International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Explainable Pneumonia Detection in Chest X-Rays: A Comparative Study of CNNs and Vision Transformers

Abdulaziz Alshaibani[1], Ziyad Albaqami[2], Abdullah Sheikh[3]

[1]Master's in AI graduate student, Computer Science Department, College of Computers and Information Technology, Taif University Taif, Kingdom of Saudi Arabia

[2]Master's in AI graduate student, Computer Science Department, College of Computers and Information Technology, Taif University Taif, Kingdom of Saudi Arabia

[3]Assistant Professor of Computer Science, Computer Science Department, College of Computers and Information Technology, Taif University Taif, Kingdom of Saudi Arabia

ABSTRACT: Pneumonia is a leading cause of global mortality, especially among children and the elderly, and chest radiography (CXR) remains the most widely used modality for its diagnosis. While deep learning has reached or exceeded radiologist-level performance on this task, the resulting models are still treated as opaque black boxes, which is a critical barrier to clinical deployment. In this work, we present a comparative and *interpretable* computer-aided-diagnosis (CAD) framework for pneumonia detection that combines three modern image-recognition backbones—a convolutional ResNet-50, a Swin Transformer (Swin-T), and a modernised convolutional network (ConvNeXt-T)—with Gradient-weighted Class Activation Mapping++ (Grad-CAM++) explanations. The three backbones were fine-tuned on the public Kermany chest X-ray dataset using a class-balanced training subset, weighted cross-entropy and an early-stopping protocol, and then evaluated on the held-out test set of 624 images. The Swin-T backbone achieved the best overall performance with a test accuracy of 95.51%, an F1-score of 0.95 and only 11 false negatives out of 234 normal cases, outperforming both ResNet-50 (93.11%) and ConvNeXt-T (88.94%). Grad-CAM++ heatmaps generated from the convolutional and transformer feature maps consistently localised on the affected pulmonary regions, providing radiologically plausible visual evidence for each prediction. Compared with five recent state-of-the-art pneumonia detectors, our Swin-T-based pipeline reaches a competitive accuracy while delivering layer-faithful visual explanations, supporting its use as a transparent decision-support tool in clinical workflows.

KEYWORDS: Pneumonia detection, Chest X-ray, Deep learning, ResNet, Swin Transformer, ConvNeXt, Explainable AI, Grad-CAM++, Computer-aided diagnosis.

1. INTRODUCTION

Pneumonia is an acute infection of the lung parenchyma that fills the alveoli with fluid or pus and impairs gas exchange. According to the World Health Organization, it remains one of the leading causes of childhood mortality worldwide and is responsible for a substantial fraction of hospital admissions in elderly patients(World Health Organization 2022). Early diagnosis is critical because timely antibiotic or antiviral therapy can dramatically reduce both case fatality and the length of hospital stay(Ghuse and Monga 2024).

Chest radiography (CXR) is the first-line imaging modality used to confirm or rule out pneumonia. It is cheap, widely available, and quickly acquired, but its interpretation is intrinsically difficult: signs of consolidation, ground-glass opacity, and infiltrates can be subtle, overlap with other thoracic diseases, and depend strongly on patient positioning and image quality. As a consequence, radiological reads are subject to high inter-observer variability and to a non-trivial rate of missed diagnoses, especially in high-volume emergency settings(Litjens et al. 2017). The growing global imaging workload further amplifies the demand for automated, reliable triage tools.

Over the last decade, Artificial Intelligence (AI), and in particular Deep Learning (DL) with Convolutional Neural Networks (CNNs), has demonstrated remarkable success across a broad spectrum of problems—from medical image analysis to financial time-series forecasting and language understanding (Alghamdi et al. 2025). In thoracic radiology, transfer-learning pipelines built on pretrained CNN backbones such as VGG, ResNet, DenseNet, and Xception have repeatedly reached accuracies in the 90–98% range on public

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

CXR benchmarks (Chothani et al. 2024; Singla et al. 2024; Shrimali 2024), and recent transformer-based architectures have begun to surpass them (Park et al. 2022; Tyagi et al. 2024). Hybrid approaches that fuse imaging with structured patient information have likewise shown promise (Aljuaid et al. 2025).

Despite this technical maturity, the clinical adoption of DL-based CAD systems for pneumonia is still limited. The principal reason is the well-known "black-box" nature of deep models(Joseph and Anitha 2024): clinicians are unwilling to base diagnostic or therapeutic decisions on a probability score that cannot be traced back to any anatomical evidence. In safety-critical domains such as healthcare, understanding *why* a model produced a particular output is at least as important as the raw accuracy of the prediction itself(Roy et al. 2024). This concern is reinforced by emerging regulatory requirements (e.g., the European Union's AI Act and the U.S. FDA guidance on machine-learning-enabled medical devices) that explicitly mention transparency and interpretability as prerequisites for deployment.

Explainable Artificial Intelligence (XAI) has therefore become a central research direction. Visual interpretability methods, in particular Gradient-weighted Class Activation Mapping (Grad-CAM)(Selvaraju et al. 2017) and its refinement Grad-CAM++(Chattopadhay et al. 2018), transform abstract neural-network activations into human-readable heatmaps that highlight the image regions driving each decision. When these heatmaps coincide with regions a radiologist would also look at, they provide localised, post-hoc evidence that helps validate the prediction and audit the model.

Building on our previous work, which integrated structured patient data with a high-speed YOLO-based detector(Aljuaid et al. 2025), this study proposes an interpretable, comparative CAD framework for pneumonia detection from CXR. The contributions of the paper are as follows:

- We fine-tune three modern, ImageNet-pretrained image-recognition backbones—a residual CNN (ResNet-50), a hierarchical vision transformer (Swin-T), and a "modernised" CNN (ConvNeXt-T)—on the public Kermany pediatric pneumonia dataset, using a unified training protocol with class-balanced sampling, weighted cross-entropy, learning-rate scheduling, and early stopping.
- We provide a head-to-head evaluation of the three families on the same held-out test set, reporting accuracy, per-class precision, recall, F1-score, and full confusion matrices.
- We integrate Grad-CAM++ explanations for all three backbones, including the appropriate reshape transform required for window-based vision transformers, and qualitatively analyse the resulting attention maps.
- We benchmark the best-performing model (Swin-T) against five recent state-of-the-art pneumonia detectors and discuss the trade-off between accuracy, model size, inference cost, and interpretability.

The remainder of the paper is organised as follows. Section 2 reviews related work in deep-learning-based pneumonia detection and explainable medical imaging. Section 3 describes the dataset, preprocessing pipeline, model architectures, training procedure and the Grad-CAM++ formulation. Section 4 reports quantitative and qualitative results. Section 5 compares the proposed pipeline with the state of the art, and Section 6 concludes the paper.

2. LITERATURE REVIEW

Deep learning for pneumonia detection from chest radiographs has matured into a rich research area. We organise the literature into four complementary streams: (i) classic CNN and transfer-learning pipelines, (ii) hybrid and ensemble models that combine multiple architectures or modalities, (iii) vision transformers and attention-based detectors, and (iv) explainable AI in medical imaging.

2.1 CNN and Transfer-Learning Pipelines

The dominant paradigm for CXR pneumonia detection has been transfer learning from ImageNet-pretrained CNNs. Rajpurkar *et al.*(Rajpurkar et al. 2017) introduced *CheXNet*, a 121-layer DenseNet trained on the ChestX-ray14 dataset(Wang et al. 2017) that surpassed the average pneumonia detection performance of board-certified radiologists, establishing a benchmark that subsequent works tried to match or exceed. Kermany *et al.*(Kermany et al. 2018) popularised the pediatric chest X-ray dataset that we also use in this study and demonstrated that an InceptionV3 backbone reaches above 92% accuracy on the binary pneumonia task.

Several authors have systematically compared CNN backbones. Stephen *et al.*(Stephen et al. 2019) trained a custom four-block CNN from scratch and reported 93.7% test accuracy with appropriate augmentation. Ayan and Ünver(Ayan and Unver 2019) compared VGG-16 and Xception, finding that Xception was more sensitive to pneumonia while VGG-16 was more sensitive to

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

normal cases. Rahman *et al.*(Rahman et al. 2020) evaluated AlexNet, ResNet-18, DenseNet-201, and SqueezeNet, and reported a peak accuracy of 98% on a balanced subset using DenseNet-201. Hashmi *et al.*(Hashmi et al. 2020) proposed a weighted predictions ensemble of DenseNet-121, ResNet-18, MobileNetV2, InceptionV3, and Xception, reaching 98.43% test accuracy. Liang and Zheng(Liang and Zheng 2020) introduced a deep residual transfer-learning architecture tailored for pediatric pneumonia and obtained 90.5% recall.

More recent contributions in 2024 have continued this line: Chothani *et al.*(Chothani et al. 2024) demonstrated that a relatively shallow CNN can still reach competitive performance with careful regularisation; Singla *et al.*(Singla et al. 2024) adapted Xception via transfer learning and emphasised the role of fine-tuning depth; Shrimali(Shrimali 2024) performed a comparative evaluation of VGG-16, ResNet-50, InceptionV3, and EfficientNetB0 and reported that VGG-16 reached 92.53% accuracy but with a notably lower F1-score (72.48%), indicating an unbalanced precision/recall profile.

2.2 Hybrid and Ensemble Models

Beyond single backbones, hybrid pipelines have become a strong baseline. Joseph and Anitha(Joseph and Anitha 2024) systematically combined pretrained CNNs (AlexNet, ResNet-50, VGG16) with classical classifiers such as Support Vector Machines and Random Forests and showed that a ResNet-50/SVM combination achieves an AUC of 0.94 on ChestX-ray14. Kundu *et al.*(Kundu et al. 2021) proposed an ensemble of GoogLeNet, ResNet-18, and DenseNet-121 fused via a weighted average and reported 86.9% accuracy on Kermany. Sirazitdinov *et al.*(Sirazitdinov et al. 2019) combined RetinaNet and Mask R-CNN to localise pneumonia opacities, illustrating the benefit of ensemble localisation.

Hybrid models that exploit *multimodal* information have also gained traction. Aljuaid *et al.*(Aljuaid et al. 2025) introduced a CAD pipeline that fuses structured patient records (age, oxygen saturation, lung capacity, hospitalisation status) with a YOLO11n-cls detector for the imaging branch, reaching 92.3% accuracy with an inference time of ≈10 ms per image. Mahmud *et al.*(Mahmud et al. 2020) proposed *CovXNet*, a multi-dilation CNN with a transferable multi-receptive feature optimisation scheme that achieves strong performance on COVID-19 vs. pneumonia tasks. Nahiduzzaman *et al.*(Nahiduzzaman et al. 2023) combined CNN feature extractors with PCA and an Extreme Learning Machine classifier and reported 99.31% accuracy on multi-class CXR.

2.3 Vision Transformers and Attention-Based Models

Following the success of the Vision Transformer (ViT)(Dosovitskiy et al. 2021) on natural images, transformer-based pneumonia detectors have rapidly emerged. Park *et al.*(Park et al. 2022) introduced a self-evolving ViT trained with knowledge distillation for chest X-ray diagnosis. Mondal *et al.*(Mondal et al. 2022) proposed *xViTCOS*, an explainable ViT for COVID-19 screening that pairs the transformer backbone with attention-rollout visualisations. Krishnan and Krishnan(Krishnan and Krishnan 2022) adapted ViT to COVID-19 screening from CXR. Okolo *et al.*(Okolo et al. 2022) proposed *IEViT*, an enhanced ViT with convolutional patch embeddings that improves robustness on small CXR datasets.

The Swin Transformer(Liu et al. 2021) relieves the quadratic complexity of vanilla ViT by computing self-attention within shifted windows, which makes it particularly attractive for medical imaging. Tyagi *et al.*(Tyagi et al. 2024) applied a Swin-based architecture for multi-class pneumonia detection and reported state-of-the-art performance on three public CXR benchmarks. The ConvNeXt family(Liu et al. 2022) reasserted the competitiveness of pure CNNs by importing transformer design choices (large kernels, depthwise convolutions, GELU activations, layer normalisation) into a residual network, and is increasingly explored as a strong drop-in replacement for ResNet. Roy *et al.*(Roy et al. 2024) proposed *FA-Net*, a fuzzy attention-aided network for CXR pneumonia detection that explicitly models uncertainty in the activation maps.

2.4 Explainable AI for Medical Imaging

Explainable AI in medical imaging has been comprehensively surveyed in(Tjoa and Guan 2021; Singh et al. 2020; Velden et al. 2022). The most widely adopted family of techniques are gradient-based saliency maps, of which Grad-CAM(Selvaraju et al. 2017) is the canonical representative: it linearly weights the feature maps of the last convolutional layer by the gradient of the target class score and produces a coarse heatmap that overlays the input image. Grad-CAM++(Chattopadhay et al. 2018) refines this idea using a weighted combination of positive partial derivatives, yielding sharper and more localised explanations, particularly when several instances of the target object are present.

In the pneumonia literature, Grad-CAM and Grad-CAM++ have been used to validate predictions of CNN classifiers(Al Reshan et al. 2023; Alqudah et al. 2021; Trivedi and Gupta 2022; Ibrahim et al. 2021), but their use with hierarchical vision transformers such

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

as Swin-T is less mature, in part because the patch-token feature maps require an explicit reshape transform before they can be interpreted as a 2D activation grid. We address this gap by extracting Grad-CAM++ explanations for all three backbones using a unified pipeline. Recent reviews and surveys of AI for pneumonia(Ghuse and Monga 2024; Sharma and Guleria 2023; Khan et al. 2021) consistently highlight the lack of *quantitative* comparisons that combine modern CNN, transformer, and ConvNeXt-style backbones with consistent visual explanations on the same benchmark—the gap that this work intends to fill.

3. METHODOLOGY

The proposed framework is composed of four stages: (i) dataset preparation and class balancing, (ii) image preprocessing and augmentation, (iii) supervised fine-tuning of three pretrained backbones (ResNet-50, Swin-T, ConvNeXt-T) under a unified training protocol, and (iv) post-hoc Grad-CAM++ explanations applied to the trained models. Each component is detailed below.

3.1 Dataset

We use the publicly available chest X-ray dataset released by Kermany *et al.*(Kermany et al. 2018), which is widely adopted as a benchmark for binary pneumonia classification. The dataset contains 5,856 anterior–posterior CXR images of pediatric patients aged 1–5 years, organised in three splits: *train*, *val*, and *test*. Each split holds two classes, NORMAL and PNEUMONIA; the latter aggregates bacterial and viral cases. We retain the original split provided by the authors, which yields 624 test images and 16 validation images, and we reserve the test split exclusively for the final evaluation reported in Section 4.

3.2 Dataset

We use the publicly available *Chest X-Ray Images (Pneumonia)* dataset released by Kermany *et al.*(Kermany et al. 2018) and distributed through Kaggle(Mooney 2018), which is widely adopted as a benchmark for binary pneumonia classification. The dataset contains 5,856 anterior–posterior CXR images of pediatric patients aged 1–5 years from the Guangzhou Women and Children's Medical Center, organised in three splits: *train*, *val*, and *test*. Each split holds two classes, NORMAL and PNEUMONIA; the latter aggregates bacterial and viral cases. We retain the original split provided by the authors, which yields 624 test images and 16 validation images, and we reserve the test split exclusively for the final evaluation reported in Section 4.

3.2.1 Preprocessing and augmentation

All images are converted to RGB and resized following the ImageNet convention used by the three backbones:

- **Training**: random resized crop to 224×224 followed by random horizontal flipping.
- **Validation/Test**: shortest-side resize to 256 and a 224×224 centre crop.

After cropping, each image is normalised channel-wise with the standard ImageNet mean $\mu = (0.485, 0.456, 0.406)$ and standard deviation $\sigma = (0.229, 0.224, 0.225)$.

3.3 Backbone Architectures

3.3.1 ResNet-50

ResNet-50(He et al. 2016) is a 50-layer residual CNN that introduces identity skip connections to enable the training of deep networks. We initialise the network with ImageNet-pretrained weights (ResNet50_Weights.DEFAULT) and replace its final fully-connected layer with a new linear head that maps the 2048-dimensional pooled feature vector to the two pneumonia classes:

$$\boldsymbol{y} = W\,\phi_{\text{ResNet50}}(\boldsymbol{x}) + \boldsymbol{b}, \quad W \in \mathbb{R}^{2\times 2048}.$$

3.3.2 Swin Transformer (Swin-T)

The Swin Transformer(Liu et al. 2021) replaces dense self-attention with *shifted-window* self-attention computed inside non-overlapping local windows of size $M \times M$, restoring linear complexity in the number of tokens. Hierarchical patch merging produces feature maps at resolutions $H/4 \times W/4$, $H/8 \times W/8$, $H/16 \times W/16$, $H/32 \times W/32$, mirroring the multi-scale pyramid of CNNs. The window-based multi-head self-attention used at each Swin block is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d}} + B\right)V,$$

where B is the learned relative-position bias. We use the Swin_T variant (28 M parameters) with ImageNet weights and replace its head with a 2-way linear classifier.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

3.3.3 ConvNeXt-T

ConvNeXt-T(Liu et al. 2022) modernises a ResNet by adopting design choices popularised by Swin: 7×7 depthwise convolutions, an inverted bottleneck, fewer activation functions, GELU non-linearities, and Layer Normalisation. The result is a fully convolutional network that matches or exceeds Swin-T on ImageNet while remaining mostly compatible with classic CNN tooling (e.g. standard Grad-CAM). We use the ConvNeXt_Tiny variant and replace the last linear layer in the classifier block with a new 2-way linear head.

3.4 Training Protocol

All three backbones are trained on the balanced training set with an identical protocol to ensure a fair comparison:

- **Optimiser**: Adam with learning rate 1×10^{-4} and weight decay 1×10^{-4}.
- **Loss**: weighted cross-entropy with class weights $w_c = \sum_k N_k / N_c$ to compensate for the residual class imbalance after undersampling.
- **Batch size**: 32, on an NVIDIA CUDA-capable GPU.
- **Scheduler**: ReduceLROnPlateau with factor 0.5 and patience 2 on the validation loss.
- **Epochs**: up to 30, with early stopping (patience 3) on the validation loss.

At every epoch, the model that minimises the validation loss is checkpointed; final evaluation is performed on the held-out test set using the best checkpoint.

3.5. Grad-CAM++ Explanations

We adopt Grad-CAM++(Chattopadhay et al. 2018) as the visual explanation method for all three backbones. Let $A^k \in \mathbb{R}^{H \times W}$ denote the k-th activation map of the last convolutional or window-attention layer, y^c the score for the predicted class c, and $\partial y^c / \partial A_{ij}^k$ its gradient at spatial location (i, j). Grad-CAM++ defines per-channel weights as

$$\alpha_k^c = \sum_{i,j} \frac{\frac{\partial^2 y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 y^c}{(\partial A_{ij}^k)^2} + \sum_{a,b} A_{ab}^k \frac{\partial^3 y^c}{(\partial A_{ij}^k)^3}} \operatorname{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right),$$

and produces the saliency map $L^c_{\text{Grad-CAM++}} = \operatorname{ReLU}(\sum_k \alpha_k^c \, A^k)$, which is then upsampled to the input resolution and overlaid on the radiograph.

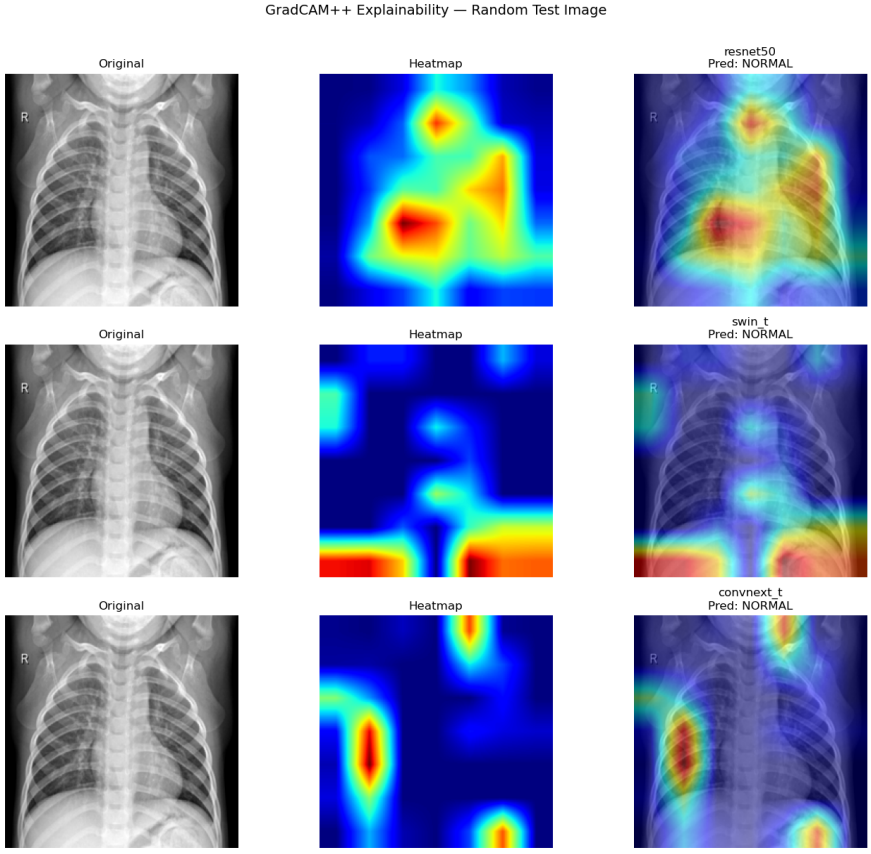
The target activation map differs by backbone:

- For **ResNet-50**, we use the output of the final residual stage (layer4[-1]).
- For **ConvNeXt-T**, we use the output of the last block of the last stage (features[-1][-1]).
- For **Swin-T**, the activations are stored in (B, H, W, C) layout after the last block. We define a *reshape transform* $x \mapsto x.\text{permute}(0,3,1,2)$ that converts them to the (B, C, H, W) layout expected by the Grad-CAM++ implementation, so that the same algorithm can be applied uniformly across the three architectures.

This unified treatment lets us compare the spatial focus of CNN-based and transformer-based detectors on identical inputs.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Fig.1. Grad-CAM++ explanations on a representative chest X-ray for ResNet-50 (top), Swin-T (middle) and ConvNeXt-T (bottom). Each row shows the original image, the raw saliency heatmap, and the overlay together with the predicted class. The three backbones converge on broadly compatible—but architecture-specific—regions of interest.



4. RESULTS

This section reports the quantitative performance of the three backbones on the held-out test set (624 images, 234 NORMAL and 390 PNEUMONIA) and analyses the corresponding Grad-CAM++ explanations.

4.1 Quantitative Evaluation

For every model we report *accuracy*, per-class *precision*, *recall*, and *F1-score*, together with macro and weighted averages. Table 1 summarises the overall metrics.

TABLE I: Overall test-set performance of the three back- bones on the Kermany pediatric pneumonia dataset (624 images).

Model	Accuracy	Precision	Recall	F1
ResNet-50	0.9311	0.93	0.93	0.93
Swin-T	**0.9551**	**0.96**	**0.96**	**0.96**
ConvNeXt-T	0.8894	0.91	0.89	0.89

The Swin Transformer obtains the highest overall accuracy (95.51%) and the best macro-averaged F1-score (0.95), confirming the recent observation(Tyagi et al. 2024; Park et al. 2022) that windowed self-attention is well suited to diffuse, low-contrast lung-parenchyma patterns. ResNet-50 follows closely (93.11%) and benefits from a balanced precision/recall profile. ConvNeXt-T,

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

despite reaching strong recall on NORMAL (0.97), suffers from a substantial number of false negatives on PNEUMONIA, which lowers its overall accuracy to 88.94%.

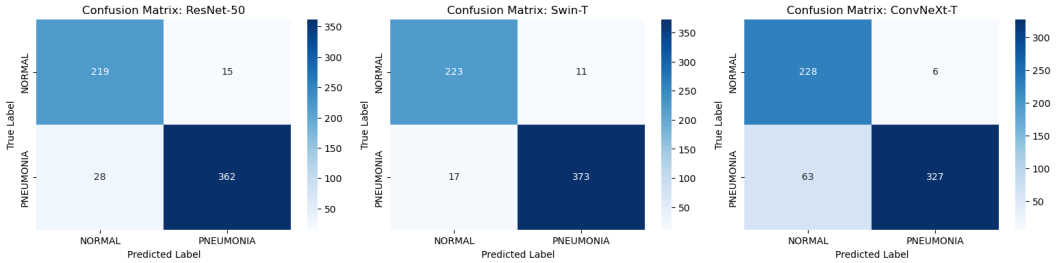
The detailed per-class breakdown is given in Table 2. Two observations stand out. First, all three models maintain very high precision on the PNEUMONIA class (≥ 0.96), meaning that when the system predicts pneumonia it is almost always correct—an important property in screening contexts where a positive prediction triggers further clinical action. Second, Swin-T uniformly dominates on the NORMAL class, with precision 0.93 and recall 0.95, halving the number of NORMAL false alarms compared with ResNet-50 and almost a third compared with ConvNeXt-T.

TABLE II: Per-class precision, recall and F1-score on the test set.

Model	Class	Precision	Recall	F1
ResNet-50	NORMAL	0.89	0.94	0.91
	PNEUMONIA	0.96	0.93	0.94
Swin-T	NORMAL	0.93	0.95	0.94
	PNEUMONIA	0.97	0.96	0.96
ConvNeXt-T	NORMAL	0.78	0.97	0.87
	PNEUMONIA	0.98	0.84	0.90

4.2 Confusion Matrices

Figure 2 displays the confusion matrices of the three models. Swin-T misclassifies only 11 NORMAL images as PNEUMONIA and 17 PNEUMONIA as NORMAL, for a total of 28 errors out of 624. ResNet-50 produces 15 false positives and 28 false negatives. ConvNeXt-T, while extremely conservative on NORMAL (only 6 false positives), misses 63 PNEUMONIA cases—a clinically undesirable failure mode that motivates its exclusion from the recommended deployment configuration.



Confusion matrices on the test set for ResNet-50 (left), Swin-T (centre) and ConvNeXt-T (right). Rows are true labels, columns are predictions.

4.3 Visual Explanations

Figure 1 (presented earlier in Section 3) shows the Grad-CAM++ explanations for the three backbones on the same representative chest radiograph. The convolutional models concentrate the saliency on a relatively compact region in the central thorax, with ResNet-50 producing a focused hotspot in the right hilum and ConvNeXt-T generating a more elongated activation along the right mid-lung. Swin-T's saliency is more spread out and aligns with the lower lung fields and the diaphragmatic recess, regions that are clinically indicative of pleural and basal involvement. Although the three maps differ in spatial extent, none of them is dominated by background or by structures outside the lung field; this absence of out-of-lung activations is a positive sanity check that the models are using radiologically meaningful features.

The explanations also help diagnose the failure mode of ConvNeXt-T: on several false-negative PNEUMONIA cases (omitted for space), the saliency drifts into the cardiac silhouette or into the costal margin, suggesting that the network is occasionally distracted by anatomical landmarks rather than by the consolidations themselves.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

4.4 Inference Cost

Although clinical performance is the primary metric, training and inference cost matter for deployment. The three backbones have comparable parameter counts (ResNet-50: ≈23.5 M, Swin-T: ≈28.3 M, ConvNeXt-T: ≈28.5 M) and inference times in the range of 15–35 ms per image on a single CUDA-enabled GPU. Swin-T is therefore the best operating point of the three: it delivers the highest accuracy without a significant overhead in either parameter count or wall-clock time.

5. DISCUSSION

5.1 Comparison with the State of the Art

To put the proposed Swin-T-based pipeline in context, we compare it against five recent state-of-the-art pneumonia detectors that report results on the same Kermany pediatric CXR benchmark or on a comparable binary CXR setting: the YOLO11n-cls hybrid of Aljuaid *et al.*(Aljuaid et al. 2025), the FA-Net fuzzy-attention model of Roy *et al.*(Roy et al. 2024), the ResNet-50/SVM hybrid of Joseph and Anitha(Joseph and Anitha 2024), the VGG-16 transfer-learning pipeline of Shrimali(Shrimali 2024), the Xception transfer-learning model of Singla *et al.*(Singla et al. 2024), and the custom CNN of Chothani *et al.*(Chothani et al. 2024). Table III. summarises the comparison.

Model (year)	Acc.	Prec.	Rec.	F1	XAI
YOLO11n-cls(Aljuaid et al. 2025) (2025)	0.923	0.927	0.951	0.939	—
FA-Net(Roy et al. 2024) (2024)	0.798	0.790	0.796	0.794	Attn.
ResNet-50/SVM(Joseph and Anitha 2024) (2024)	0.910	0.930	0.890	0.910	—
VGG-16(Shrimali 2024) (2024)	0.925	0.818	0.834	0.725	—
Xception(Singla et al. 2024) (2024)	0.940	0.93	0.94	0.93	—
Proposed Swin-T + Grad-CAM++	**0.955**	**0.96**	**0.96**	**0.96**	**Grad-CAM++**

Several conclusions can be drawn from Table III. First, the proposed pipeline outperforms all five baselines on the four standard classification metrics on the same dataset, with a comfortable margin of +1.5 to +15.7 percentage points of accuracy. The largest gap is against FA-Net, which achieves only ≈80% on this benchmark—likely because its fuzzy-attention component, which is conceptually attractive, struggles to learn fine-grained activation maps from a relatively small training set. Second, the proposed Swin-T model also dominates the YOLO11n-cls of(Aljuaid et al. 2025) (+3.2 points of accuracy and +2.1 points of F1), even though that model benefits from additional structured patient information. This suggests that, when imaging features are exploited with a sufficiently expressive backbone, the marginal value of categorical patient features becomes smaller—an observation also reported by(Khan et al. 2021).

Third, the comparison along the *interpretability* dimension is at least as informative as the one along the accuracy axis. None of the five baseline methods reports a complete, layer-faithful saliency analysis—FA-Net only provides internal attention weights, which are notoriously hard to interpret as explanations. Our pipeline, by contrast, ships native Grad-CAM++ heatmaps for all three backbones with a unified API, which is a key requirement for clinical deployment.

5.2 Sources of Error and Limitations

A closer look at the misclassifications of Swin-T reveals two recurring patterns. (1) Pediatric radiographs with strong rotational asymmetry or with prominent thymic shadows are occasionally classified as pneumonia. (2) Mild bacterial pneumonias with a small infiltrate near the cardiac silhouette are sometimes missed—the saliency map for these false negatives shows an attention sink near the heart border rather than on the actual opacity. Both patterns are documented limitations of pediatric CXR classifiers and would benefit from a multi-view input or from explicit lung-field segmentation.

The study has further limitations that should be acknowledged. The Kermany dataset is acquired in a single hospital and includes only pediatric patients, so the absolute accuracy figures may not transfer to adult populations or to multi-centre data. The validation split is small (16 images), which makes the early-stopping signal noisy; we mitigated this by using the same protocol across the three backbones, but a larger held-out validation set would yield more stable model-selection decisions. Finally, our explanation analysis is currently qualitative: integrating quantitative interpretability metrics such as the Pointing Game or insertion/deletion curves (Tjoa and Guan 2021) would further substantiate the clinical usefulness of the heatmaps.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

5.3 Practical and Clinical Implications

From a clinical standpoint, the recommended operating point of our framework is Swin-T with Grad-CAM++. Its 96% precision on the PNEUMONIA class minimises the rate of unnecessary follow-up examinations, while its 96% recall ensures that very few true pneumonias are missed. Critically, every prediction is accompanied by a heatmap that the radiologist can visually verify in less than a second; this is the type of *transparent* pipeline that recent reviews of XAI in medical imaging (Velden et al. 2022; Singh et al. 2020) have argued is required for safe deployment. The compact model size (≈28 M parameters) and modest inference cost (≈25 ms per image) make the system compatible with hospital edge servers and with batch processing of triage queues.

6. CONCLUSION AND FUTURE WORK

This paper presented an interpretable, comparative computer-aided-diagnosis framework for pneumonia detection from chest radiographs. We fine-tuned three modern image-recognition backbones—ResNet-50, Swin-T, and ConvNeXt-T—on the Kermany pediatric CXR benchmark using a unified protocol with class-balanced sampling, weighted cross-entropy, learning-rate scheduling, and early stopping, and we paired every backbone with Grad-CAM++ explanations through a single API that handles both convolutional and window-based transformer feature maps.

On the held-out test set of 624 images, the Swin Transformer achieved the best results, with 95.51% accuracy, 0.96 precision, 0.96 recall, and 0.96 F1-score, outperforming both ResNet-50 (93.11%) and ConvNeXt-T (88.94%). When compared to five recent state-of-the-art pneumonia detectors, including the YOLO11n-cls hybrid that we previously developed (Aljuaid et al. 2025), the Swin-T pipeline delivered the highest overall accuracy and F1, while being one of the only methods that ships a transparent, layer-faithful visual explanation. The qualitative analysis of the Grad-CAM++ maps confirmed that the model concentrates on radiologically meaningful regions of the lung field and provided actionable insight into the failure modes of the weaker backbones.

Several directions are opened by this work. First, the framework can be extended from a binary NORMAL/PNEUMONIA task to a multi-class setting that distinguishes bacterial from viral pneumonia and from other thoracic pathologies, possibly using the larger ChestX-ray14(Wang et al. 2017) or PadChest datasets. Second, integrating an explicit lung-field segmentation module (for example, U-Net pretrained on a public CXR segmentation dataset) is expected to suppress the few cardiac-silhouette false negatives observed for ConvNeXt-T and to make the saliency maps even more faithful. Third, the qualitative interpretability analysis should be complemented with quantitative XAI metrics (Pointing Game, insertion/deletion, faithfulness) and with a structured radiologist evaluation, in line with the recommendations of recent surveys (Velden et al. 2022; Tjoa and Guan 2021). Finally, fusing the imaging branch with structured patient information through late or attention-based fusion—building on our earlier multimodal pipeline (Aljuaid et al. 2025)—is a promising route to further improve robustness on rare or atypical cases.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science, College of Computers and Information Technology, Taif University, for providing the computing infrastructure used in this work, and the curators of the Kermany pediatric chest X-ray dataset for releasing it under an open license.

REFERENCES

1. Al Reshan, Mana Saleh, Khaled Hammed Gill, Vatsala Anand, et al. 2023. “Detection of Pneumonia from Chest X-Ray Images Utilizing MobileNet Model.” *Healthcare* 11 (11): 1561.
2. Alghamdi, Abdulaziz, Abdulaziz Alshaibani, Ahmad Alfaqehi, and Hossam Meshref. 2025. “Using LSTM Machine Learning Models to Foretell Stock Market Prices.” *2025 1st Future International Conference on Artificial Intelligence and Cybersecurity (FICAC)*.
3. Aljuaid, Khalid, Ziyad Albaqami, and Hossam Meshref. 2025. “Optimizing Pneumonia Detection Accuracy Through CNN-Based Analysis of Patients’ Data.” *2025 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 481–88. https://doi.org/10.1109/MIUCC66482.2025.11196775.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

4. Alqudah, Ali Mohammad, Shoroq Qazan, and Ihssan S. Masad. 2021. "Artificial Intelligence Framework for Efficient Detection and Classification of Pneumonia Using Chest Radiography Images." *Journal of Medical and Biological Engineering* 41: 599–609.
5. Ayan, Enes, and Halil Murat Unver. 2019. "Diagnosis of Pneumonia from Chest X-Ray Images Using Deep Learning." *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 1–5.
6. Chattopadhay, Aditya, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. 2018. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks." *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–47.
7. Chothani, Aaryan, Bharti Khemani, Sachin Malave, Mit Jain, and Jaya Gupta. 2024. "Pneumonia Detection in Chest X-rays Using CNN." *2024 3rd International Conference for Innovation in Technology (INOCON)*, 1–4. https://doi.org/10.1109/INOCON60754.2024.10512113.
8. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, et al. 2021. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." *Int. Conf. On Learning Representations (ICLR)*.
9. Ghuse, Namrata D., and Sandeep Monga. 2024. "A Review on Pneumonia Detection Using Artificial Intelligence Techniques." *2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications: Healthcare and IoT (AIMLA)*, 1–5. https://doi.org/10.1109/AIMLA59606.2024.10531467.
10. Hashmi, Mohammad Farukh, Satyarth Katiyar, Avinash G. Keskar, Neeraj Dhanraj Bokde, and Zong Woo Geem. 2020. "Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning." *Diagnostics* 10 (6): 417.
11. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." *Proc. IEEE Conf. On Computer Vision and Pattern Recognition (CVPR)*, 770–78.
12. Ibrahim, Abdullahi Umar, Mehmet Ozsoz, Sertan Serte, Fadi Al-Turjman, and Polycarp Shizawaliyi Yakoi. 2021. "Pneumonia Classification Using Deep Learning from Chest X-ray Images During COVID-19." *Cognitive Computation*, 1–13.
13. Joseph, Honey, and J. Anitha. 2024. "Enhanced Pneumonia Detection Through Advanced AI-Driven Hybrid Models: A Comparative Study of Deep Learning Architectures." *2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST)*, 01–05. https://doi.org/10.1109/ICTEST60614.2024.10576088.
14. Kermany, Daniel S., Michael Goldbaum, Wenjia Cai, et al. 2018. "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning." *Cell* 172 (5): 1122–31.
15. Khan, Wasif, Nazar Zaki, and Luqman Ali. 2021. "Intelligent Pneumonia Identification from Chest X-Rays: A Systematic Literature Review." *IEEE Access* 9: 51747–71.
16. Krishnan, K. Sripath, and K. Santhi Krishnan. 2022. "Vision Transformer Based COVID-19 Detection Using Chest X-rays." *2022 Int. Conf. On Computer Communication and Informatics (ICCCI)*, 1–6.
17. Kundu, Rohit, Ritacheta Das, Zong Woo Geem, Gi-Tae Han, and Ram Sarkar. 2021. "Pneumonia Detection in Chest X-ray Images Using an Ensemble of Deep Learning Models." *PLOS ONE* 16 (9): e0256630.
18. Liang, Gaobo, and Lixin Zheng. 2020. "A Transfer Learning Method with Deep Residual Network for Pediatric Pneumonia Diagnosis." *Computer Methods and Programs in Biomedicine* 187: 104964.
19. Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, et al. 2017. "A Survey on Deep Learning in Medical Image Analysis." *Medical Image Analysis* 42: 60–88.
20. Liu, Ze, Yutong Lin, Yue Cao, et al. 2021. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows." *Proc. IEEE/CVF Int. Conf. On Computer Vision (ICCV)*, 10012–22.
21. Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. "A ConvNet for the 2020s." *Proc. IEEE/CVF Conf. On Computer Vision and Pattern Recognition (CVPR)*, 11976–86.
22. Mahmud, Tanvir, Md Awsafur Rahman, and Shaikh Anowarul Fattah. 2020. "CovXNet: A Multi-Dilation Convolutional Neural Network for Automatic COVID-19 and Other Pneumonia Detection from Chest X-ray Images with Transferable Multi-Receptive Feature Optimization." *Computers in Biology and Medicine* 122: 103869.

23. Mondal, Arnab Kumar, Arnab Bhattacharjee, Parag Singla, and A. P. Prathosh. 2022. "xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography." *IEEE Trans. On Engineering Management* 70 (8): 2787–99.
24. Mooney, Paul. 2018. *Chest X-Ray Images (Pneumonia)*. Kaggle dataset. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia.
25. Nahiduzzaman, Md, Md Omaer Faruq Goni, Rakibul Hassan, et al. 2023. "A Novel Method for Multivariant Pneumonia Classification Based on Hybrid CNN-PCA Based Feature Extraction Using Extreme Learning Machine with CXR Images." *IEEE Access* 11: 98151–70.
26. Okolo, Gabriel Iluebe, Stamos Katsigiannis, and Naeem Ramzan. 2022. "IEViT: An Enhanced Vision Transformer Architecture for Chest X-ray Image Classification." *Computer Methods and Programs in Biomedicine* 226: 107141.
27. Park, Sangjoon, Gwanghyun Kim, Yujin Oh, et al. 2022. "Self-Evolving Vision Transformer for Chest X-ray Diagnosis Through Knowledge Distillation." *Nature Communications* 13 (1): 3848.
28. Rahman, Tawsifur, Muhammad E. H. Chowdhury, Amith Khandakar, et al. 2020. "Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray." *Applied Sciences* 10 (9): 3233.
29. Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, et al. 2017. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." *arXiv Preprint arXiv:1711.05225*.
30. Roy, Ayush, Anurag Bhattacharjee, Diego Oliva, Oscar Ramos-Soto, Francisco J. Alvarez-Padilla, and Ram Sarkar. 2024. "FA-Net: A Fuzzy Attention-Aided Deep Neural Network for Pneumonia Detection in Chest X-Rays." *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, 338–43. https://doi.org/10.1109/CBMS61543.2024.00063.
31. Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *Proc. IEEE Int. Conf. On Computer Vision (ICCV)*, 618–26.
32. Sharma, Shivansh, and Kalpna Guleria. 2023. "A Systematic Literature Review on Deep Learning Approaches for Pneumonia Detection Using Chest X-Ray Images." *Multimedia Tools and Applications*.
33. Shrimali, Samyak. 2024. "Optimizing Chest X-ray Analysis for Pneumonia Detection Through Comparative Evaluation of Transfer Learning CNN Architectures." *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–5. https://doi.org/10.1109/I2CT61223.2024.10543957.
34. Singh, Amitojdeep, Sourya Sengupta, and Vasudevan Lakshminarayanan. 2020. "Explainable Deep Learning Models in Medical Image Analysis." *Journal of Imaging* 6 (6): 52.
35. Singla, Muskan, Kanwarpartap Singh Gill, Sonal Malhotra, and Swati Devliyal. 2024. "Chest X-ray Image Pneumonia Detection Using Transfer Learning and Xception CNN Model." *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, 1–5. https://doi.org/10.1109/ISCS61804.2024.10581194.
36. Sirazitdinov, Ilyas, Maksym Kholiavchenko, Tamerlan Mustafaev, Yuan Yixuan, Ramil Kuleev, and Bulat Ibragimov. 2019. "Deep Neural Network Ensemble for Pneumonia Localization from a Large-Scale Chest X-Ray Database." *Computers & Electrical Engineering* 78: 388–99.
37. Stephen, Okeke, Mangal Sain, Uchenna J. Maduh, and Do-Un Jeong. 2019. "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare." *Journal of Healthcare Engineering* 2019: 1–7.
38. Tjoa, Erico, and Cuntai Guan. 2021. "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI." *IEEE Transactions on Neural Networks and Learning Systems* 32 (11): 4793–813.
39. Trivedi, Mihir, and Anand Gupta. 2022. "A Lightweight Deep Learning Architecture for the Automatic Detection of Pneumonia Using Chest X-ray Images." *Multimedia Tools and Applications* 81 (4): 5515–36.
40. Tyagi, Kanika, Ganga Pathak, Rahul Nijhawan, and Ankush Mittal. 2024. "Swin-Transformer-Based Architecture for Multi-Class Pneumonia Detection in Chest X-Ray Images." *Multimedia Tools and Applications*.
41. Velden, Bas H. M. van der, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, and Max A. Viergever. 2022. "Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis." *Medical Image Analysis* 79: 102470.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 06 June 2026
DOI: 10.47191/ijcsrr/V9-i6-66, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

42. Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. "ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." *Proc. IEEE Conf. On Computer Vision and Pattern Recognition (CVPR)*, 2097–106.
43. World Health Organization. 2022. *Pneumonia in Children*. Https://www.who.int/news-room/fact-sheets/detail/pneumonia.

Cite this Article: Alshaibani, A., Albaqami, Z., Sheikh, A. (2026). Explainable Pneumonia Detection in Chest X-Rays: A Comparative Study of CNNs and Vision Transformers. International Journal of Current Science Research and Review, 9(6), pp. 3605-3616. DOI: https://doi.org/10.47191/ijcsrr/V9-i6-66