd International Con- ference on Advances in Computation, Communication and Informa- tion Technology (ICAICCIT), 2024, pp. 842–847, doi: 10.1109/ICAIC- CIT64383.2024.10912224.
14. S. K. Behera, K. M. Gopal, and S. B. Punuri, "A Deep Learning- Based Pneumonia Detection System with Explainable AI for Medical Decision Support," in 2025 11th International Conference on Com- munication and Signal Processing (ICCSP), 2025, pp. 694–699, doi: 10.1109/ICCSP64183.2025.11089357.
15. P. V. K. Pandey, S. Chakravarty, S. S. Sahu, and C. Chin, "CNN- LSTM-Based Lung Sound Analysis for Pneumonia Detection," in SoutheastCon 2025, 2025, pp. 1072–1077, doi: 10.1109/SOUTHEAST-CON56624.2025.10971692.

16. G. D'Souza, N. V. S. Reddy, and K. N. Manjunath, "Localization of lung abnormalities on chest X-rays using self-supervised equivariant attention," Biomedical Engineering Letters, vol. 13, no. 1, pp. 21–30, Feb. 2023, doi: 10.1007/s13534-022-00249-5.
17. A. Manaf and N. Mughal, "AI-ENHANCED PEDIATRIC PNEUMO- NIA DETECTION: A CNN-BASED APPROACH USING DATA AUG- MENTATION AND GENERATIVE ADVERSARIAL NETWORKS (GANS)," arXiv preprint arXiv:2507.09759, 2025.
18. E. Ayan, B. Karabulut, and H. M. U¨ nver, "Diagnosis of pediatric pneumonia with ensemble of deep convolutional neural networks in chest X-ray images," Arabian Journal for Science and Engineering, vol. 47, no. 2, pp. 2123–2139, Feb. 2022.
19. H. H. Pham, N. H. Nguyen, T. T. Tran, T. N. M. Nguyen, and H. Q. Nguyen, "PediCXR: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children," Scientific Data, vol. 10, art. no. 240, Apr. 2023.
20. M. Salehi, R. Mohammadi, H. Ghaffari, N. Sadighi, and R. Reiazi, "Automated detection of pneumonia cases using deep transfer learning with paediatric chest X-ray images," British Journal of Radiology, vol. 94, no. 1121, art. no. 20201263, May 2021.
21. V. Arya and T. Kumar, "Ensemble classifier for pediatric pneumonia diagnosis in chest X-ray images," in Proc. International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), Faridabad, India, Nov. 2023, pp. 1340–1345.
22. R. G. de Luna et al., "High accuracy diagnosis of pediatric pneumonia: Achieved using convolutional neural network," in Proc. IEEE Interna- tional Conference on Imaging Systems and Techniques (IST), Batangas, Philippines, Oct. 2024, pp. 1–6.
23. A. B. Godbin and S. G. Jasmine, "Pediatric pneumonia detection in chest X-ray images: A deep feature analysis approach enhanced with LightGBM," in Proc. IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Chennai, India, Aug. 2024, pp. 1–6.
24. R. Kumar, V. Srirama, K. Chadaga, M. H, N. Sampathila, S. Prabhu, and R. Chadaga, "Using explainable machine learning methods to predict the survivability rate of pediatric respiratory diseases," IEEE Access, vol. 12, pp. 189515–189528, Dec. 2024.
25. I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," Scientific Reports, vol. 9, art. no. 6381, Apr. 2019.