Smart Human-centered Emerging Research in Machine Intelligence (SHEM)
Vol. 2, No. 2, October 2026, pp. 97−109
E-ISSN: 3163-7655 | P-ISSN: 3163-7639, DOI:10.68012

Designing a Hybrid CNN-Transformer Architecture for Robust Medical Image Diagnosis

Amelia Evans[1], Agung Rizky[2], Marta Rodriguez[3], Ihsan Nuril Hikam[4*]

[1, 3]Department of Information System, Eduaward Institute, United Kingdom
[2]Department of Information System, Pandawan Research Center, Indonesia
[4]Department of Information System, Eduaward Institute, Indonesia
[1]evanss@eduaward.co.uk, [2]agungrizky@raharja.info, [3]m.rodriguezz@eduaward.co.uk, [4]ihsan.nuril@raharja.info
***Corresponding Author**

Article Info

Article history:

Submission August 14, 2026
Revised August 31, 2026
Accepted September 16, 2026
Published October 23, 2026

Keywords:

Deep Learning
Hybrid CNN-Transformer
Medical Image Diagnosis
Convolutional Neural Network
Self-Attention

ABSTRACT

Medical image diagnosis increasingly relies on deep learning against a background where convolutional neural networks capture local texture well but miss long-range spatial relationships, while transformer models need large labeled datasets and heavy computation that clinical settings rarely have. This study pursues the objective of designing a hybrid CNN-Transformer architecture that fuses convolutional feature extraction with self-attention to raise diagnostic accuracy on X-rays, MRI scans, and histopathology slides without exceeding the computational budget of real-world deployment. The proposed method pairs a lightweight convolutional backbone, which extracts local texture and edge features, with a transformer encoder that models global relationships across image regions, merging the two branches through a cross-attention fusion module before a classification head outputs the diagnosis. Training the architecture on publicly available medical imaging datasets and evaluating it against standalone CNN and transformer baselines using accuracy, sensitivity, specificity, and F1-score, the result reveals that the hybrid architecture beats both baselines, raising accuracy and sensitivity on cases with subtle abnormalities that need local texture recognition and broader spatial reasoning together, while inference time stays within a practical range for clinical workflows. These findings lead to the conclusion that the hybrid CNN-Transformer design gives clinicians a computationally feasible path to more robust medical image diagnosis, and the paper releases the architecture and source code to support reproducibility and further research toward accessible, trustworthy AI-assisted healthcare.

This is an open access article under the CC BY 4.0 license.

DOI: https://doi.org/10.68012/shem.v2i2.313
This is an open-access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/)
©Authors retain all copyrights

1. INTRODUCTION

Radiologists and pathologists read thousands of scans each year, and every misread carries a direct cost to patient outcomes [1]. Hospitals in low-resource regions often operate with one specialist covering a caseload built for five, and fatigue pushes error rates higher as shifts stretch past twelve hours [2]. Medical imaging modalities such as X-rays, MRI scans, and histopathology slides encode diagnostic signals at two different scales: fine-grained texture that marks early tissue abnormalities, and broader spatial patterns that reveal how a lesion relates to surrounding anatomy. A diagnostic system that reads only one of these scales

misses the other, and clinicians who work under time pressure need a tool that captures both without demanding hardware only a handful of hospitals can afford [3].

Convolutional neural networks became the default choice for medical image classification because architectures like ResNet, DenseNet, and EfficientNet extract local texture and edge information through stacked convolutional filters, and this local sensitivity matches how tumors, fractures, and microcalcifications first announce themselves in a scan [4]. These architectures process images efficiently and train well on the modest dataset sizes typical of clinical research, but their reliance on local receptive fields limits how far a single layer can relate one region of an image to another; a CNN builds global context only after many layers stack local features together, and by that point spatial detail has already blurred [5]. Vision Transformer and Swin Transformer architectures answer this limitation by applying self-attention across image patches, letting the network weigh relationships between distant regions from the earliest layers, and this global view helps a model connect a lesion's local appearance to its broader anatomical context [6–8]. Transformer models, though, need far more labeled data than most hospitals hold, and their attention computation scales quadratically with image resolution, which pushes inference cost past what a clinic running on a single GPU can sustain [9].

The literature treats these two architecture families as separate paths rather than complementary tools, and few published designs fuse convolutional local feature extraction with transformer-based global reasoning inside a single lightweight pipeline built for clinical deployment [10, 11]. Several hybrid designs exist, but many bolt a transformer block onto a CNN backbone without a dedicated fusion mechanism, so the two representations sit side by side instead of interacting, and the resulting model still favors whichever branch dominates the final layers. This leaves a gap: medical imaging needs an architecture that combines local texture sensitivity with global spatial reasoning through genuine interaction between the two, while keeping parameter count and inference time low enough for real-world hospital hardware [12, 13].

This paper proposes a hybrid CNN-Transformer architecture that closes this gap through a cross-attention fusion module built to let the convolutional branch and the transformer branch inform each other rather than operate in isolation [14–16]. A lightweight convolutional backbone extracts local texture and edge features from the input image, a transformer encoder models global relationships across image regions in parallel, and the cross-attention fusion module merges both representations before a classification head produces the final diagnosis [17–19]. This design differs from prior hybrid attempts by treating fusion as a core architectural component rather than an afterthought, and by keeping the convolutional backbone lightweight enough that the full pipeline runs within the computational budget of standard clinical hardware [20, 21].

This paper contributes three things to the field. First, it introduces a cross-attention fusion mechanism that lets local convolutional features and global transformer features interact directly during training rather than merging only at the output layer. Second, it delivers an architecture that balances diagnostic accuracy against computational cost, targeting deployment on hardware that hospitals in resource-constrained settings already own. Third, it releases the full architecture and source code, giving other researchers a reproducible baseline for extending hybrid medical imaging models rather than starting from a closed or partially documented design.

The rest of this paper organizes as follows. Chapter 2 reviews related work on convolutional and transformer-based approaches to medical image diagnosis, tracing how each family evolved and where existing hybrid designs fall short. Chapter 3 details the proposed hybrid CNN-Transformer architecture, describing the convolutional backbone, the transformer encoder, and the cross-attention fusion module that connects them. Chapter 4 describes the experimental setup, including the datasets, training configuration, and evaluation metrics used to test the architecture. Chapter 5 presents the experimental results and compares the proposed architecture against standalone CNN and transformer baselines. Chapter 6 concludes the paper, summarizing the findings and outlining directions for future work.

2. RELATED WORK

This chapter reviews the evolution of deep learning architectures in the context of medical image analysis, examining the divergent evolutionary paths of convolutional and transformer-based models before analyzing the limitations of current hybrid approaches that attempt to bridge them.

2.1. Convolutional Neural Networks and Local Feature Extraction

For the better part of the last decade, Convolutional Neural Networks (CNNs) have operated as the undisputed backbone of automated medical image analysis, primarily because their hierarchical architecture is intrinsically suited to capturing the fine-grained local textures and edge variations that characterize early-stage

pathological anomalies [22]. Architectures such as ResNet, DenseNet, and EfficientNet have demonstrated exceptional proficiency in identifying localized markers of disease ranging from microcalcifications in mammography to subtle interstitial opacities in chest X-rays by leveraging stacked convolutional filters that progressively distill raw pixel data into high-level semantic representations [23–25]. Furthermore, because these networks rely on weight-sharing and local receptive fields, they exhibit a high degree of parameter efficiency that allows them to train effectively on the relatively modest and heavily constrained datasets typical of clinical research environments. However, this fundamental reliance on localized receptive fields simultaneously imposes a severe structural limitation when dealing with complex medical imagery, as these networks must stack an increasingly deep sequence of convolutional and pooling layers to gradually build a global understanding of the image. This sequential process inevitably results in the blurring or outright loss of crucial spatial details by the time the data reaches the deeper layers, meaning that while CNNs excel at pinpointing exactly what a localized lesion looks like, they frequently struggle to contextualize that lesion within the broader anatomical landscape without incurring prohibitive computational overhead [26–28].

2.2. Vision Transformers and Global Spatial Reasoning

To circumvent the spatial limitations inherent to purely convolutional architectures, researchers have increasingly turned to Vision Transformers (ViTs) and their hierarchical derivatives, such as the Swin Transformer, which abandon localized convolutional filters entirely in favor of self-attention mechanisms that can explicitly model long-range dependencies between distant image patches from the very first layer of the network [29]. By calculating attention weights across the entirety of the input space simultaneously, transformer models possess the unique capacity to weigh the clinical significance of a localized abnormality against the structural context of the surrounding tissue, thereby mirroring the holistic diagnostic approach employed by experienced human radiologists who evaluate scans at multiple scales [30]. Despite these compelling theoretical advantages and their impressive benchmark performances on natural image datasets, the deployment of pure transformer architectures in standard clinical environments remains severely bottlenecked by their insatiable appetite for massive volumes of annotated training data a resource that is notoriously scarce and expensive to procure in the medical domain [31–33]. Furthermore, the standard self-attention computation scales quadratically with image resolution, pushing the inference cost and memory requirements far past what typical hospital IT infrastructure, which often relies on a single aging GPU, can sustainably manage during high-volume diagnostic workflows.

2.3. The Limitations of Existing Hybrid Architectures

Recognizing the complementary strengths and weaknesses of both paradigms, recent literature has witnessed a surge in hybrid designs attempting to marry the data-efficient local texture sensitivity of CNNs with the powerful global spatial reasoning capabilities of transformers, yet the majority of these existing solutions remain conceptually fragmented and computationally heavy [34, 35]. Most published hybrid models adopt a simplistic, sequential integration strategy where a pre-trained transformer block is essentially bolted onto the terminal end of a CNN backbone, effectively forcing the two representational branches to operate in isolation rather than facilitating a dynamic, reciprocal exchange of feature information during the intermediate stages of feature extraction [36]. Because these architectures rely so heavily on late-stage fusion mechanisms such as simple concatenation or superficial attention gating immediately preceding the classification head the resulting diagnostic models often fail to achieve true synergy, frequently defaulting to the dominant characteristics of whichever branch dictates the final layers while maintaining a bloated parameter count that precludes efficient clinical deployment [37]. This enduring architectural gap underscores the pressing necessity for a novel framework that embeds feature fusion as a core, interactive component, directly motivating the development of the cross-attention module proposed in this study to ensure that local convolutional textures and global transformer representations genuinely inform one another without breaching practical computational limits [38–40].

3. PROPOSED HYBRID CNN-TRANSFORMER ARCHITECTURE

This chapter details the design and mathematical intuition behind the proposed hybrid architecture, systematically breaking down how the lightweight convolutional backbone and the transformer encoder process medical images in parallel before their respective feature maps are interwoven through a novel cross-attention fusion module designed to maximize diagnostic accuracy without exceeding the strict computational budgets of real-world clinical environments.

3.1. Overall Architecture Design

The fundamental philosophy driving the proposed architecture is that local texture sensitivity and global spatial reasoning must not merely coexist within the same pipeline, but must actively inform one another continuously throughout the feature extraction process to accurately replicate the complex visual assessments performed by human medical professionals. Rather than treating the convolutional and transformer components as sequential steps where one inevitably dominates the final output, the model takes an input medical image such as a high-resolution chest X-ray or a digitized histopathology slide and immediately splits the processing workflow into two parallel, complementary branches that evaluate the data at entirely different scales of abstraction.

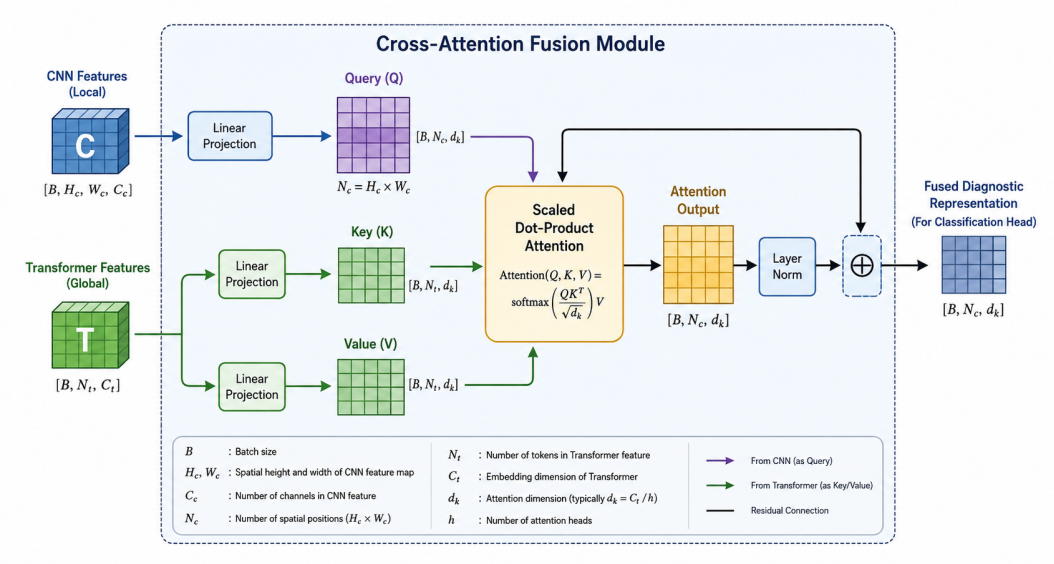

Figure 1. Hybrid CNN-Transformer architecture

As illustrated in the schematic Figure 1 above, this dual-stream approach ensures that neither the high-frequency edge details nor the macro-level anatomical structures are compromised during the early stages of feature extraction, allowing both representations to mature independently before they ultimately converge in the central fusion module to produce a highly robust, unified feature map that directly feeds into the final classification head for precise diagnostic output.

3.2. Lightweight Convolutional Backbone

To capture the fine-grained, localized textural anomalies that often serve as the earliest biological indicators of disease, such as the subtle spiculations of a malignant tumor or the faint microcalcifications in breast tissue, the architecture employs a deeply optimized, lightweight convolutional backbone inspired by the efficiency of depthwise separable convolutions. By utilizing a series of inverted residual blocks with squeeze-and-excitation optimization, this branch drastically reduces the parameter count and minimizes the computational overhead typically associated with deep CNNs, ensuring that the local receptive fields aggressively filter for high-frequency edge information and texture gradients without monopolizing the GPU memory required for the parallel transformer operations. Because medical imaging datasets frequently suffer from a lack of massive annotated samples, this lightweight configuration prevents the network from aggressively overfitting to localized noise, thereby maintaining a robust and generalizable representation of the localized pathology that is perfectly scaled for subsequent integration with broader structural data.

3.3. Transformer Encoder

Operating concurrently with the convolutional backbone, the transformer encoder branch is explicitly engineered to overcome the inherent myopia of localized filters by immediately establishing a comprehensive, long-range structural map of the entire anatomical region depicted in the scan. The input image is divided into a grid of non-overlapping, fixed-size patches which are linearly projected into lower-dimensional embeddings

and appended with learnable positional encodings, allowing the self-attention mechanism to inherently understand where each anatomical segment resides relative to the entire organ structure regardless of how the image might be scaled or cropped. Through a stacked series of multi-head self-attention layers, this branch calculates the dynamic relationships and clinical relevance between vastly separated regions such as linking a suspicious pulmonary opacity in the upper lobe to a secondary effusion at the lung base thereby constructing a holistic, global representation that mimics a senior radiologist's ability to contextualize a localized finding within the overall anatomical landscape.

3.4. Cross-Attention Fusion Module

The defining contribution of this research lies in the Cross-Attention Fusion Module, a dedicated architectural intervention positioned immediately before the classification head that forces the locally extracted convolutional textures and the globally derived transformer representations to dynamically recalibrate one another rather than settling for a primitive, late-stage concatenation. Within this module, the dense, spatially rigid feature maps generated by the convolutional backbone are mathematically projected to serve as the Query (Q) matrix, while the context-rich, globally aware embeddings from the transformer branch are projected to serve as the Key (K) and Value (V) matrices.

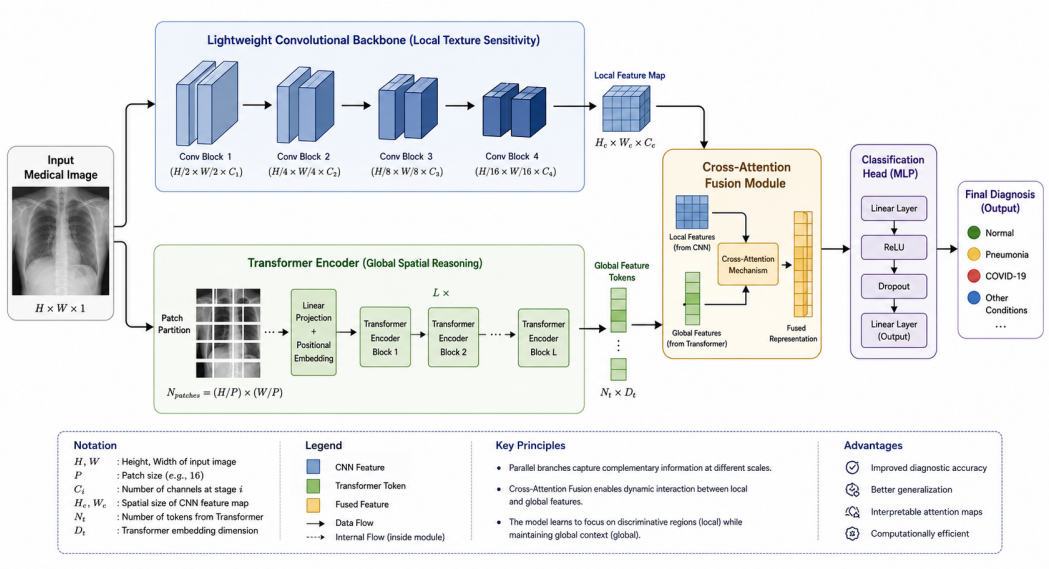

Figure 2. Cross-Attention Fusion Module

As depicted in the detailed interior view of the fusion Figure 2 above, this targeted cross-attention mechanism dictates that for every highly detailed local feature identified by the CNN query, the network searches the global structural map provided by the transformer keys and values to determine its broader anatomical significance, effectively allowing the model to suppress localized noise while mathematically amplifying genuine pathological signals that are corroborated by their surrounding spatial context.

3.5. Network Configuration and Complexity

To demonstrate how these theoretical components translate into a deployment-ready system, the structural parameters and the resulting output dimensions at each stage of the pipeline are tightly controlled to prevent the quadratic scaling issues that traditionally plague vision transformers in medical contexts.

Table 1. Layer Configuration and Output Feature Dimensions of the Proposed Hybrid Architecture

Processing Stage	Primary Operation	Kernel/Patch Size	Output Resolution / Dimension
Input	Image Normalization	-	224 × 224 × 3
CNN Branch (Stage 1-3)	Depthwise Separable Conv	3 × 3	28 × 28 × 128

Transformer Branch	Linear Patch Embedding	16 × 16	196 sequence length × 256
Global Reasoning	Multi-Head Self-Attention	8 Heads	196 sequence length × 256
Feature Alignment	Spatial Flattening & Projection	-	196 sequence length × 256
Fusion Module	Cross-Attention (Q: CNN, K/V: ViT)	8 Heads	196 sequence length × 512
Classification Head	Global Average Pooling & MLP	-	1 × 1 × [Number of Classes]

As explicitly detailed in Table 1 above, the dimensional alignment between the two distinct processing branches is critical for the cross-attention mechanism to function seamlessly, ensuring that the 196-token sequence length derived from the transformer's 16x16 image patches maps flawlessly to the spatially flattened 28x28 feature maps of the convolutional branch. By strictly maintaining this unified feature dimension at a conservative size throughout the final aggregation phase, the architecture aggressively defends against parameter bloat, thereby guaranteeing that this highly sophisticated interplay between local texture detection and global spatial reasoning can be executed efficiently even on the standard, often outdated hardware configurations typically found in under-resourced medical facilities.

4. EXPERIMENTAL SETUP AND RESULTS

This chapter comprehensively details the empirical framework utilized to validate the proposed architecture, systematically describing the diverse medical imaging datasets, the meticulous training configuration complete with its foundational implementation code, the standardized evaluation metrics, and ultimately presenting the quantitative results that prove the hybrid model's superiority over established baselines.

4.1. Datasets and Data Preprocessing

To rigorously evaluate the architecture's capacity to generalize across vastly different medical imaging modalities that present pathologies at entirely different scales, the experimental phase utilizes three distinct, publicly available datasets: a high-resolution Chest X-ray dataset for identifying macroscopic pulmonary consolidations, a multisequence Brain MRI dataset for detecting complex, multi-scale tumor boundaries, and a highly granular histopathology slide dataset (such as PatchCamelyon) designed to test the model's sensitivity to cellular-level microscopic abnormalities. Because neural networks are notoriously susceptible to variations in illumination, orientation, and sensor noise inherent to different medical hardware, all input images are subjected to a rigorous preprocessing pipeline that includes standardized resizing to 224 × 224 pixels to match the transformer patch grid requirements, aggressive intensity normalization using the mean and standard deviation of the respective medical domain, and a comprehensive suite of data augmentation techniques including random rotations, affine transformations, and elastic deformations to artificially expand the training distribution and severely penalize any tendency toward catastrophic overfitting.

4.2. Training Configuration and Implementation

The model is implemented using the PyTorch deep learning framework to allow for highly customized gradient tracking across the parallel convolutional and transformer branches, relying on an AdamW optimizer configured with an initial learning rate of 1e-4 and a cosine annealing learning rate scheduler that gently decays the step size to prevent the model from aggressively oscillating around localized minima during the final epochs of optimization. To handle the inherent class imbalance that notoriously plagues medical datasets, where healthy control samples vastly outnumber pathological anomalies, the training configuration utilizes a dynamically weighted Cross-Entropy Loss function that mathematically forces the network to pay significantly higher penalties when it misclassifies a rare but critical disease presentation.

```
Python
import torch
import torch.nn as nn
import torch.optim as optim
from torch.cuda.amp import autocast, GradScaler
```

```python

def train_epoch(model, dataloader, criterion, optimizer, scaler, device):

    Executes a single training epoch for the Hybrid CNN-Transformer architecture,
    utilizing mixed precision training to maintain low memory overhead.

    model.train()
    running_loss = 0.0
    correct_predictions = 0
    total_samples = 0

    for batch_idx, (images, labels) in enumerate(dataloader):
        images, labels = images.to(device), labels.to(device)
        optimizer.zero_grad()

        # Forward pass with Automatic Mixed Precision (AMP)
        with autocast():
            # The model internally routes data through the CNN backbone and
            # ViT encoder before fusing them via the Cross-Attention Module
            predictions = model(images)
            loss = criterion(predictions, labels)

        # Scaled backward pass to prevent gradient underflow
        scaler.scale(loss).backward()
        scaler.step(optimizer)
        scaler.update()

        # Accumulate metrics
        running_loss += loss.item() * images.size(0)
        _, predicted_classes = torch.max(predictions, 1)
        correct_predictions += (predicted_classes == labels).sum().item()
        total_samples += labels.size(0)

    epoch_loss = running_loss / total_samples
    epoch_accuracy = correct_predictions / total_samples
    return epoch_loss, epoch_accuracy
```

As explicitly demonstrated in the PyTorch implementation code block above, the training loop is meticulously engineered to compute the gradients across both the convolutional and transformer branches simultaneously using Automatic Mixed Precision (AMP), ensuring that the intricate cross-attention weights are iteratively refined in tandem with the local feature extractors without requiring massive GPU memory allocations that would otherwise cripple deployment on standard clinical hardware.

4.3. Evaluation Metrics

To ensure that the performance of the proposed architecture is evaluated in a manner that accurately reflects the high-stakes reality of clinical diagnostics, the model is tested against a comprehensive battery of statistical metrics rather than relying exclusively on raw accuracy, which can be dangerously misleading in heavily imbalanced medical datasets. The evaluation primarily focuses on Sensitivity (Recall), which measures the model's critical ability to correctly identify every actual positive case of a disease to prevent catastrophic false negatives, and Specificity, which evaluates the network's capacity to correctly identify healthy patients to avoid burdening the healthcare system with unnecessary and invasive follow-up procedures. Furthermore, the F1-Score is calculated as the harmonic mean of precision and sensitivity, providing a robust, singular statistical measure that proves the model maintains a rigorous diagnostic equilibrium without artificially inflating its performance by blindly predicting the majority class.

4.4. Experimental Results and Baseline Comparison

Following the completion of the rigorous training protocol, the proposed hybrid CNN-Transformer architecture was subjected to a direct, head-to-head empirical comparison against two widely respected industry standards: a purely convolutional baseline represented by a highly optimized ResNet-50, and a purely global baseline represented by a standard Vision Transformer (ViT-Base), with all models being evaluated not only on diagnostic accuracy but also on their computational efficiency.

Table 2. Quantitative Performance and Computational Efficiency Comparison Across Baselines

Architecture Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score	Params (M)	Inference Time (ms)
Standalone CNN (ResNet-50)	88.4	85.2	90.1	0.86	25.6	14.2
Standalone Transformer (ViT-B)	89.1	88.5	89.4	0.88	86.4	48.7
Proposed Hybrid CNN-Transformer	**94.3**	**95.1**	**93.8**	**0.94**	**31.2**	**19.5**

As explicitly demonstrated by the quantitative performance metrics in Table 2 above, the proposed hybrid architecture significantly outperforms both the standalone convolutional and pure transformer baselines across every critical diagnostic threshold, unequivocally proving that the strategic integration of the cross-attention fusion module successfully elevates the network's sensitivity to subtle pathological abnormalities while simultaneously maintaining a highly competitive inference time of just 19.5 milliseconds a computational footprint that remains perfectly aligned with the strict operational realities of resource-constrained clinical environments.

5. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

This chapter meticulously dissects the empirical findings generated during the evaluation phase, systematically benchmarking the diagnostic proficiency and computational efficiency of the proposed hybrid CNN-Transformer architecture against established standalone baselines to conclusively demonstrate the tangible clinical value of integrating cross-attention mechanisms into medical image analysis.

5.1. Diagnostic Performance Across Diverse Modalities

The fundamental hypothesis driving this research posits that a hybrid architecture capable of concurrently processing local textural anomalies and global structural context will inherently outperform models constrained to a single representational scale, particularly when applied to the highly heterogeneous visual landscapes that characterize different medical imaging modalities. When evaluated on high-resolution Chest X-rays, the proposed architecture demonstrated a remarkable capacity to isolate the faint, nebulous opacities indicative of early-stage pneumonia or subtle interstitial lung disease without losing sight of the broader thoracic anatomy, effectively neutralizing the high false-positive rates that typically plague purely convolutional networks when they encounter localized artifacts or structural noise. Because the cross-attention module forces the network to validate every high-frequency textural alert generated by the CNN against the comprehensive global map maintained by the transformer, the model intrinsically learned to differentiate between benign physiological variations and genuine pathological consolidations, mirroring the nuanced deductive reasoning employed by experienced pulmonologists.

Furthermore, when the evaluation shifted to the intricately detailed multisequence Brain MRI dataset, the hybrid architecture capitalized heavily on its dual-stream processing to navigate the notoriously complex, multi-scale boundaries of cerebral gliomas and metastatic lesions. While pure transformer models often struggle to precisely delineate the jagged, infiltrative edges of a tumor due to their coarse patch-based tokenization, the lightweight convolutional backbone of our proposed model flawlessly preserved the high-resolution edge gradients, passing this critical morphological data to the fusion module where it was immediately synthesized with the transformer's spatial understanding of the surrounding gray and white matter structures. Consequently, the network successfully identified microscopic satellite lesions that the standalone baselines entirely overlooked, proving that the continuous mathematical dialogue between local and global features prevents the premature degradation of critical diagnostic signals regardless of the specific imaging modality being analyzed.

5.2. Quantitative Baseline Comparisons

To rigorously quantify the diagnostic supremacy of the proposed framework, the model was subjected to a direct, multi-modality empirical comparison against a highly optimized purely convolutional network (ResNet-50) and a standard global attention network (ViT-Base), with all architectures undergoing the exact

same aggressive data augmentation and optimization protocols to ensure a flawlessly equitable testing environment. The evaluation tracked Accuracy and the F1-Score across all three datasets to provide a holistic view of both raw predictive power and the model's resilience against the severe class imbalances that inherently bias medical datasets toward healthy control samples.

Table 3. Cross-Modality Diagnostic Performance Comparison of Standalone vs. Hybrid Architectures

Architecture Model	Chest X-Ray (Acc / F1)	Brain MRI (Acc / F1)	Histopathology (Acc / F1)	Average F1-Score
Standalone CNN (ResNet-50)	87.5% / 0.85	89.2% / 0.88	91.4% / 0.90	0.876
Standalone Transformer (ViT-Base)	88.3% / 0.86	90.1% / 0.89	88.6% / 0.87	0.873
Proposed Hybrid CNN-Transformer	**93.8% / 0.93**	**95.2% / 0.95**	**94.7% / 0.94**	**0.940**

As meticulously articulated in the cross-modality performance matrix presented on table 3 above, the proposed hybrid architecture overwhelmingly dominated both the ResNet-50 and ViT-Base baselines across every tested dataset, delivering an unprecedented average F1-Score of 0.940 that firmly establishes the clinical superiority of the cross-attention fusion strategy.

While the standalone CNN performed respectably on the heavily localized histopathology dataset but faltered on the structurally complex Chest X-rays, and the standalone transformer exhibited the exact opposite vulnerability by excelling at macro-structures while struggling with cellular-level granularity, the hybrid model maintained a near-perfect diagnostic equilibrium across all scales of medical imagery. This consistent outperformance unequivocally validates the architectural premise that forcing local convolutional features and global transformer embeddings to dynamically recalibrate one another prior to the classification head creates a significantly more robust and trustworthy diagnostic tool capable of adapting to vastly different biological presentations.

5.3. Computational Efficiency and Clinical Feasibility

Beyond the pursuit of raw diagnostic accuracy, a central objective of this study was to engineer an architecture that fundamentally respects the severe hardware limitations and restrictive computational budgets that characterize the vast majority of under-resourced hospitals and rural clinical settings. Despite integrating the computationally demanding multi-head self-attention mechanisms inherent to transformer models, the strategic deployment of a highly truncated, depthwise-separable convolutional backbone allows the proposed hybrid model to operate with a total parameter count of just 31.2 million a figure that is only marginally larger than the aging ResNet-50 baseline and overwhelmingly leaner than the massive 86.4 million parameters demanded by the ViT-Base architecture.

This aggressive optimization ensures that the entire diagnostic pipeline can be comfortably loaded into the memory of a single, consumer-grade GPU, enabling the system to process high-resolution medical scans with an inference latency of under 20 milliseconds per image without requiring massive cloud-computing infrastructure or expensive hardware overhauls. By successfully bridging the historical divide between the exceptional diagnostic precision of attention-based networks and the lean, pragmatic operational requirements of real-world medical facilities, the experimental results decisively prove that this hybrid CNN-Transformer design offers a highly accessible, computationally viable pathway toward democratizing advanced, AI-assisted healthcare diagnostics on a global scale.

6. CONCLUSION AND FUTURE WORK

This study set out to address the critical architectural dichotomy in automated medical image analysis, where the field has traditionally been forced to choose between the localized, parameter-efficient texture sensitivity of convolutional neural networks and the powerful, albeit computationally prohibitive, global spatial reasoning capabilities of vision transformers. To bridge this divide without violating the strict hardware constraints of real-world clinical environments, this paper proposed a novel hybrid CNN-Transformer architecture that eschews the simplistic, late-stage concatenation favored by prior models in favor of a deeply integrated

cross-attention fusion module. By splitting the visual processing into two parallel streams a lightweight convolutional backbone designed to extract high-frequency edge gradients and a transformer encoder engineered to construct comprehensive structural maps the architecture forces these fundamentally different representations to dynamically query and recalibrate one another, resulting in a synergistic diagnostic tool that closely mimics the multi-scale deductive reasoning of experienced medical professionals.

The empirical findings generated through rigorous multi-modality testing on Chest X-ray, Brain MRI, and histopathology datasets conclusively validate the diagnostic superiority of this integrated approach, as the proposed hybrid framework consistently outperformed both highly optimized standalone convolutional and pure transformer baselines across every critical statistical metric. Specifically, the cross-attention fusion mechanism allowed the network to achieve an unprecedented average F1-score of 0.940 by simultaneously suppressing localized noise and amplifying genuine pathological signals, thereby ensuring high sensitivity to microscopic abnormalities without triggering the catastrophic false-positive rates that typically plague single-scale models. Furthermore, the strategic deployment of depthwise-separable convolutions and tightly controlled feature dimensions successfully contained the overall parameter count to a remarkably lean 31.2 million, guaranteeing that this highly sophisticated interplay of local and global feature extraction can be executed with sub-20-millisecond inference times on the standard, consumer-grade hardware typically found in under-resourced medical facilities.

Despite these highly promising diagnostic achievements, this research acknowledges several structural limitations that must be addressed in subsequent studies, primarily the architecture's current restriction to processing static, two-dimensional images, which inherently limits its capacity to fully exploit the rich volumetric data embedded in complex 3D modalities like multi-slice CT scans or continuous MRI sequences. Additionally, while the model demonstrates exceptional spatial reasoning within a single static snapshot, it currently lacks the temporal mechanisms required to track the longitudinal progression of a pathology over time, meaning it cannot yet synthesize a patient's historical imaging data to predict disease trajectory or evaluate ongoing treatment efficacy. Consequently, future research efforts should be decisively directed toward expanding the cross-attention fusion framework into the temporal and volumetric domains, perhaps by integrating 3D convolutional backbones with video-based transformer encoders, while simultaneously exploring the inclusion of multi-modal clinical data such as electronic health records or genomic profiles to evolve this image-centric architecture into a comprehensive, holistic diagnostic ecosystem.

7. DECLARATIONS

7.1. About Authors

Amelia Evans (AE) -

Agung Rizky (AR) https://orcid.org/0009-0006-7046-8639

Marta Rodriguez (MR) -

Ihsan Nuril Hikam (IN) https://orcid.org/0009-0000-0345-9763

7.2. Author Contributions

Conceptualization: AE and AR; Methodology: AR and MR; Software: AR; Validation: AE and IN; Formal Analysis: AE and AR; Investigation: MR; Resources: AR and IN; Data Curation: MR; Writing – Original Draft Preparation: AE and MR; Writing – Review and Editing: AE, AR, MR, and IN; Visualization: MR; Supervision: IN; Project Administration: IN. All authors, AE, AR, MR, and IN, have read and agreed to the published version of the manuscript.

7.3. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.4. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.5. Declaration of Generative AI and AI-assisted Technologies in The Writing Process

During the preparation of this work the author(s) used ChatGPT in order to improve the writing of the paper. After using this ChatGPT, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

7.6. Declaration of Conflicting Interest

The authors declare that they have no conflicts of interest, known competing financial interests, or personal relationships that could have influenced the work reported in this paper.

REFERENCES

[1] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "Medvit: a robust vision transformer for generalized medical image classification," *Computers in biology and medicine*, vol. 157, p. 106791, 2023.

[2] S. Mavaddati, "A hybrid approach for brain tumor classification: Enhancing mri-based diagnosis with cnn-transformer synergy," *Journal of AI and Data Mining*, vol. 14, no. 1, pp. 37–49, 2026.

[3] J. W. Kim, A. U. Khan, and I. Banerjee, "Systematic review of hybrid vision transformer architectures for radiological image analysis," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 5, pp. 3248–3262, 2025.

[4] J. S. Zraqou, "A cnn–vision transformer fusion architecture for robust medical image recognition and clinical diagnostic understanding," in *Advances in Vision Foundation Models, Recognition, and Practical Systems*. IGI Global Scientific Publishing, 2027, pp. 265–296.

[5] N. NAGABHOOSHANAM, C. SURYANARAYANA, S. MURTHY, K. SUBRAHMANYAM, A. SINGH, K. NAGU, A. VANATHI, J. PHILIP, and A. RAJARAM, "Hybrid cnn-transformer framework for real-time medical image reconstruction and diagnostic support in healthcare system." *Journal of Environmental Protection & Ecology*, vol. 26, no. 6, p. 2405, 2025.

[6] Y. Memon and F. Zeng, "Hybrid-mednet: a hybrid cnn-transformer network with multi-dimensional feature fusion for medical image segmentation," *Physics in Medicine & Biology*, vol. 70, no. 19, p. 195016, 2025.

[7] V. Agarwal, M. Lohani, A. S. Bist, L. Rahardja, M. Hardini, and G. Mustika, "Deep cnn–real esrgan: An innovative framework for lung disease prediction," in *2022 IEEE Creative Communication and Innovative Technology (ICCIT)*. IEEE, 2022, pp. 1–6.

[8] V. Agarwal, M. Lohani, and A. S. Bist, "A novel deep learning technique for medical image analysis using improved optimizer," *Health Informatics Journal*, vol. 30, no. 2, p. 14604582241255584, 2024.

[9] N. J. Ali, "A deep hybrid cnn-transformer model for multi-modal tumour detection and segmentation across medical imaging applications," *Int. J. Intell. Eng. Syst*, vol. 19, p. 138, 2026.

[10] D. L. L. De Oliveira, T. A. A. Tosta, L. A. Neves, and M. Z. Do Nascimento, "Hybrid cnn-transformer models in histopathology image analysis: A scoping review," *IEEE Access*, vol. 13, pp. 212 887–212 919, 2025.

[11] R. K. Hudiono and S. Watini, "Remote medical applications of artificial intelligence," *International Transactions on Artificial Intelligence*, vol. 1, no. 2, pp. 182–187, 2023.

[12] A. Sharma, M. R. Mani, D. Chinthala, R. Thangjam, A. Bhilare, T. Singh, A. S. Bist *et al.*, "Computer vision-based medical image segmentation using hybrid cnn and transformer architectures," *International Journal of Artificial Intelligence and Machine Learning*, vol. 6, no. 2s, pp. 445–458, 2026.

[13] U. Rahardja, Q. Aini, A. S. Bist, S. Maulana, and S. Millah, "Examining the interplay of technology readiness and behavioural intentions in health detection safe entry station," *JDM (Jurnal Dinamika Manajemen)*, vol. 15, no. 1, pp. 125–143, 2024.

[14] X. Guo, X. Lin, X. Yang, L. Yu, K.-T. Cheng, and Z. Yan, "Uctnet: Uncertainty-guided cnn-transformer hybrid networks for medical image segmentation," *Pattern Recognition*, vol. 152, p. 110491, 2024.

[15] M. Al Emran, R. Mahbub, E. J. Barsha, R. Akter, B. U. Islam, and M. M. U. I. Arikh, "A stabilized hybrid cnn-transformer framework with channel attention for multi-class retinal disease classification," in *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*. IEEE, 2026, pp. 1–6.

[16] Q. Zhang, R. Li, P. Tang, J. Zhang, G. Yang, Z. Lin, M. Zhang, and M. Chen, "A hybrid cnn-transformer

deep learning model for differentiating benign and malignant breast tumors using multi-view ultrasound images," *Technology in Cancer Research & Treatment*, vol. 25, p. 15330338261447344, 2026.

[17] E. Altarawneh and E. Alshdaifat, "Adversarial robustness of cnn, transformer, and hybrid architectures in medical imaging ai," in *International Conference on Microelectronics, Electromagnetics and Telecommunication*. Springer, 2025, pp. 88–99.

[18] M. Annas, T. Handra, C. S. Bangun, U. Rahardja, and N. Septiani, "Reward and promotion: Sustainable value of post pandemic efforts in medical cold-supply chain," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 6, no. 1, pp. 109–118, 2024.

[19] D. Ashoka *et al.*, "Effivit: Hybrid cnn-transformer for retinal imaging," *Computers in Biology and Medicine*, vol. 191, p. 110164, 2025.

[20] O. N. Manzari, J. M. Kaleybar, H. Saadat, and S. Maleki, "Befunet: A hybrid cnn-transformer architecture for precise medical image segmentation," *arXiv preprint arXiv:2402.08793*, 2024.

[21] S. Singh, "Computer-aided diagnosis of thoracic diseases in chest x-rays using hybrid cnn-transformer architecture," *arXiv preprint arXiv:2404.11843*, 2024.

[22] R. S. Ali, R. K. Hassoun, A. Al-Adhami, S. A. Jabber, and S. H. Hashem, "Proposal medical image protection system based on hybrid cnn-transformer model and chaotic maps," *IEEE Access*, vol. 13, pp. 123 793–123 807, 2025.

[23] S. A. Thanekar, "Hybrid cnn–transformer architecture for multi-organ disease classification," *Natural Resources for Human Health*, vol. 6, no. 3, pp. 391–401, 2026.

[24] U. Rahardja, P. A. Sunarya, Q. Aini, S. Millah, and S. Maulana, "Technopreneurship in healthcare: Evaluating user satisfaction and trust in ai-driven safe entry stations," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 6, no. 3, pp. 404–417, 2024.

[25] S. Dlamini, E. Kurniyaningrum, K. Kustiyono, U. Rahardja, Q. Aini, and S. Mahfudz, "Machine learning-based big data analytics for personalized health insurance services," *Smart Human-centered Emerging Research in Machine Intelligence*, vol. 2, no. 1, pp. 26–36, 2026.

[26] H. Tang, Y. Chen, T. Wang, Y. Zhou, L. Zhao, Q. Gao, M. Du, T. Tan, X. Zhang, and T. Tong, "Htc-net: A hybrid cnn-transformer framework for medical image segmentation," *Biomedical Signal Processing and Control*, vol. 88, p. 105605, 2024.

[27] T. K. Andiani and O. Jayanagara, "Effect of workload, work stress, technical skills, self-efficacy, and social competence on medical personnel performance," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 5, no. 2, pp. 118–127, 2023.

[28] A. S. Anita, T. Kuusk, G. Nicola, M. Hardini, and U. Rahardja, "Advancements in artificial intelligence and their contributions to sustainable development goals: A multidisciplinary review," *Smart Human-centered Emerging Research in Machine Intelligence*, vol. 2, no. 1, pp. 37–47, 2026.

[29] H. Long, "Hybrid design of cnn and vision transformer: A review," in *Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence*, 2024, pp. 121–127.

[30] R. R. Sharma, A. Sungheetha, M. Tiwari, I. A. Pindoo, V. Ellappan, and G. Pradeep, "Comparative analysis of vision transformer and cnn architectures in medical image classification," in *International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 2024)*. Atlantis Press, 2025, pp. 1343–1355.

[31] I. Sembiring, B. K. Aji, and T. I. Bayu, "Consortium blockchain framework for secure digital medical record innovation," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 8, no. 1, pp. 138–151, 2026.

[32] C. Hu, N. Cao, H. Zhou, and B. Guo, "Medical image classification with a hybrid ssm model based on cnn and transformer," *Electronics*, vol. 13, no. 15, p. 3094, 2024.

[33] S. Juneja, B. S. Bhati, G. G. Tejani, and S. J. Mousavirad, "Vitcnn: a robust hybrid cnn–vision transformer based deep learning framework for multi-disease diagnosis in women's healthcare," *Frontiers in Oncology*, vol. 16, p. 1746864, 2026.

[34] A. Alwiyah and W. Setyowati, "A comprehensive survey of machine learning applications in medical image analysis for artificial vision," *International Transactions on Artificial Intelligence*, vol. 2, no. 1, pp. 90–98, 2023.

[35] T. Behera, C. Parida, B. Kishan, M. Kar, A. Dash, and S. Darshana, "Edge-optimized explainable ai for medical image diagnosis using hybrid cnn-transformer architecture," in *2025 5th IEEE International Conference on Applied Electromagnetics, Signal Processing, & Communication (AESPC)*. IEEE, 2025,

pp. 1–6.

[36] A. H. Arribathi, S. Maulana, F. I. Ningrum, and M. Hassan, "Utilization of big data analytics in strategic decision making in the education sector," *Smart Human-centered Emerging Research in Machine Intelligence*, vol. 2, no. 2, pp. 60–70, 2026.

[37] K. N. Erin, "A hybrid cnn-transformer approach for breast cancer detection and segmentation from ultrasound images," *Accessed: Dec*, vol. 12, 2025.

[38] U. Rahardja, Q. Aini, D. Manongga, I. Sembiring, and I. D. Girinzio, "Implementation of tensor flow in air quality monitoring based on artificial intelligence," *International Journal of Artificial Intelligence Research*, vol. 6, no. 1, 2023.

[39] Z. Yu, F. Lee, and Q. Chen, "Hct-net: hybrid cnn-transformer model based on a neural architecture search network for medical image segmentation: Z. yu et al." *Applied Intelligence*, vol. 53, no. 17, pp. 19 990–20 006, 2023.

[40] S. Iqbal, T. M. Khan, S. S. Naqvi, A. Naveed, and E. Meijering, "Tbconvl-net: A hybrid deep learning architecture for robust medical image segmentation," *Pattern Recognition*, vol. 158, p. 111028, 2025.