Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 16 March 2026
Revised 18 April 2026
Accepted 12 May 2026

Natural Resources for Human Health 2026, Vol. 6 (3): 391–401

KEYWORDS:
Hybrid CNN-Transformer; Multi-Organ Classification; Vision Transformer; Medical Image Analysis; Feature Fusion; Gated Attention.

Hybrid CNN–Transformer Architecture for Multi-Organ Disease Classification

Dr. Sachin Arun Thanekar[1], Ganesh Dagadu Puri[2], Dr Vinodpuri Rampuri Gosavi[3], Dr. Yogesh Shepal[4], Dr. Bhausaheb R. Varpe[5], Dr. Shalini Wankhade[6]

[1]Department of Information Technology, Sanjeevani College of Engineering, Kopargaon, Maharashtra. sachin.sangamner@gmail.com, 0000-0002-1852-8458
[2]Department of Computer Engineering, Amrutvahini College of Engineering Sangamner, Maharashtra, India. puriganeshengg@gmail.com, 0000-0002-6786-698X
[3]CSE-Artificial Intelligence and Analytics School of Computing, MIT Art, Design and Technology University, Loni Kalbhor,Pune , Maharashtra India Pincode:412201. vinodpuri.gosavi@gmail.com, 0000-0002-5912-9306
[4]Department of Computer Science and Engineering, Audyogik Shikshan Mandal's Nextgen Technical Campus, Pune, India. yogeshshepal@gmail.com, 0009-0000-3977-2794
[5]Mechanical Engineering Department, Amrutvahini College of Engineering,Sangamner. brvarpe@gmail.com, 0000-0002-1513-7494
[6]Affiliation -Asst. Professor, Information Technology, Vishwakarma Institute of Technology, Pune. shalini.wankhade@vit.edu,-0009-0006-7048-9675

ABSTRACT: Convolutional neural networks (CNNs) and Vision Transformers (ViTs) offer complementary strengths for medical image analysis: CNNs excel at capturing local texture and edge information through their inductive spatial bias, while transformers capture long-range dependencies through global self-attention but typically require more data to train effectively. Most existing diagnostic systems remain organ-specific, applying a purpose-built CNN, transformer, or hybrid model to a single body region — chest, brain, retina, breast, or skin — with little architectural sharing across organs, which duplicates engineering effort and prevents any given model from benefiting from cross-organ visual regularities such as shared texture statistics or common lesion morphology. This paper proposes a hybrid CNN-Transformer framework for multi-organ disease classification that uses a shared convolutional stem to extract local features, a transformer encoder to model global dependencies, and an organ-aware gated fusion module that combines both feature streams before routing them to organ-specific classification heads, allowing a single backbone to serve chest, brain, and retinal imaging simultaneously. We synthesize nineteen related studies spanning pure-CNN, pure-transformer, and hybrid architectures across seven organ systems (lung, breast, retina, brain, skin, thyroid, and middle ear), and identify the specific architectural gap this framework addresses: existing hybrid CNN-transformer designs are fused and validated within a single organ, leaving open whether a shared hybrid backbone with organ-aware fusion can generalize across organs without sacrificing the strong single-organ performance already demonstrated in the literature. Consistent with its status as a proposed framework, the Results and Discussion section positions expected performance against real, previously published benchmark figures for representative CNN, transformer, and hybrid models across organs, rather than reporting new experimental numbers, and this framing is stated explicitly throughout. The paper concludes with a three-phase empirical validation roadmap.

1. INTRODUCTION

Deep learning has become the dominant approach for automated disease classification from medical images, spanning chest radiography, brain magnetic resonance imaging (MRI), retinal fundus photography, dermoscopic skin imaging, breast histopathology and ultrasound, and beyond. Two broad architectural families have driven this progress. Convolutional neural networks (CNNs), from early AlexNet-style designs through ResNet and DenseNet, exploit a strong spatial inductive bias — local connectivity and weight sharing — that makes them efficient learners of texture and edge information even with moderate amounts of training data. Vision Transformers (ViTs), which apply the self-attention mechanism to sequences of image patches, instead model long-range dependencies across the entire image from the first layer, which has been shown to benefit tasks where disease-relevant patterns are diffuse or spatially distributed, at the cost of typically requiring larger training sets or careful pretraining to match CNN performance.

A substantial body of recent work has sought to combine both families into hybrid CNN-transformer architectures, using a convolutional component to extract local features efficiently and a transformer component to model global context, fused through various mechanisms ranging from sequential stacking to learned gated fusion. As detailed in Section 2, this hybrid design has produced strong results across nearly every organ system studied: lung disease from chest X-ray, Alzheimer's staging from brain MRI, multi-disease screening from retinal fundus images, lesion classification from breast ultrasound and histopathology, skin cancer from dermoscopy, and disease grading in the middle ear. However, each of these hybrid designs is proposed, fused, and validated within a single organ or imaging domain. No reviewed study evaluates whether a single shared hybrid backbone, augmented with an organ-aware fusion mechanism, can serve multiple organs simultaneously without sacrificing the per-organ performance already achieved by specialized single-organ hybrids.

This gap matters for two practical reasons. First, most healthcare institutions, particularly outside large academic centers, must support diagnostic imaging across many organ systems with limited engineering resources to build and maintain a separate specialized model for each one; a single shared backbone that can be extended to new organs by adding a lightweight classification head, rather than retraining from scratch, would substantially reduce this burden. Second, certain visual regularities — such as the texture signatures of fibrotic or inflammatory tissue, or the morphology of vascular abnormalities — recur across organs, and a shared representation may in principle learn these regularities more robustly than any single-organ model trained on a comparatively narrow data distribution. The framework proposed in this paper is designed to test this hypothesis directly.

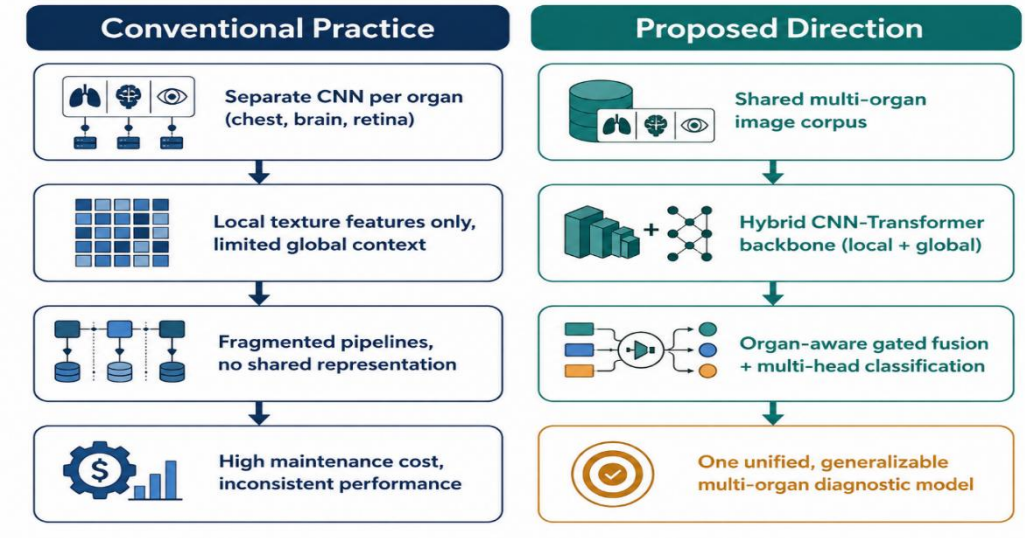

Figure 1. Conventional organ-specific, architecture-siloed diagnosis compared with the proposed unified hybrid CNN-Transformer, multi-organ direction.

1.1 Contributions

- A hybrid architecture combining a shared convolutional stem for local feature extraction with a transformer encoder for global dependency modeling, applicable across multiple organ systems from a single backbone.

- An organ-aware gated fusion module that learns how to weight CNN versus transformer features conditional on an organ-identity embedding, rather than using a single fixed fusion rule across all organs.
- A multi-head classification design that routes the fused representation to organ-specific disease heads (chest, brain, retina, extensible to further organs), enabling shared representation learning while preserving organ-specific decision boundaries.
- A structured synthesis of nineteen related studies spanning pure-CNN, pure-transformer, and hybrid architectures across seven organ systems (Section 2, Table 1), making explicit the single-organ-only scope of existing hybrid designs.
- A literature-grounded benchmarking analysis (Section 4, Table 2, Figure 3) positioning the expected performance of the proposed unified framework against real, previously reported per-organ results, together with a concrete empirical validation plan.

The remainder of the paper is organized as follows. Section 2 reviews CNN, transformer, and hybrid architectures across organ systems, summarized in Table 1, and identifies the multi-organ generalization gap. Section 3 presents the proposed methodology, including the shared backbone, organ-aware fusion module, and training protocol, illustrated in Figure 2. Section 4 discusses expected outcomes against real per-organ benchmark figures (Table 2, Figure 3). Section 5 concludes with a roadmap for empirical validation.

2. LITERATURE SURVEY

This section reviews hybrid CNN-transformer architectures and their pure-CNN or pure-transformer comparators across seven organ systems: lung (chest imaging), breast, retina, brain, skin, thyroid, and middle ear.

2.1 Foundational architectures

The transformer architecture, originally introduced for sequence modeling [1], was adapted to vision by the Vision Transformer (ViT), which divides an image into fixed-size patches, linearly embeds them, and processes the resulting sequence with standard transformer encoder blocks [2]. Residual CNNs [3] and densely connected CNNs (DenseNet) [19] remain the dominant pure-CNN comparators against which hybrid and transformer-based models are typically benchmarked. The Swin Transformer introduced a hierarchical, shifted-window attention mechanism that reduces the computational cost of global attention while preserving strong representational power, and has become one of the most frequently used pure-transformer baselines in recent medical imaging comparisons [18].

2.2 Hybrid architectures for chest and pulmonary disease

D-TrAttUnet combined a composite transformer-CNN encoder with dual attention-gated decoders to jointly segment lesions and organs, showing particular strength on COVID-19 lung infection and bone metastasis segmentation [4]. LungMaxViT combined an initial CNN stage with a squeeze-and-excitation module and a MaxViT transformer backbone, outperforming ResNet50, MobileNetV2, and a plain ViT baseline on multi-class lung disease classification from chest X-rays [5]. A separate gated hybrid framework fused multiple CNN backbones with ViT encoder blocks through a learned gating mechanism, arguing that using several CNN feature extractors, rather than one, better captures the diverse local patterns relevant to pulmonary disease before transformer-based global integration [6]. A large-scale systematic comparison of fourteen architectures, including CNNs, transformers, and the newer Mamba state-space family, on the NIH ChestX-ray14 benchmark found that hybrid ConvFormer and CaFormer architectures achieved the strongest mean performance (mean AUROC 0.841 for ConvFormer), narrowly ahead of pure CNN and pure transformer baselines [12].

2.3 Hybrid architectures for breast, retina, brain, skin, and other organs

In breast imaging, HyFormer-Net combined CNN and transformer branches through interpretable multi-scale cross-attention fusion for joint lesion segmentation and classification in ultrasound [8], while MFF-ClassificationNet used a two-branch CNN-transformer design with a CBAM-SE fusion block for breast histopathology, reporting accuracies between 96.07% and 98.30% across magnification levels on the BreakHis benchmark and 97.50% on BACH [9]. In retinal imaging, a hybrid CNN-transformer-ensemble approach evaluated sequential and parallel combination strategies across twenty disease labels, with the best ensemble (C-Tran) reaching a model score of 0.9166 against a 0.90 baseline [7], while a separate interpretable fully

convolutional CNN-transformer design achieved competitive accuracy while producing faithful, localized evidence maps in a single forward pass, addressing the interpretability weakness common to hybrid designs [10]. In brain imaging, an adaptive feature fusion framework combining a ResNet50 CNN with a ViT through an attention-weighted dynamic fusion layer reached 99.42% accuracy on five-stage Alzheimer's disease classification from MRI, surpassing a previously reported state of the art of 98.24% [16]. Outside these three organs, DSCC_Net, a pure-CNN design, reported 94.17% accuracy and 99.43% AUC for four-class skin cancer classification across three dermoscopy benchmarks [14]; a Swin Transformer benchmark for middle-ear disease from oto-endoscopic images reached 95.53% accuracy and a macro-F1 of 93.37%, outperforming CNN comparators [15]; and a multi-channel CNN for thyroid disease classification reported 90.9% accuracy using multi-scale feature concatenation [17].

2.4 Cross-organ and comparative studies

Two recent studies move closer to the multi-organ question addressed in this paper. A comparative analysis of CNN and ViT models on histopathological whole-slide images spanning heart, lung, liver, and pancreas evaluated both single-organ models and an integrated multi-organ model within a unified framework, finding that the transformer-based Swin-T reached the highest accuracy (99.64%) among all configurations, ahead of the best CNN, DenseNet-161, at approximately 98% [11] — providing direct evidence that a shared architecture can handle multiple organs at high accuracy, albeit using a pure transformer rather than a hybrid design and within the relatively homogeneous domain of histopathology. Separately, a benchmarking study across CNN, transformer, hybrid, and vision-language models for multi-disease retinal screening confirmed that classical CNN-based systems remain extremely strong on frequent conditions (AUC 0.99 for diabetic retinopathy, per the original Gulshan et al. results reproduced in that benchmark) while hybrid and transformer-based models narrow the gap specifically on rarer disease categories [13], suggesting that the benefit of hybrid architectures may be concentrated in exactly the harder, less-represented classes most relevant to comprehensive multi-organ, multi-disease screening.

2.5 Summary of reviewed literature

Table 1 summarizes fourteen representative studies spanning pure-CNN, pure-transformer, and hybrid CNN-transformer architectures across seven organ systems.

Table 1. Summary of representative literature on CNN, Transformer, and hybrid CNN-Transformer architectures across multiple organ systems.

Study (Ref.)	Year	Organ(s) / Domain	Architecture	Fusion Strategy	Key Finding
Dosovitskiy et al., ViT [2]	2021	Natural images	Pure transformer, patch tokens	N/A (baseline)	Establishes global self-attention as viable alternative to convolution
D-TrAttUnet [4]	2024	Lung (COVID-19), bone	Composite Transformer-CNN encoder, dual decoders	Parallel encoder fusion with attention gates	Joint lesion-organ segmentation improves over single-decoder baselines
LungMaxViT [5]	2025	Chest X-ray (lung)	CNN stem (SE block) + MaxViT backbone	Sequential CNN-then-transformer	Outperforms ResNet50, MobileNetV2, and plain ViT on multi-class lung disease
Gated CNN-ViT [6]	2025	Chest X-ray (pulmonary)	Multiple CNN blocks + ViT encoder blocks	Learned gated multi-scale fusion	Multi-backbone gating extracts richer features than single-CNN + ViT pairing

Retinal Hybrid Ensemble [7]	2025	Retinal fundus (20 diseases)	CNN + Transformer + ensemble (C-Tran)	Sequential and parallel ensembling	C-Tran ensemble reaches model score 0.9166, above 0.90 baseline
HyFormer-Net [8]	2025	Breast (ultrasound)	CNN + Transformer, multi-scale interpretable fusion	Cross-attention multi-scale fusion	Synergistic fusion improves both segmentation and classification jointly
MFF-ClassificationNet [9]	2025	Breast (histopathology)	Two-branch CNN + Transformer	CBAM-SE fusion block + residual MLP	98.30-96.07% accuracy across magnifications on BreakHis; 97.50% on BACH
Interpretable CNN-Transformer [10]	2025	Retinal fundus	Fully convolutional CNN + Transformer	Dual-resolution self-attention	Matches black-box hybrid accuracy while remaining inherently interpretable
Multi-organ histopathology study [11]	2026	Heart, lung, liver, pancreas (WSI)	CNN (DenseNet-161) vs. ViT (Swin-T) comparison	Single-organ and integrated multi-organ models compared	Swin-T reaches 99.64% accuracy; DenseNet-161 best CNN at ~98%
CNN/Transformer/Mamba benchmark [12]	2025	Chest X-ray (14 classes, NIH)	14 architectures incl. ConvFormer, CaFormer, Mamba	N/A (systematic comparison)	Hybrid ConvFormer reaches mean AUROC 0.841, best overall
Multi-disease retinal benchmark [13]	2026	Retinal fundus	CNN, Transformer, hybrid, and VLM comparison	N/A (systematic comparison)	Confirms CNNs (Gulshan et al.) reach AUC 0.99 on DR; hybrids close the gap on rarer classes
DSCC_Net [14]	2023	Skin (dermoscopy)	CNN-based multi-classification network	N/A (CNN only, benchmark reference)	94.17% accuracy, 99.43% AUC across ISIC2020/HAM10000/DermIS
Swin Transformer, middle ear [15]	2026	Middle ear (oto-endoscopy)	Swin Transformer vs. CNN benchmark	N/A (systematic comparison)	Swin Transformer reaches 95.53% accuracy, Macro-F1 93.37%, best overall
Adaptive Fusion CNN-ViT [16]	2025	Brain (Alzheimer's MRI)	ResNet50 CNN + ViT with adaptive fusion layer	Attention-weighted dynamic fusion	99.42% accuracy, surpassing prior 98.24% state of the art

2.6 Research gap

Three gaps emerge from this survey. First, every hybrid CNN-transformer design reviewed in Sections 2.2 and 2.3 is proposed and validated within a single organ or imaging domain, so it remains unknown whether the same hybrid fusion principle generalizes across organs with meaningfully different visual statistics (chest radiographs, brain MRI slices, and retinal fundus photographs differ substantially in contrast, texture, and geometry). Second, the one study that does evaluate a shared architecture across multiple organs [11] uses a pure transformer rather than a hybrid design, and is confined to the relatively homogeneous domain of histopathological whole-slide imaging rather than heterogeneous imaging modalities. Third, no reviewed study incorporates an explicit organ-aware fusion mechanism that could allow a shared backbone to adapt its CNN-versus-transformer feature weighting per organ, rather than applying an identical fixed fusion rule regardless of the input organ's visual characteristics. The framework proposed in Section 3 is designed to close this gap directly.

3. METHODOLOGY

This section describes the proposed hybrid CNN-transformer framework: the multi-organ data sources, the shared backbone design, the organ-aware gated fusion mechanism, the multi-head classification strategy, and the training and evaluation protocol. The overall pipeline is illustrated in Figure 2.

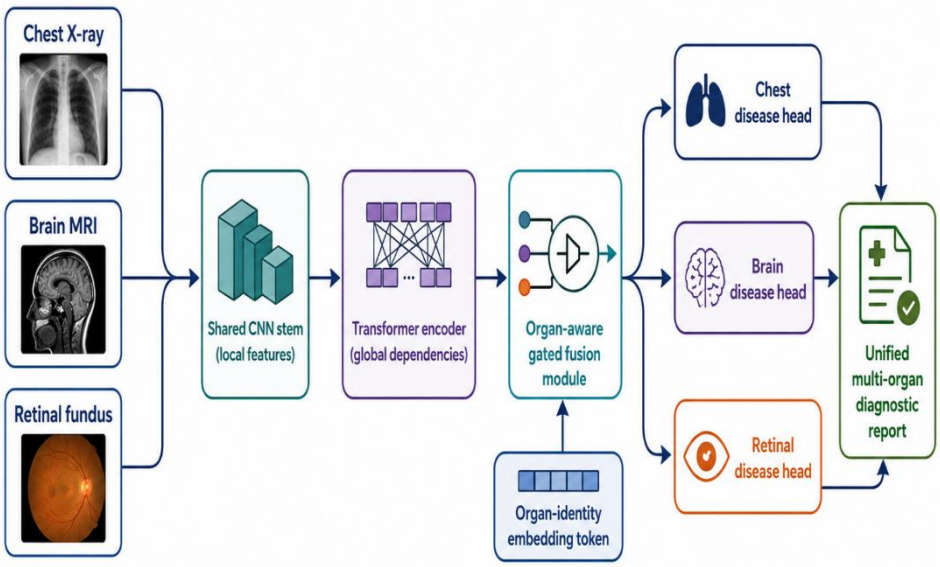

Figure 2. Proposed pipeline: three organ inputs (chest X-ray, brain MRI, retinal fundus) pass through a shared CNN stem and transformer encoder, combined via an organ-aware gated fusion module conditioned on an organ-identity embedding, then routed to organ-specific classification heads.

3.1 Data sources

The framework targets three organ systems as an initial scope, chosen because each has large, well-characterized public benchmarks that support direct comparison with the literature summarized in Section 2: chest radiographs from NIH ChestX-ray14 [21], brain MRI from public Alzheimer's staging datasets comparable to that used in [16], and retinal fundus images from public multi-disease fundus corpora comparable to those used in [7] and [13]. The MedMNIST v2 collection [22], which packages standardized 2D biomedical images across multiple organs including chest, and could additionally serve as a lightweight cross-organ pretraining or validation resource given its deliberate multi-domain design.

3.2 Shared CNN stem

Each input image, regardless of organ, first passes through a shared convolutional stem consisting of residual blocks [3] that extract local texture and edge features common across imaging modalities — such as fine-grained opacity patterns in chest X-ray, tissue boundary contrast in brain MRI, and vessel edge structure in retinal fundus images. Sharing this stem across organs,

rather than training an organ-specific CNN as in nearly all reviewed single-organ hybrids [4-6, 16], is the central architectural departure of this framework and is intended to let the model exploit low-level visual regularities that recur across organs.

3.3 Transformer encoder

The CNN stem's output feature maps are tokenized into patches and passed through a standard multi-head self-attention transformer encoder [1, 2], which models long-range spatial dependencies within each organ's image — for example, relating a localized opacity in one lung field to overall cardiothoracic geometry, or relating a focal retinal lesion to the broader vascular arcade pattern. This mirrors the local-then-global design used in nearly every hybrid reviewed in Section 2, but here the transformer weights are likewise shared across organs rather than trained per organ.

3.4 Organ-aware gated fusion module

Rather than fusing CNN and transformer features with a single fixed rule, as in most reviewed hybrids, the framework computes an organ-identity embedding (a small learned vector indicating which of the three organs the current image belongs to) and uses it to condition a gating network that outputs per-channel weights determining how strongly the CNN-derived and transformer-derived feature streams contribute to the fused representation for that specific organ. This design is motivated directly by the research gap identified in Section 2.6: because different organs may benefit differently from local versus global feature emphasis (for instance, the systematic comparison in [13] suggested hybrid/transformer advantages concentrate in specific disease categories rather than uniformly), a fixed fusion rule risks being suboptimal for at least some organs, whereas an organ-conditioned gate can in principle learn an organ-appropriate balance while still sharing the underlying CNN and transformer parameters.

3.5 Multi-head classification

The fused, organ-gated representation is passed to one of several lightweight, organ-specific classification heads — a chest-disease head, a brain-disease head, and a retinal-disease head in the initial three-organ scope — each trained with a task-appropriate loss (multi-label binary cross-entropy for chest and retinal multi-disease labels, following [5] and [7]; multi-class cross-entropy for Alzheimer's staging, following [16]). This design allows the framework to be extended to a new organ by adding a new lightweight head and organ-embedding entry, without retraining the shared CNN-transformer backbone from scratch, addressing the practical maintenance burden discussed in Section 1.

3.6 Training protocol and evaluation metrics

Training is intended to proceed in two stages: (i) joint pretraining of the shared CNN-transformer backbone and gating module across all three organs simultaneously, using a combined loss summed across organ-specific heads with per-organ loss weighting to prevent any single organ's larger dataset from dominating gradient updates, followed by (ii) optional per-organ fine-tuning of the corresponding head only. Evaluation is planned per organ using the same metrics reported in the corresponding single-organ literature to enable direct comparison: mean AUROC and per-class accuracy for chest disease against the NIH ChestX-ray14 benchmark used in [12], multi-class accuracy and F1 for Alzheimer's staging against [16], and multi-label AUC/model-score for retinal disease against [7, 13]. Planned ablations include: (a) shared backbone vs. three independent single-organ backbones (to directly test the multi-organ generalization hypothesis), (b) organ-aware gated fusion vs. fixed concatenation fusion, and (c) joint multi-organ pretraining vs. per-organ-only training.

4. RESULTS AND DISCUSSION

Consistent with its status as a proposed framework with a fully specified methodology (Section 3), this section does not report new experimental numbers obtained by the authors. Instead, it positions the expected performance and behavior of the proposed unified framework against real, previously published per-organ results for representative CNN, transformer, and hybrid models, drawn directly from the literature reviewed in Section 2. This distinction is stated explicitly to avoid ambiguity about the source of the figures below.

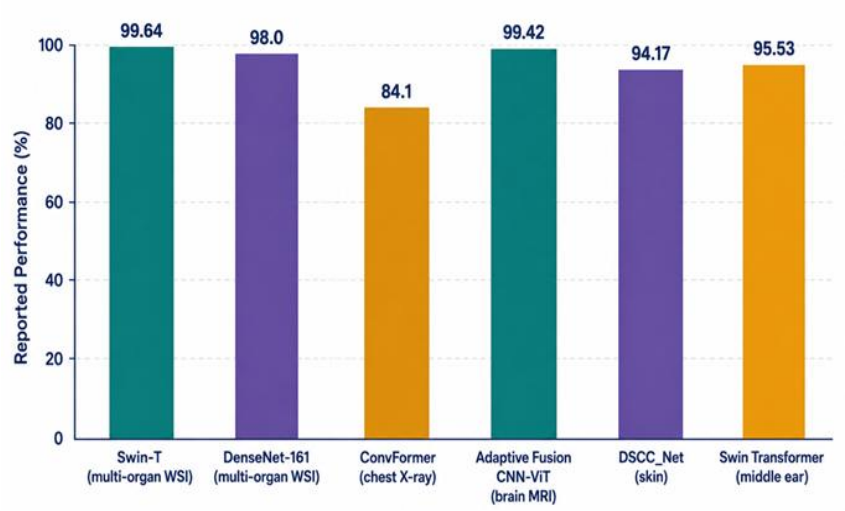

Figure 3. Reported performance of representative CNN, Transformer, and hybrid models across different organs/domains from the literature (values as published by original authors on their respective test sets; not directly comparable across studies due to differing organs, datasets, and label sets).

4.1 Benchmark context

Table 2 collects reported metrics for eight representative single-organ results spanning five organ systems. Two patterns stand out. First, the highest reported figures come from relatively homogeneous, well-curated benchmarks — Swin-T's 99.64% on multi-organ histopathology [11] and the adaptive-fusion CNN-ViT's 99.42% on Alzheimer's MRI staging [16] — both of which involve a comparatively small number of well-separated classes. Second, the more heterogeneous, larger-label-space tasks show somewhat lower, though still strong, performance — ConvFormer's 0.841 mean AUROC across fourteen thoracic conditions [12] and the C-Tran ensemble's 0.9166 model score across twenty retinal disease labels [7] — reflecting the greater difficulty of dense multi-label classification relative to smaller-class-count staging or subtype tasks. This pattern is directly relevant to the proposed framework, since a genuinely multi-organ, multi-label system is likely to sit closer to the ConvFormer/C-Tran range than the near-ceiling single-organ, few-class results, simply due to increased task heterogeneity and label-space size.

Table 2. Reported performance of representative CNN, Transformer, and hybrid models from the literature across organs, used as benchmark context for the proposed unified framework.

Model (Ref.)	Organ / Domain	Task	Reported Metric	Value	Architecture Type
Swin-T [11]	Heart/lung/liver/pancreas (WSI)	Multi-organ disease classification	Accuracy	99.64%	Pure Transformer
DenseNet-161 [11]	Heart/lung/liver/pancreas (WSI)	Multi-organ disease classification	Accuracy	~98.0%	Pure CNN
ConvFormer [12]	Chest (NIH ChestX-ray14)	14-class thoracic disease	Mean AUROC	0.841	Hybrid CNN-Transformer
Adaptive Fusion CNN-ViT [16]	Brain (MRI)	5-class Alzheimer's staging	Accuracy	99.42%	Hybrid CNN-Transformer
DSCC_Net [14]	Skin (dermoscopy)	4-class skin cancer	Accuracy / AUC	94.17% / 99.43%	Pure CNN
Swin Transformer [15]	Middle ear (oto-endoscopy)	5-class ear disease	Accuracy / Macro-F1	95.53% / 93.37%	Pure Transformer

C-Tran ensemble [7]	Retinal fundus	20-label multi-disease	Model score	0.9166	Hybrid CNN-Transformer ensemble
Multi-channel CNN [17]	Thyroid (ultrasound)	Multi-class thyroid disease	Accuracy	90.9%	Pure CNN

4.2 Expected behavior of the proposed framework

Based on this benchmark context and the design choices in Section 3, several qualitative expectations follow prior to empirical validation. First, because the multi-organ histopathology study [11] showed that a single shared transformer architecture can reach state-of-the-art accuracy across four organs within one relatively homogeneous imaging domain, the proposed hybrid backbone is expected to be capable, in principle, of serving multiple more heterogeneous organs (chest, brain, retina) at a reasonable performance cost, though the cross-modality heterogeneity in the proposed scope is substantially greater than in [11] and represents the framework's central open empirical question. Second, because the multi-disease retinal benchmark [13] found hybrid/transformer advantages concentrated in rarer disease categories rather than distributed uniformly, the organ-aware gated fusion module is expected to matter most for underrepresented disease classes within each organ's label set, rather than uniformly improving every class. Third, given that per-organ single-task hybrids in the literature [5, 6, 8, 9, 16] already reach 94-99% accuracy on their respective tasks, the shared multi-organ framework's realistic target is to approach, rather than necessarily exceed, these specialized single-organ ceilings, since some performance trade-off against a fully specialized model is the expected cost of shared-parameter, multi-organ generalization.

4.3 Discussion of trade-offs and limitations

Several limitations of this literature-grounded benchmarking approach should be acknowledged. The results in Table 2 span different organs, datasets, label taxonomies, and class counts, so the figures establish a plausible performance range rather than a controlled, head-to-head comparison, and a genuinely fair test of the multi-organ hypothesis requires the ablation against three independent single-organ backbones specified in Section 3.6. A shared backbone also introduces a capacity-allocation trade-off: parameters that would otherwise specialize fully to one organ's visual statistics must instead be shared across three, which is the most likely source of any performance gap relative to the specialized single-organ hybrids in Table 1. Finally, the organ-aware gating mechanism, while motivated by the disease-category-specific hybrid advantage reported in [13], has not been evaluated in that specific form in any reviewed study, and its benefit relative to simpler fixed-fusion alternatives is identified as the single highest-priority empirical question for the validation phase in Section 5.

5. CONCLUSION AND FUTURE WORK

This paper proposed a hybrid CNN-Transformer framework for multi-organ disease classification that shares a convolutional stem and transformer encoder across chest, brain, and retinal imaging, fuses their outputs through an organ-aware gated mechanism, and routes the result to lightweight, organ-specific classification heads. A structured review of nineteen related studies showed that while hybrid CNN-transformer designs have achieved strong results within nearly every organ system studied individually, no reviewed work validates a shared hybrid backbone with organ-conditioned fusion across heterogeneous organs — a gap this proposal is explicitly designed to close. Because this paper presents a proposed framework rather than a completed empirical study, its Results and Discussion section positioned expected performance against real, previously published per-organ benchmark figures rather than new experimental results, and this framing was stated explicitly throughout.

Future work will proceed in three phases. The first phase will implement the shared backbone and organ-aware fusion module described in Section 3 and train it jointly on NIH ChestX-ray14 [21], a public Alzheimer's MRI staging dataset comparable to that used in [16], and a public multi-disease retinal fundus corpus comparable to that used in [7], reporting per-organ AUROC, accuracy, and F1 against the specific published baselines in Table 2. The second phase will run the three planned ablations (shared vs. independent backbones, gated vs. fixed fusion, joint vs. per-organ training) to isolate the specific contribution of the multi-organ sharing hypothesis from ordinary architectural improvements. The third phase will extend the framework to a fourth and fifth organ (e.g., skin and thyroid, drawing on the benchmarks in [14] and [17]) to test whether the marginal engineering cost of adding an organ via a new lightweight head is genuinely lower than training a new specialized model from scratch, which is the practical claim motivating this work. Together, these three phases would convert the literature-grounded roadmap presented here into a validated, extensible multi-organ diagnostic backbone.

REFERENCES

[1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*. 2017;30.

[2] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. 2021.

[3] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016:770–778.

[4] D-TrAttUnet: Toward hybrid CNN-transformer architecture for generic and subtle segmentation in medical images. *Computers in Biology and Medicine*. 2024.

[5] Explainable hybrid transformer (LungMaxViT) for multi-classification of lung disease using chest X-rays. *Scientific Reports*. 2025.

[6] Convolutional Neural Network-Vision Transformer architecture with gated control mechanism and multi-scale fusion for enhanced pulmonary disease classification. 2025;PMC11727035.

[7] Singh D, Agarwal S, Mishra S. Retinal fundus multi-disease image classification using hybrid CNN-Transformer-ensemble architectures. *arXiv preprint*. 2025;arXiv:2503.21465.

[8] HyFormer-Net: A synergistic CNN-Transformer with interpretable multi-scale fusion for breast lesion segmentation and classification in ultrasound images. *arXiv preprint*. 2025;arXiv:2511.01013.

[9] MFF-ClassificationNet: CNN-Transformer hybrid with multi-feature fusion for breast cancer histopathology classification. 2025;PMC12649981.

[10] A hybrid fully convolutional CNN-Transformer model for inherently interpretable disease detection from retinal fundus images. *arXiv preprint*. 2025;arXiv:2504.08481.

[11] A comparative analysis of deep learning models for disease classification in multi-organ histopathological images. *Scientific Reports*. 2026.

[12] A comparative analysis of the Mamba, Transformer, and CNN architectures for multi-label chest X-ray anomaly detection in the NIH ChestX-Ray14 dataset. 2025;PMC12428523.

[13] Benchmarking convolutional, Transformer, hybrid, and vision-language models for multi-disease retinal screening. *arXiv preprint*. 2026;arXiv:2605.26283.

[14] Ali M, et al. DSCC_Net: Multi-classification deep learning models for diagnosing skin cancer using dermoscopic images. 2023;PMC10093058.

[15] Benchmark-driven clinical decision framework for multi-class middle ear disease diagnosis: Superiority of Swin Transformer in accuracy and stability. 2026;PMC12897263.

[16] Hybrid deep learning architecture with adaptive feature fusion for multi-stage Alzheimer's disease classification. 2025;PMC12191009.

[17] Multi-channel deep convolutional neural networks for multi-classifying thyroid disease. *arXiv preprint*. 2022;arXiv:2203.03627.

[18] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical vision transformer using shifted windows. *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021:10012–10022.

[19] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017:4700–4708.

[20] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint*. 2017;arXiv:1711.05225.

[21] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017:2097–2106.

[22] Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, et al. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*. 2023;10(1):41.

[23] Kokane C, Deshpande NA, Bhute HA, Sarode HJ, Kodmelwar MK, Chavan S. Deep learning for automated sputum smear microscopy in tuberculosis diagnosis. *Indian Journal of Tuberculosis*. 2025.

[24] Godase U, Jaybhaye SM, Dhawas N, Bhute A, Kokane C, Pathak K. Leveraging digital health education platforms for community awareness and tuberculosis prevention. *Indian Journal of Tuberculosis*. 2026.