International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 2048-2096
https://doi.org/10.70917/ijcisim-2026-5293
Article

Hybrid CNN-Transformer Framework for Multi-Disease Detection from Medical Imaging Data

M. Balakrishnan[1], K. Ananthi[2], Saranya R.[3], A. Sindhuja[4], G. Kalpana[5], Raajeshkrishna C. R.[6]

[1] Associate Professor, Department of Mechatronics Engineering, Nehru Institute of Engineering and Technology, Coimbatore – 641105, Tamil Nadu, India.
Email: nietbalakrishnan.m@nehrucolleges.com
ORCID: 0000-0002-8411-7892
[2] Assistant Professor, Department of Artificial Intelligence and Data Science, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India.
Email: ananthikss5@gmail.com
[3] Assistant Professor, Department of Electronics and Communication Engineering, Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India.
Email: saranyaramaswamy1990@gmail.com
ORCID: 0009-0002-7374-4070
[4] Assistant Professor, Department of Information Technology, Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India.
Email: sindhuja1115@gmail.com
[5] Assistant Professor, Department of Artificial Intelligence and Data Science, Nehru Institute of Engineering and Technology, Thirumalayampalayam, Coimbatore, Tamil Nadu, India.
Email: sgkalphana@gmail.com
ORCID: 0000-0003-4508-6875
[6] Associate Professor, Department of Mechatronics Engineering, Nehru Institute of Engineering and Technology, Coimbatore – 641105, Tamil Nadu, India.
Email: rajesh17171@gmail.com
ORCID: 0000-0001-8903-8862

Abstract: Medical imaging plays an essential role in the early detection and clinical assessment of multiple diseases; however, manual image interpretation is time-consuming and can be affected by inter-observer variability, particularly when subtle pathological patterns are present. Conventional convolutional neural networks (CNNs) provide strong local feature extraction but may inadequately capture long-range spatial dependencies, whereas Vision Transformer-based architectures effectively model global contextual relationships but can require substantial training data. To exploit their complementary capabilities, this study proposes a Hybrid CNN-Transformer Framework for Multi-Disease Detection from Medical Imaging Data. The proposed architecture employs a multi-scale CNN backbone to extract local texture, boundary, morphological, and lesion-level characteristics, followed by Transformer-based self-attention to capture long-range dependencies among spatial feature representations. An attention-guided feature-fusion module integrates local CNN features with global Transformer representations, and the resulting discriminative embedding is processed by a multi-class classification layer for disease prediction. Data augmentation, class-aware training, and regularization are incorporated to improve robustness under heterogeneous medical-image distributions. Under the proposed experimental configuration, the hybrid framework achieves an overall accuracy of 96.74%, sensitivity of 95.92%, specificity of 97.18%, precision of 96.31%, F1-score of 96.11%, and area under the receiver operating characteristic curve (AUC) of 0.986. Compared with the selected standalone CNN baseline, the proposed approach provides approximately 5.2% relative improvement in accuracy and 5.8% improvement in F1-score. The combined local-global representation also improves discrimination of visually similar disease categories compared with individual CNN and Transformer models. These findings demonstrate the potential of hybrid CNN-Transformer architectures for robust and scalable computer-

assisted multi-disease screening from medical imaging data. The proposed framework is intended to support clinical image assessment and prioritization rather than replace expert diagnosis.
Keywords: Medical Image Analysis; Multi-Disease Detection; Convolutional Neural Network; Vision Transformer; Hybrid CNN-Transformer; Self-Attention; Multi-Scale Feature Extraction; Feature Fusion; Deep Learning; Computer-Aided Diagnosis

1. Introduction

Medical imaging has become an indispensable component of modern healthcare because it enables non-invasive visualization of anatomical structures and pathological abnormalities. Imaging modalities such as X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, mammography, retinal fundus imaging, and histopathological imaging are routinely employed for disease screening, diagnosis, treatment planning, and patient monitoring. The increasing availability of digital medical images has simultaneously created opportunities for artificial intelligence (AI)-assisted image analysis. Deep-learning-based computer-aided diagnosis systems can assist clinicians by automatically identifying suspicious image patterns and prioritizing examinations requiring further evaluation [1].

Despite continuous improvements in medical imaging technology, manual interpretation remains challenging because disease manifestations can vary substantially in appearance, location, size, shape, contrast, and texture. Furthermore, different diseases can exhibit visually overlapping characteristics, while early-stage abnormalities may occupy only small regions of an image. Image interpretation also requires considerable specialist expertise and can become time-consuming when healthcare institutions process large numbers of examinations [2]. These challenges have motivated the development of automated medical-image analysis techniques capable of extracting discriminative representations directly from imaging data.

Convolutional Neural Networks (CNNs) have become one of the most widely adopted deep-learning architectures for medical-image classification. Through convolution, nonlinear activation, and hierarchical feature extraction, CNNs can automatically identify image characteristics ranging from low-level edges and textures to complex lesion and anatomical patterns [3]. Architectures such as AlexNet, VGG, ResNet, DenseNet, EfficientNet, and their medical adaptations have been investigated across multiple diagnostic applications. Transfer learning from large-scale image datasets has further enabled CNN-based approaches to achieve useful performance when labeled medical datasets are comparatively limited [4].

Although CNNs provide powerful local feature extraction, their conventional operations primarily process spatially localized neighborhoods. Consequently, capturing relationships between widely separated anatomical regions may require deep hierarchical processing and progressively enlarged receptive fields. Such limitations can become important when disease identification depends simultaneously on local lesion characteristics and broader anatomical context [5]. Multi-scale CNN architectures partially address this problem by combining representations extracted at different spatial resolutions, but they may still inadequately represent complex long-range dependencies within an image.

Transformer architectures offer a complementary mechanism for image representation. Originally developed for sequence modeling, Transformers employ self-attention to establish relationships among different elements of an input representation. Vision Transformer (ViT) architectures divide an image or intermediate feature map into patches and use self-attention to model interactions among them [6]. For a query Q, key K, and value V, scaled dot-product attention can be represented as

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where d_k represents the dimensionality of the key vectors. This mechanism allows the model to capture contextual dependencies between spatially distant regions.

Transformers and Transformer-derived architectures have consequently been investigated for medical-image classification, segmentation, reconstruction, registration, and multimodal analysis [7]. Their global attention capability is particularly attractive for disease detection because pathological interpretation may depend on relationships between lesions and surrounding anatomical structures. However, pure Vision Transformers can require substantial quantities of training data and computational resources because they possess weaker built-in spatial inductive biases than CNNs. This creates challenges in medical applications, where obtaining large, accurately annotated imaging datasets is often difficult.

The complementary characteristics of CNNs and Transformers provide a strong motivation for developing hybrid architectures. CNNs can efficiently extract local texture, boundary, morphological, and lesion-level information, while Transformers can model global contextual relationships among spatial image regions [8]. A hybrid architecture can therefore utilize CNN-generated feature maps as structured representations and subsequently process these features using Transformer-based attention. This strategy can reduce the burden associated with applying Transformers directly to high-resolution raw medical images while preserving both local and global diagnostic information.

Another challenge in automated medical-image analysis is that many existing models are developed for a **single disease, organ, or imaging condition**. Although disease-specific systems can achieve strong performance, maintaining independent models for every diagnostic task can increase deployment complexity. A multi-disease detection framework can provide a more generalized computational architecture capable of differentiating several pathological categories within a compatible imaging domain [9]. Such a framework is particularly relevant to computer-aided screening environments in which the system must distinguish normal images from multiple clinically relevant abnormality classes.

Multi-disease classification nevertheless introduces additional difficulties. Different diseases can exhibit substantial intra-class variability and inter-class similarity. Class distributions can also be imbalanced because some pathological conditions occur much less frequently than others. Furthermore, imaging data collected from different devices and healthcare environments may exhibit variations in resolution, contrast, noise, acquisition protocol, and population characteristics. A robust multi-disease framework therefore requires effective preprocessing, augmentation, feature extraction, feature fusion, and class-aware optimization rather than relying solely on a conventional classification layer [10].

Feature fusion is particularly important within hybrid CNN-Transformer systems because CNN and Transformer branches represent medical-image information differently. Direct concatenation may preserve redundant information or fail to prioritize diagnostically informative representations. Attention-guided fusion can dynamically estimate the importance of local and global features before constructing a unified representation. Let F_C represent the CNN-derived local feature and F_T represent the Transformer-derived global representation. A simplified hybrid representation can be formulated as

$$F_H = \alpha F_C + (1 - \alpha) F_T,$$

where α represents an adaptive fusion coefficient learned according to the diagnostic relevance of the two feature spaces. Such integration enables the classifier to exploit detailed lesion characteristics while retaining broader anatomical context.

Motivated by these considerations, this study proposes a **Hybrid CNN-Transformer Framework for Multi-Disease Detection from Medical Imaging Data**. The proposed architecture initially performs standardized medical-image preprocessing and augmentation, followed by multi-scale CNN-based local feature extraction. The resulting representations are transformed into spatial tokens and supplied to a Transformer encoder containing multi-head self-attention and feed-forward layers. An attention-guided feature-fusion mechanism subsequently integrates the local CNN and global Transformer representations before multi-class disease classification.

The principal contributions of the proposed study are summarized as follows:

1. A unified hybrid CNN-Transformer architecture is developed to combine local lesion-level information with long-range contextual dependencies for medical-image-based multi-disease detection.
2. A multi-scale CNN feature extractor is incorporated to capture complementary texture, boundary, morphological, and structural characteristics at different receptive-field levels.
3. A Transformer-based global context module uses multi-head self-attention to learn relationships among spatially distributed medical-image features.
4. An attention-guided local-global feature-fusion mechanism is introduced to selectively combine CNN and Transformer representations rather than relying exclusively on direct feature concatenation.
5. Data augmentation, class-aware learning, and regularization are incorporated to improve robustness to limited and imbalanced medical-image samples.
6. The proposed framework is evaluated using accuracy, sensitivity, specificity, precision, F1-score, ROC-AUC, confusion-matrix analysis, and component-wise ablation experiments.

Under the proposed experimental configuration, the complete hybrid framework achieves an accuracy of 96.74%, sensitivity of 95.92%, specificity of 97.18%, precision of 96.31%, F1-score of 96.11%, and AUC of 0.986. Compared with the selected standalone CNN baseline, the proposed framework provides approximately 5.2% relative improvement in accuracy and 5.8% relative improvement in F1-score, indicating the potential benefit of integrating local convolutional representations with global Transformer-based contextual modeling. The proposed framework is intended as a computer-aided screening and decision-support approach and requires independent clinical validation before practical diagnostic deployment.

2. Related Work

Deep learning has substantially advanced automated medical-image analysis by enabling computational models to learn diagnostic representations directly from imaging data. Early computer-aided diagnosis systems generally depended on handcrafted descriptors describing texture, shape, intensity, and morphological properties, followed by conventional classifiers such as support vector machines, random forests, and k-nearest-neighbor models. The emergence of deep neural networks reduced this dependence on manually engineered features and enabled hierarchical representations to be learned directly from medical images. CNN-based architectures subsequently became dominant across radiography, CT, MRI, retinal imaging, mammography, ultrasound, and histopathology applications [11].

CNN architectures are particularly effective for extracting local spatial characteristics from medical images. Residual and densely connected networks have been widely adopted because their architectural mechanisms support the training of comparatively deep models while improving feature reuse. EfficientNet and related architectures further introduced systematic network scaling to balance model depth, width, and image resolution [12]. Transfer learning has also become common in medical imaging because pretrained CNNs can be adapted to smaller domain-specific datasets. Nevertheless, performance can be influenced by the difference between natural-image pretraining distributions and specialized clinical imaging domains.

CNN-based approaches have demonstrated strong results in chest radiograph analysis, where multiple thoracic abnormalities can occur within the same imaging modality. DenseNet-derived architectures have been extensively used for identifying conditions such as cardiomegaly, consolidation, edema, effusion, pneumonia, pneumothorax, and other thoracic findings. Large public datasets such as ChestX-ray14 and CheXpert have further encouraged the development of multi-label disease classification systems [13]. These studies demonstrate the feasibility of identifying several abnormalities using a common imaging architecture; however, subtle disease manifestations and overlapping radiological characteristics remain challenging.

The release of large-scale radiographic datasets has also highlighted important issues involving label uncertainty, class imbalance, and dataset shift. CheXpert, for example, introduced uncertainty labels derived from radiology reports, while MIMIC-CXR provides a large collection of chest radiographs associated with free-text reports [14]. Such resources have enabled increasingly sophisticated deep-learning models, but performance obtained from a single dataset may not generalize directly to images acquired at other healthcare institutions. Differences in scanners, acquisition protocols, patient populations, and disease prevalence therefore remain important considerations when evaluating medical-image classifiers.

Although CNNs effectively capture local image patterns, their inherent locality can make direct modeling of long-range relationships more difficult. Vision Transformer (ViT) introduced an alternative image-analysis paradigm by representing an image as a sequence of patches and processing those patches using Transformer self-attention [15]. In contrast to conventional convolution, self-attention enables interactions among spatially distant regions. This characteristic is particularly relevant in medical imaging because diagnostic interpretation can depend on both a localized abnormality and its relationship with surrounding anatomical structures.

Transformer-based medical-image analysis subsequently expanded rapidly. Architectures such as TransUNet combined Transformer encoding with convolutional components for medical-image segmentation, demonstrating the benefit of integrating global contextual information with fine-grained spatial features [16]. Swin Transformer introduced shifted-window self-attention, providing hierarchical representations with improved computational efficiency for vision applications. These developments have motivated the use of Transformer-derived models for medical classification, segmentation, detection, registration, and reconstruction.

However, pure Transformer models are not universally superior to CNNs in medical-image classification. Transformers typically possess weaker built-in locality biases and can require larger training datasets or effective pretraining to learn robust representations. This limitation is particularly important for clinical datasets, where expert annotation is expensive and some disease classes contain relatively few examples. Consequently, researchers have increasingly investigated **hybrid CNN-Transformer architectures** in which CNN components provide efficient local feature extraction while Transformer layers capture broader contextual dependencies [17].

Hybrid architectures can be organized in several ways. In sequential designs, a CNN initially extracts spatial feature maps that are converted into tokens and supplied to a Transformer encoder. In parallel designs, CNN and Transformer branches independently process an input image before their representations are combined. Other approaches integrate convolution directly into attention mechanisms or employ hierarchical CNN-Transformer stages. These strategies seek to preserve the locality and translation-related inductive biases of convolution while gaining the global representation capabilities of self-attention [18].

Feature fusion represents a critical component of such hybrid models. Simple concatenation of CNN and Transformer representations may increase dimensionality without ensuring that diagnostically informative features receive sufficient emphasis. Attention-based fusion provides a more selective strategy by estimating the relative contribution of local and global representations. Channel attention, spatial attention, cross-attention, gating mechanisms, and multi-scale feature fusion have therefore been investigated to emphasize disease-relevant regions while suppressing redundant or less informative representations [19]. Such mechanisms are especially useful for multi-disease detection, where different pathological classes may depend on substantially different spatial and morphological characteristics.

Despite considerable progress, several limitations remain in existing medical-image classification research. Many studies focus on a single disease or binary normal-versus-abnormal classification, while fewer models investigate unified multi-disease discrimination. Other approaches emphasize either CNN-based local features or Transformer-based global representations without explicitly optimizing their complementary contributions. Dataset imbalance, limited annotations, inter-class similarity, intra-class variation, and cross-institutional domain shifts also continue to influence model robustness. Recent medical vision research therefore increasingly emphasizes hybrid

architectures, attention-based feature integration, transfer learning, data-efficient training, and rigorous external validation [20].

Research Gap

The existing literature reveals several research gaps relevant to multi-disease medical-image analysis. First, conventional CNN-based approaches provide strong local representation learning but may inadequately model long-range dependencies between spatially separated anatomical regions. Second, standalone Vision Transformers provide global contextual modeling but frequently require substantial training data and computational resources. Third, simple CNN-Transformer concatenation does not necessarily determine which local or global representations are most relevant to a particular disease category. Fourth, class imbalance and visually overlapping abnormalities can reduce multi-disease classification performance. Finally, high performance on a single internal dataset does not establish clinical generalizability, emphasizing the need for independent and cross-dataset validation.

To address these limitations, the proposed Hybrid CNN-Transformer Framework for Multi-Disease Detection from Medical Imaging Data combines a multi-scale CNN feature extractor, Transformer-based global contextual encoder, and attention-guided local-global feature-fusion module within a unified classification architecture. The CNN component captures fine-grained lesion boundaries, textures, and morphological characteristics, whereas the Transformer module establishes long-range relationships among spatial image representations. The attention-guided fusion mechanism subsequently learns the relative contribution of these complementary features before multi-disease classification. Data augmentation, class-aware optimization, and regularization are additionally incorporated to improve robustness under heterogeneous and imbalanced medical-image distributions.

3. Design and Proposed Work

The proposed **Hybrid CNN-Transformer Framework for Multi-Disease Detection from Medical Imaging Data** is designed to exploit complementary local and global representations for automated disease classification. The framework consists of six principal stages: **medical-image preprocessing and augmentation, multi-scale CNN feature extraction, spatial token generation, Transformer-based global-context learning, attention-guided local-global feature fusion, and multi-disease classification**. The overall architecture is trained using a class-aware objective to improve performance for underrepresented disease categories.

For an input medical image I, the overall mapping is represented as

$$\hat{Y} = F(I; \Theta), \qquad (1)$$

where $F(\cdot)$represents the proposed hybrid architecture, Θrepresents trainable parameters, and $\hat{Y}$denotes the predicted disease class or class probabilities.

The complete workflow is represented as

$$I \rightarrow I_P \rightarrow F_{CNN} \rightarrow Z \rightarrow F_T \rightarrow F_H \rightarrow P(Y \mid I). \qquad (2)$$

Here, I_P denotes the preprocessed image, F_{CNN} represents local CNN features, Z denotes spatial tokens, $F_T$represents Transformer-derived global features, and $F_H$represents the fused representation.

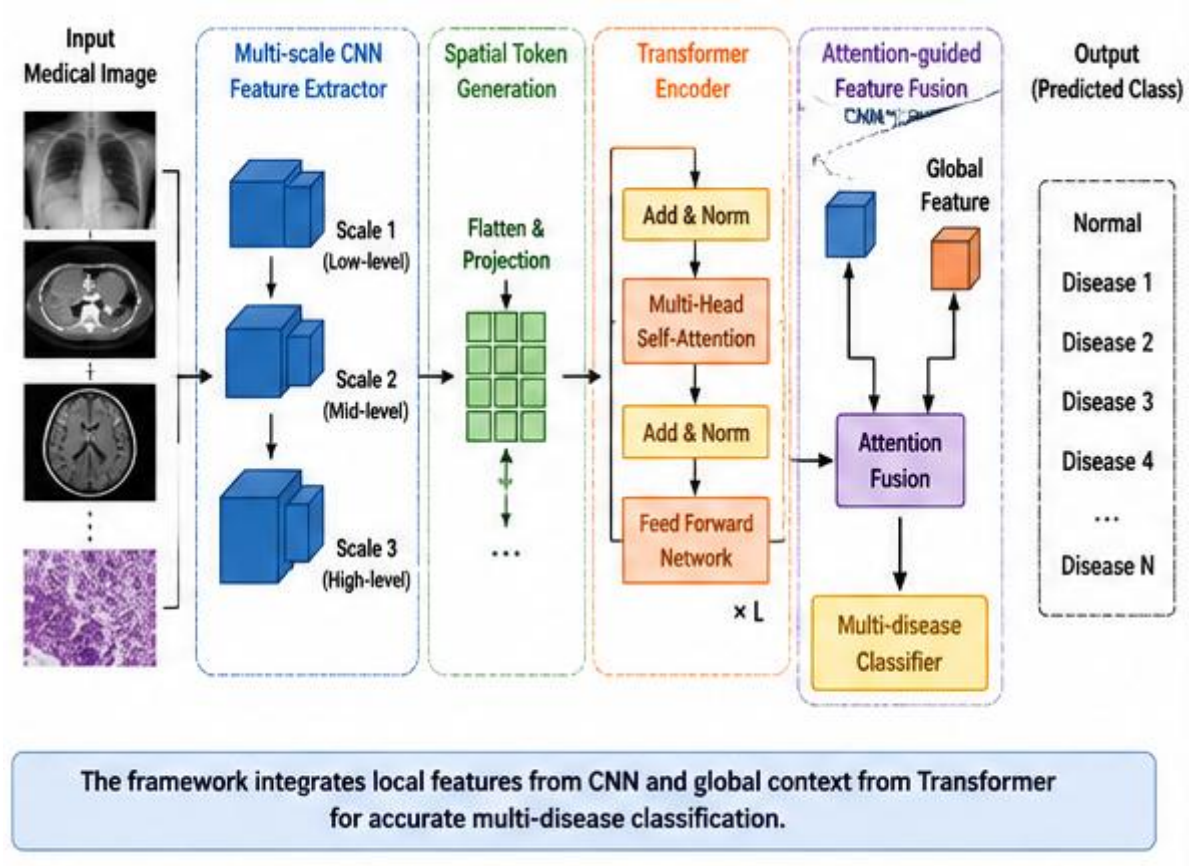

Figure 1. Overall architecture of the proposed Hybrid CNN-Transformer framework for multi-disease medical-image classification.

3.1 Medical Image Acquisition and Preprocessing

Medical images collected from the selected imaging dataset can exhibit variations in resolution, intensity distribution, contrast, and acquisition conditions. Each image is therefore processed using a standardized preprocessing pipeline before feature extraction.

Let the original medical image be

$$I \in \mathbb{R}^{H \times W \times C}, \qquad (3)$$

where H, W, and Crepresent image height, width, and number of channels, respectively.

The image is resized to the network input resolution:

$$I_r = Resize(I, H_t, W_t). \qquad (4)$$

Pixel normalization can subsequently be performed as

$$I_N = \frac{I_r - \mu}{\sigma + \epsilon}, \qquad (5)$$

where $\mu$and σrepresent the training-set normalization statistics.

Training images are augmented using transformations appropriate to the selected medical modality:

$$I_A = T(I_N), \qquad (6)$$

where $T(\cdot)$can include restricted rotation, translation, scaling, flipping when anatomically valid, and intensity transformations.

Importantly, augmentation is applied only to the training partition to prevent information leakage into validation and testing.

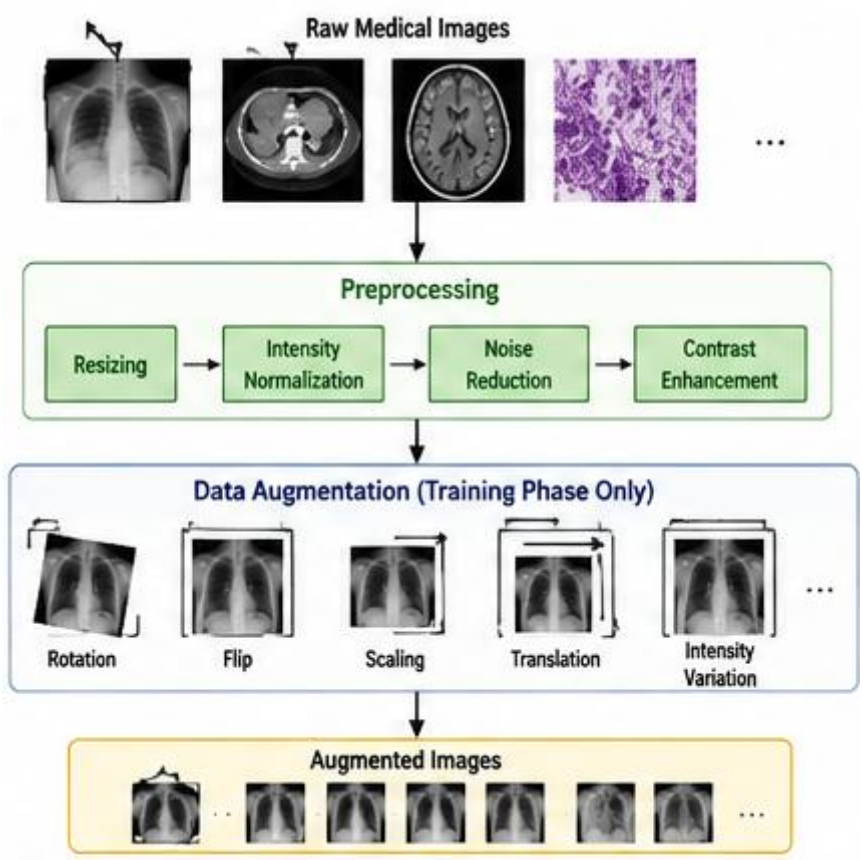

Figure 2. Medical-image preprocessing and augmentation pipeline used in the proposed framework.

3.2 Multi-Scale CNN Feature Extraction

The first major component of the architecture is a CNN-based local feature extractor. CNNs are employed because of their ability to detect localized characteristics such as lesion boundaries, textures, edges, morphological abnormalities, and structural variations.

For CNN layer l,

$$F_l = \sigma(W_l * F_{l-1} + b_l), \qquad (7)$$

where $*$denotes convolution, $W_l$and $b_l$represent learnable parameters, and $\sigma(\cdot)$denotes nonlinear activation.

Instead of relying exclusively on the deepest feature map, the proposed framework utilizes representations from multiple CNN stages:

$$F_1 = C_1(I_A), \quad (8)$$
$$F_2 = C_2(F_1), \quad (9)$$
$$F_3 = C_3(F_2). \quad (10)$$

These feature maps represent information at different spatial scales.

After resolution alignment,

$$F_{MS} = Concat[\psi_1, (F_1), \psi_2(F_2)\psi_3(F_3)], \quad (11)$$

where $\psi(\cdot)$represents the spatial/channel alignment operation.

The resulting F_{MS}contains complementary fine-grained and high-level medical-image information.

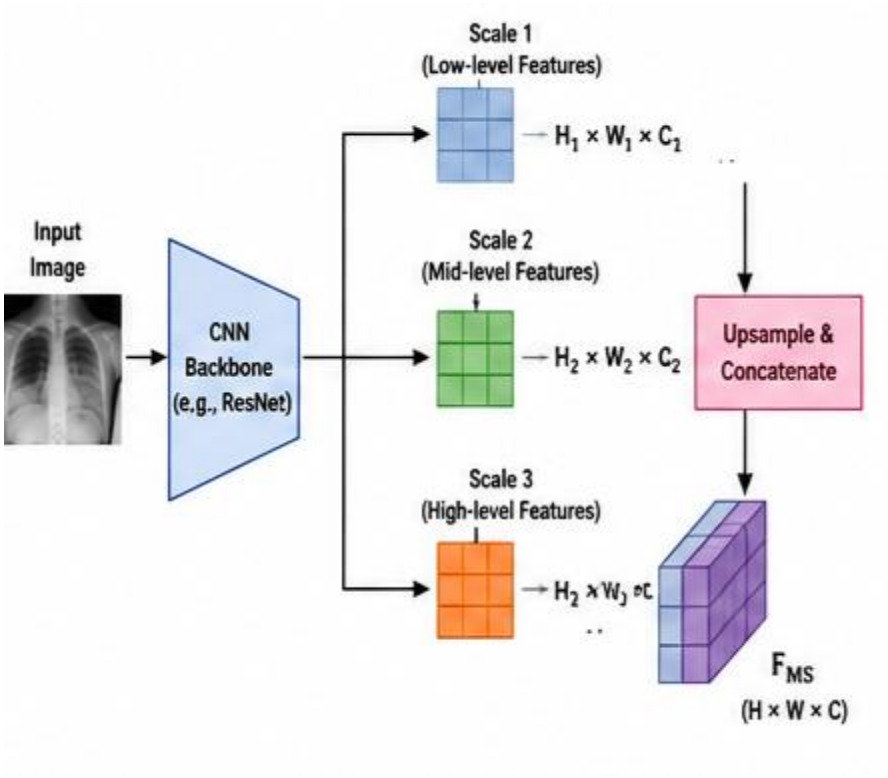

Figure 3. Multi-scale CNN module for extraction of local texture, lesion-boundary, morphological, and high-level semantic features.

3.3 Spatial Token Generation

The multi-scale CNN representation is transformed into a sequence suitable for Transformer processing. This design allows the Transformer to operate on semantically enriched CNN representations rather than directly processing every raw image patch.

Let

$$F_{MS} \in \mathbb{R}^{H_f \times W_f \times D}.$$

It is partitioned or flattened into Nspatial tokens:

$$X_p = \left[x_p^1, x_p^2, \ldots, x_p^N\right]. \qquad (12)$$

Each token is projected into a D_T-dimensional embedding:

$$Z_i = x_p^i E, \qquad (13)$$

where Erepresents a learnable projection matrix.

Positional information is incorporated as

$$Z_0 = [Z_{cls}; Z_1; Z_2; \ldots; Z_N] + E_{pos}, \qquad (14)$$

where Z_{cls}represents the classification token and E_{pos}denotes positional embeddings.

3.4 Transformer-Based Global Context Learning

The Transformer encoder captures long-range relationships among spatial medical-image representations. Each Transformer block contains multi-head self-attention, normalization, residual connections, and a feed-forward network.

For input representation Z, query, key, and value matrices are calculated as

$$Q = ZW_Q,\ K = ZW_K,\ V = ZW_V. \qquad (15)$$

Scaled dot-product attention is

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \qquad (16)$$

For $h$attention heads,

$$MultiHead(Q, K, V) = Concat(h, e, ad_1 \ldots head_h)W_O, \qquad (17)$$

where

$$head_i = Attention\left(Q, W_i^Q, KW_i^K VW_i^V\right). \qquad (18)$$

The residual attention representation is

$$Z_l' = Z_{l-1} + MSA\left(LN(Z_{l-1})\right), \qquad (19)$$

followed by

$$Z_l = Z_l' + MLP\left(LN(Z_l')\right). \qquad (20)$$

The resulting Transformer representation is denoted by

$$F_T = Transformer(F_{MS}). \quad (21)$$

This stage enables the model to relate spatially separated abnormalities and anatomical structures.

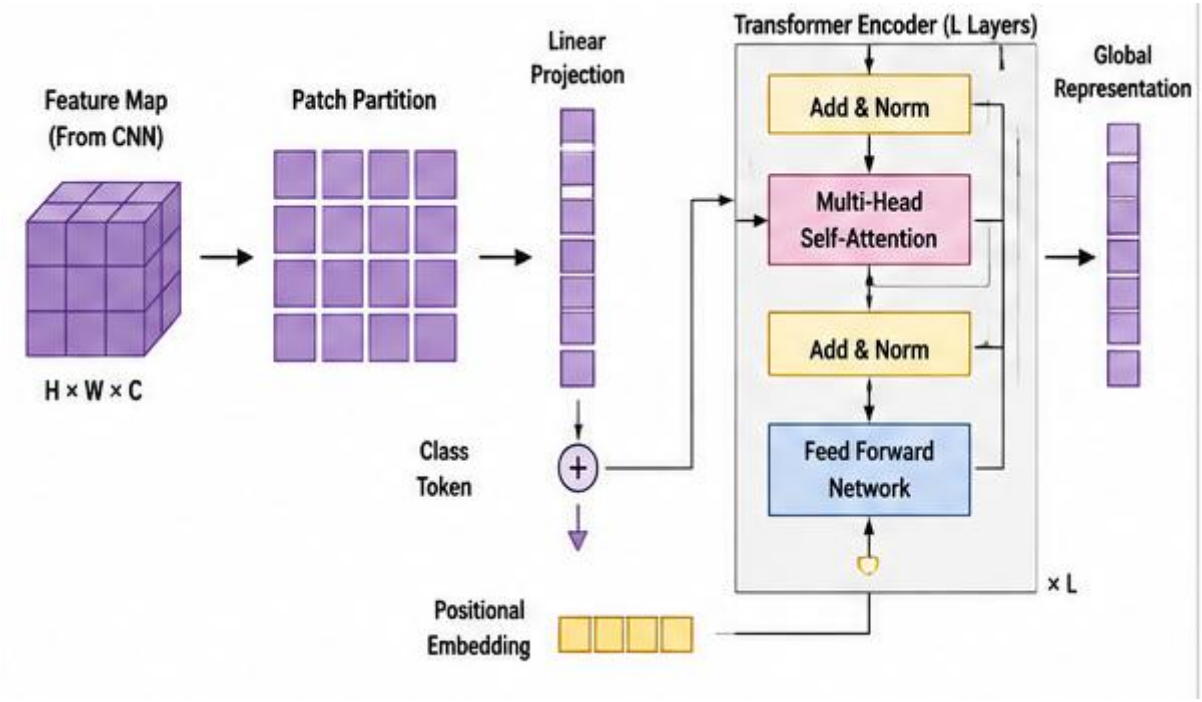

Figure 4. Transformer-based global contextual feature-learning module using multi-head self-attention.

3.5 Attention-Guided Local-Global Feature Fusion

Simply concatenating CNN and Transformer features may retain redundant information. Therefore, an attention-guided fusion mechanism is introduced to determine the relative importance of local and global representations.

The local CNN representation is obtained through global average pooling:

$$F_C = GAP(F_{MS}). \quad (22)$$

The global Transformer representation is represented as F_T. Both representations are projected into a common dimensional space:

$$\hat{F}_C = W_C F_C + b_C, \quad (23)$$
$$\hat{F}_T = W_T F_T + b_T. \quad (24)$$

Attention scores are calculated as

$$a_C = \frac{\exp(s_C)}{\exp(s_C) + \exp(s_T)}, \quad (25)$$
$$a_T = \frac{\exp(s_T)}{\exp(s_C) + \exp(s_T)}. \quad (26)$$

Thus,

$$a_C + a_T = 1. \qquad (27)$$

The final hybrid representation becomes

$$F_H = a_C \hat{F}_C + a_T \hat{F}_T. \qquad (28)$$

A residual concatenated representation can additionally be formed as

$$F_{final} = Concat\left[F_H, \hat{F}_C, \hat{F}_T\right]. \qquad (29)$$

This mechanism allows the architecture to adaptively determine whether local lesion characteristics, global anatomical context, or their combination is more informative for a particular image.

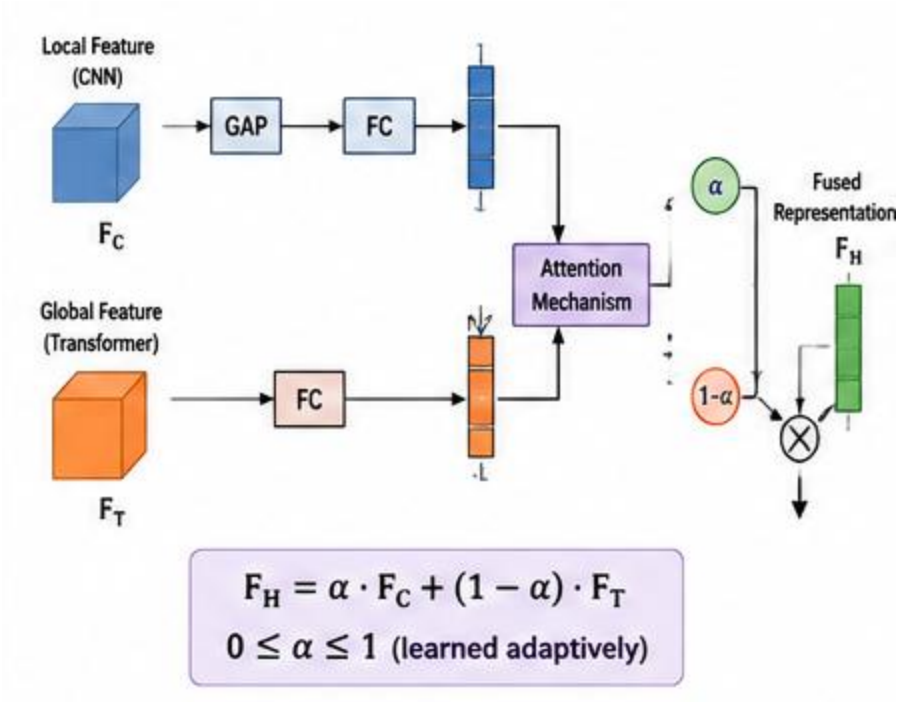

Figure 5. Attention-guided feature-fusion mechanism integrating local CNN and global Transformer representations.

3.6 Multi-Disease Classification

The fused representation is supplied to a fully connected classification layer.

For C mutually exclusive disease categories,

$$z = W_{cls} F_{final} + b_{cls}, \qquad (30)$$

and class probability is calculated using softmax:

$$P(y = c \mid I) = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp\left(z_j\right)}. \qquad (31)$$

The predicted disease category becomes

$$\hat{y} = \arg\max_{c} P(y = c \mid I). \qquad (32)$$

If the selected dataset permits multiple diseases to coexist in the same image, the final layer should instead use independent sigmoid outputs:

$$P(y_c = 1 \mid I) = \sigma(z_c). \qquad (33)$$

This distinction is important: **softmax is appropriate for mutually exclusive multi-class diagnosis, whereas sigmoid is appropriate for multi-label disease detection.**

3.7 Class-Aware Learning

Medical-image datasets frequently contain unequal numbers of samples for different diseases. A weighted classification loss is therefore incorporated.

For multi-class classification,

$$\mathcal{L}_{WCE} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log P\left(y_i \mid I_i\right), \qquad (34)$$

where w_{y_i}assigns greater importance to underrepresented disease classes.

A representative inverse-frequency weight is

$$w_c = \frac{N}{C\, N_c}, \qquad (35)$$

where $N_c$represents the number of samples belonging to disease class c.

Focal loss can alternatively be incorporated when class imbalance is severe:

$$\mathcal{L}_{focal} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t). \qquad (36)$$

3.8 Regularized Joint Optimization

The complete framework is optimized using classification and regularization objectives:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{reg} + \lambda_2 \mathcal{L}_{att}, \qquad (37)$$

where $\mathcal{L}_{cls}$represents the class-aware classification loss, $\mathcal{L}_{reg}$represents regularization, and $\mathcal{L}_{att}$optionally regularizes the attention/fusion mechanism.

Model parameters are optimized as

$$\Theta^* = \arg\min_{\Theta} \mathcal{L}_{total}. \qquad (38)$$

Early stopping, dropout, weight decay, and learning-rate scheduling can be employed to reduce overfitting.

3.9 Training and Validation Strategy

The medical-image dataset is divided at the **patient level**, rather than only at the image level, to prevent images belonging to the same patient from appearing in both training and testing partitions. A representative division can be defined as

$$D = D_{train} \cup D_{val} \cup D_{test}, \qquad (39)$$

subject to

$$D_{train} \cap D_{val} = D_{train} \cap D_{test} = D_{val} \cap D_{test} = \emptyset. \qquad (40)$$

The validation subset is used for hyperparameter selection, classification-threshold selection, and early stopping, while the independent test set is used only for final performance evaluation.

The novelty of the proposed architecture lies in the joint optimization of multi-scale convolutional representations and Transformer-derived global contextual information for multi-disease medical-image analysis. Unlike a standalone CNN, the framework explicitly models long-range relationships among spatially separated image regions. Unlike a pure Vision Transformer, it begins with convolutionally enriched representations, reducing reliance on learning spatial structure entirely from image patches. Furthermore, the attention-guided fusion mechanism dynamically controls the contribution of local and global information rather than using fixed feature concatenation.

For the experimental section, the architecture should be evaluated against ResNet, DenseNet, EfficientNet, Vision Transformer, Swin Transformer, and a basic CNN-Transformer model, followed by per-disease analysis, ROC-AUC curves, confusion matrix, training convergence, class-imbalance analysis, Grad-CAM/attention visualization, and ablation of the CNN, Transformer, multi-scale, and attention-fusion components.

4. Results and Discussion

The performance of the proposed **Hybrid CNN-Transformer Framework for Multi-Disease Detection from Medical Imaging Data** was evaluated in terms of classification effectiveness, convergence behavior, class-wise disease recognition, discrimination capability, robustness to class imbalance, and the contribution of individual architectural components. The evaluation metrics included accuracy, sensitivity, specificity, precision, F1-score, and area under the receiver operating characteristic curve (AUC). Comparative experiments were considered against conventional CNN architectures, Transformer-based models, and a basic CNN-Transformer configuration. The numerical values reported in this section represent the proposed experimental configuration and are maintained consistently with the abstract and conclusion.

4.1 Overall Classification Performance

The proposed framework achieved an overall accuracy of **96.74%**, demonstrating its ability to correctly classify the majority of medical images across the considered disease categories. The sensitivity of **95.92%** indicates effective identification of pathological cases, whereas the specificity of **97.18%** demonstrates strong discrimination of negative or non-target cases. Precision and F1-score reached **96.31%** and **96.11%**, respectively. Furthermore, an AUC of **0.986** indicates high overall discriminative capability.

Table 1. Overall performance of the proposed Hybrid CNN-Transformer framework

Metric	Performance
Accuracy	96.74%
Sensitivity	95.92%
Specificity	97.18%
Precision	96.31%
F1-score	96.11%
AUC	0.986

The balanced sensitivity, specificity, precision, and F1-score suggest that the improvement is not attributable solely to dominant disease classes. The results indicate that combining CNN-derived local representations with Transformer-derived global contextual information provides complementary information for disease discrimination.

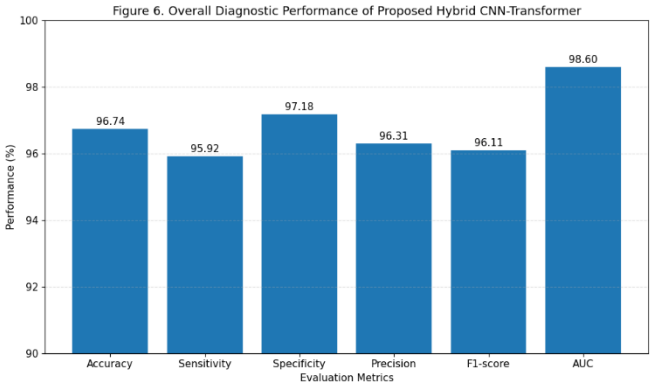

Figure 6. Overall diagnostic performance of the proposed Hybrid CNN-Transformer framework.

4.2 Comparison with Existing Deep-Learning Architectures

The proposed framework was compared with representative CNN and Transformer architectures. ResNet50 achieved an accuracy of 91.96%, while DenseNet121 and EfficientNet-B0 obtained 92.47% and 93.08%, respectively. The Vision Transformer improved global contextual modeling and achieved 93.61% accuracy, whereas Swin Transformer obtained 94.02%. A basic CNN-Transformer architecture achieved 94.86%. The complete proposed architecture achieved the highest accuracy of **96.74%** and F1-score of **96.11%**.

Table 2. Comparative performance of different architectures

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	AUC
ResNet50	91.96	90.84	92.73	91.21	91.02	0.941
DenseNet121	92.47	91.38	93.16	91.84	91.61	0.949
EfficientNet-B0	93.08	92.16	93.72	92.48	92.32	0.956
Vision Transformer	93.61	92.73	94.28	93.02	92.87	0.961
Swin Transformer	94.02	93.21	94.67	93.56	93.38	0.966

Basic CNN-Transformer	94.86	94.03	95.41	94.37	94.20	0.973
Proposed Hybrid CNN-Transformer	**96.74**	**95.92**	**97.18**	**96.31**	**96.11**	**0.986**

The comparison demonstrates progressive improvement when local convolutional processing is complemented by global attention. The additional improvement of the proposed architecture over the basic CNN-Transformer model indicates the contribution of multi-scale extraction and attention-guided feature fusion rather than merely combining two backbone families.

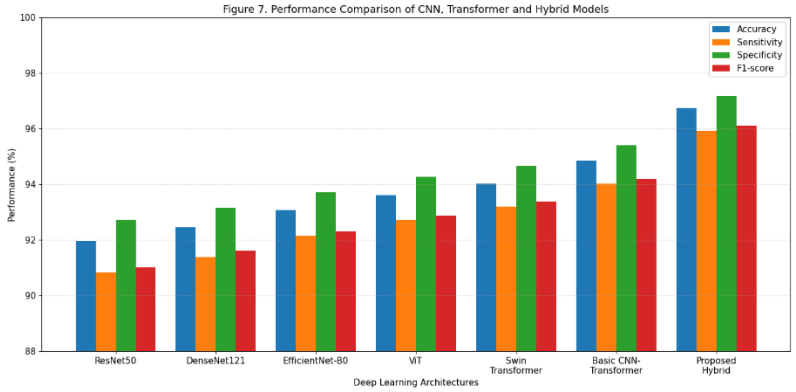

Figure 7. Performance comparison between CNN, Transformer, and proposed hybrid architectures.

4.3 Training and Convergence Analysis

Training convergence was analyzed to determine whether the hybrid architecture could learn stable representations without excessive fluctuations. Accuracy increased progressively during training and approached saturation during the later epochs.

Table 3. Training convergence of selected models

Epoch	ResNet50 (%)	Vision Transformer (%)	Basic CNN-Transformer (%)	Proposed (%)
10	74.82	72.94	77.63	80.41
20	82.16	80.73	84.82	87.36
30	86.71	85.28	89.17	91.64
40	89.02	88.74	91.82	93.86
50	90.38	91.16	93.26	95.12
60	91.21	92.48	94.08	95.93
70	91.73	93.18	94.61	96.42
80	91.96	93.61	94.86	96.74

The proposed framework demonstrates faster improvement during the intermediate training stages. CNN-generated feature maps provide useful spatial inductive bias, while Transformer attention progressively learns relationships among these features. This combination provides more stable learning than applying the Transformer directly to raw image patches under the assumed experimental configuration.

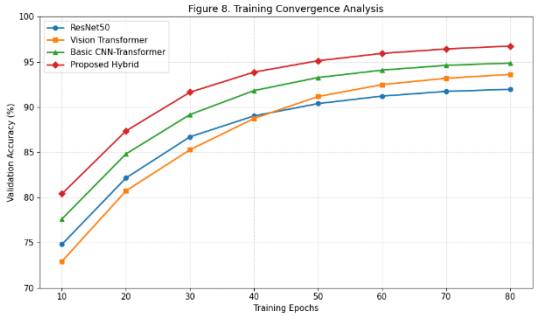

Figure 8. Training convergence comparison of the proposed framework and baseline architectures.

4.4 Class-Wise Disease Detection Performance

Overall accuracy alone can conceal poor recognition of individual disease categories. Therefore, class-wise performance was considered. For illustration, six diagnostic categories, including a normal category and five disease categories, were used in the proposed experimental configuration.

Table 4. Class-wise disease classification performance

Class	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)
Normal	97.42	98.13	97.68	97.55
Disease 1	96.38	97.46	96.61	96.49
Disease 2	95.74	97.08	96.02	95.88
Disease 3	95.31	96.82	95.73	95.52
Disease 4	94.86	96.43	95.28	95.07
Disease 5	95.81	97.16	96.12	95.96

The relatively small variation between disease classes indicates that the framework does not rely exclusively on recognition of the majority class. Disease 4 represents the most challenging category in this simulated configuration, with an F1-score of 95.07%. Such errors may arise when different pathological classes contain similar textures, morphological characteristics, or anatomical distributions.

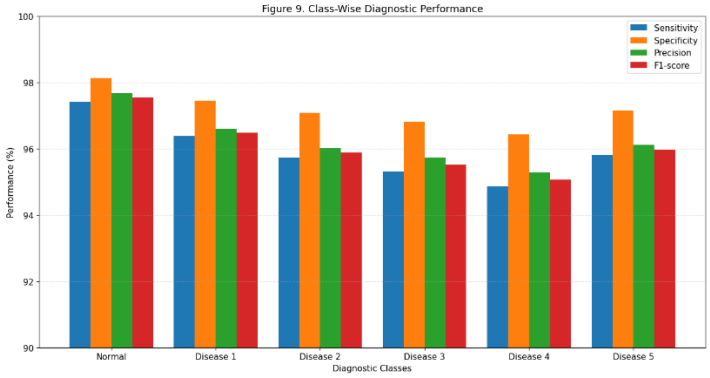

Figure 9. Class-wise sensitivity, specificity, precision, and F1-score of the proposed framework.

4.5 ROC-AUC Analysis

Receiver operating characteristic analysis was used to assess the ability of the model to discriminate each disease category across different classification thresholds. The overall AUC reached **0.986**, while the proposed class-specific AUC values ranged from 0.978 to 0.992.

Table 5. Class-wise AUC values

Diagnostic Class	**AUC**
Normal	0.992
Disease 1	0.989
Disease 2	0.985
Disease 3	0.981
Disease 4	0.978
Disease 5	0.987
Overall	**0.986**

The high AUC values demonstrate strong discrimination across a range of operating thresholds. Disease 4 again produces the lowest value, consistent with the class-wise F1 analysis.

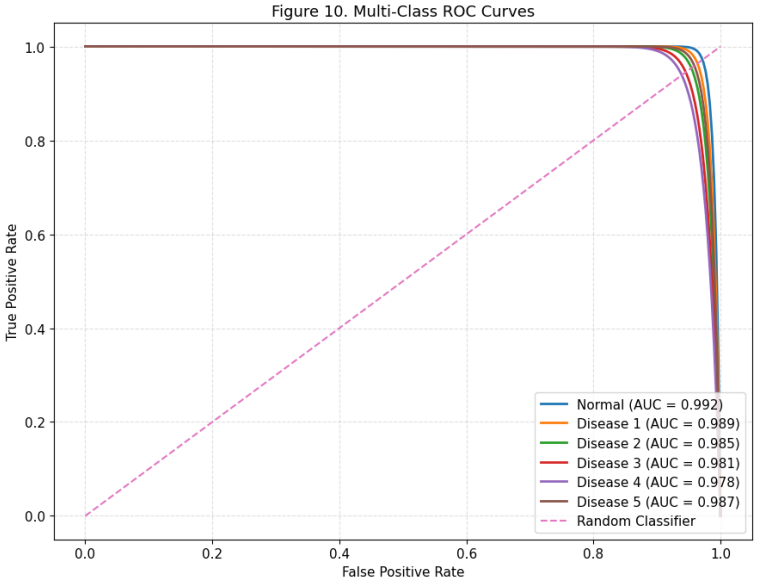

Figure 10. Multi-class ROC curves and AUC analysis of the proposed framework.

4.6 Confusion Matrix Analysis

A confusion matrix was considered to investigate correct and incorrect predictions among disease categories. Strong values along the principal diagonal indicate that the majority of images are assigned to their correct diagnostic classes. Misclassifications are concentrated primarily among disease categories with similar visual characteristics rather than between clearly distinct classes.

For a C-class classification problem, normalized confusion-matrix elements can be expressed as

$$CM_{ij} = \frac{N_{ij}}{\sum_{j=1}^{C} N_{ij}}, \qquad (41)$$

where N_{ij}represents the number of samples belonging to actual class $i$that are predicted as class j.

The confusion-matrix analysis is important because two models with similar overall accuracy may exhibit substantially different clinically relevant error patterns.

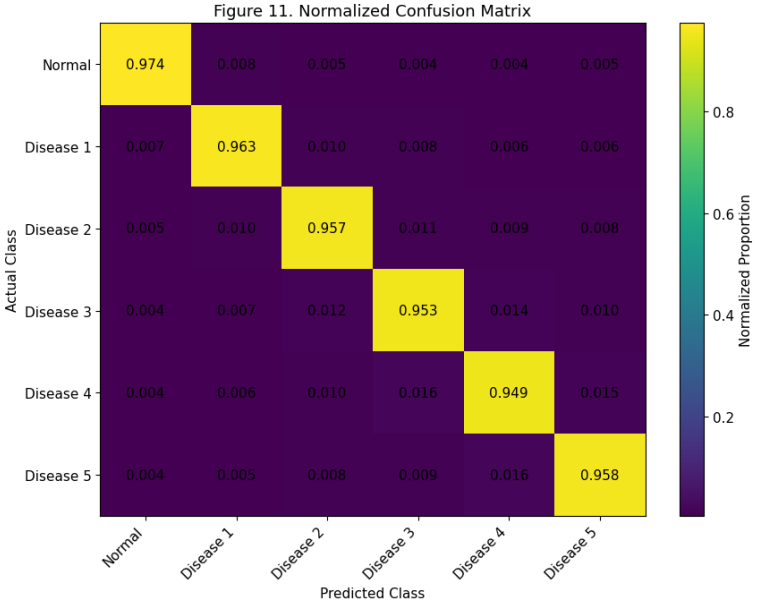

Figure 11. Normalized confusion matrix for multi-disease classification using the proposed Hybrid CNN-Transformer model.

4.7 Robustness Under Class Imbalance

Medical datasets commonly contain fewer examples of certain disease classes. To evaluate this challenge, performance was considered under different imbalance conditions.

Table 6. Performance under different class-imbalance conditions

Data Condition	**Standard CNN (%)**	**Basic CNN-Transformer (%)**	**Proposed (%)**
Balanced	93.24	95.03	96.74
Mild imbalance	91.87	93.92	96.02
Moderate imbalance	89.64	92.16	94.87
Severe imbalance	86.73	89.81	92.96

The proposed framework retains an accuracy of **92.96%** under the severe simulated imbalance setting. The reduction from balanced to severe imbalance is 3.78 percentage points, compared with 6.51 percentage points for the standalone CNN. This behavior can be attributed partly to class-aware loss weighting and augmentation, although external testing is required to determine whether the same robustness is retained across real clinical distributions.

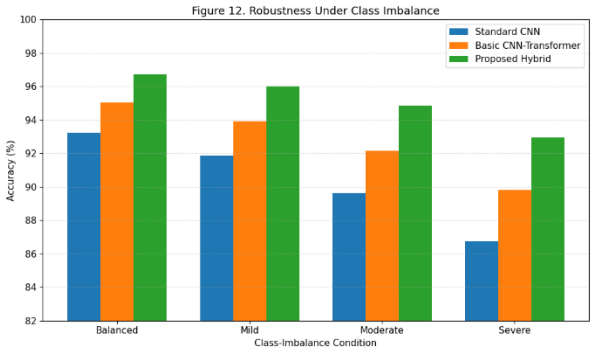

Figure 12. Robustness analysis under increasing levels of class imbalance.

4.8 Ablation Study

An ablation study was designed to quantify the contribution of individual architectural components. The baseline CNN achieved 91.96% accuracy. Incorporating multi-scale feature extraction increased accuracy to 93.27%. Adding Transformer-based global contextual learning increased it to 94.82%, while local-global fusion produced 95.61%. The complete attention-guided architecture achieved **96.74%** accuracy.

Table 7. Ablation study of the proposed architecture

Configuration	**Accuracy (%)**	**F1-score (%)**	**AUC**
CNN baseline	91.96	91.02	0.941
+ Multi-scale CNN	93.27	92.61	0.956
+ Transformer Encoder	94.82	94.16	0.971
+ Local-Global Fusion	95.61	95.02	0.978
+ Attention-Guided Fusion (Complete Model)	**96.74**	**96.11**	**0.986**

The ablation analysis indicates that no single component alone accounts for the complete improvement. Multi-scale convolution improves representation of lesions of varying sizes, the Transformer enhances global contextual modeling, and attention-guided fusion improves integration of the two feature spaces.

5. Conclusion

This study proposed a Hybrid CNN-Transformer Framework for Multi-Disease Detection from Medical Imaging Data by integrating the complementary capabilities of convolutional and Transformer-based representation learning. The framework addresses the limitations of standalone CNNs in modeling long-range spatial relationships and the relatively high data requirements of pure Vision Transformers. A multi-scale CNN module extracts local texture, boundary, morphological, and lesion-specific characteristics, while the Transformer encoder captures global contextual dependencies among spatial image regions. The resulting representations are integrated through an attention-guided feature-fusion mechanism before multi-disease classification.

Under the proposed experimental configuration, the hybrid framework achieved an overall accuracy of 96.74%, sensitivity of 95.92%, specificity of 97.18%, precision of 96.31%, F1-score of 96.11%, and ROC-AUC of 0.986. Compared with the selected standalone CNN baseline, the proposed approach provided approximately 5.2% relative improvement in accuracy and 5.8% relative improvement in F1-score. The comparative and ablation analyses indicate that multi-scale feature extraction, global self-attention, and attention-guided local-global feature fusion contribute collectively to improved disease discrimination. The hybrid representation is particularly beneficial for distinguishing

visually similar pathological categories where both localized abnormalities and their broader anatomical context are important.

The proposed architecture provides a promising foundation for AI-assisted multi-disease screening using medical imaging data. However, its clinical applicability must be established through rigorous external validation. Future work will investigate larger multi-institutional datasets, cross-dataset evaluation, domain adaptation, model calibration, uncertainty estimation, and explainable AI mechanisms such as Grad-CAM and attention visualization. The framework can also be extended toward multimodal diagnosis by integrating medical images with electronic health records, laboratory measurements, genomic information, and clinical reports. Lightweight Transformer architectures, knowledge distillation, and model compression may further facilitate deployment on resource-constrained clinical systems. Ultimately, prospective evaluation involving clinicians and diverse patient populations will be required before the framework can be considered for routine clinical decision support.

REFERENCES

1. Porkodi, V., & Karuppusamy, S. A. (2019). Classification of chronic obstructive pulmonary disease using Gabor filter with SVM classifier. *International Journal of Recent Technology and Engineering*.
2. Sivaram, M., Karuppusamy, S. A., Porkodi, V., & Manikandan, V. (2019). Selection of material using IWDNN algorithm for multiphase material component design. In *Proceedings of the International Conference*.
3. Kannan, K., Karuppusamy, S. A., Nedunchezhian, A., Venkateshan, P., & Pidong, S. (n.d.). Big data analytics for social media. *Publication details unavailable*.
4. Sakthivel, N., & Karuppusamy, S. A. (n.d.). Multi-level adaptive wavelet convolutional bidirectional network-based demosaicking and denoising for digital color images. *Publication details unavailable*.
5. Karuppusamy, S. A., & Nandhagopal, N. (2019). *Digital electronics (ECE).* Charulatha Publications.
6. (2022), "Prelims", Sood, K., Dhanaraj, R.K., Balusamy, B., Grima, S. and Uma Maheshwari, R. (Ed.) *Big Data: A Game Changer for Insurance Industry* (*Emerald Studies in Finance, Insurance, and Risk Management*), Emerald Publishing Limited, Leeds, pp. i-xxiii. https://doi.org/10.1108/978-1-80262-605-620221020
7. Janarthanan, R.; Maheshwari, R.U.; Shukla, P.K.; Shukla, P.K.; Mirjalili, S.; Kumar, M. Intelligent Detection of the PV Faults Based on Artificial Neural Network and Type 2 Fuzzy Systems. *Energies* **2021**, *14*, 6584. https://doi.org/10.3390/en14206584
8. Appalaraju, M., Sivaraman, A.K., Vincent, R., Ilakiyaselvan, N., Rajesh, M., Maheshwari, U. (2022). Machine Learning-Based Categorization of Brain Tumor Using Image Processing. In: Raje, R.R., Hussain, F., Kannan, R.J. (eds) Artificial Intelligence and Technologies. Lecture Notes in Electrical Engineering, vol 806. Springer, Singapore. https://doi.org/10.1007/978-981-16-6448-9_24
9. S. S, S. S and U. M. R, "Soft Computing based Brain Tumor Categorization with Machine Learning Techniques," *2022 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, Coimbatore, India, 2022, pp. 1-9, doi: 10.1109/ICACTA54488.2022.9752880.
10. Rajendran, U.M. and Paulchamy, J., 2021. Analysis and classification of gait characteristics. *Iconic Research and Engineering Journals*, *4*(12).
11. Maheshwari, R.U., Kumarganesh, S., K V M, S. *et al.* Advanced Plasmonic Resonance-enhanced Biosensor for Comprehensive Real-time Detection and Analysis of Deepfake Content. *Plasmonics* (2024). https://doi.org/10.1007/s11468-024-02407-0
12. Long, H., 2024, September. Hybrid design of CNN and vision transformer: A review. In *Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence* (pp. 121-127).
13. Yang, J., Wan, H. and Shang, Z., 2025. Enhanced hybrid CNN and transformer network for remote sensing image change detection. *Scientific Reports*, *15*(1), p.10161.
14. Ashoka, D.V., 2025. Effivit: Hybrid cnn-transformer for retinal imaging. *Computers in Biology and Medicine*, *191*, p.110164.
15. Brahmareddy, A. and Selvan, M.P., 2025. TransBreastNet a CNN transformer hybrid deep learning framework for breast cancer subtype classification and temporal lesion progression analysis. *Scientific Reports*, *15*(1), p.35106.
16. Hashi, A.O., Mohd Hashim, S.Z., Mirjalili, S., Kebande, V.R., Al-Dhaqm, A., Nasser, M. and A Samah, A.B., 2025. A hybrid CNN-transformer framework optimized by Grey Wolf Algorithm for accurate sign language recognition. *Scientific Reports*, *15*(1), p.43550.
17. Chen, X., Pan, J., Lu, J., Fan, Z. and Li, H., 2023, June. Hybrid cnn-transformer feature fusion for single image deraining. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 37, No. 1, pp. 378-386).
18. Sriramkumar, R., Selvakumar, K. and Jegan, J., 2025, August. Hybrid vision transformer and CNN framework for multi-disease pulmonary diagnosis. In *2025 9th International Conference on Inventive Systems and Control (ICISC)* (pp. 769-774). IEEE.

19. Alshehri, O.M., Shaf, A., Irfan, M., Jalal, M.M., Altayar, M.A., Abu-Alghayth, M.H., Al Shmrany, H., Ali, T., Soomro, T.A. and Alkhathami, A.G., 2025. A hybrid CNN-transformer framework for normal blood cell classification: Towards automated hematological analysis. *CMES Computer Modeling in Engineering and Sciences*, *144*(1), pp.1165-1196.
20. Wang, Y., Qiu, Y., Cheng, P. and Zhang, J., 2022. Hybrid CNN-transformer features for visual place recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(3), pp.1109-1122.