International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

NanoRay V2: Bridging the Gap Between Transformers and Edge AI via Cross-Architecture Distillation

Kabir Thayani
Independent Researcher

Abstract. While Vision Transformers (ViTs) achieve state-of-the-art performance in medical image analysis, their massive computational cost makes them unsuitable for edge deployment in resource-constrained environments. This study introduces **NanoRay V2**, a lightweight 2.5M-parameter MobileNetV3 distilled from an 86M-parameter Vision Transformer (ViT-Base). By leveraging a soft-target distillation objective ($\alpha = 0.25$, $T = 4.0$), we transfer global attention behavior from the Transformer into the compact CNN. The distilled model achieves **84.19%** accuracy, surpassing both its teacher (**83.96%**) and a baseline CNN trained from scratch (**83.29%**) on the RSNA Pneumonia dataset. Grad-CAM analysis confirms that NanoRay V2 inherits structure-aware global attention while maintaining inference speeds suitable for CPU-native mobile hardware. This work is intended strictly for research purposes and is not a clinical diagnostic system.

1 INTRODUCTION

Pneumonia remains a leading cause of mortality worldwide, particularly in regions where access to expert radiologists is limited. Deep learning models have shown promise in automating diagnosis from chest X-rays (CXRs), yet a fundamental tradeoff persists between accuracy and deployability.

Convolutional Neural Networks (CNNs) such as ResNet and MobileNet are computationally efficient but exhibit strong local inductive bias, often focusing on texture rather than global lung structure. Vision Transformers (ViTs), by contrast, leverage self-attention to capture long-range dependencies and provide superior contextual reasoning, but their computational demands make them impractical for edge deployment.

This work proposes **NanoRay V2**, a cross-architecture knowledge distillation framework that transfers global attention from a Transformer teacher into a compact CNN student, achieving server-grade intelligence on edge-grade hardware.

2 METHODOLOGY

2.1 Dataset

We utilize the RSNA Pneumonia Detection Challenge dataset. All images are resized to 224 × 224 resolution to ensure compatibility with ViT patch embeddings. The dataset is split into 80% training, 10% validation, and 10% testing partitions with no patient overlap.

2.2 Architecture Design

A cross-architecture teacher–student paradigm is employed:

- **Teacher:** ViT-Base-16 (86M parameters), pretrained on ImageNet and fine-tuned on RSNA.
- **Student:** MobileNetV3-Small (2.5M parameters), optimized for CPU-based inference.

2.3 Distillation Objective

The student is trained using a dual-objective loss:

$$L = \alpha T^2 \cdot L_{KL}(p_s, p_t) + (1 - \alpha) \cdot L_{CE}(y, \hat{y}) \qquad (1)$$

where L_{KL} denotes Kullback–Leibler divergence between softened student and teacher logits, L_{CE} is the standard cross-entropy loss, $T = 4.0$ is the temperature parameter, and $\alpha = 0.25$ balances imitation and ground-truth supervision.

3 RESULTS

3.1 Quantitative Metrics

Table 1 summarizes model performance on the test set.

Table 1: Performance Comparison on Test Set

Model	Params	Accuracy	F1-Score
ViT Teacher	86M	83.96%	0.8307
MobileNet (Base)	2.5M	83.29%	0.8236
NanoRay V2 (Distilled)	**2.5M**	**84.19%**	**0.8268**

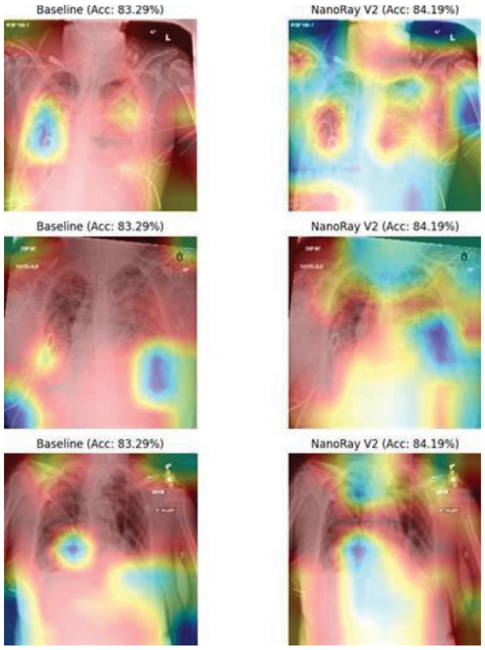

Figure 1: Grad-CAM comparison between the baseline MobileNet (left) and NanoRay V2 (right). The baseline model exhibits fragmented and noise-sensitive attention, often focusing on bone edges, while NanoRay V2 demonstrates smoother, structure-aware activation con- centrated on pulmonary opacities.

3.2 Qualitative Analysis

Grad-CAM visualizations reveal that the distilled model attends to anatomically coherent lung regions rather than spurious high- frequency structures such as ribs and edges. This confirms effective transfer of global contextual reasoning from the Trans- former teacher to the CNN student.

4 CONCLUSION

We present NanoRay V2, a cross-architecture distillation framework that compresses the global reasoning ability of Vision Transformers into a mobile-ready CNN. Despite a 97% reduc- tion in parameter count, the distilled model surpasses both its teacher and a baseline CNN, demonstrating that reliable medical image analysis can be achieved on edge hardware. Future work will explore quantization and hardware-aware optimization to further reduce latency and memory footprint.

REFERENCES

[1] A. Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR, 2021.

[2] G. Hinton, O. Vinyals, J. Dean, *Distilling the Knowledge in a Neural Network*, arXiv:1503.02531, 2015.

[3] A. Howard et al., *Searching for MobileNetV3*, ICCV, 2019.