JOURNAL
India Publication

Journal of Computer Science and Intelligent Systems

Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

Medfusion-Net: A Dual-Path CNN-Transformer Model with Multi-Scale Attention for Medical Image Analysis

*H. Srimathi**

ABSTRACT

Purpose

Medical image analysis increasingly relies on deep learning for classification and decision-support tasks, but models must balance local texture recognition, global contextual reasoning, computational efficiency, and interpretability. This paper proposes MedFusion-Net, a new dual-path deep learning architecture designed for medical image classification. The model is intended to combine convolutional feature extraction with transformer-based global context modelling and multi-scale attention so that clinically relevant local and global patterns can be represented jointly.

Design/Methodology/Approach

The proposed framework uses a CNN branch for hierarchical local feature extraction and a lightweight transformer branch for long-range dependencies. Multi-scale channel-spatial attention is introduced before feature fusion, followed by a gated fusion module and classification head. The proposed model is designed for reproducible evaluation using publicly available medical-image benchmarks such as MedMNIST v2, with patient-level or officially supplied train-validation-test splits preserved where applicable. Evaluation is specified using accuracy, precision, recall, F1-score, ROC-AUC, parameter count, inference cost, and calibration/error analysis. Baselines include conventional CNNs and representative CNN-transformer architectures.

Findings

This manuscript presents the architecture, training protocol, evaluation design, and analytical framework for validating MedFusion-Net. No numerical experimental results are fabricated in this manuscript. The proposed contribution is therefore a model and a reproducible validation protocol; numerical claims should be inserted only after the experiments have been executed on the selected dataset(s). The design is motivated by evidence that transformer-based approaches can capture long-range relationships that are difficult for purely local convolutional operators, while CNNs remain effective for fine-grained spatial patterns.

Practical Implications

The framework can support researchers developing lightweight medical-image decision-support systems and can serve as a baseline for investigating the trade-off between accuracy, computational cost, and explainability. The proposed attention and visualization components can also be used to inspect whether predictions are driven by clinically meaningful image regions.

* *Corresponding Author: Assistant Professor, School of Computing, E.S. Arts and Science College, Villupuram, Tamil Nadu, India, [Email: srimathi0696@gmail.com]*

Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

Originality/Value

The paper proposes a unified dual-path architecture in which local convolutional representations and global transformer representations are selectively fused using multi-scale channel-spatial attention and a gated fusion mechanism. The principal novelty is the architectural integration and the explicit evaluation of accuracy, efficiency, calibration, and explanation quality rather than accuracy alone.

Keywords: Medical image analysis; deep learning; CNN; vision transformer; attention mechanism; image classification

JEL Codes: C45; I10; O33

---

1. Introduction

Medical image analysis is an important application area of computer vision because radiographs, dermoscopic images, histopathology images, retinal images, ultrasound images, computed tomography (CT), and magnetic resonance imaging (MRI) contain complex visual patterns that can support screening, diagnosis, segmentation, and treatment planning. Deep learning has substantially changed this field by allowing models to learn task-specific representations directly from image data. However, the performance of a model is influenced not only by its feature extractor but also by its ability to capture small discriminative structures, global anatomical context, class imbalance, acquisition variability, and uncertainty.

Convolutional neural networks (CNNs) have traditionally been effective because convolutional operators provide strong inductive biases for spatial locality and translation-related visual patterns. U-Net and related encoder-decoder architectures further demonstrated the value of multi-scale feature extraction and skip connections in biomedical image analysis (Ronneberger et al., 2015). More recently, transformer architectures have attracted substantial attention because self-attention can model long-range relationships. Reviews of transformers in medical imaging report applications across classification, segmentation, detection, registration, synthesis, and other tasks (Takahashi et al., 2024; Azad et al., 2024).

The coexistence of CNN and transformer approaches suggests that medical-image analysis may benefit from hybrid architectures. CNNs can retain strong local inductive biases, whereas transformer blocks can model broader contextual relationships. Nevertheless, simply

Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

concatenating CNN and transformer features can introduce redundant representations and increase computational cost. A more selective fusion mechanism is therefore required.

This paper addresses the following research question: Can a lightweight dual-path architecture that selectively fuses multi-scale CNN and transformer representations provide a robust and computationally conscious framework for medical image classification?

The objectives are: (i) to design MedFusion-Net, a new CNN-transformer architecture for medical image classification; (ii) to introduce multi-scale channel-spatial attention for selective feature refinement; (iii) to develop a gated fusion mechanism for combining local and global representations; and (iv) to define a reproducible evaluation protocol covering predictive performance, efficiency, calibration, robustness, and explainability.

The paper is deliberately structured as a model-development and validation framework. Since no experimental dataset or executed training logs were supplied for this manuscript, numerical performance values are not claimed. This distinction is important for research integrity: actual results must be obtained before submission.

2. Literature Review

2.1 CNN-based medical image analysis

CNNs learn hierarchical representations through repeated convolution, nonlinear activation, normalization, and downsampling operations. Their local receptive fields make them effective at recognizing edges, textures, shapes, and other spatial patterns. U-Net introduced a contracting path for context extraction and an expanding path for precise localization, demonstrating the effectiveness of multi-scale representations in biomedical segmentation (Ronneberger et al., 2015). Although the present study focuses on classification, the same principle of preserving information at multiple resolutions motivates the proposed multi-scale feature design.

2.2 Transformer-based medical image analysis

Transformers were introduced into medical imaging to address limitations associated with purely local receptive fields. Vision-transformer-based methods can capture long-range dependencies and

JOURNAL India Publication
Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

global spatial relationships. Swin UNETR, for example, uses a hierarchical shifted-window transformer encoder with multi-resolution features for 3D brain-tumour segmentation (Hatamizadeh et al., 2022). Recent reviews identify transformer applications in medical image segmentation, detection, classification, restoration, synthesis, registration, and clinical report generation (Shamshad et al., 2023).

2.3 CNN-transformer hybridization

Comparative evidence indicates that both CNNs and vision transformers remain important in medical imaging, with transformer approaches showing considerable promise but also requiring careful consideration of pre-training, computational requirements, data scale, and robustness (Takahashi et al., 2024). The proposed architecture follows a complementary representation hypothesis: local convolutional features should be preserved while global contextual features are learned through attention.

2.4 Research gap

Three gaps motivate the proposed framework. First, hybrid models may combine representations without sufficiently controlling redundancy. Second, medical datasets often vary in scale, modality, class balance, and image quality, requiring efficient architectures and robust validation. Third, accuracy alone is insufficient for a medical decision-support setting; calibration, class-specific sensitivity, computational cost, and explanation quality should also be examined. MedMNIST v2 provides standardized 2D and 3D biomedical image datasets and fixed splits that can support reproducible benchmarking, although it is explicitly not intended for clinical use (Yang et al., 2023).

The proposed contribution addresses these gaps through selective multi-scale attention, gated feature fusion, and a validation protocol that treats predictive performance and computational/interpretability measures as complementary outcomes.

JOURNAL India Publication
Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

3. Proposed MedFusion-Net Model

3.1 Design principles

MedFusion-Net is designed around four principles: local feature preservation, global contextual reasoning, selective multi-scale refinement, and efficient feature fusion. The architecture contains five major components: input normalization and augmentation, CNN local branch, transformer global branch, multi-scale attention fusion block, and classification head.

3.2 Architecture

The CNN branch contains three stages of convolutional feature extraction. Each stage uses convolution, normalization, nonlinear activation, and downsampling. The transformer branch receives a projected representation of the input and applies lightweight self-attention blocks. Rather than performing full-resolution self-attention at every stage, the proposed design operates on reduced feature maps to limit computational complexity.

The multi-scale attention block receives feature maps from the CNN branch at selected resolutions. Channel attention estimates the importance of feature channels, while spatial attention estimates the importance of image regions. The refined features are passed to a projection layer so that their dimensionality is compatible with the transformer representation.

3.3 Gated feature fusion

Let F_c represents the CNN feature vector and F_t represent the transformer feature vector. The fusion gate is defined as:

$$g = \text{sigmoid}(W_g [F_c \,||\, F_t] + b_g) \quad (1)$$

where || denotes concatenation, W_g and b_g are learnable parameters, and g controls the relative contribution of the two representations. The fused feature is:

$$F_f = g \odot F_c + (1 - g) \odot F_t \quad (2)$$

where $\odot$ denotes element-wise multiplication. This mechanism allows the network to learn whether local or global information should dominate for a given input.

JOURNAL India Publication
Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

3.4 Multi-objective training loss

For class-imbalanced medical data, the training objective can combine weighted cross-entropy and a focal component. A general formulation is:

$$L_{total} = \lambda_1 L_{CE} + \lambda_2 L_{Focal} + \lambda_3 L_{reg} \quad (3)$$

where L_{CE} is class-weighted cross-entropy, L_{Focal} reduces the influence of easy examples, and L_{reg} represents regularization. The coefficients λ_1, λ_2, and λ_3 should be selected using the validation set rather than tuned on the test set.

3.5 Interpretability

Grad-CAM or a comparable attribution method can be applied to the CNN feature maps to generate class-discriminative localization maps. For transformer representations, attention rollout or token attribution can be investigated. Explanations should be evaluated qualitatively and, where annotations are available, quantitatively using overlap or localization measures. An explanation is not treated as proof of clinical validity.

4. Methodology

4.1 Research design

The study follows a quantitative experimental research design. The independent variable is the model architecture, while the dependent variables are predictive performance and efficiency metrics. The experimental design compares MedFusion-Net with established CNN and CNN-transformer baselines under identical preprocessing, data splits, augmentation policy, and evaluation criteria.

4.2 Dataset

A recommended primary benchmark is MedMNIST v2 because it provides standardized biomedical image datasets with predefined train, validation, and test splits. The collection includes 2D and 3D datasets covering diverse biomedical modalities and task types (Yang et al., 2023). For a focused study, PneumoniaMNIST or another clinically meaningful subset can be selected. The

Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

selected dataset, version, class distribution, licensing conditions, and official split must be recorded in the final experimental manuscript.

MedMNIST is suitable for algorithmic benchmarking but is not a substitute for clinical validation. Therefore, any paper using it should avoid claims that the model is clinically deployable unless additional clinical datasets and appropriate validation are performed.

4.3 Pre-processing

Images should be resized to the input resolution required by the selected model, normalized consistently, and augmented only within medically reasonable transformations. Candidate augmentations include small rotations, horizontal flipping where anatomically appropriate, mild translation, random cropping, and intensity normalization. Transformations that can alter pathology or anatomical meaning should be excluded.

4.4 Training protocol

The proposed model should be implemented using PyTorch or an equivalent deep-learning framework. AdamW may be used as the optimizer with an initial learning rate selected through validation. A cosine learning-rate schedule and early stopping may be employed. Training should use fixed random seeds and record the software version, hardware, batch size, epochs, optimizer, learning rate, augmentation settings, and model parameter count.

4.5 Baselines

The minimum baseline set should contain a conventional CNN such as ResNet-18, a stronger CNN such as EfficientNet-B0, and a transformer or hybrid architecture appropriate to the chosen dataset. The objective is not merely to outperform one baseline, but to establish whether each proposed component provides measurable benefit.

Model	Role	Key comparison
ResNet-18	CNN baseline	Local hierarchical representation
EfficientNet-B0	Efficient CNN baseline	Accuracy-efficiency trade-off
Vision Transformer / Swin variant	Transformer baseline	Global contextual modelling

JOURNAL India Publication
Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

MedFusion-Net	Proposed model	CNN + transformer + selective fusion

Source: Proposed experimental design.

5. Research Results and Discussion

5.1 Evaluation metrics

The final experiment should report accuracy, precision, recall, macro-F1, class-wise sensitivity, specificity where appropriate, ROC-AUC for binary or suitable multi-class settings, confusion matrices, parameter count, inference time, and peak memory usage. For imbalanced datasets, macro-F1 and class-wise recall should receive particular attention rather than relying on accuracy alone.

Accuracy is defined as:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

Precision, recall, and F1-score are:

$$\text{Precision} = TP / (TP + FP) \quad (5)$$

$$\text{Recall} = TP / (TP + FN) \quad (6)$$

$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (7)$$

5.2 Required comparative results

Model	Accuracy	Macro-F1	ROC-AUC	Params	Inference time
ResNet-18	To be measured	To be measured	To be measured	To be measured	To be measured
EfficientNet-B0	To be measured	To be measured	To be measured	To be measured	To be measured
Transformer baseline	To be measured	To be measured	To be measured	To be measured	To be measured
MedFusion-Net	To be measured	To be measured	To be measured	To be measured	To be measured

Source: Experimental results must be populated from the actual executed runs.

JOURNAL
India Publication
Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

5.3 Ablation study

An ablation study is essential to establish whether the proposed components contribute independently. The recommended configurations are: CNN branch only; transformer branch only; CNN + transformer without attention; CNN + transformer with attention but without gated fusion; and the complete MedFusion-Net. Performance differences should be reported with confidence intervals or repeated-run variability where feasible.

Configuration	CNN	Transformer	Attention	Gated fusion	Result
A	✓	—	—	—	To be measured
B	—	✓	—	—	To be measured
C	✓	✓	—	—	To be measured
D	✓	✓	✓	—	To be measured
E: MedFusion-Net	✓	✓	✓	✓	To be measured

Source: Proposed ablation protocol.

5.4 Discussion framework

The discussion should test four hypotheses. H1: the hybrid representation improves macro-F1 relative to a CNN-only baseline. H2: multi-scale attention improves minority-class sensitivity and reduces irrelevant feature activation. H3: gated fusion improves the hybrid model compared with simple concatenation. H4: the proposed architecture provides an acceptable accuracy-efficiency trade-off relative to larger transformer models.

Because actual training results are not available in the supplied materials, these hypotheses are not presented as confirmed findings. Once experiments are executed, the discussion should report effect sizes, confidence intervals where possible, failure cases, confusion matrices, and representative explanation maps. The paper should avoid replacing measured evidence with expected values.

JOURNAL India Publication
Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

5.5 Error analysis

Error analysis should group incorrect predictions by image quality, lesion or abnormality size, class imbalance, acquisition characteristics, and visually ambiguous cases. The analysis should also check whether attention maps focus on relevant anatomical regions or on artifacts such as borders, text labels, markers, or acquisition devices. This is particularly important because deep learning models can exploit unintended shortcuts.

5.6 Robustness and fairness

If demographic or acquisition metadata are available, subgroup evaluation should be conducted. Medical-image models can exhibit fairness problems when training data are not representative, and recent systematic evidence highlights fairness as an important issue in deep-learning medical image analysis (Xu et al., 2024). The final study should therefore report limitations in dataset diversity and avoid claiming universal generalization.

6. Conclusions, Proposals and Recommendations

This paper proposes MedFusion-Net, a dual-path CNN-transformer model for medical image classification. The architecture is designed to combine CNN-based local representations with transformer-based global contextual information through multi-scale channel-spatial attention and gated feature fusion. The paper also specifies a reproducible evaluation framework that includes predictive performance, computational efficiency, calibration, robustness, and interpretability.

The central contribution is architectural: rather than relying exclusively on convolutional locality or transformer global attention, the proposed model learns to combine both forms of representation selectively. The approach is intended as a research framework and requires empirical validation before any superiority claim can be made.

Recommendations are as follows: (i) evaluate the model on at least one standardized benchmark and one external dataset when available; (ii) perform repeated experiments and statistical testing; (iii) include complete ablation studies; (iv) report model size and inference cost; (v) evaluate

JOURNAL India Publication
Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

explanation maps for shortcut learning; and (vi) conduct external and clinically relevant validation before considering deployment.

7. Theoretical Implications

The proposed framework contributes to computer vision research by treating local and global visual representations as complementary rather than mutually exclusive. It operationalizes the hypothesis that medical-image classification can benefit from selective fusion of hierarchical convolutional features and attention-derived contextual representations. The gated fusion mechanism provides an explicit learnable interface for this complementarity.

8. Practical Implications

For researchers and educators, MedFusion-Net provides a modular architecture that can be implemented and evaluated using publicly available datasets. For healthcare AI developers, the framework emphasizes that performance should be accompanied by efficiency, calibration, robustness, and explanation analysis. The model is not a medical device and should not be interpreted as a clinical diagnostic system without appropriate clinical validation and regulatory assessment.

9. Social Implications

Medical AI can potentially support earlier screening and reduce workload, but poorly validated models may amplify disparities or produce misleading confidence. The proposed validation framework therefore treats subgroup performance, dataset representativeness, and explainability as important research concerns. Responsible use requires human oversight and independent clinical evaluation.

10. Industry Implications

The proposed architecture may be relevant to organizations developing AI-assisted radiology, pathology, dermatology, or screening tools. The explicit attention to parameter count and inference cost can help assess whether a model is appropriate for resource-constrained environments.

JOURNAL India Publication
Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

However, industrial adoption requires external validation, security and privacy controls, workflow integration, regulatory compliance, and monitoring for performance drift.

11. Limitations & Scope for Future Research

The principal limitation of the current manuscript is that it presents a proposed architecture and validation protocol rather than completed empirical evidence. Numerical performance values are intentionally omitted to prevent fabricated results. The proposed model also requires systematic tuning, external validation, and comparison against current state-of-the-art methods.

Future work should include multi-center datasets, higher-resolution images, 3D medical imaging, self-supervised or foundation-model pretraining, uncertainty estimation, federated learning, subgroup fairness analysis, and prospective clinical evaluation. Additional work can investigate knowledge distillation and quantization to reduce computational cost for edge or near-patient deployment.

Conflicts of Interest Declaration

The author declares that there are no conflicts of interest regarding the publication of this research.

Informed Consent Declaration

The proposed experimental design uses publicly available, de-identified benchmark data. If a future version of the study uses newly collected human-participant data, appropriate institutional ethics approval, informed consent, privacy safeguards, and data-governance procedures must be obtained before data collection and analysis.

References

Azad, R., Kazerouni, A., Heidari, M., Khodapanah Aghdam, E., Molaei, A., Jia, Y., Jose, A., Roy, R. and Merhof, D. 2024. Advances in medical image analysis with vision Transformers: A comprehensive review. Medical Image Analysis, 91, 103000.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. and Xu, D. 2022. Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images. arXiv preprint arXiv:2201.01266.

Ronneberger, O., Fischer, P. and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer, pp. 234–241.

Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S. and Fu, H. 2023. Transformers in medical imaging: A survey. Medical Image Analysis, 91, 102772.

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., Machino, H., Kobayashi, K., Asada, K., Komatsu, M., Kaneko, S., Sugiyama, M. and Hamamoto, R. 2024. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. Journal of Medical Systems, 48(1), 84.

Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M. and Zhou, S.K. 2024. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. npj Digital Medicine, 7, 286.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H. and Ni, B. 2023. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data, 10, 41.