International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

A Hybrid CNN–Transformer Deep Learning Architecture for Automated Pneumonia Detection from Chest Radiographs

Abdulazeez Danjuma[1], M. S. Aliyu[2], Zaharradeen S. Iro, Usman Abdullahi Musa, Abdulrrazaq A Umar, Abdullahi Ahmed Talba
[1,2,3,45,] Federal University Dutse
[6]Federal College of Education (Technical) Potiskum

Abstract — Pneumonia is a prominent preventable cause of death in children, leading to 14% of all fatalities under five and over 700,000 paediatric deaths annually. Chest radiography is the principal diagnostic tool, but interpretation depends on radiologist availability and inter-observer variability, causing severe bottlenecks in low- and middle-income countries.A hybrid CNN–Transformer architecture with convolutional local feature extraction and multi-head self-attention for binary pneumonia classification on chest radiographs was designed, implemented, and evaluated. Studies used a publicly available chest radiograph dataset of 5,856 pictures (4,273 pneumonia, 1,583 normal). Five convolutional blocks (32→64→64→128→256 filters) with batch normalisation and dropout (0.3–0.5) generate a 6,400-dimensional feature vector, which is reshaped into a token sequence using positional encoding and passed through two Transformer encoder layers (8 attention heads, feed-forward dimension 512 The hybrid model had 92.0% accuracy, 96.1% precision, 92.8% recall, 94.4% F1-score, and 0.998 AUC. Confusion-matrix analysis on the held-out test partition (n = 879; 641 pneumonia, 238 normal) gave 595 true positives, 214 true negatives, 24 false positives, and 46 false negatives Self-attention and convolutional feature extraction increase discriminative performance over CNN-only baselines, with precision outperforming accuracy. External multi-institutional validation, multi-class subtyping, and attention-based interpretability are needed before clinical application.

Keywords— Artificial Intelligence, Branding, Brand Strategy, Consumer Behaviour, Personalization, Customer Engagement, Digital Marketing

I. INTRODUCTION

Pneumonia is an acute lower respiratory infection that continues to impose a disproportionate burden on the youngest and oldest segments of the global population. The World Health Organization (WHO, 2021) attributes approximately 14% of all deaths in children under five years of age to pneumonia, corresponding to more than 700,000 paediatric deaths each year. The burden is concentrated in low- and middle-income countries, where diagnostic and therapeutic capacity is weakest relative to need. In Nigeria, pneumonia is the single largest infectious cause of childhood death, contributing over 162,000 fatalities annually (UNICEF, 2021), a toll driven less by the intrinsic severity of the disease than by inadequate health infrastructure, shortages of trained personnel and constrained diagnostic resources in rural and peri-urban settings (Wonodi et al., 2020).

Chest radiography remains the principal imaging modality for pneumonia confirmation because it is inexpensive, widely deployable and rapid. Its diagnostic value, however, is realised only through expert interpretation. In settings where radiologists are scarce, radiographs may be acquired but not read in a clinically actionable timeframe, and where they are read, agreement between readers is imperfect. Inter-observer variability in the identification of pulmonary opacities means that the same radiograph may yield different diagnostic conclusions depending on the reader (Pham et al., 2022). The consequence is a diagnostic pipeline whose weakest link is human interpretive capacity rather than image acquisition, and this is precisely the link that automated analysis is positioned to strengthen.

Deep learning has consequently attracted sustained attention as a means of automating radiographic triage. Convolutional neural networks in particular have demonstrated the ability to learn discriminative representations directly from pixel data without hand-engineered features, in several reported settings approaching or matching expert-level performance on constrained classification tasks (Abdou et al., 2022; Nessipkhanov et al., 2023). A substantial literature now documents CNN-based pneumonia detection using architectures such as CheXNet (Rajpurkar et al., 2017; Babu et al., 2024), fine-tuned VGG-16 backbones (Beri & Sharma, 2024) and EfficientNet variants adapted through transfer learning (Singla & Gupta, 2024).

© 2026 IJSRET

Three recurring limitations temper this progress. The first is an accuracy ceiling that falls short of clinical requirements: Morcos et al. (2023), applying the TorchXRayVision framework to paediatric imaging, reported 76.5% accuracy, and Bukhari et al. (2025), using a multi-scale CNN with atrous spatial pyramid pooling, reported 83.8% — both improvements on earlier work but neither sufficient for autonomous deployment. The second is instability and overfitting: Singla and Gupta (2024) documented pronounced oscillation in validation accuracy and loss across epochs, and Beri and Sharma (2024) identified class imbalance as a direct determinant of degraded reliability, indicating that even well-established backbones generalise poorly on modest, skewed medical datasets. The third, and the one that motivates the present work, is architectural: the convolution operator aggregates information within a bounded receptive field, so a CNN builds global understanding only indirectly, through depth. Pneumonic consolidation, however, is frequently diffuse, bilateral or asymmetric, and its radiographic signature is partly relational — defined by how opacity in one lung zone compares with the contralateral zone and with the overall parenchymal pattern. Purely convolutional models capture such long-range dependencies inefficiently.

The Transformer encoder offers a direct remedy. Its multi-head self-attention mechanism computes pairwise interactions across all positions in a sequence in a single operation, giving every element access to global context regardless of spatial separation (Vaswani et al., 2017), a property subsequently transferred to image recognition (Dosovitskiy et al., 2021). Pure vision transformers, however, are notably data-hungry and lack the inductive biases of locality and translation equivariance that make CNNs efficient on datasets of the size typically available in medical imaging. This suggests a hybrid rather than a substitutive design: convolutional layers to extract local texture and edge structure economically, followed by attention layers to reason over the resulting feature set globally. This study develops and evaluates such a hybrid CNN–Transformer network for binary pneumonia classification on chest radiographs. The specific objectives are: (i) to design an architecture in which convolutional feature extraction, Transformer-based global attention and an explicit fusion stage operate in sequence; (ii) to implement and train this architecture on a class-imbalanced public chest radiograph dataset with a reproducible augmentation and partitioning protocol; (iii) to quantify performance using accuracy, precision, recall, F1-score, confusion-matrix error decomposition and ROC analysis; and (iv) to position the resulting performance against published results for comparable tasks. The contribution is a concrete, fully specified hybrid configuration together with an error analysis oriented toward the asymmetric clinical costs of false-negative and false-positive radiographic reads.

II. MATERIALS AND METHODS

1. Study Design

This was a retrospective computational study using a de-identified, publicly available chest radiograph dataset. Because the study involved no prospective recruitment, no intervention and no access to identifiable patient information, institutional ethical approval was not required; the study was conducted in accordance with the terms of use of the source dataset. The work followed a two-phase framework of analysis and design. The analysis phase characterised the diagnostic problem, the data requirements and the operating constraints of the target deployment setting; the design phase specified the network architecture, the data-processing pipeline and the evaluation protocol reported below.

2. Dataset

The dataset comprised 5,856 frontal chest radiographs labelled at the image level as either pneumonia-positive or normal. Of these, 4,273 images (73.0%) were pneumonia cases and 1,583 (27.0%) were normal, giving a class-imbalance ratio of approximately 2.7:1 in favour of the positive class. This distribution reflects the referral-enriched composition typical of curated radiographic corpora rather than the prevalence of pneumonia in a screening population, and it carries a known risk of biasing an untreated classifier toward the majority class. The imbalance was therefore addressed explicitly at both the partitioning and augmentation stages, as described in Sections 2.3 and 2.4.

Table 1. Composition of the chest radiograph dataset by class and partition.

Partition	Images (n)	Proportion of dataset
Training	4,099	70%
Validation	878	15%
Test	879	15%
Total	5,856	100%
Class — Pneumonia	4,273	73.0%
Class — Normal	1,583	27.0%
Imbalance ratio	2.7:1	pneumonia : normal

© 2026 IJSRET

3. Preprocessing and data partitioning

All radiographs were resized to a uniform resolution of 224 × 224 pixels and pixel intensities were rescaled to the interval [0, 1]. Class labels were binary-encoded, with pneumonia assigned the positive class. The dataset was then divided by stratified random sampling into training (70%, n = 4,099), validation (15%, n = 878) and test (15%, n = 879) partitions. Stratification preserved the 73:27 class ratio within each subset, ensuring that validation and test performance were measured against the same class prior encountered during training and that no partition was inadvertently enriched or depleted of positive cases. The test partition was held out entirely from model selection and used only for the single final evaluation reported in Section 3.

4. Data augmentation

Real-time augmentation was applied to the training partition only; validation and test images were passed to the model in preprocessed but otherwise unmodified form. The augmentation operators were random rotation within ±30°, random zoom of up to 20% inward or outward, random horizontal and vertical translation of up to 10% of the corresponding image dimension, and random horizontal flipping. Vertical flipping was deliberately excluded because an inverted thorax is anatomically implausible and would introduce a transformation the model can never encounter at inference. These operators were selected to emulate the acquisition variability that arises in practice from differences in patient positioning, source-to-detector distance and detector alignment, rather than to synthesise new pathology. Because augmentation was applied stochastically at each epoch, the model was exposed to a continuously varying realisation of the training set without any increase in the number of underlying unique studies. Over the 12 training epochs this corresponds to 49,188 augmented sample presentations (4,099 unique training images × 12 epochs). This figure quantifies training exposure only; the number of unique underlying studies remained 4,099 throughout and no synthetic increase in effective sample size is claimed.

5. Proposed hybrid architecture

The proposed network comprises three functional stages executed in sequence: a convolutional feature extractor, a Transformer encoder, and a fusion and classification head. Figure 1 presents the overall architecture.

Figure 1: Hybrid CNN–Transformer architecture

Convolutional feature extraction

The convolutional backbone transforms the raw radiograph into a high-dimensional spatial representation. It consists of five convolutional blocks with progressively increasing filter counts of 32, 64, 64, 128 and 256. Each block applies convolution followed by batch normalisation, which stabilises the distribution of layer inputs and permits a higher effective learning rate, and dropout regularisation in the range 0.3 to 0.5, which suppresses co-adaptation of feature detectors. Max-pooling between blocks reduces spatial resolution, expanding the effective receptive field of subsequent layers while containing computational cost. The hierarchy captures increasingly abstract structure: early blocks respond to edges and local intensity gradients corresponding to rib margins and vascular markings, while deeper blocks respond to the coarser texture and opacity patterns characteristic of consolidation. The output of the final block is flattened into a 6,400-dimensional feature vector.

Transformer encoder

The flattened convolutional output is reshaped into a sequence of embeddings and augmented with positional encodings, which reinstate the spatial ordering that flattening discards — a necessary step, since self-attention is permutation-invariant and would otherwise treat the feature set as an unordered collection. The sequence is passed through two Transformer encoder layers. Each layer applies multi-head self-attention with eight heads, followed by residual connection and layer normalisation, and then a position-wise feed-forward sub-layer of dimension 512 with a further residual connection and normalisation.

Self-attention computes, for each element of the sequence, a weighted sum over all other elements, with weights derived from the compatibility of learned query and key projections:

© 2026 IJSRET

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k})\, V \qquad (1)$$

where Q, K and V are the query, key and value matrices and d_k is the key dimension, the scaling factor $\sqrt{d_k}$ preventing the dot products from entering regions of vanishingly small softmax gradient. Employing eight parallel heads allows the model to attend to several distinct relational patterns simultaneously — for example, contralateral zone comparison and apex-to-base gradient — rather than collapsing all such relationships into a single attention distribution. This mechanism is the source of the architecture's capacity to represent long-range dependencies that convolution captures only through accumulated depth.

Fusion and classification

The convolutional feature representation and the attention-enhanced representation are combined in an explicit fusion layer by concatenation followed by dense projection. This design preserves both streams rather than allowing the attention pathway to overwrite the convolutional one, so that fine-grained local texture and global contextual reasoning both contribute to the decision. The fused representation passes through a fully connected layer of 128 units with rectified linear unit activation, and then to a single output unit with sigmoid activation producing a calibrated probability of pneumonia. A decision threshold of 0.5 was applied to convert this probability to a binary label.

6. Implementation and training configuration

The convolutional backbone was initialised with ImageNet-pretrained weights and fine-tuned on the radiographic data, while the Transformer encoder layers were trained from random initialisation. Optimisation used the Adam algorithm (Kingma & Ba, 2015) with an initial learning rate of 1×10^{-4} and binary cross-entropy as the loss function:

$$L = -(1/N) \Sigma \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right] \qquad (2)$$

where y_i is the ground-truth label and $\hat{y}_i$ the predicted probability for the i-th image across N samples. The batch size was 32 and training ran for 12 epochs. Two callbacks governed convergence: early stopping, which terminated training when validation loss ceased to improve and thereby limited overfitting to the training distribution, and reduction of the learning rate on plateau, which lowered the step size when progress stalled to permit finer convergence. Table 2 summarises the complete configuration. The RNN and LSTM baselines reported in Section 3.2 were trained under an identical protocol: the same stratified partitions, preprocessing, augmentation pipeline, optimiser, learning rate, loss function, batch size, epoch budget, callbacks and decision threshold, differing only in the sequence-modelling stage applied to the flattened convolutional features.

Table 2. Architectural and training hyperparameters of the proposed hybrid model.

Component	Parameter	Value
CNN backbone	Convolutional blocks	5
	Filters per block	32, 64, 64, 128, 256
	Regularisation	Batch normalisation; dropout 0.3–0.5
	Flattened output dimension	6,400
Transformer	Encoder layers	2
	Attention heads	8
	Feed-forward dimension	512
	Positional encoding	Applied
Fusion / head	Fusion strategy	Concatenation + dense projection
	Dense layer	128 units, ReLU
	Output layer	1 unit, sigmoid
Training	Optimiser	Adam
	Learning rate	1×10^{-4}
	Loss function	Binary cross-entropy
	Batch size	32
	Epochs	12
	Callbacks	Early stopping; ReduceLROnPlateau
	Input resolution	224 × 224 × 3
	Decision threshold	0.50

7. Evaluation metrics

Model performance was quantified using four complementary metrics derived from the confusion matrix, in which TP and TN denote correctly classified pneumonia and normal cases respectively, and FP and FN denote normal cases misclassified as pneumonia and pneumonia cases misclassified as normal. Accuracy expresses the proportion of all cases classified correctly:

© 2026 IJSRET

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

Because accuracy is dominated by the majority class in an imbalanced dataset, it was interpreted alongside rather than in place of the class-specific metrics. Sensitivity, equivalently recall or the true positive rate, expresses the proportion of actual pneumonia cases the model identifies:

$$Sensitivity = Recall = TP / (TP + FN) \quad (4)$$

This metric carries the greatest clinical weight, since a false negative withholds treatment from a patient who requires it. Precision, or positive predictive value, expresses the proportion of positive predictions that are correct:

$$Precision = TP / (TP + FP) \quad (5)$$

The F1-score is the harmonic mean of precision and recall, penalising models that achieve one at the expense of the other:

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (6)$$

In addition, receiver operating characteristic curves were constructed by sweeping the decision threshold across its full range, and the area under each curve (AUC) was computed as a threshold-independent measure of discriminative capacity.

8. Computing environment

Experiments were executed on a workstation equipped with an NVIDIA RTX 3090 graphics processing unit, an Intel Core i7 central processing unit, 16 GB of system memory and solid-state storage, running Ubuntu 20.04 LTS. The implementation used Python 3.8 with TensorFlow and PyTorch, supported by OpenCV, NumPy, Pandas, SciPy and Matplotlib for image handling, numerical computation and visualisation.

III. RESULTS

1. Dataset characteristics and augmentation

The class distribution of the source dataset, summarised in Table 1, showed a 2.7:1 excess of pneumonia over normal cases. Stratified partitioning preserved this ratio across the training, validation and test subsets. Augmentation applied to the training partition generated a continuously varying stream of transformed images across epochs — 49,188 augmented sample presentations over 12 epochs, drawn from the same 4,099 unique training images — while leaving the number of unique underlying studies unchanged; the effect was therefore an increase in the diversity of training signal rather than an increase in independent information. Figure 2 illustrates the class distribution before and after augmentation.

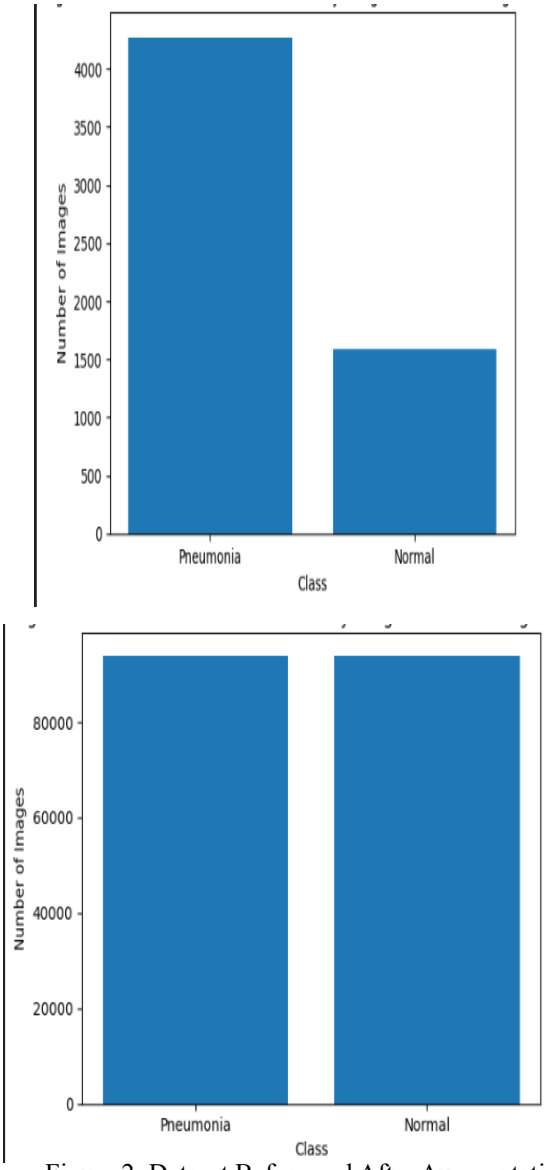

Figure 2: Dataset Before and After Augmentation

2. Classification performance

The hybrid CNN–Transformer model was evaluated on the held-out test partition alongside two recurrent baselines trained under an identical protocol. Table 3 reports the results.

© 2026 IJSRET

Table 3. Classification performance of the proposed hybrid model and recurrent baselines.

Model	Accuracy	Precision	Recall	F1-score
RNN	90.0%	93.7%	92.5%	93.1%
LSTM	91.0%	94.9%	92.7%	93.8%
Hybrid CNN–Transformer	92.0%	96.1%	92.8%	94.4%

The hybrid model achieved the highest value on every metric. Its accuracy of 92.0% exceeded the LSTM baseline by one percentage point and the RNN baseline by two. The margin was wider in precision, where the hybrid model reached 96.1% against 94.9% for the LSTM and 93.7% for the RNN, a gain of 1.2 to 2.4 percentage points. Recall was 92.8% for the hybrid model, marginally above the LSTM at 92.7% and the RNN at 92.5%. The F1-score of 94.4% was 0.6 points above the LSTM and 1.3 points above the RNN. The pattern is consistent: the hybrid architecture's advantage is concentrated in its ability to avoid false positives rather than in a uniform improvement across all error types.

3. Error decomposition

The confusion matrix for the hybrid model is presented in Table 4. The test partition contained 879 radiographs, which stratification composed of 641 pneumonia and 238 normal cases. The model correctly classified 595 pneumonia cases and 214 normal cases. It produced 46 false negatives, that is pneumonia-positive radiographs assigned a normal label, and 24 false positives, normal radiographs assigned a pneumonia label. The four cells sum to 879, the size of the test partition, and yield an accuracy of 92.0%, a recall of 92.8%, a precision of 96.1% and a specificity of 89.9%. False negatives therefore outnumbered false positives by a factor of approximately 1.9, indicating that the residual error is weighted toward missed disease rather than over-calling.

Table 4. Confusion matrix for the hybrid CNN–Transformer model on the held-out test partition (n = 879; 641 pneumonia, 238 normal).

	Predicted: Pneumonia	Predicted: Normal
Actual: Pneumonia	595 (TP)	46 (FN)
Actual: Normal	24 (FP)	214 (TN)

4. Discrimination analysis

Receiver operating characteristic analysis, shown in Figure 3, yielded an AUC of 0.998 for the hybrid CNN–Transformer model, 0.992 for the LSTM and 0.973 for the RNN. All three models discriminated far above chance. The absolute differences between them were small, and the ordering of models by AUC matched the ordering by accuracy and F1-score, but the AUC values were sufficiently compressed toward the upper bound that they conveyed less information about relative model quality than the threshold-dependent metrics in Table 3.

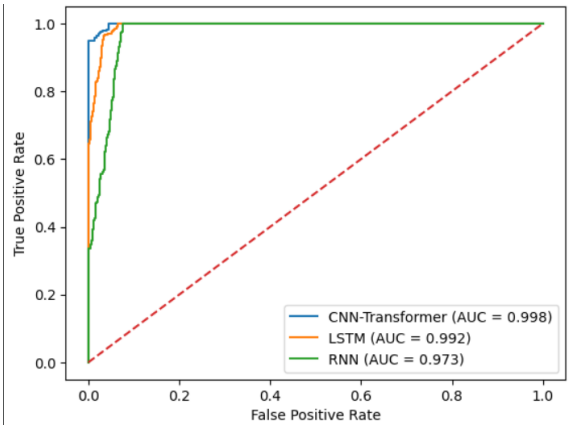

Figure 3: ROC Curve Comparison of Models for Pneumonia Detection

5. Qualitative assessment

Sample predictions on individual test radiographs are shown in Figure 5, with the predicted class displayed against the ground-truth label.

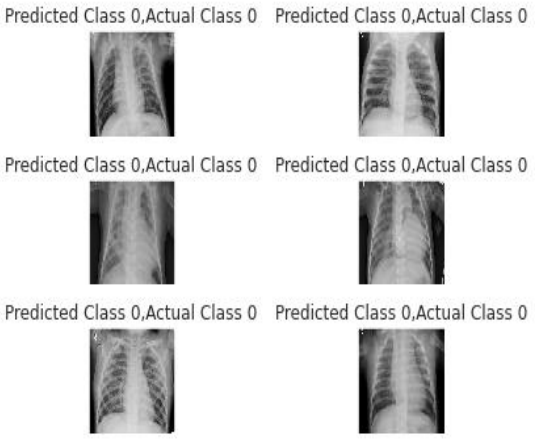

Figure 4: Sample Model Predictions

© 2026 IJSRET

In the illustrated cases the model assigned the normal class in agreement with the reference label, demonstrating expected behaviour on unambiguous studies. Qualitative inspection of this kind indicates operating behaviour on representative images but does not substitute for the aggregate metrics reported above.

6. Comparison with published studies

Table 5 positions the present results against three published approaches to radiographic pneumonia detection.

Table 5. Comparison of the proposed approach with previously published methods.

Study	Method	Reported performance
Morcos et al. (2023)	TorchXRayVision (adult-trained model applied to paediatric imaging)	Accuracy 76.54%; sensitivity 79.83%; specificity 67.66%; PPV 86.95%; NPV 55.41%; F1 0.83
Zunaed et al. (2024)	Contrastive learning with embedding similarity	AUROC 0.8464
Bukhari et al. (2025)	Multi-scale CNN with atrous spatial pyramid pooling	Accuracy 83.75%
Present study	Hybrid CNN–Transformer	Accuracy 92.0%; precision 96.1%; recall 92.8%; F1 94.4%; AUC 0.998

The proposed model reported higher headline performance than each comparator. Morcos et al. (2023) recorded 76.54% accuracy with markedly asymmetric error behaviour — specificity of 67.66% and negative predictive value of 55.41% indicate substantial difficulty in confirming pneumonia-negative cases. Bukhari et al. (2025) improved on this with 83.75% accuracy through multi-scale contextual aggregation. Zunaed et al. (2024) reported an AUROC of 0.8464 using contrastive representation learning. Section 4.4 addresses the methodological caveats that qualify these comparisons.

IV. DISCUSSION

1. Principal findings

The hybrid CNN–Transformer architecture achieved 92.0% accuracy, 96.1% precision, 92.8% recall, 94.4% F1-score and an AUC of 0.998 on held-out chest radiographs, outperforming recurrent baselines trained under an identical protocol on every metric. The more informative observation, however, concerns the shape rather than the magnitude of the improvement. The gain in accuracy over the baselines was 1.0 to 2.0 percentage points, whereas the gain in precision was 1.2 to 2.4 points and the gain in recall only 0.1 to 0.3 points. The hybrid model's advantage therefore lies less in overall correctness than in the cleanliness of the separation it draws between classes: it commits fewer false-positive errors relative to its true-positive volume than either recurrent comparator, and the near-ceiling AUC is consistent with a decision function whose positive and negative score distributions overlap only marginally.

2. Architectural interpretation

The performance ordering is consistent with the architectural rationale set out in Section 1. Both recurrent baselines process flattened image features as an ordered sequence, propagating information step by step. As sequence length grows, early context is attenuated, and a flattened radiographic feature map is a long sequence whose original two-dimensional adjacency structure has been destroyed by flattening. The recurrent models are consequently applied to spatial data in a form that suits them poorly. Self-attention avoids this bottleneck: it relates every position to every other in one operation, with no sequential decay and no dependence on the linear ordering of the flattened vector beyond what positional encoding restores.
The concatenation fusion strategy is likely to contribute as well. By preserving the convolutional representation alongside the attention-enhanced one, the classification head retains access to fine-grained textural evidence — the local opacity and margin characteristics on which the radiographic diagnosis of consolidation substantially rests — while gaining the contextual comparison that attention supplies. It should nonetheless be acknowledged that these mechanistic accounts are inferred from aggregate performance, not demonstrated directly. No ablation study isolating the contribution of the Transformer encoder, the fusion layer or the number of attention heads was performed, and such an ablation would be required to attribute the observed gains to specific architectural components with confidence.

3. Clinical implications of the error profile

Aggregate metrics obscure a distinction that matters at the bedside: the two error types carry different costs. A false negative returns a normal report on a patient with pneumonia, delaying antibiotic therapy and permitting clinical deterioration. A false positive returns a pneumonia report on a healthy patient, prompting unnecessary follow-up imaging or antimicrobial exposure. The first is generally the graver error in an acute respiratory presentation, and it is the error the

© 2026 IJSRET

present model makes more often — 46 false negatives against 24 false positives.

Recall of 92.8% means the model identifies the large majority of true pneumonia cases, which is the property most relevant to a screening or triage role. Precision of 96.1% means it rarely raises a false alarm, which limits the downstream burden of unnecessary investigation. Taken together these figures suggest a model reasonably balanced for a triage application in which its output is reviewed rather than acted upon autonomously. The 46 missed cases nevertheless represent a residual clinical risk that must be communicated explicitly to any deploying institution: this is a measured error rate, not a solved problem, and any deployment protocol should specify how discordant or clinically suspicious cases are escalated to human review irrespective of model output. It is also worth noting that the decision threshold of 0.5 is a default rather than an optimum. Because sensitivity and precision trade off against one another as the threshold moves, an institution prioritising case capture over specificity could lower the threshold to convert some false negatives into false positives — a legitimate configuration choice that the ROC analysis in Figure 3 characterises but that this study did not itself optimise.

4. Comparison with prior work

The proposed model reported higher performance than each of the comparator studies in Table 5, but the comparison requires qualification, because the studies differ in dataset, population and task in ways that affect intrinsic difficulty.

Morcos et al. (2023) evaluated an adult-trained model on paediatric radiographs. Their reported 76.54% accuracy reflects a domain-shift problem — the mismatch between adult-derived priors and paediatric thoracic anatomy — which is a different and arguably harder challenge than the within-domain classification addressed here. Their low negative predictive value of 55.41% is diagnostic of that mismatch rather than of the underlying architecture. Bukhari et al. (2025) targeted multi-scale lesion detection across both COVID-19 and pneumonia, a broader and more heterogeneous task than binary pneumonia classification, so part of the accuracy gap reflects task scope rather than architectural merit. Zunaed et al. (2024) applied contrastive learning and embedding similarity, methods typically selected for low-label-data regimes; their AUROC of 0.8464 may therefore reflect a deliberately data-scarce evaluation setting rather than an inferior method. Reported metrics are not directly commensurable across studies that differ in these respects, and the present results should be read as competitive within their own evaluation setting rather than as establishing architectural superiority in general.

5. Limitations

Several limitations qualify the findings. First, the dataset originates from a single public source, and the pronounced class imbalance means the model observed substantially more positive than negative examples. Although stratified splitting and augmentation mitigate the consequences, performance on an independent, multi-institutional dataset acquired under different equipment, protocols and patient demographics may differ materially, and no such external validation was performed.

Second, training ran for a comparatively small number of epochs under early stopping. This constrains overfitting but may also mean the Transformer component — trained from random initialisation rather than pretrained — did not fully converge relative to what a longer schedule or a pretrained vision-transformer backbone might achieve.

Third, the task formulation is binary. The model does not distinguish bacterial from viral pneumonia, a clinically consequential distinction that governs antibiotic decisions and would require additional labelled data and a multi-class output. Fourth, no interpretability mechanism was implemented. Clinicians cannot see which regions of a radiograph drove a given prediction, a gap that bears directly on clinical trust and on regulatory acceptability. The attention weights within the Transformer layers are in principle extractable and offer a natural route to saliency visualisation, but this was not undertaken here.

Fifth, no ablation study was conducted, no comparison was made against modern CNN or pure vision-transformer baselines under matched conditions, and results are reported from a single training run without confidence intervals or repeated-seed variance estimates. Differences of one to two percentage points between models should therefore be interpreted with caution, as they may fall within run-to-run variability.

6. Conclusion

This study designed, implemented and evaluated a hybrid CNN–Transformer architecture for pneumonia detection on chest radiographs, coupling five convolutional blocks for local feature extraction with two multi-head self-attention encoder layers for global contextual reasoning and an explicit concatenation fusion stage. The model achieved 92.0% accuracy, 96.1% precision, 92.8% recall, an F1-score of 94.4% and an AUC of 0.998, exceeding recurrent baselines and published comparator results, with its clearest advantage in precision.

© 2026 IJSRET

The findings support the design premise that convolution and attention are complementary rather than competing operators for radiographic classification, and that combining them yields a better-separated decision function than either paradigm applied alone at this data scale. Translating this into clinical utility will require external validation on multi-institutional data, extension to bacterial-versus-viral subtyping, attention-based saliency mapping to make predictions inspectable, threshold calibration to the error asymmetry of the intended deployment setting, and prospective evaluation of the model as a decision aid within a real diagnostic workflow.

REFERENCES

1. Abdou, M. A., et al. (2022). Literature review: Efficient deep neural networks techniques for medical image analysis. Neural Computing and Applications, 34(8), 5791–5812.
2. Babu, T., Naik, P. K., Nair, R. R., & Pallavi, K. (2024). A robust convolutional neural network architecture for automated pneumonia detection from chest X-ray images.
3. Beri, M., & Sharma, N. (2024). VGG-16 model-based pneumonia detection with chest X-ray images.
4. Bukhari, M. A. S., Bukhari, F., Asif, M., Aljuaid, H., & Iqbal, W. (2025). A multi-scale CNN with atrous spatial pyramid pooling for enhanced chest-based disease detection.
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. Proceedings of the International Conference on Learning Representations (ICLR).
6. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of the International Conference on Learning Representations (ICLR).
7. Morcos, G., Yi, P. H., & Jeudy, J. (2023). Applying artificial intelligence to pediatric chest imaging: Reliability of leveraging adult-based artificial intelligence models.
8. Nessipkhanov, D., et al. (2023). Deep learning approaches for the detection of lung diseases from chest X-ray images.
9. Pham, H. H., et al. (2022). Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing, 437, 186–194.
10. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.
11. Singla, S., & Gupta, R. (2024). Pneumonia detection from chest X-ray images using transfer learning with EfficientNetB1.
12. UNICEF. (2021). Pneumonia in children: Nigeria country profile. United Nations Children's Fund. https://data.unicef.org/topic/child-health/pneumonia/
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
14. Wonodi, C., et al. (2020). Pneumonia in Nigeria: Epidemiology, burden and control.
15. World Health Organization. (2021). Pneumonia in children [Fact sheet]. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/pneumonia
16. Zunaed, M., Hasan, A., & Hasan, T. (2024). Improving pediatric pneumonia diagnosis with adult chest X-ray images utilizing contrastive learning and embedding similarity.

© 2026 IJSRET