A COMPARATIVE ANALYSIS OF VISION TRANSFORMERS FOR IMAGE CLASSIFICATION AND HEALTHCARE APPLICATIONS

[1]D.N.Kuldeep Shamgar, [2]K.Poojitha,
[1]Assistant Professor, [2] Assistant Professor
[1] Department of ECE, , [2] Department of ECE
[1]S K U College of Engg. & Technology, Anantapuramu, INDIA

Abstract: Vision Transformers (ViTs) have emerged as a transformative architecture in computer vision, challenging the long-standing dominance of Convolutional Neural Networks (CNNs). This paper presents a comprehensive comparative analysis of ViT architectures, examining their theoretical foundations, architectural variants, performance benchmarks, and application domains. We synthesize findings from multiple recent survey papers and evaluate a controversial case study—the retracted ECG-ViT paper—to illustrate critical issues in model validation and research integrity. Our analysis reveals that while ViTs achieve state-of-the-art performance on large-scale datasets, they face significant challenges including data efficiency, computational complexity, and inductive bias limitations. We provide quantitative comparisons across 30+ transformer variants and 15+ CNN baselines, analyse the discrepancy between reported results in retracted versus validated research, and propose methodological guidelines for robust ViT evaluation. This paper serves as both a technical reference for practitioners and a cautionary examination of quality standards in deep learning research.

Keywords: **Vision Transformer, Image Classification, Knowledge Distillation, ECG Classification, Comparative Analysis, Research Integrity**

1: Introduction

1.1 *Background and Motivation*

The Transformer architecture, introduced by Vaswani in 2017, fundamentally reshaped natural language processing through its self-attention mechanism. Unlike recurrent neural networks that process sequences sequentially, Transformers enable parallel processing and capture long-range dependencies through attention-based computations. This architectural breakthrough achieved state-of-the-art performance in machine translation, language modelling, and numerous NLP benchmarks. The natural question emerged: could this architecture be adapted for computer vision? Convolutional Neural Networks (CNNs) had dominated visual recognition since AlexNet's breakthrough in 2012, leveraging inductive biases of locality, translation equivariance, and hierarchical feature learning. However, CNNs possess inherent limitations: their receptive fields grow slowly through successive layers, making global context modelling computationally expensive, and their fixed kernel designs cannot dynamically adapt to input content. The Vision Transformer (ViT), proposed by Dosovitskiy et al. in 2020, demonstrated that a pure Transformer architecture—applied directly to sequences of image patches—could achieve competitive or superior performance to CNNs when pre-trained on sufficiently large datasets. This discovery catalysed an explosion of research activity, producing hundreds of ViT variants targeting improved data efficiency, reduced computational complexity, enhanced feature representation, and domain-specific adaptations.

Table 1.1: Timeline of Major Vision Transformer Developments

Year	Model	Key Contribution	Params (M)	Top-1 (%)
2020	ViT	First pure transformer for vision	86–307	76.5–77.9
2021	DeiT	Data-efficient training	6–86	72.2–81.8
2021	T2T-ViT	Token aggregation	6.8–64.1	76.5–82.3
2021	CvT	Convolutional token embedding	20–277	81.6–87.7
2021	Swin	Hierarchical shifted windows	29–197	83.5–86.4
2022	UniFormer	Unified conv-attention	–	86.3

1.2 Scope and Contributions

This paper provides a systematic comparative analysis of Vision Transformer architectures across multiple dimensions:

Theoretical Comparison: We analyze the mathematical foundations of self-attention versus convolution, examining the relationship between attention mechanisms and learnable receptive fields.

Architectural Taxonomy: We develop a comprehensive categorization of ViT variants based on their improvement strategies: data efficiency, CNN hybridization, lightweight design, depth scaling, and attention mechanism innovation.

Quantitative Benchmarking: We compile and normalize performance metrics (accuracy, parameters, FLOPs, throughput) across 30+ transformer models and 15+ CNN baselines on ImageNet, CIFAR, and downstream transfer tasks.

Critical Case Study: We conduct a detailed examination of the retracted ECG-ViT paper, analyzing the discrepancy between its reported results (99.7% accuracy) and the systematic manipulation indicators that led to retraction. This provides methodological lessons for evaluating biomedical AI research.

Application Domain Analysis: We compare ViT performance across medical imaging, remote sensing, video analysis, and multimodal tasks, identifying domain-specific challenges and successful adaptations.

2: **Theoretical Foundations**

2.1 *The Self-Attention Mechanism*

The self-attention mechanism forms the computational core of all Transformer architectures. Given an input sequence self-attention computes three matrices through learned linear projections:

The softmax function normalizes these similarities into attention probabilities, which weight the corresponding value vectors.

Computational Complexity: For image classification with \(n = H \times W / P^2\) patches, this quadratic scaling becomes prohibitive for high-resolution inputs—a primary motivation for hierarchical and sparse attention variants.

Relationship to Convolution: Cordonnier et al. demonstrated that multi-head self-attention with sufficient heads can express any convolutional layer, and conversely, convolutions with kernel size can be expressed as self-attention with relative position bias restricted to a neighborhood. This theoretical equivalence suggests that CNNs and Transformers occupy different points on a spectrum of architectural flexibility rather than fundamentally distinct paradigms.

2.2 *Multi-Head Attention*

Single-head attention limits the model's capacity to attend to multiple representation subspaces. Multi-head attention addresses this by computing \(h\) parallel attention functions:

Each head learns different projection matrices, enabling the model to focus on different types of relationships (e.g., local texture patterns, long-range dependencies, object-part relationships). The outputs are concatenated and linearly projected, typically maintaining the original dimension.

2.3 *Positional Encoding*

Self-attention is permutation-invariant—it processes sets rather than sequences. To inject spatial information, Transformers require positional encodings. The original Transformer used sinusoidal functions

ViT adopted learnable 1D positional embeddings added to patch embeddings. However, subsequent research identified limitations: fixed-length encodings require interpolation for different resolutions, and they fail to capture the 2D structure of images. This motivated conditional positional encodings (CPVT) generated through convolutions and relative position biases (LeViT, Swin).

2.4 *Vision Transformer Architecture*

The Vision Transformer represents a minimal adaptation of the original Transformer encoder for image classification. The architecture proceeds through four stages:

1. *Patch Embedding:* An image is divided into non-overlapping patches of size. Each patch is flattened to dimensions and linearly projected to dimension.

2. *Token Construction*: A learnable class token is prepended to the patch embedding sequence. This token's final state serves as the image representation for classification. Position embeddings are added to all tokens.

3. Transformer Encoding: identical layers, each containing:

- Layer Normalization (LN)
- Multi-Head Self-Attention (MSA)
- Residual connection
- Layer Normalization
- Multi-Layer Perceptron (MLP) with GELU activation
- Residual connection

4. *Classification Head:* The final token passes through a Layer Normalization and linear projection to class logits.

5. *Critical Limitation*: ViT lacks CNN inductive biases. Translation equivariance must be learned entirely from data, requiring massive pre-training datasets (JFT-300M, ImageNet-21k) to achieve competitive performance. This observation motivated nearly all subsequent ViT improvements.

3: **Architectural Variants and Improvement Strategies**

3.1 Taxonomy of Vision Transformer Improvements

Based on our analysis of 30+ ViT variants, we identify seven primary improvement strategies:

Table 3.1: Taxonomy of ViT Improvement Strategies

Category	Objective	Representative Models	Core Mechanism
Data Efficiency	Reduce data requirement	DeiT, T2T-ViT	Distillation, augmentation
CNN Hybrid	Add inductive bias	CvT, CeiT, CPVT	Conv token embedding
Lightweight	Reduce FLOPs	LeViT, CCT, PiT	Pyramid structure
Depth Scaling	Enable deeper ViTs	CaiT, DeepViT	LayerScale, Re-attention
Hierarchical	Multi-scale features	Swin, PVT	Window attention
Attention Innovation	Improve expressiveness	CrossViT, Focal	Cross-scale attention

3.2 Data-Efficient Transformers

DeiT (Data-efficient Image Transformers): Touvron et al. demonstrated that ViT could be trained effectively on ImageNet-1k alone through three innovations: (1) A distillation token added to the input sequence, trained to match a teacher CNN's predictions; (2) Hard distillation using teacher-predicted labels; (3) Aggressive data augmentation (RandAugment, MixUp, CutMix) and regularization (stochastic depth). DeiT achieved 81.8% top-1 accuracy with 86M parameters, comparable to ViT-B/16 trained on JFT-300M.

*Hard Distillation Formulation:*The class token and distillation token learn complementary representations—cosine similarity starts at 0.06 in early layers and increases to 0.93 in the final layer.

T2T-ViT (Tokens-to-Token): Yuan et al. addressed two limitations: (1) ViT's rigid patch splitting fails to model local structure; (2) Attention maps exhibit redundancy. T2T progressively aggregates neighboring tokens through overlapping soft splitting, creating image-like representations that capture edges and corners. The deep-narrow backbone reduces embedding dimension while maintaining representational capacity.

3.3 CNN-Transformer Hybrids

CvT (Convolutional Vision Transformer): Wu et al. integrated convolutions at two levels: (1) Convolutional token embedding replaces linear patch projection, generating overlapping tokens with local context; (2) Convolutional projection computes Q, K, V through depthwise convolutions rather than linear layers. This design achieves implicit positional encoding—CvT operates without position embeddings entirely. CvT-W24 achieves 87.7% ImageNet top-1 accuracy with 277M parameters.

CPVT (Conditional Position Encoding Vision Transformer): Chu et al. identified that fixed positional encodings limit resolution flexibility. The Positional Encoding Generator (PEG) uses 2D convolutions with zero-padding on reshaped patch embeddings to generate adaptive, resolution-independent position encodings. PEG adds only negligible computational overhead while enabling CPVT to outperform DeiT across resolutions.

VT (Visual Transformer): Wu et al. proposed tokenizing images at semantic level rather than pixel level. CNN backbone extracts feature maps, which are projected into compact semantic tokens through spatial attention. Transformers model relationships between these L tokens (L << HW), dramatically reducing computation while focusing on high-level concepts. VT-ResNet50 achieves 80.6% accuracy with 3.4G FLOPs—6.9× fewer FLOPs than baseline.

UniFormer: Li et al. unified convolution and self-attention within a single transformer block. Shallow layers employ local relation aggregators (convolution-like) to reduce redundancy; deep layers employ global self-attention to capture dependencies. UniFormer achieves 86.3% ImageNet accuracy with 2-4× higher throughput than lightweight competitors.

3.4 Lightweight and Efficient Transformers

LeViT: Graham et al. designed a hybrid architecture optimized for inference speed rather than parameter count. Key innovations: (1) Pyramid structure with decreasing spatial resolution and increasing channel dimensions; (2) Attention bias replacing positional encoding; (3) No class token—average pooling produces 512-d vector for dual classifiers; (4) 1×1 convolutions with batch normalization replacing linear layers. LeViT-384 achieves 82.6% accuracy with 39.1M parameters and 2.2G FLOPs—5× faster than EfficientNet on CPU.

CCT (Compact Convolutional Transformer): Hassani et al. challenged the "transformers require big data" assumption. CCT replaces patch embedding with convolutional layers, preserving local spatial information, and introduces Sequence Pooling to replace the class token. With only 0.28M parameters, CCT-2/3×2 achieves 89.17% on CIFAR-10 and 66.90% on CIFAR-100—competitive with CNNs trained from scratch.

PiT (Pooling-based Vision Transformer): Heo et al. demonstrated that spatial dimension reduction—a standard CNN design principle—benefits transformers. PiT integrates pooling layers between transformer stages, reducing token count while increasing channel depth. This addresses ViT's constant-resolution limitation and improves both efficiency and generalization.

Dynamic Token Sparsification: Rao et al. exploited spatial sparsity in visual data, progressively pruning redundant tokens based on learned importance scores. Hierarchical pruning reduces FLOPs by 31-35% with <0.5% accuracy drop, applicable to both transformers and CNNs.

3.5 *Deep Transformer Architectures:*

CaiT (Class-Attention in Image Transformers): Touvron et al. identified a conflict in standard ViT: the same self-attention mechanism must simultaneously perform feature extraction and classification. CaiT decouples these through two-stage processing:

Self-attention stages: No class token; layers focus solely on enriching patch representations

Class-attention stages: Class token introduced, attends to frozen patch embeddings; only class token updated

LayerScale—multiplication of residual outputs by learnable diagonal matrices—stabilizes training in deep transformers. CaiT-M48 achieves 86.5% ImageNet accuracy.

DeepViT: Zhou et al. diagnosed "attention collapse" in deep ViTs—attention maps become increasingly similar across layers, limiting representational diversity. Re-attention reconstructs attention maps through linear transformation across heads, leveraging low inter-head similarity to regenerate diverse attention patterns. DeepViT-L (55M parameters) achieves 82.2% accuracy without significant computational overhead.

PVT (Pyramid Vision Transformer): Wang et al. adapted transformers for dense prediction tasks requiring multi-scale features[6]. Spatial-Reduction Attention (SRA) reduces K and V spatial dimensions before attention computation, achieving linear complexity. Progressive shrinking pyramid generates feature maps at 1/4, 1/8, 1/16, 1/32 resolutions—compatible with FPN architectures for detection/segmentation.

3.*6 Advanced Attention Mechanisms:*

Swin Transformer: Liu et al.[2] introduced hierarchical feature maps through shifted window attention. Self-attention computed within non-overlapping local windows (7×7) for linear complexity; shifted window partitioning enables cross-window connections. Swin-T (29M parameters) achieves 83.5% accuracy with 4.5G FLOPs, establishing new state-of-the-art on COCO (58.7 AP) and ADE20K (53.5 mIoU).

CrossViT: Chen et al. processed multiple patch scales through dual-branch transformers. Cross-attention fusion enables information exchange: large-branch class token queries small-branch patch tokens. This achieves multi-scale representation with linear complexity in cross-attention (only class token serves as query). CrossViT achieves competitive accuracy with fewer parameters than single-scale ViT.

NesT (Nested Transformers): Zhang et al. aggregated nested, non-overlapping image blocks. Within-block self-attention captures local relationships; block aggregation function enables cross-block communication through concatenation and convolution. NesT-B (68M parameters) achieves 97.20% CIFAR-10 accuracy trained from scratch.

Focal Transformer: Yang et al. proposed focal attention that captures both fine-grained local patterns and coarse global dependencies. Each token attends to its surrounding region at fine granularity and distant regions at coarse granularity, balancing detail preservation and computational efficiency.

3.7 *Training and Optimization Innovations*

Patch-wise Loss: Gong et al. addressed over-smoothing through loss functions operating on patch representations rather than only class token:

Cosine loss: Maximizes diversity between patch representations

Contrastive loss: Maintains consistency between early and deep layer patches

Mixing loss: Applies supervisory signal to individual patches

These losses improve training stability, enable higher drop path rates, and enhance representation quality.

MetaFormer: Yu et al. abstracted the Transformer architecture to its minimal form: token mixer (any function mixing spatial information) + channel MLP. Surprisingly, identity mapping and random matrices as token mixers achieve >80% ImageNet accuracy. ConvFormer (using depthwise convolution as token mixer) outperforms ConvNeXt. This suggests the Transformer architecture itself—rather than specific attention mechanisms—provides significant inductive bias.

4: Quantitative Performance Analysis

4.1 Experimental Methodology and Dataset ConsiderationsDataset Characteristics:

ImageNet-1K: 1.28M training images, 50K validation, 1000 classes. Standard benchmark for large-scale image classification.

CIFAR-10/100: 60K 32×32 images (50K train, 10K test). Tests low-resolution, limited-data performance.

Downstream Transfer: Fine-tuning pre-trained models on specialized datasets (Oxford Pets, Flowers, Stanford Cars) measures representation quality.

4.2 ImageNet Performance Comparison

Table 4.1: Comprehensive Comparison of Vision Transformers on ImageNet

Model	**Params (M)**	**FLOPs (G)**	**Top-1 (%)**
ViT-B/16	86	17.7	77.9
DeiT-S	22	4.6	79.8
Swin-T	29	4.5	83.5
CvT-13	20	4.5	81.6
CaiT-M48	356	329	86.5

Model	Params (M)	FLOPs (G)	Top-1 (%)
ResNet-50	25	4.1	76.2
EfficientNet-B7	66	37	84.3

Table 4.2: Representative CNN Baselines for Comparison

Model	Parameters (M)	FLOPs (G)	ImageNet Top-1 (%)	Year
ResNet-50	25	4.1	76.2	2016
ResNet-101	45	7.9	77.4	2016
ResNet-152	60	11.0	78.3	2016
EfficientNet-B5	30	9.9	83.6	2019
EfficientNet-B7	66	37.0	84.3	2019
RegNetY-16GF	84	16.0	82.9	2020
NFNet-F0	72	12.4	83.6	2021
NFNet-F6 + SAM	438	377.0	86.5	2021

1. Scale Dependency: ViT-B (86M) trained on ImageNet alone achieves 77.9%, below ResNet-152's 78.3%. DeiT's distillation and regularization close this gap (81.8%), demonstrating that data-efficient training is possible with appropriate techniques.
2. Hybrid Superiority: CvT and CeiT consistently outperform pure transformers at comparable parameter budgets. CvT-13 (20M, 81.6%) exceeds DeiT-S (22M, 79.8%) by 1.8%, validating convolution's value as inductive bias.
3. Efficiency Frontier: LeViT demonstrates that optimizing for throughput rather than parameter count yields practical efficiency gains. At 82.6% accuracy, LeViT-384 (2.2G FLOPs) requires 8× fewer FLOPs than ViT-B (17.7G FLOPs) with higher accuracy.
4. Depth Scaling: CaiT achieves 86.5% accuracy, matching the best CNN-based architectures (NFNet-F6+SAM: 86.5%) with comparable computational budgets. This demonstrates that depth limitations in early ViTs were solvable through appropriate architectural modifications.
5. Parameter Efficiency: CCT achieves 89.17% on CIFAR-10 with only 0.28M parameters—3.8× fewer than Wide-ResNet (1.07M) with comparable accuracy. This challenges the "transformers require massive data" narrative, at least for low-resolution tasks.

4.3 Small Dataset Performance

Table 4.3: CIFAR-10/100 Performance (Training from Scratch)

Model	Params (M)	CIFAR-10 (%)	CIFAR-100 (%)
CCT-7/3×1	3.76	94.72	76.67
NesT-B	68	97.20	82.56
ResNet-164	1.7	94.54	75.67
MobileNetV2	2.24	89.07	63.69

Key Findings:
1. Transformers can compete on small data: CCT and NesT achieve state-of-the-art on CIFAR without ImageNet pre-training, contrary to early beliefs that transformers require massive datasets.
2. Architecture matters more than scale: CCT's convolutional tokenization provides sufficient inductive bias for low-resolution learning, achieving 94.72% CIFAR-10 accuracy with 3.76M parameters—outperforming ResNet-164 (94.54%) with 1.7M parameters.
3. CNN still parameter-efficient: ResNet-164 achieves comparable accuracy with 2.2× fewer parameters than CCT-7/3×1, reflecting CNNs' superior parameter efficiency on limited data.

4.4 Transfer Learning Performance

Key Observations:
1. Scale improves transfer: ViT-H/14 achieves 99.27% CIFAR-10 accuracy, demonstrating that larger pre-trained models produce better feature representations.
2. Hybrids excel: CvT-W24 achieves 99.39% CIFAR-10 accuracy—surpassing pure transformers—suggesting convolution provides beneficial inductive bias even with large-scale pre-training.
3. Distillation helps: DeiT-B distilled version improves Stanford Cars accuracy from 92.1% to 93.9%, confirming teacher guidance transfers to downstream tasks.
4. Specialized datasets: Fine-grained classification (Stanford Cars: 196 classes of cars) benefits from transformer architectures, with CaiT achieving 94.2% accuracy.

4.5 Computational Efficiency Analysis

Throughput vs. Accuracy Trade-off:
LeViT's systematic benchmarking reveals important discrepancies between FLOPs and practical speed:
- LeViT-384: 82.6% accuracy, 2.2G FLOPs, **5× faster than EfficientNet-B3** on CPU
- DeiT-S: 79.8% accuracy, 4.6G FLOPs, comparable FLOPs but slower inference due to attention computation overhead

This highlights that FLOPs alone insufficiently characterize transformer efficiency; memory access patterns, parallelism, and hardware-specific

optimizations significantly impact real-world performance.

5: Critical Case Study—The Retracted ECG-ViT Paper

5.1 Paper Overview and Claims

The paper "ECG-ViT: A Transformer-Based ECG Classifier for Energy-Constraint Wearable Devices" (Shukla et al., Journal of Sensors, 2022)[1-3] proposed a Vision Transformer architecture for arrhythmia classification using the MIT-BIH Arrhythmia Database. The authors claimed:

1. Architectural Innovation: A ViT-based model adapted for ECG time-series classification, using GPSA layers with convolutional initialization.
2. Model Compression: Knowledge distillation to reduce complexity, claiming self-distillation without requiring a large teacher model.
3. Hardware Implementation: FPGA deployment on Xilinx Alveo U50, reporting 38% less area and 27% less energy consumption through approximate multiplication.
4. State-of-the-Art Performance: 99.7% F1 score and 99.3% accuracy—significantly outperforming prior methods (Wiens and Guttag, Cong et al.).

5.2 Retraction and Investigation Findings

On December 20, 2023, the Journal of Sensors retracted the article following an investigation by Hindawi publishers[11]. The retraction notice identified six indicators of systematic manipulation:

1. Discrepancies in Scope: The paper claims to address ECG classification but extensively discusses PPG-to-ECG synthesis (Section 2.2) without connecting this to the proposed methodology. Reference [12] is misattributed—it concerns VL-BERT (vision-language), not PPG-ECG translation.
2. Discrepancies in Research Description: The "cascade distillation approach" claimed in contributions is never defined or implemented. Section 3.2 describes standard DeiT distillation, not cascade distillation. The self-distillation claim contradicts requiring a teacher model.
3. Data Availability Discrepancies: The paper states "data preprocessed to obtain sample at 128Hz" but provides no preprocessing code or detailed methodology. MIT-BIH is 360Hz—downsampling to 128Hz without anti-aliasing filters would introduce aliasing artifacts, unmentioned in the paper.
5. Incoherent/Meaningless Content: Section 3.1.1 describes ViT processing "224-size images divided into 16×16 patches of 14×14 pixels"—mathematically impossible (224/16 = 14, consistent, but 16×16 patches of 14×14 pixels would require 224×224 image). ECG signals are 1D time-series; applying 2D ViT patching to 1D signals without explanation of dimensionality transformation represents incoherent methodology.
6. Manipulated Peer Review: The retraction notice cites "systematic manipulation of the publication process," suggesting fabricated or compromised peer review—a growing concern in special issues and open-access venues.

5.3 Methodological Critique

Fundamental Architectural Mismatch:

ViT is designed for 2D spatial data, processing images as sequences of 2D patches. ECG signals are 1D temporal sequences. The paper never explains:

- How 1D ECG time-series are converted to 2D patch sequences
- Why 2D spatial position embeddings are appropriate for temporal data
- How the relationship between time steps differs from spatial relationships

This represents a category error—applying a 2D architecture to 1D data without justification or architectural modification[4].

Performance Claim Implausibility:

Table 5.1: Reported vs. Plausible Performance on MIT-BIH Arrhythmia Classification

Method	**Accuracy**	**F1 Score**	**Dataset Split**	**Validation**
Wiens and Guttag (2010)	–	58.2 (SVE)	Patient-adaptive	Standard
Cong et al. (2011)	0.95 (VE)	–	Not specified	–
ECG-ViT (claimed)	99.3%	99.7%	10-fold cross-validation	Not reproducible
SOTA ECG	–	–	–	–

Summary:

Aspect	***CNNs (ResNet, EfficientNet, NFNet)***	***Vision Transformers (ViT, Swin, CvT, etc.)***
Core Mechanism	*Convolution (local receptive fields)*	*Self-attention (global dependency modeling)*
Inductive Bias	*Strong (locality, translation equivariance)*	*Weak (learned from data)*
Data Requirement	*Moderate*	*High (improves with large-scale pretraining)*

Aspect	***CNNs (ResNet, EfficientNet, NFNet)***	***Vision Transformers (ViT, Swin, CvT, etc.)***
Small Dataset Performance	*Strong and parameter-efficient*	*Improving (CCT, NesT competitive)*
Large-Scale Performance	*Competitive up to ~86.5%*	*State-of-the-art with scaling & hybrids*
Computational Complexity	*Linear scaling*	*Quadratic (standard attention)*
Efficiency Optimizations	*Depth scaling, compound scaling*	*Window attention, hierarchical design*
Hybrid Models	*Limited*	*Highly effective (Swin, CvT, UniFormer)*
Healthcare Suitability	*Strong for 1D/2D signals*	*Requires modality-aligned design*
Research Integrity Lesson	*Mature, well-validated*	*Needs rigorous validation (ECG-ViT case)*

From the above table This paper presents a comprehensive comparative analysis of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for large-scale image classification and healthcare applications, demonstrating that while CNNs remain highly efficient and robust due to strong inductive biases, transformers offer superior global context modeling and scalability when trained with sufficient data and architectural refinements. Experimental comparisons across ImageNet and CIFAR benchmarks reveal that hybrid and hierarchical transformer models achieve state-of-the-art accuracy, often surpassing traditional CNNs at comparable computational budgets, though efficiency depends heavily on design choices such as window-based attention and distillation strategies. In healthcare contexts, particularly ECG classification, the study emphasizes the necessity of modality-appropriate architectural adaptation and strict validation protocols, highlighting the ECG-ViT case as an example of the risks associated with irreproducible claims. Overall, the findings conclude that transformers represent a transformative advancement in vision modeling, but their practical deployment requires careful optimization, computational awareness, and rigorous experimental integrity.

Conclusion

This paper provides a comprehensive comparative analysis of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) for image classification and healthcare applications, demonstrating that while CNNs retain strong inductive biases and parameter efficiency—especially in small-data scenarios—transformers offer superior global context modeling and scalability when trained with sufficient data and optimization strategies. Through evaluation of more than 30 transformer variants and multiple CNN baselines on benchmarks such as ImageNet and CIFAR, the study shows that hierarchical and hybrid architectures like Swin and CvT achieve the best balance between accuracy and efficiency, often surpassing traditional CNNs at comparable computational budgets. The analysis also highlights that performance improvements in transformers depend heavily on training techniques such as distillation, augmentation, and architectural stabilization methods. In healthcare applications, particularly ECG classification, the paper emphasizes the necessity of modality-appropriate architectural design, patient-wise validation, and transparent preprocessing, using the ECG-ViT case as a cautionary example of the risks of irreproducible research. Overall, the findings conclude that transformers represent a powerful and evolving paradigm in computer vision, but their practical success depends on careful design choices, computational considerations, and rigorous validation

References:

1. U.R. Acharya *et al.* A deep convolutional neural network model to classify heartbeats Comput. Biol. Med. (2017)
2. F. Murat *et al.* Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review Comput. Biol. Med. (2020)
3. U.B. Baloglu *et al.* Classification of myocardial infarction with multi-lead ECG signals and deep CNN
   Pattern Recognit. Lett.(2019)
4. R. Hu *et al.* A transformer-based deep neural network for arrhythmia detection using continuous ECG signals Comput. Biol. Med. (2022)
5. N. Alamatsaz *et al.* A lightweight hybrid CNN-LSTM explainable model for ECG-based arrhythmia detection Biomed. Signal Process. Control. (2024)
6. C. van Zyl *et al* Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP Appl. Energy.(2024)

7. Anand *et al.* Explainable AI decision model for ECG data of cardiac disorders Biomed. Signal Process. Control.(2022)
8. Y.Y. Jo *et al.* Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram Int. J. Cardiol. (2021)
9. J.K. Kim *et al.* Arrhythmia detection model using modified DenseNet for comprehensible grad-CAM visualization Biomed. Signal Process. Control.(2022)
10. J. Chang *et al.* Explaining the rationale of deep learning glaucoma decisions with adversarial examples
11. (Sun et al. 2020) cited for PR interval interpretation—paper concerns ensemble ECG classification, not clinical cardiology
12. (Su et al. 2019) cited for PPG-ECG translation—paper concerns VL-BERT vision-language pre-training