Journal of Imaging Informatics in Medicine (2025) 38:3928–3971
https://doi.org/10.1007/s10278-025-01481-y

REVIEW

Vision Transformers in Medical Imaging: a Comprehensive Review of Advancements and Applications Across Multiple Diseases

Sanad Aburass[1] · Osama Dorgham[2] · Jamil Al Shaqsi[3] · Maha Abu Rumman[2] · Omar Al-Kadi[4]

Received: 27 December 2024 / Revised: 15 February 2025 / Accepted: 11 March 2025 / Published online: 31 March 2025

Abstract

The rapid advancement of artificial intelligence techniques, particularly deep learning, has transformed medical imaging. This paper presents a comprehensive review of recent research that leverage vision transformer (ViT) models for medical image classification across various disciplines. The medical fields of focus include breast cancer, skin lesions, magnetic resonance imaging brain tumors, lung diseases, retinal and eye analysis, COVID-19, heart diseases, colon cancer, brain disorders, diabetic retinopathy, skin diseases, kidney diseases, lymph node diseases, and bone analysis. Each work is critically analyzed and interpreted with respect to its performance, data preprocessing methodologies, model architecture, transfer learning techniques, model interpretability, and identified challenges. Our findings suggest that ViT shows promising results in the medical imaging domain, often outperforming traditional convolutional neural networks (CNN). A comprehensive overview is presented in the form of figures and tables summarizing the key findings from each field. This paper provides critical insights into the current state of medical image classification using ViT and highlights potential future directions for this rapidly evolving research area.

Keywords Vision transformers · Medical image analysis · Convolutional neural networks · Medical image classification · Clinical decision support

Introduction

In a world where medical imaging serves as the linchpin of modern diagnostics, harnessing the full potential of each pixel can be a matter of life and death. Convolutional neural networks, which are the main type of deep learning, have already shown that they can change this field by getting to levels of accuracy that have never been seen before, yet as we stand on the brink of a new era, vision transformers are revolutionizing the field, challenging conventional wisdom and setting new standards for performance [1–3]. One type of deep learning model that is often used in computer vision tasks is the CNN, which sorts images into groups and finds objects in them, CNNs work by convolving the input image with filters to extract local features and then aggregating these features through pooling layers, CNNs, on the other hand, have flaws, like features that are only available in set sizes and the failure to record world context information well [4–8]. ViTs are a new type of deep neural networks that have shown promise as an option to CNNs for picture-taking tasks; ViTs are inspired by the transformer architecture originally introduced for natural language processing (NLP) tasks. The

✉ Sanad Aburass
saburass@luther.edu

Osama Dorgham
o.dorgham@bau.edu.jo

Jamil Al Shaqsi
alshaqsi@squ.edu.om

Maha Abu Rumman
maha.aburumman@bau.edu.jo

Omar Al-Kadi
o.alkadi@ju.edu.jo

1 Department of Computer Science, Luther College, Decorah, IA, USA

2 Prince Abdullah Bin Ghazi Faculty of Information and Communication Technology, Al-Balqa Applied University, Al-Salt, Jordan

3 Information Systems Department, Sultan Qaboos University, Seeb, Oman

4 Artificial Intelligence Department, King Abdullah II School of Information Technology, University of Jordan, Amman 11942, Jordan

transformer design is based on self-attention processes that let the network decide on the fly how important each trait in the input chain is and gather information about the overall context [3, 9]. This makes ViTs well suited for image analysis tasks where global context information is important [10]. Vision transformers work by converting an input image into a set of fixed-length feature representations, also known as tokens, the image is divided into non-overlapping square patches, and each patch is treated as an individual feature. The tokens are then processed through multi-head self-attention mechanisms, which dynamically weigh the importance of each token in the input sequence. This allows the network to capture global context information and focus on different parts of the image based on the context [6, 10]. ViTs' main parts are tokenization, self-sttention mechanisms, feed-forward networks, and pooling and classification layers [11]. Tokenization in vision transformers is the process of breaking down a picture into a set of fixed-length feature representations, also known as tokens, which can be processed by a transformer-based model. This is typically done by encoding the image into a multi-dimensional feature map and then flattening this map into a one-dimensional vector, this vector is then passed through a linear layer to produce the final set of token representations. Tokenization is a necessary preprocessing step in vision transformers because it lets the model work with pictures. Self-attention mechanisms are a key component in the architecture of vision transformers, which are neural networks designed to process visual information, in self-attention, each element in a sequence (e.g., an image) attends to every other element in that sequence, weighing its importance in the computation of the final representation. This allows the network to dynamically focus on different parts of the image based on context and capture long-range dependencies [12]. In vision transformers, self-attention is applied over the spatial dimensions of the image, rather than the temporal dimension as in traditional transformer models. The result is a more powerful and efficient network that can handle large and complex images and is well suited for a variety of computer vision tasks, including image classification and object detection. Feed-forward networks (FFNs) are another key component in the architecture of vision transformers, which are neural networks designed to process visual information. FFNs are used to transform the intermediate representations produced by the self-attention mechanisms into a final output. The FFN consists of a series of linear transformations followed by non-linear activation functions and can be thought of as a dense layer in a traditional convolutional neural network. In vision transformers, the FFN is typically used to map the high-dimensional representations produced by the self-attention layers to a lower-dimensional space, suitable for the final prediction task (e.g., image classification). Vision transformers can get more accurate results when they use both self-attention and FFN layers together. This is because they can take in both global and local information in a picture. The pooling layer and the classification layer are used for down sampling and final prediction in vision transformers, the pooling layer is typically applied after the transformer blocks to reduce the spatial resolution of the feature maps, this helps to reduce the computational complexity and increase the robustness of the model to small translations, common pooling operations include max-pooling and average pooling. The classification layer is a fully connected layer with a SoftMax activation used to make the final prediction, the inputs to the classification layer are the flattened feature maps obtained after pooling, and the outputs are the class probabilities, the number of outputs is equal to the number of classes in the classification task. Together, the pooling and classification layers make up the last part of a vision transformer's prediction process. They take a picture as input and make a guess about the things that are in it. CNNs and ViTs are two popular architectures for computer vision tasks such as image classification. ViTs have been applied to a wide array of computer vision tasks, demonstrating their versatility and efficacy. They have achieved state-of-the-art performance in image classification on benchmark datasets such as ImageNet. In object detection, they have incorporated object-level attention mechanisms, effectively transforming the challenge into a token-level classification problem. Additionally, ViTs have made strides in image segmentation by implementing segmentation-specific attention mechanisms, further showcasing their adaptability and potential. The main difference between vision transformers and CNNs lies in their architecture and how they process the input image. CNNs extract local features through convolution and pooling operations and are limited by their reliance on fixed-sized local features. On the other hand, vision transformers use self-attention mechanisms to dynamically weigh the importance of each feature in the input sequence, allowing them to effectively capture global context information [4, 6, 13, 14]. Another difference between ViTs and CNNs is the way they handle large images, CNNs typically use pooling layers to reduce the spatial dimension of the input, which can result in the loss of essential information. Vision transformers, on the other hand, use self-attention mechanisms to dynamically focus on distinct parts of the image, making them well suited for handling large images.

While several review papers have examined the role of vision transformers in medical imaging, most existing surveys focus on general architectural discussions or applications within a limited set of medical tasks, lacking a comprehensive disease-centric perspective. This paper distinguishes itself by adopting a structured, application-driven approach, categorizing studies based on specific medical conditions such as breast cancer classification, skin lesion analysis, MRI brain tumor analysis, lung diseases, and retinal disorders. By systematically grouping research based on medical application areas rather than a general technical perspective,

this review provides a targeted and practical resource for researchers and clinicians, allowing them to navigate advancements relevant to their specific fields. Furthermore, we analyze key aspects such as data preprocessing, model architectures, performance comparisons, and the integration of ViTs with other deep learning techniques. By offering an intuitive synthesis of ViT advancements across multiple clinical domains, this paper not only highlights their transformative potential in medical imaging but also identifies current challenges and future directions, serving as a critical reference for the ongoing development and optimization of ViT-based medical imaging solutions.

To ensure a rigorous and comprehensive review of vision transformers in medical imaging, we established clear inclusion and exclusion criteria for selecting the surveyed papers. Our selection process considered peer-reviewed articles published in reputed journals and conferences, focusing on studies that applied ViTs to medical image classification across various disease domains. We prioritized works that reported quantitative performance metrics, used publicly available or clinically relevant datasets, and demonstrated methodological innovations. Studies that lacked sufficient experimental validation, were primarily theoretical, or provided only a brief mention of ViT applications without in-depth analysis were excluded. Despite the broad scope of ViT research, we aimed for a balanced and representative selection, acknowledging that new publications continue to emerge rapidly in this field.

Vision Transformer

In the following section, we will dive deep into ViTs and try to understand them from a mathematical standpoint, but before we move ahead, let us understand the basics.

Overview of Vision Transformers

Transformers were initially developed for natural language processing tasks; however, they have proven highly efficient in other domains like computer vision as well. A ViT is such a model that applies the Transformer architecture to computer vision tasks, treating images as sequences of pixels akin to a sequence of words in a sentence. This gives it a capability to model long-range dependencies between pixels and learn from global context [15–17]. Figure 1 shows the ViT architecture.

Detailed Exploration

Patch Embedding

ViT starts by splitting the input image into a fixed number of patches, typically using non-overlapping patches. Mathematically, suppose we have an input image "x" $\in \mathbb{R}^{(H \times W \times C)}$ where H, W, and C are the height, width, and the number of channels of the image, respectively. This image is divided into "(P1 × P2)" patches each of size "($S \times S \times$ C)." Here, "P1 = H/S" and "P2 = W/S" are the number of patches along the height and width of the image respectively. Each patch is then flattened and transformed to a D-dimensional patch embedding using a learnable linear transformation. This is equivalent to a fully connected layer, represented as follows:

$$E \in \mathbb{R}^{(N \times D)} = \text{patch matrix } P \in \mathbb{R}^{(N \times S2 \times C)} \times W \in \mathbb{R}^{(S2 \times C \times D)} \quad (1)$$

Here, $N = \text{P1} \times \text{P2}$ is the total number of patches, D is the dimension of the embeddings, and W is the learnable weight matrix.

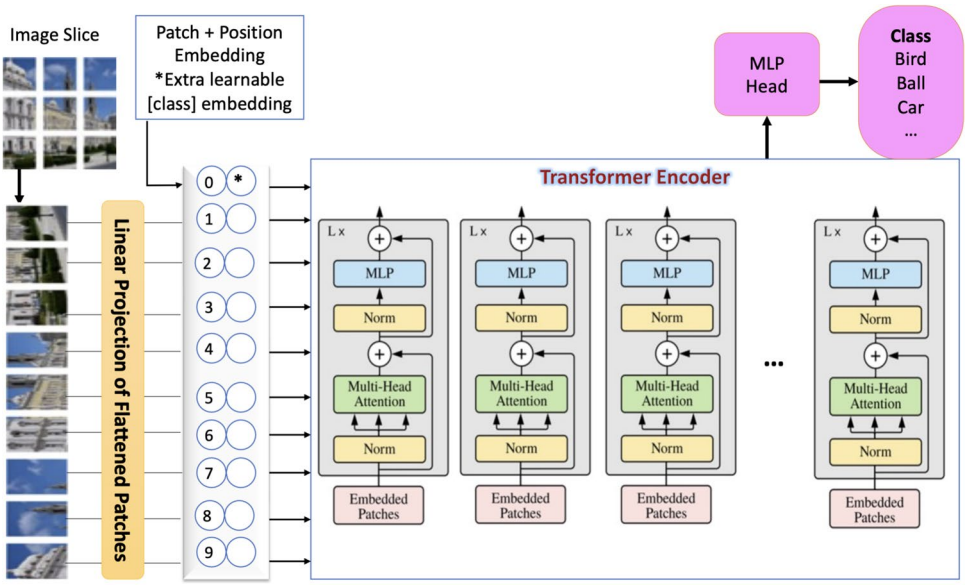


Fig. 1 The vision transforms architecture

Positional Encoding

Given that transformers do not inherently understand the sequential order of data, positional encoding is added to the patch embeddings to retain positional information about the patches. If $PE \in \mathbb{R}^{(N \times D)}$ is the positional encoding matrix, the final input to the transformer $Z \in \mathbb{R}^{(N \times D)}$ is given by:

$$Z = E + \mathrm{PE} \tag{2}$$

Transformer Encoder

The transformer encoder consists of multiple identical layers with each layer having two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. There is a residual connection around each of the two sub-layers followed by layer normalization. Let us consider the i^{th} layer and let Z_i denote the input to this layer. The output $Z_{(i+1)}$ of this layer is calculated as follows:

1. Self-Attention: First, we compute three matrices: Query (Q), Key (K), and Value (V) as follows:

$$Q = Z_i \times W_Q \tag{3}$$

$$K = Z_i \times W_K \tag{4}$$

$$V = Z_i \times W_V \tag{5}$$

Here, W_Q, W_K, and W_V are learnable weight matrices. The output of the attention mechanism A is given by:

$$\mathrm{A} = \mathrm{softmax}\left(\frac{Q_K^T}{\sqrt{D}}\right) \times \mathrm{V} \tag{6}$$

The division by sqrt(D) is done for scaling purposes.

2. Feed-forward network: It consists of two linear transformations with a ReLU activation in between:

$$FFN = ReLU(A \times W_1 + b_1) \times W_2 + b_2 \tag{7}$$

Here, W_1, W_2, b_1, and b_2 are learnable parameters.

3. Add and Normalize: Finally, layer normalization is applied on the sum of the input and the output of each sub-layer:

$$Z_{(i+1)} = LayerNorm(Z_i + A) \tag{8}$$

$$Z_{(i+1)} = LayerNorm(Z_{(i+1)}x + FFN) \tag{9}$$

Classification Head

The first output token Z_0 is taken as the representation of the whole image, and it is passed through a classification head which is a simple linear layer followed by a softmax activation function. If $W_{cls} \in \mathbb{R}^{(D \times num\_classes)}$ are the learnable weights of the linear layer, the final output probabilities $p \in \mathbb{R}^{num\_classes}$ is given by:

$$p = SoftMax(Z_0 \times W_{cls}) \tag{10}$$

This is a high-level overview of the vision transformer model. Its ability to process non-local information in images makes it a powerful model for medical image analysis where the diagnosis often relies on global features.

Efficient Vision Transformers

The emergence of ViTs has proven to be a groundbreaking development in the field of computer vision, while the original ViT architecture offers substantial performance advantages over its CNN counterparts, multiple variants have since been developed to cater to specific needs, computational constraints, and data limitations. The goal of this part is to put light on these versions by breaking down their designs, looking at how they vary from the original ViT, and talking about hyperparameter fine-tuning techniques that can be used to get the best performance.

DeiT (Data-Efficient Image Transformer)

Starting with the DeiT, one observes a focus on data efficiency while largely adhering to the vanilla ViT architecture, what sets the DeiT apart is its use of teacher-student distillation, here, a larger teacher model assists a more compact student model in learning. This method lets DeiT get great results even when working with smaller labeled datasets. The original ViT does not have this kind of data efficiency because it usually needs a lot of data to work at its best. When tuning hyperparameters, you may want to focus on improving the knowledge distillation parameters to get the most out of the DeiT architecture [18].

DeiT follows the standard Transformer-based architecture for image classification as introduced in ViT. Given an input image $I \in R^{H \times W \times C}$, it is divided into N non-overlapping patches of size $P \times P$, forming a sequence of flattened patch embeddings:

$$\mathrm{X} \in \mathrm{R}^{\mathrm{N} \times \mathrm{D}}, \mathrm{N} = \frac{H \times W}{P^2} \tag{11}$$

Each patch is linearly projected into an embedding space of dimension D, producing X. The model further appends a

learnable class token x_{cls} and employs positional encodings E_{pos} to preserve spatial information:

$$Z_0 = [x_{cls}, X] + E_{pos}, Z_0 \in \mathrm{R}^{(N+1)\times D} \tag{12}$$

This sequence is processed through L transformer blocks, where each block consists of Multi-Head Self-Attention (MSA) and a feed-forward network (FFN):

$$\begin{aligned} Z' &= \mathrm{MSA}(\mathrm{LN}(Z_{l-1})) + Z_{l-1} \\ Z_l &= \mathrm{FFN}(\mathrm{LN}(Z')) + Z' \end{aligned} \tag{13}$$

where layer normalization (LN) is applied before each sublayer.

Finally, the class token representation Z_L^{cls} is passed through a linear classifier for prediction.

Knowledge Distillation in DeiT To improve training efficiency, DeiT employs a novel distillation token $\mathrm{x}_{\text{distill}}$ that interacts with the patch and class tokens through self-attention. This mechanism replaces standard distillation, ensuring the model learns from a teacher network (typically a CNN like RegNet) using hard-label distillation:

$$\mathrm{L}_{\text{hardDistill}} = \frac{1}{2}\mathrm{L}_{\text{CE}}(\mathrm{p_s}, \mathrm{y}) + \frac{1}{2}\mathrm{L}_{\text{CE}}(\mathrm{p_s}, \mathrm{y_t}) \tag{14}$$

where:

LCE is the cross-entropy loss,

$\mathrm{p_s}$ Softmax (Z_L^{cls})is the student prediction,

y is the ground-truth label,

yt arg maxc Zt(c) is the hard label from the teacher

Alternatively, soft-label distillation is also considered using Kullback–Leibler (KL) divergence:

$$\mathrm{L}_{\text{softDistill}} = (1-\lambda)\mathrm{L}_{\text{CE}}(\mathrm{p_s}, \mathrm{y}) + \lambda\tau^2\mathrm{D}_{\text{KL}}(p_s^{\tau} \parallel p_t^{\tau}) \tag{15}$$

where p_s^{τ}=Softmax($\mathrm{Z_s}/\tau$) and p_t^{τ}=Softmax($\mathrm{Z_t}/\tau$), with temperature τ.

The final loss combines both class token and distillation token outputs:

$$L_{total} = L_{CE}(Z_L^{cls}, \mathrm{y}) + L_{CE}(Z_L^{\text{distill}}, \mathrm{y}_t) \tag{16}$$

Training Efficiency and Regularization DeiT employs data augmentation and regularization techniques such as Mixup, CutMix, Random Erasing, and stochastic depth to enhance performance without requiring large datasets. Training uses AdamW optimization with weight decay λ, and cosine learning rate scheduling:

$$\eta_t = \eta_{\max} \times 0.5(1 + \cos(\frac{t}{T}\pi)) \tag{17}$$

where η_t is the learning rate at epoch t, and T is the total number of epochs.

Swin Transformer

The Swin Transformer is another noteworthy variant that distinguishes itself through a hierarchical self-attention mechanism, unlike the original ViT, which applies global self-attention across all patches, Swin Transformer uses window-based, local self-attention. To get a better sense of the whole picture, these small windows are moved around it. This window-based method makes for a more computationally efficient model that can still understand both local and global situations. To fine-tune Swin, you would have to find the best window sizes and moving mechanisms for each job [19].

Computational Complexity of Self-Attention Self-attention in the traditional vision transformer has a quadratic complexity with respect to the number of patches N:

$$O(N^2.d) \tag{18}$$

where d is the feature dimension.

In contrast, Swin Transformer partitions the image into local windows of fixed size M×M. The self-attention is then computed within each window instead of across the entire image. The complexity is:

$$\mathrm{O(hwC^2 + M^2hwC)} \tag{19}$$

where

h, w are the height and width of the image,

C is the channel dimension,

M is the window size.

Since M is fixed, Swin Transformer has linear complexity relative to image size O(N), unlike ViT, which has quadratic complexity.

Shifted Window Mechanism The shifted window approach improves information exchange between neighboring windows. The update rule for tokens in layer l and $l+1$ is:

$$z^{(l+1)} = SW - MSA(LN(z^{(l)})) + z^{(l)} \quad (20)$$

$$z^{(l+2)} = MLP(LN(z^{(l+1)})) + z^{(l+1)} \quad (21)$$

where

SW-MSA is shifted window multi-head self-attention,

LN(·) is Layer Normalization,

MLP(·) is a feed-forward network.

This alternating window mechanism maintains efficiency while allowing global interactions.

Relative Position Bias in Attention The attention mechanism includes a relative positional bias matrix B, added to the self-attention scores:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (22)$$

This enhances spatial relationships while maintaining efficiency.

Hierarchical Feature Representation Swin Transformer constructs a multi-scale representation through patch merging. At each stage, feature maps are downsampled:

$$H_{l+1} = \frac{H_l}{2}, W_{l+1} = \frac{W_l}{2}, C_{l+1} = \frac{C_l}{2} \quad (23)$$

which increases the channel dimension while reducing spatial resolution, allowing Swin Transformer to efficiently model coarse-to-fine representations.

ConViT (Convolutional Vision Transformer)

ConViT offers a unique blend of convolutional and Transformer architectures, while the original ViT depends solely on self-attention mechanisms to process image patches, ConViT introduces convolutional layers in the initial stages of the network. These layers are great at capturing local features, which go well with the global feature capturing abilities of the following Transformer layers. To fine-tune ConViT, one could try increasing the kernel sizes of the convolutional layers and the attention heads of the following Transformer layers to better capture a wider range of features [20].

Gated Positional Self-Attention (GPSA) The ConViT modifies the standard self-attention formulation by incorporating relative positional encodings to introduce soft convolutional biases. The GPSA layer combines content-based and position-based attention through a gating mechanism:

$$A_h^{GPSA} = (1 - \sigma(\lambda h)).\ \text{softmax}(Q_h K_h^T) + \sigma(\lambda_h).\ \text{softmax}(v_h^\top r_{ij}) \quad (24)$$

where

r_{ij} represents relative positional encodings based on the spatial distance between patches i and j,

v_h is a trainable embedding that determines the attention span,

$\sigma(\lambda_h)$ is a sigmoid gating function that interpolates between content-based self-attention and positional attention.

By initializing GPSA layers with strong locality constraints (high σ(λh)), ConViT ensures that early layers behave like convolutional layers. As training progresses, the network gradually reduces σ(λh), allowing it to escape locality constraints and transition into a fully flexible attention mechanism.

The Convolutional Bias ConViT's attention heads can approximate a convolutional kernel of size $\sqrt{N_h}x\sqrt{N_h}$. The initialization enforces locality via:

$$v_h^\top r_\delta = -\alpha h(1, -2\Delta_h^1, -2\Delta_h^2, 0, \ldots, 0) \quad (25)$$

where

$\Delta h = (\Delta_h^1, \Delta_h^2)$ defines the position of the receptive field for head h,

αh is a locality strength parameter controlling the attention spread.

This ensures that at initialization, attention heads act as discrete convolutional filters, allowing ConViT to inherit CNNs' sample efficiency.

LeViT

LeViT aims to bring vision transformers closer to convolutional neural networks in terms of inference speed, without sacrificing much in terms of performance, unlike

the original ViT which solely relies on attention mechanisms, LeViT incorporates convolutional layers in a way that makes it more efficient for real-time inference tasks. It is designed to offer a trade-off between computational efficiency and performance, making it suitable for applications where both speed and accuracy are crucial [21]. By understanding these architectural nuances and fine-tuning strategies, practitioners can make more informed choices about which vision transformer variant to employ, depending on their specific needs and constraints.

Patch Embedding with Convolutional Reduction Instead of using a fixed-size patch embedding as in ViT, LeViT employs a series of 3×3 convolutional layers with stride 2 to reduce the spatial resolution while increasing the number of channels:

$$X_0 = f_\theta(\text{Image}) \tag{26}$$

where $\mathrm{X}_0 \in \mathrm{RC}_0 \times H_0 \times W_0$ is the resulting feature map.

Multi-head Self-Attention with Attention Bias Each LeViT transformer block consists of a self-attention mechanism with a learned per-head translation-invariant bias:

$$A^h_{(\mathrm{x},\mathrm{y}),(x',y')} = Q_{(x,y)} \cdot K^T_{(x',y')} + B^h_{|\mathrm{x}-x'|,|\mathrm{y}-y'|} \tag{27}$$

where

Q, K, and V are the query, key, and value matrices,

$B^h_{|\mathrm{x}-\mathrm{x}'|,|\mathrm{y}-\mathrm{y}\prime|}$ is a learned spatial bias that replaces traditional positional embeddings.

The attention operation is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \tag{28}$$

where d_k is the dimensionality of the key vectors.

Pyramid Pooling for Resolution Reduction LeViT incorporates a multi-stage pyramid structure where resolution is progressively reduced using attention-based downsampling:

$$\mathrm{X}_{l+1} = \mathrm{Attn}(\mathrm{Q}_l, \mathrm{K}_l, \mathrm{V}_l) \tag{29}$$

$$X^{\downarrow}_{l+1} = \text{Downsample}(\mathrm{X}_{l+1}) \tag{30}$$

where the downsampling step reduces spatial dimensions $\mathrm{H}_l \times \mathrm{W}_l \rightarrow \mathrm{H}_{l+1} \times \mathrm{W}_{l+1}$, typically halving the resolution.

Efficient MLP Blocks LeViT modifies the traditional MLP head in transformers by reducing the expansion factor:

$$Y = \sigma(W_2 \cdot \text{Hardswish}(\mathrm{W}_1\mathrm{X})) \tag{31}$$

where

W_1 and W_2 are the weight matrices of the MLP layers,

σ is batch normalization,

Hardswish is used instead of GELU for computational efficiency.

Classification Head Instead of using a class token, LeViT applies global average pooling over the final feature map:

$$\hat{y} = \mathrm{softmax}(\mathrm{W}_{\mathrm{cls}} \cdot \mathrm{GAP}(\mathrm{X}_\mathrm{L})) \tag{32}$$

where GAP is global average pooling.

Recent Advancements in Medical Image Classification

In the section titled, we aim to present a comprehensive review and analysis of the studied papers, focusing on the innovative applications of vision transformer in the classification of medical images across various fields, this section has been systematically organized by the medical field to offer a clearer view of the developments and achievements within each specific area. The wide array of medical fields addressed includes, but is not limited to, breast cancer classification, skin lesion analysis, MRI brain tumor analysis, and coronavirus 2019 classification. Through this synthesis of recent advancements, we aim to explore the latest in medical image classification with ViT and the challenges and solutions in each domain.

Breast Cancer Classification

Gheflati and Rivaz [22] propose a method for breast US image classification based on transfer learning of pre-trained ViT models. They use two different datasets of breast US images and evaluate their results using the metrics of classification accuracy and AUC. The ViT architecture consists of a Transformer encoder with self-attention modules, and the authors also compare their results with CNN-based models. Fine-tuning of the models is done using the Adam optimizer for CNNs and SGD (Stochastic Gradient Descent) with a weighted cross-entropy loss for ViTs. The experiment is

conducted on 85% of the data for training and 15% for testing, with fivefold cross-validation for final evaluation, and they have achieved an accuracy of 86.7 on BUSI + B dataset.

Tummala, Kim, et al. [23] investigated the ability of a variant of vision transformer called SwinT to classify benign and malignant breast cancer subtypes using histopathology images. The researchers used an openly available dataset of 7909 images taken at different zoom levels. The results showed that an ensemble of SwinTs, including varied sizes, had an average test accuracy of 96% for eight-class classification and 99.6% for two-class classification, outperforming previous works. This method could lead to relief for pathologists. Future research could include the development of a real-time decision-support system to assist pathologists in making faster and more accurate diagnoses. Figure 2 shows the proposed approach.

Abimouloud et al. [24] proposed a hybrid TokenMixer architecture that integrates convolutional neural networks and vision transformers to enhance breast cancer classification using histopathology images. Their approach addresses the high computational cost and complexity of ViTs by reducing the number of input patches through convolutional layers while leveraging transformer encoder layers for fast and accurate classification. The BreakHis dataset was used for evaluation, and their model achieved 97.02% accuracy in binary classification and 93.29% accuracy in multi-class classification across different magnification levels (×40, ×100, ×200, ×400). Figure 3 shows the proposed approach.

Wang et al. [25] propose a model for breast cancer detection using two datasets: the breast ultrasound images (BUSI) dataset and the Breast Cancer Histopathological Image Classification (BreakHis) dataset. The BUSI dataset was collected and annotated at the Baheya hospital, with a total of 780 images after preprocessing, in grayscale with a resolution of 1280 × 1024. The BreakHis dataset contains 7909 microscopic images of breast tissue, including 2480 benign and 5429 malignant samples, with 4 magnifying factors, 700 × 460 pixels and stored in PNG format. The method uses an Attention Transformer-ViT (ATS-ViT) model, trained using supervised and consistency training, with three types of losses combined to optimize the neural network. The architecture of the ATS-ViT model is based on the transformer architecture, which uses an attention mechanism to capture semantic meanings and relationships among the input tokens. The experimental results show that they have achieved an accuracy of 98.12% compared to other methods. Future directions include improving the accuracy and generalization ability of the model. Limitations include the need for more labeled data and the difficulty of interpreting the model. Figure 4 shows the proposed approach.

Chen et al. [26] present a study that uses ViTs for breast cancer detection in mammograms. The dataset contains 3796 mammograms from 949 patients, where some were diagnosed as benign, and others were confirmed to have malignant lesions. The mammograms are processed using the multi-view vision transformer (MVT) model, which includes two parts: local transformer blocks and global transformer blocks. The local transformer blocks process information from individual view images while the global blocks process information from all view images together. The model includes self-attention and multi-head attention and is trained using patch and position embeddings. The authors subsampled the original mammograms and reduced

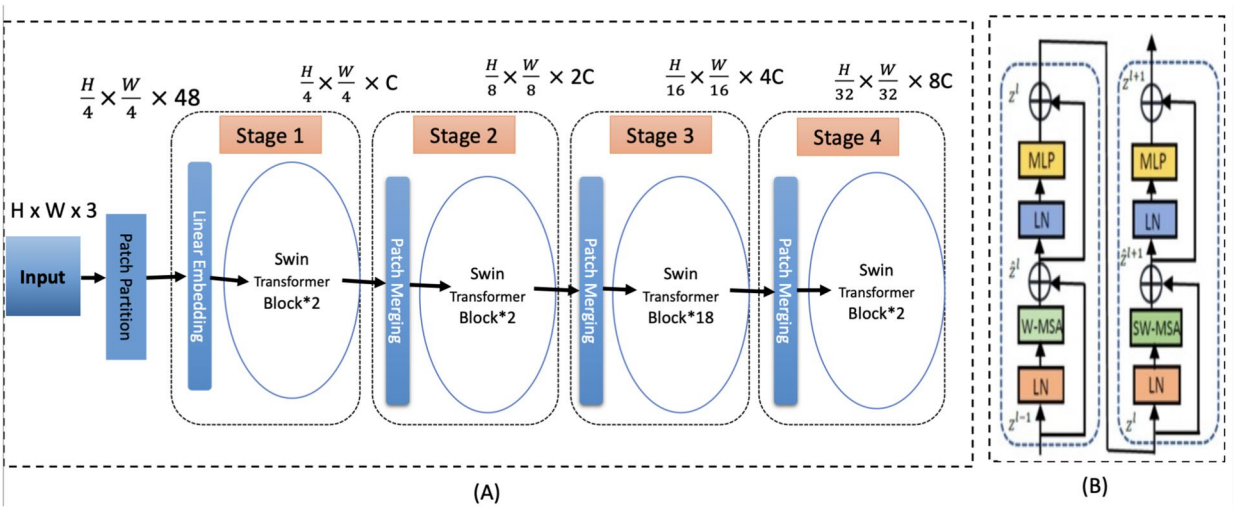


Fig. 2 Illustrating the hierarchical Swin Transformer architecture. **A** Overview of the four-stage feature extraction process with patch embedding and Swin Transformer blocks. **B** Detailed structure of the Swin Transformer block with multi-layer perceptron (MLP), layer normalization (LN), and window-based self-attention (W-MSA/SW-MSA) mechanisms [23]

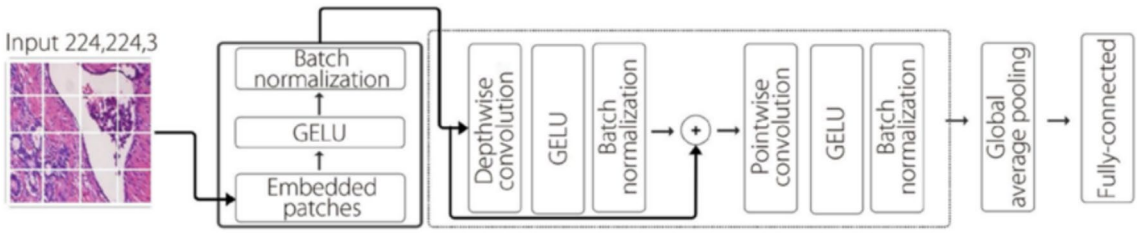


Fig. 5 Convmixer classification architecture

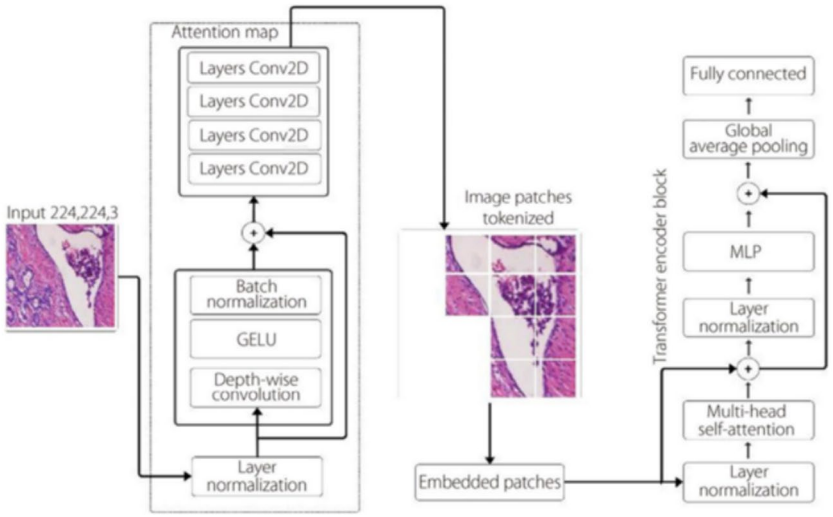


Fig. 3 Overview of the TokenMixer architecture for breast cancer classification. The top diagram illustrates the ConvMixer model, which processes embedded image patches through convolutional layers, batch normalization, and GELU activation. The bottom diagram represents the TokenMixer pipeline, integrating an attention map for token selection, convolution-based feature extraction, and a transformer encoder block for classification [24]

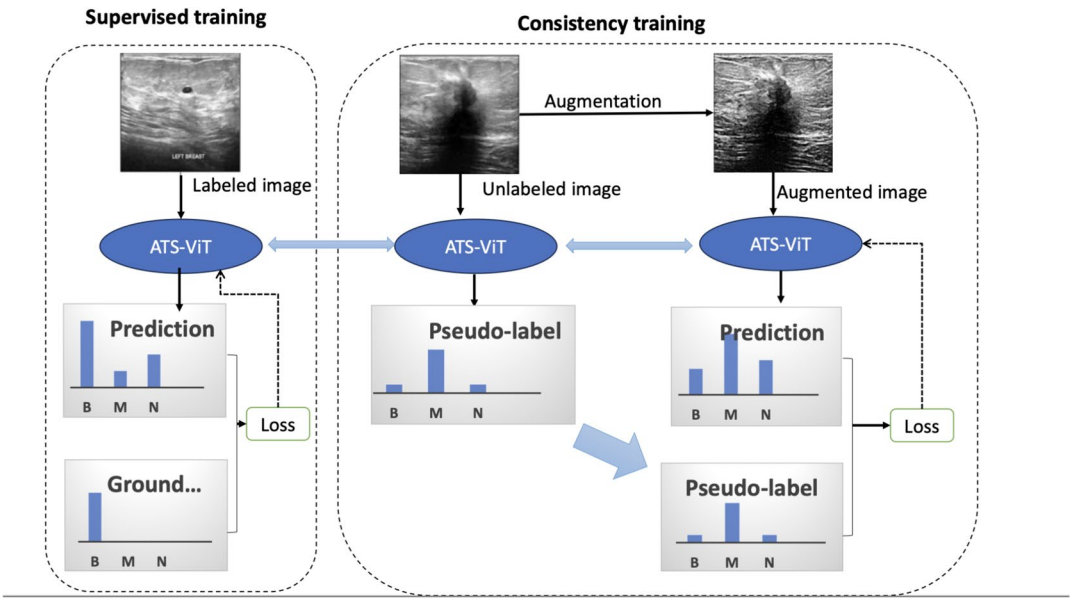


Fig. 4 The approach incorporates both supervised training with labeled images and consistency training with unlabeled images. Supervised training minimizes the loss between predictions and ground truth labels, while consistency training generates pseudo-labels for augmented unlabeled images, ensuring prediction consistency [25]

their sizes to improve computation time. This study achieved a case classification performance with an area under ROC (receiver operating characteristic curve of 0.818 ± 0.039, which outperforms the state-of-the-art multi-view CNNs ($AUC = 0.784 \pm 0.016$) and two one-view-two-side models ($AUC = 0.724 \pm 0.013$ and 0.769 ± 0.036). Limitations include the need for larger datasets and the difficulty of interpreting the model. Future directions include investigating the use of different transformer architectures and exploring the potential for transfer learning.

In the work of He et al. [27], a new Deconv-Transformer (DecT) network is introduced that integrates color deconvolution in the form of convolution layers and self-attention mechanism. It combines information from RGB and HED color space images and uses a residual connection to prevent information loss. The model's parameters are optimized in two stages, and color jitter is used for data augmentation to avoid overfitting. The DecT model shows an average accuracy of 93.02% and F1 score of 0.9389 on the BreakHis dataset and an average accuracy of 79.06% and 81.36% on the BACH and UC datasets. Limitations include the need for more diverse datasets and the difficulty of interpretability. Future directions include exploring the potential of using the model for other types of cancer and optimizing the model's performance for different tasks.

Mehta et al. [28] introduce HATNet which unifies the separate components of histopathological image analysis into a single neural network and allows for end-to-end training and inference. HATNet is based on the transformer architecture and uses a bag-of-words approach with attention mechanisms to learn inter-word and inter-bag representations. The network comprises several modules including word-to-word attention, word-to-bag attention, bag-to-bag attention, bag-to-image attention, and image classification. These modules enable HATNet to identify important words and bags in images and produce pathologist-level performance. Limitations include the need for more diverse datasets and the difficulty of interpretability. Future directions include investigating the use of different transformer architectures and exploring the potential for transfer learning. Figure 5 shows the proposed approach.

Kaczmarek et al. [29] propose an approach for classifying cancer using patient-matched miRNA and mRNA data from the TCGA. The patient data is preprocessed and mapped as a bipartite graph with miRNA and mRNA expressions as node features. The graph classification is performed through a network structure of a linear layer, two Graph Transformer Layers (GTLs), and a Multilayer Perceptron (MLP). The GTLs use self-attention mechanisms to update node features based on the expressions of neighboring nodes and the final features are input to the MLP for classification. The attention of the trained GTL is then used to identify significant molecules and pathways for different cancers. The GTN achieved a 93.56% accuracy in classifying 12 types of cancers, outperforming the graph convolutional network, random forest, support vector machine, and multilayer perceptron. Limitations include the need for more diverse datasets and the difficulty of interpretability. Future directions include exploring the use of different types of multi-omics data and improving the model's performance on specific cancer types. Figure 6 shows the proposed approach.

He et al. [30] proposed a vision transformer network enhanced with wavelet-based features for breast ultrasound classification. The approach integrates the Discrete Wavelet

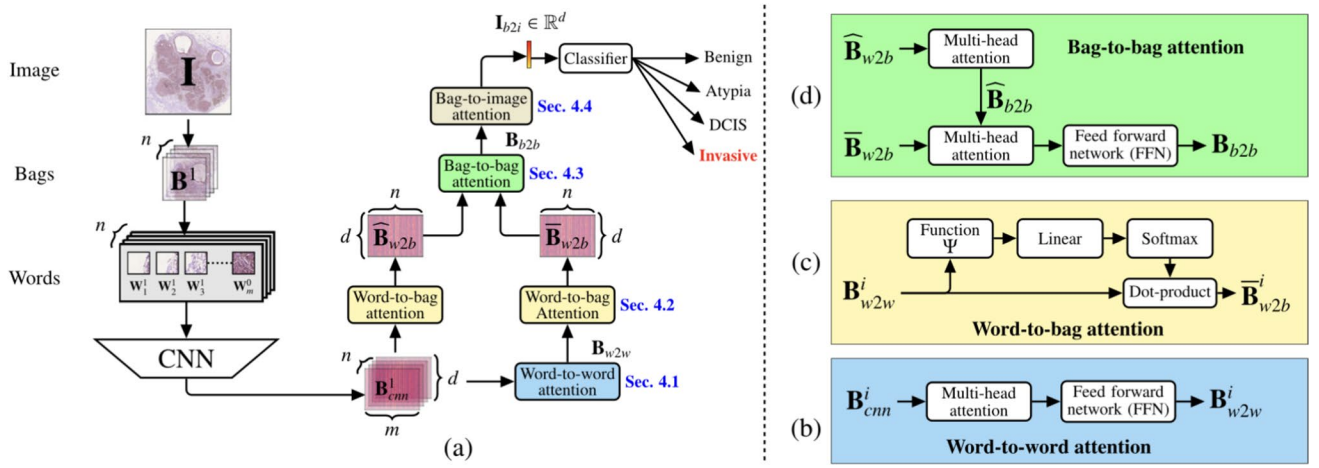


Fig. 5 **a** The hierarchical multi-instance learning framework processes whole-slide images (WSIs) using a CNN for feature extraction at the word level, followed by word-to-word, word-to-bag, and bag-to-bag attention mechanisms. **b** Word-to-word attention module enhances local feature interactions. **c** Word-to-bag attention aggregates word features into bag representations. **d** Bag-to-bag attention refines the final bag representations for classification. The final feature vector is used for clustering and classification into benign, atypia, and DCIS categories [28]

Transform (DWT) at the input stage to extract significant frequency-domain features, improving the receptive field of the transformer network. This enables better recognition of intricate breast tissue characteristics and more accurate classification of malignant and benign tumors. The study focused on breast ultrasound images, utilizing two datasets: 780 cases from Baheya Hospital in Egypt and 267 cases from the UDIAT Diagnostic Centre in Spain. Experimental results demonstrated that the proposed model outperformed conventional deep learning techniques, achieving an area under the curve (AUC) of 0.984 and 0.968 on the two datasets, respectively.

Baroni et al. [31] proposed a self-attention vision transformer model optimized for classifying breast cancer in histopathology images. Their approach investigates the impact of various training strategies, including pretraining, data augmentation, color normalization, patch configurations, and tile overlap on model performance. The primary modality used in this study is histopathology images of breast cancer, with experiments conducted on the BACH dataset and additional evaluations on BRACS and AIDPATH datasets to assess generalization. Their best-performing ViT model, pretrained on ImageNet and fine-tuned on breast cancer histology images with geometric and color augmentation, achieved an accuracy of 0.91 on the BACH dataset, surpassing previous BACH challenge winners (0.87). The model also demonstrated 0.74 accuracy on BRACS and 0.92 accuracy in binary tumor classification on AIDPATH.

Table 1 shows a summary of the above approaches in breast cancer classification.

Skin Lesion Analysis

Sarker and Moreno-García [32] introduce a transformer-based model for skin lesion classification. The model employs a bidirectional encoder representation from dermatoscopic images to classify lesions. Experiments using the public HAM10000 dataset yield promising results, with accuracy, precision, recall, and F1 score of 90.22%, 99.54%, 94.05%, and 96.28% respectively. These results suggest potential for further exploration of transformer-based approaches in addressing challenges in medical image classification, specifically generalization to diverse samples. Limitations include the need for more diverse datasets and the difficulty of interpretability. Future directions include exploring the use of additional data modalities and improving the model's performance on rare skin lesions.

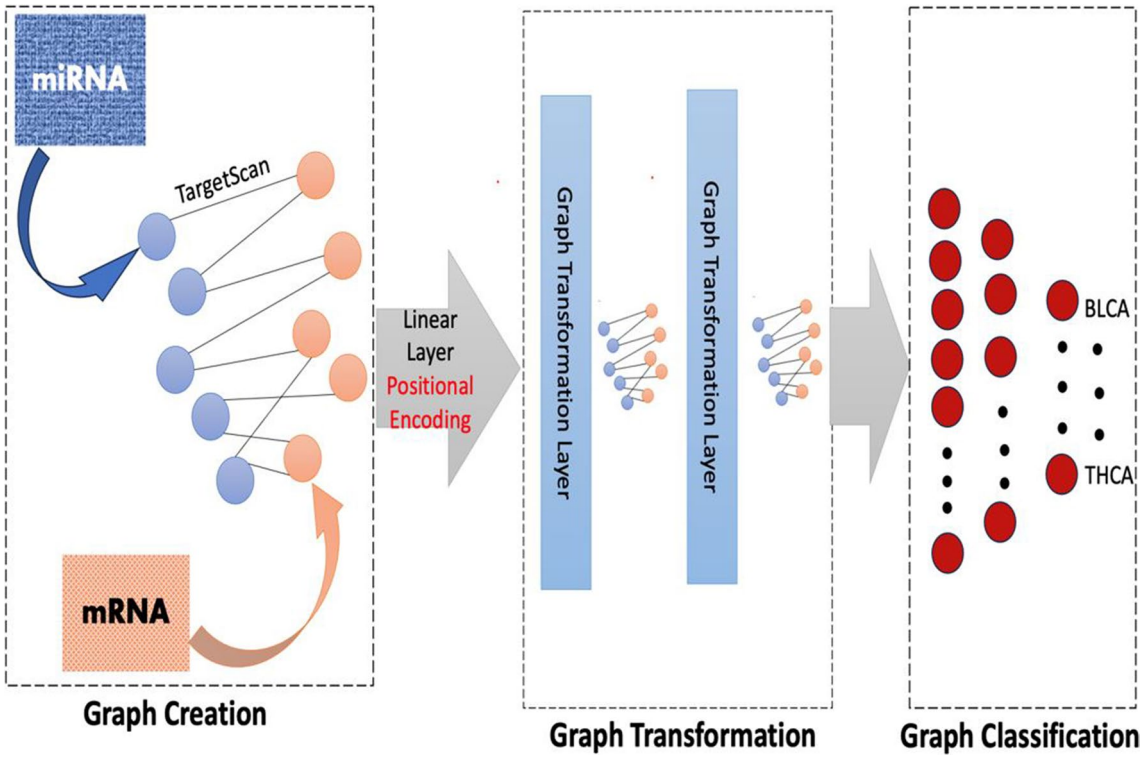


Fig. 6 The process consists of three main stages: (1) graph creation, where miRNA-mRNA interactions are modeled as a graph using TargetScan predictions; (2) graph transformation, where the graph undergoes transformation through multiple layers incorporating linear layers and positional encoding; and (3) graph classification, where the transformed graph is classified into different cancer types (e.g., BLCA and THCA) [29]

Table 1 Comparative overview of vision transformer studies in breast cancer classification

	Gheflati and Rivaz [22]	Tummala, Kim, et al. [23]	Abimouloud et al. [24]	Wang et al. [25]	X. Chen et al. [26]	He et al. [27]	Mehta et al. [28]	Kaczmarek et al. [29]	He et al. [30]	Baroni et al. [31]
Imaging level	Cells and tissues	Tissues	Tissues	Tissues	Organs	Tissues	Tissues	Cells, tissues, and organs	Tissues	Tissues
Tumor type	Malignant	Malignant	Malignant	Malignant	Malignant	Malignant	Benign and malignant tumors	Malignant	Malignant	benign
Modality specificity or medical image analysis in general	Breast ultrasound images	Histopathology	Histopathology	Histopathology	Mammography	Histopathology	Histopathology	Multi-omic data	Breast ultrasound images	Histopathology
Dimensionality	2D	2D	2D	2D	Multi-view mammograms	2D	2D	2D	2D	2D
Architecture type	Full-scale vision transformer architecture	Full architectures	Full architectures	Semi-supervised vision transformer with adaptive token sampling	Transformer-based architecture	Deconv-Transformer architecture	End-to-end transformer-based architecture	Multi-omic graph transformer architecture	Full architectures	Full architectures

Nie et al. [33] propose a hybrid neural network for classifying skin lesions in medical diagnostics. The hybrid model combines CNN with ViT components. The CNN component uses ResNet-50 as the feature extractor and the ViT component consists of an embedding layer, a transformer encoder, and a multi-layer perceptron (MLP) head. The transformer encoder uses multi-head attention (MHA) and MLP blocks to process the results of the MHA. To handle the class imbalance, cross-entropy, weighted cross-entropy, and FL functions have been included. This paper conducted 6 experiments on dermoscopic images in the ISIC 2018 dataset and improved the classification accuracy from 74 to 89% using 6 models to classify skin cancer into 7 categories. Figure 7 shows the proposed approach.

He et al. [34] propose an FTN, a Fully Transformer Network, to analyze skin lesions by learning long-range contextual information. FTN is a hierarchical Transformer that computes features using the Spatial Pyramid Transformer (SPT). SPT has linear computational complexity and reduces computation and memory usage through a spatial pyramid pooling (SPP) module in the multi-head attention (MHA). The effectiveness and efficiency of FTN are verified through extensive skin lesion analysis experiments using the ISIC 2018 dataset. Results show that FTN surpasses other state-of-the-art CNNs in terms of computational efficiency and the number of tunable parameters, thanks to the efficient SPT and hierarchical network structure. The authors discusses the need for larger datasets and exploring ways to incorporate clinical data.

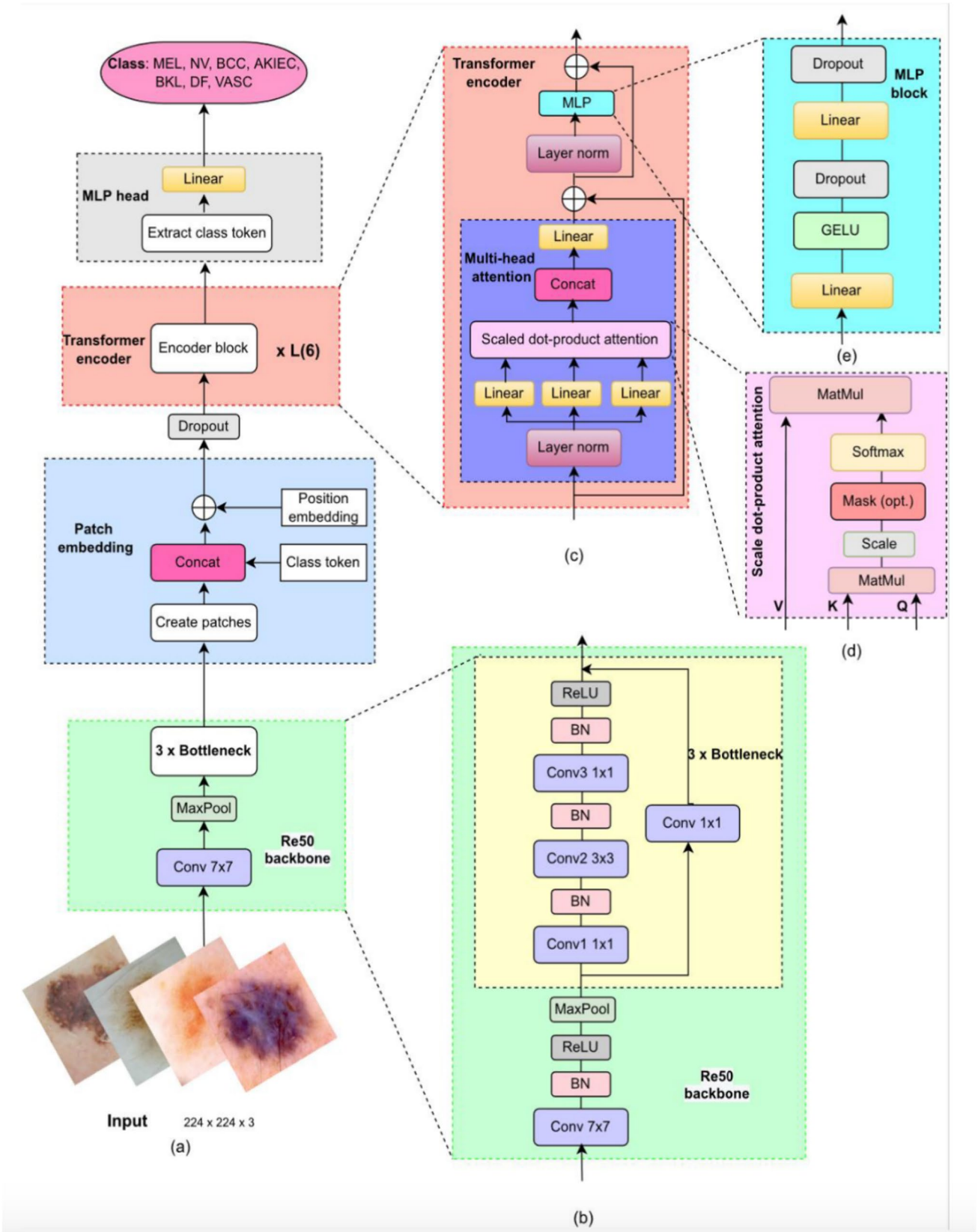
Lima and Krohling [35] investigate 20 skin lesion classification models with 30 million parameters or less and GPU (Graphics Processing Unit) memory requirements less than 12 GB. The models are selected from the 20 best models in the TIMM pre-trained collection based on ImageNet-Real, with a fresh set of labels intended to improve on the original annotations. The input data consists of a clinical image and clinical information, and the clinical image features are extracted by a pre-trained backbone model. A fusion block is used to add clinical information to the reformatted image features, and four alternatives to the fusion block are analyzed. The models are fine-tuned on a dataset of 2298 skin lesion samples with 6 types of skin lesions and 21 clinical features, which was chosen because it deals with multimodal data for skin lesion diagnosis. The study found that distillation and transformer architectures (PiT, CoaT, and ViT) can improve the performance in skin lesion diagnosis task on PAD- UFES-20 dataset, with the average AUC of "pit_s_distilled_224" and "coat_lite_small" models with MetaBlock fusion achieving competitive results (0.941 ± 0.006 and 0.940 ± 0.002 respectively). The approach is limited to skin lesions, potential to improve with larger datasets and more advanced architectures.

Wu et al. [36] introduce a method called "Scale-Aware Transformers" (ScATNet) for learning representations from histopathological images at multiple scales in an end-to-end manner. The method is based on the transformer architecture which has two modules: self-attention and feed-forward. The method involves three main steps: (1) patch-wise embeddings are learned using a CNN for each input scale, (2) patch embeddings are fed to a transformer to learn inter-patch relationships, (3) contextualized scale embeddings are learned across multiple input scales using transformers. The method overcomes the limitations of patch-based CNNs by allowing the system to learn both local and global representations. The software will be made available. The experiments show that this model achieved an accuracy of 64% on skin biopsy dataset. This approach is limited to melanocytic lesions, future work to explore the method's generalizability to other skin lesions. Figure 8 shows the proposed approach.

Zhao [37] proposes a method for skin cancer classification consisting of data pre-processing and model construction. Data pre-processing involves median frequency balancing to handle class imbalance, cropping images to fit the model, data augmentation, and normalization. The models used for skin cancer classification include VGG16, ResNet18, ViT, and DeepViT. VGG16 is a 16-layer CNN with five Max pooling layers and three fully connected layers. ResNet18 is a 72-layer architecture based on VGGNet with residual connections to mitigate vanishing and bursting gradients. ViT splits the input into tokens and pushes them into the transformer encoder. DeepViT replaces the self-attention layer within the transformer block with a reattention module to improve accuracy and it achieves an accuracy of 84% and a loss of 67% on HAM10000 dataset. Limitations include small dataset size and potential overfitting; future directions include incorporating additional data sources and optimizing hyperparameters.

Zhou et al. [38] present a neural network model for skin lesion diagnosis that utilizes both images and textual information. They extract fine-grained features from both data types using pre-trained deep learning models and introduce a Mutual Attention Transformer (MAT) approach to combine the features. MAT incorporates self-attention blocks and guided-attention blocks for concurrent interaction between features from both modalities. A fusion mechanism integrates the represented features before classification. Their method outperforms existing state-of-the-art techniques based on experimental results on the HAM10000 dataset. Future directions include exploring multi-modal and explainable AI approaches. The method is limited by dataset bias and lack of interpretability.

Himel et al. [39] proposed a ViT-based approach for skin cancer segmentation and classification using dermatoscopic images. The study utilizes the HAM10000 dataset, which contains 10,015 skin lesion images categorized as benign


Class: MEL, NV, BCC, AKIEC, BKL, DF, VASC
MLP head
Linear
Extract class token
Transformer encoder
Encoder block
x L(6)
Dropout
Patch embedding
Position embedding
Concat
Class token
Create patches
3 x Bottleneck
MaxPool
Conv 7x7
Re50 backbone
Input
224 x 224 x 3
(a)
Transformer encoder
MLP
Layer norm
Multi-head attention
Linear
Concat
Scaled dot-product attention
Linear
Linear
Linear
Layer norm
(c)
Dropout
MLP block
Linear
Dropout
GELU
Linear
(e)
Scale dot-product attention
MatMul
Softmax
Mask (opt.)
Scale
MatMul
V
K
Q
(d)
ReLU
BN
Conv3 1x1
BN
Conv2 3x3
BN
Conv1 1x1
3 x Bottleneck
Conv 1x1
MaxPool
ReLU
BN
Conv 7x7
Re50 backbone
(b)

◂**Fig. 7** The diagram illustrates a hybrid deep learning framework integrating a ResNet50 backbone for feature extraction and a vision transformer for classification. The process involves patch embedding, transformer encoders with multi-head self-attention mechanisms, and an MLP head for final classification. The approach is designed to classify skin lesions into multiple categories, including melanoma (MEL), nevus (NV), basal cell carcinoma (BCC), and other skin conditions [33]

or malignant. The approach incorporates the Segment Anything Model (SAM) for precise segmentation of cancerous regions, achieving an IOU of 96.01% and a Dice coefficient of 98.14%. For classification, they experimented with multiple ViT-based models, including ViT-Google, ViT-MAE, ViT-ResNet50, ViT-VAN, ViT-BEiT, and ViT-DiT. The ViT-Google (patch-32) model demonstrated superior performance, attaining 96.15% accuracy with a low false negative rate.

Xin et al. [40] proposed an improved vision transformer ViTmodel named SkinTrans for skin cancer classification using dermoscopic images. The approach enhances ViT with multi-scale overlapping sliding windows for better feature extraction and contrastive learning to improve classification performance by making similar skin cancer data encode similarly while maximizing differences between different classes.The study focuses on dermoscopic images, utilizing the HAM10000 dataset and a clinical dataset collected through dermoscopy. The SkinTrans model achieves 94.3% accuracy on HAM10000 and 94.1% on the clinical dataset.

Table 2 shows a summary of the above approaches in skin lesion classification.

MRI Brain Tumor Analysis

Tummala, et al. [41] propose a ViT architecture for classifying MRI images from a dataset of 3064 slices from 233 patients with meningiomas, gliomas, or pituitary tumors. The dataset has images with different spatial resolutions in sagittal, coronal, and axial directions. The ViT model works by treating image patches as words and using a transformer encoder to classify the image. Pretrained and fine-tuned ViT base and large models were used on the ImageNet-21 k and ImageNet-1 k datasets. The MRI images were resized to meet the input requirements of the models. A learnable embedding was concatenated to the sequence of patch embeddings to produce the final output and achieving an accuracy of 98.7% on a brain tumor dataset from figshare.

Lyu et al. [42] developed a deep-learning method for precise, non-invasive digital histology using whole-brain MRI data. A total of 1582 contrast-enhanced T1-weighted and fast spoiled gradient echo MRI exams were processed and used in the deep-learning workflow for tumor segmentation, modality transfer, and primary site classification into 5 classes. Results showed an overall AUC of 0.878 (95% CI: 0.873–0.883) through tenfold cross-validation, indicating that whole-brain imaging features are sufficient for accurate diagnosis of primary malignancy sites. This end-to-end deep radiomic approach has potential to classify metastatic tumor types from whole-brain MRI images and may become a valuable clinical tool for faster primary cancer site identification and better treatment outcomes with further improvement. Figure 9 shows the proposed approach.

Reddy et al. [43] proposed a fine-tuned vision transformer (FTVT) approach for multi-class brain tumor classification using MRI scans. The study introduced four FTVT models—FTVT-b16, FTVT-b32, FTVT-l16, and FTVT-l32—while comparing their performance against conventional deep learning models such as ResNet-50, MobileNet-V2, and EfficientNet-B0. The dataset used contained 7,023 MRI images categorized into four classes: glioma, meningioma, pituitary, and no tumor. The experimental results demonstrated that the FTVT-l16 model achieved the highest classification accuracy of 98.70%, outperforming other FTVT variants (FTVT-b16: 98.09%, FTVT-b32: 96.87%, and FTVT-l32: 98.62%) as well as deep learning baselines (ResNet-50: 96.5%, EfficientNet-B0: 95.1%, and MobileNet-V2: 94.9%).

Zhang et al. [44] proposed AugTransU-Net, a novel Augmented Transformer U-Net for MRI brain tumor segmentation. The approach addresses the limitations of existing transformer-based U-Net models by incorporating Augmented Shortcuts into transformer blocks at the bottleneck to enhance feature diversity. Additionally, Paired Attention (PA) modules are introduced at both the encoder and decoder stages to establish long-range spatial and channel relationships.The method was evaluated on the BraTS2019-2021 brain tumor segmentation datasets using MRI scans as the imaging modality. The experimental results demonstrated the effectiveness and competitiveness of AugTransU-Net, achieving Dice similarity coefficients (DSC) of 89.7%, 78.2%, and 80.4% for whole tumor (WT), enhancing tumor (ET), and tumor core (TC) segmentation, respectively, on BraTS2019-2020 validation sets.

Table 3 shows a summary of the above approaches in brain tumor classification.

Lung Cancer/Disease Classification

Sun and Pang [45] present a framework for recognizing early lung cancer using the Swin Transformer. The Swin Transformer framework consists of three phases: image processing, attention blocking, and downstream operations. Image processing involves converting RGB images into non-overlapping patches, which are then transformed into feature matrices. The attention blocking phase consists of stacking several blocks of the Swin Transformer, which uses shift window-based multi-head self-attention and multi-layer

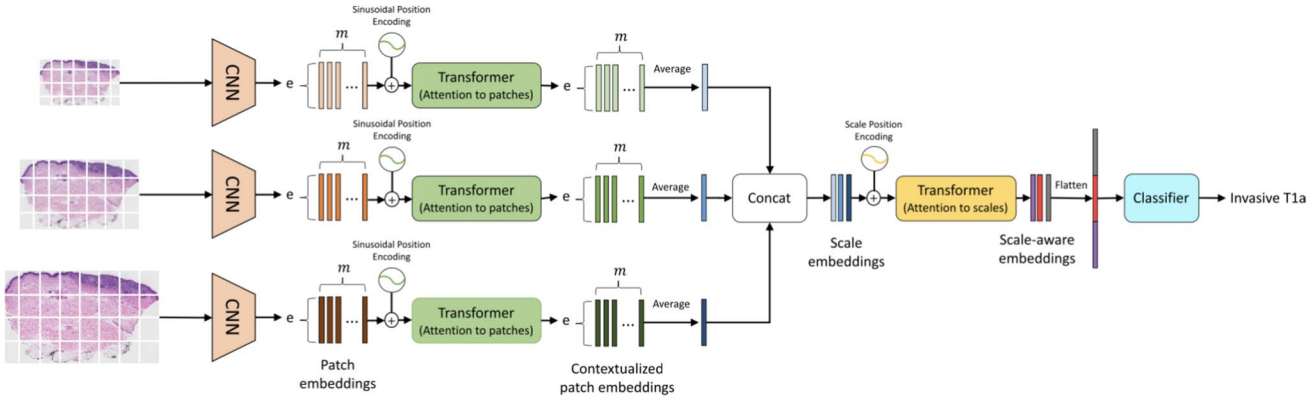


Fig. 8 The approach integrates hierarchical feature extraction, scale-aware embeddings, and a classification module to enhance lesion classification performance [36]

Table 2 Comparative overview of vision transformer studies in skin lesion classification

	Sarker and Moreno-García [32]	Nie et al. [33]	X. He et al. [34]	Lima and Krohling [35]	Wu et al. [36]	Zhao [37]	Zhou et al. [38]	Himel et al. [34]	Xin et al. [35]
Types of diseases included	Skin lesions	Skin lesions	Skin lesions (benign and malignant)	Malignant and benign skin tumors	Melanocytic lesions	Malignant skin lesions (melanoma and non-melanoma)	Benign and malignant tumors	Benign and malignant tumors	Malignant skin lesions
Full architectures versus lightweight models	Transformer-based architecture	Deep CNN transformer hybrid model	Full architecture	Full architectures	Full architectures	Compared full architectures and a lightweight model	Full architectures	Full architectures	Full architectures

perceptron to learn image features. The shifted window partitioning technique is used to connect adjacent non-overlapping windows, which increases the perceptual field of view. The model achieved a maximum accuracy of 82.3% on the classification task and over 95% on the segmentation task after pre-training. The authors suggest further investigation on the use of the proposed method for other medical image analysis tasks and the development of a more efficient algorithm for real-time application. Figure 10 shows the proposed approach.

Khademi et al. [46] propose a framework for lung nodule classification called CAET-SWin. The framework uses a self-attention mechanism, which is a key component of the transformer architecture. The input data is first pre-processed with a U-Net-based lung segmentation model. The extracted lung areas are then passed into two parallel paths of the CAET-SWin framework. The multi-head self-attention mechanism computes weighted averages of the input sequence's instances and allows the model to jointly attend to information from different representation subspaces. The method achieved an accuracy of 82.65%, sensitivity of 83.66%, and specificity of 81.66% using tenfold cross-validation, improving the reliability of the invasiveness prediction task. The authors suggest further investigation on the use of the proposed method for other medical image analysis tasks and the exploration of more efficient training strategies for 3D + time medical image analysis. Figure 11 shows the proposed approach.

Chen et al. [47] collected 347 images of lung washout cells from 10 patients, which were labeled and counted by pathologists resulting in 2473 lung cells, including 143 cancerous, 724 normal, and 1606 noisy cells. The paper introduced NucleAIzer, a deep learning framework for cell nucleus segmentation which was used to segment each lung

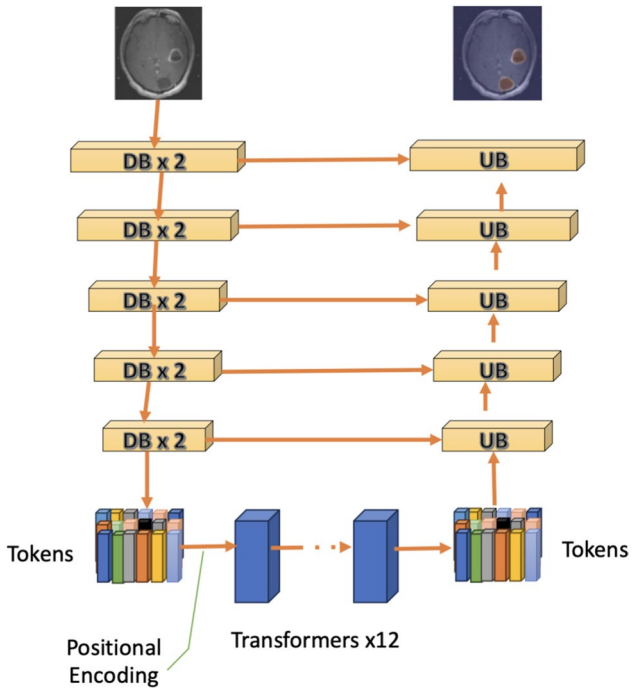


Fig. 9 The architecture processes MRI images using tokenization and positional encoding, followed by multiple transformer layers. A hierarchical decoder with downsampling blocks (DB) and upsampling blocks (UB) refines the feature representations, enabling accurate tumor segmentation and analysis [42]

cell image. To maximize the use of each data, the individual segmented cell was treated as separate data. The data was enhanced by Gaussian blurring, rotation, inversion, and adding noise to increase robustness of the model. The final enhanced data consisted of 3106 images with 500 cancerous, 1000 normal, and 1606 noisy cells. The Swin Transformer structure used for lung cancer cytopathology image classification involved multiple stages of Patch Merging and Swin Transformer Block. The Swin Transformer used W-MSA and SW-MSA instead of MSA to reduce computational complexity while maintaining global information. The experiments show that the model achieved an accuracy of 96.16% on lung cells dataset, future work should focus on more comprehensive experiments with larger datasets and multiple modalities. Figure 12 shows the proposed approach.

Sun et al. [48] present a compact convolutional transformer (CCT) model for reducing class token and embedding requirements in computer vision tasks. The model has a lightweight transformer structure, including a convolution module, an embedding layer, a transformer encoder, and a residual structure of ResNet50. The convolution module uses normalization, convolution, ReLU (Rectified Linear Unit), and max pooling to extract features. The transformer encoder uses multiple self-attention layers, which can extract powerful features by focusing on global content information.

Table 3 Comparative overview of vision transformer studies in brain tumor classification

	Kadry et al. [41]	Lyu et al. [42]	Reddy et al. [43]	Zhang et al. [39]
Types of diseases included	Tumor: benign or malignant	Brain metastases	Tumor	Tumor
Modality specificity	MRI	Whole-brain MRI	MRI	MRI
Computational complexity and memory requirements of the method	High computational complexity and memory requirements	The proposed transformer-based model has a low computational cost and memory requirements compared to other deep learning models	Require significant computational resources due to their transformer-based architecture	High computational complexity and memory requirements
Full architectures versus lightweight models	Full architectures	The proposed transformer-based model is a full architecture model	Fine-tuned vision transformer	Augmented Transformer
Limitations and future directions of the method	Limited sample size and future work can explore ensembling with other modalities and architectures	The authors suggest further investigation on the use of transformer-based models for other medical image analysis tasks and the incorporation of clinical data for improved classification performance	It is prone to overfitting on smaller datasets and lack interpretability, which is critical for medical decision-making	The limitations of AugTransU-Net include the high computational cost associated with the 3D transformer layers, making it resource-intensive and challenging to train on limited hardware

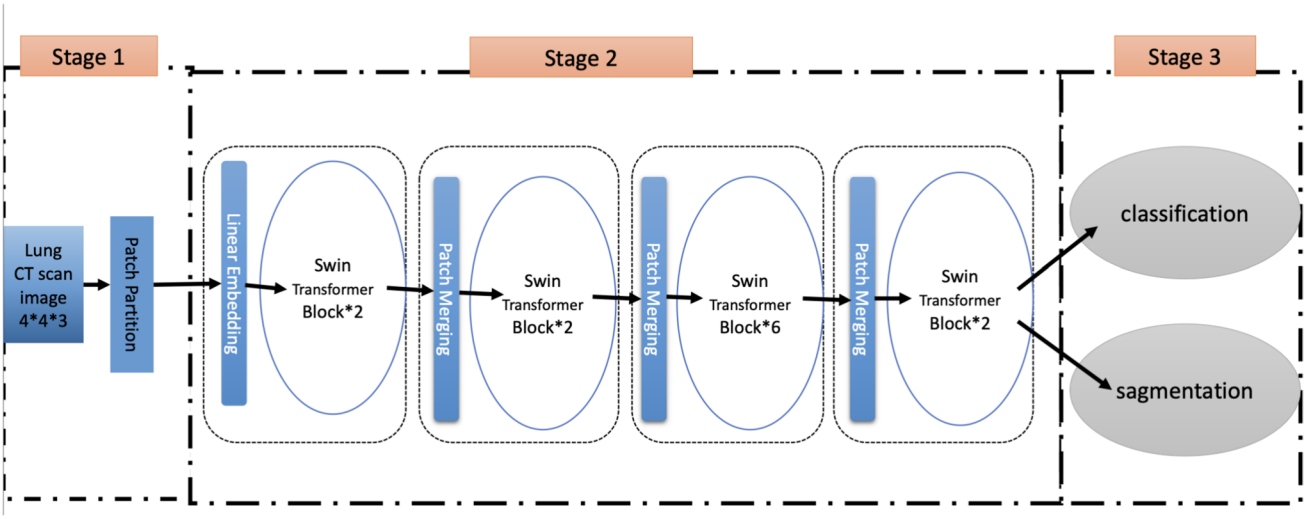


Fig. 10 The model consists of three main stages: (1) patch extraction from Lung CT scan images, (2) feature extraction using Swin Transformer blocks, and (3) dual-task classification and segmentation for disease diagnosis [45]

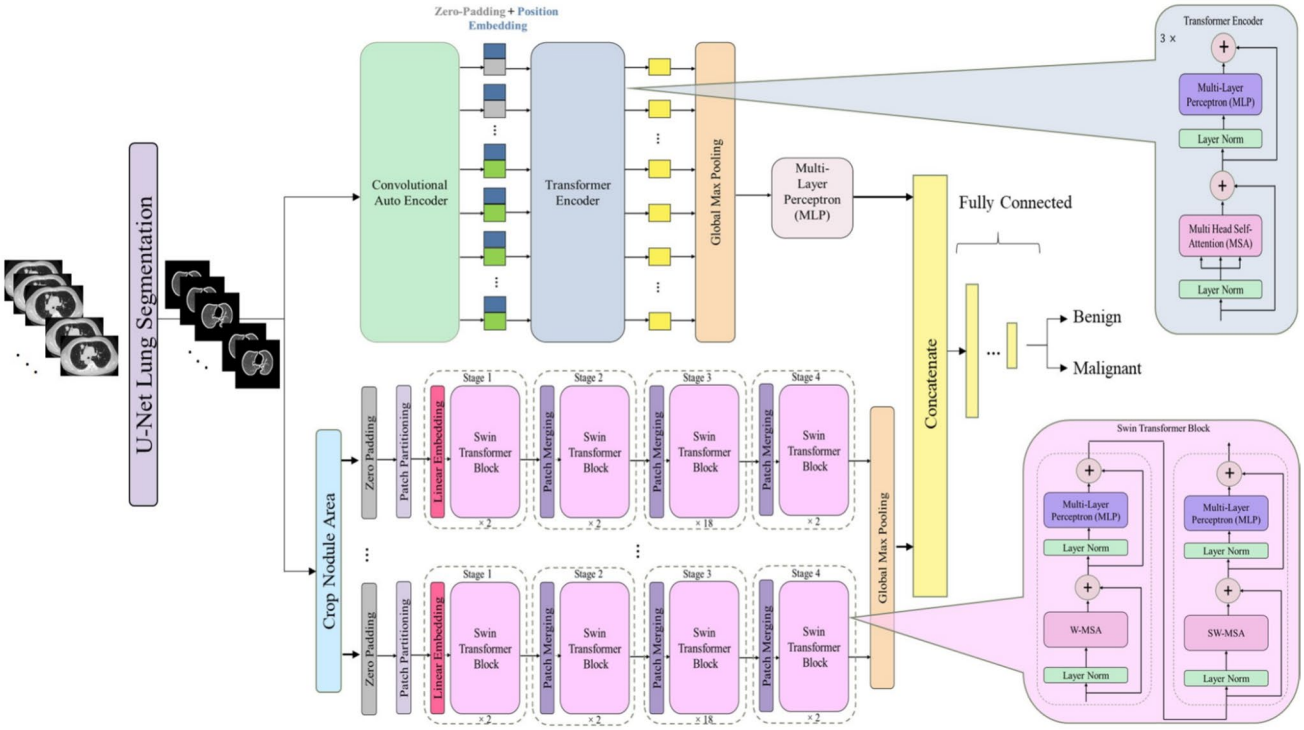


Fig. 11 The model integrates U-Net lung segmentation, a convolutional autoencoder, a transformer encoder, and multi-layer perceptron (MLP) classifiers to distinguish between benign and malignant cases [46]

The self-attention mechanism is divided into two 1D mechanisms to improve efficiency, and a position code is added to the query mechanism to make it more responsive to position information. The model achieved an accuracy of 98.5% and a sensitivity of 98.6% on CC-CCII dataset. Figure 13 shows the proposed approach.

Wu et al. [49] propose a ViT model that classifies emphysema subtypes using CT images. Large patches are cropped and divided into smaller patches, which are converted to patch embeddings and fed into a transformer encoder for final representation. The pre-trained ViT model on ImageNet has an average accuracy of 95.95% in

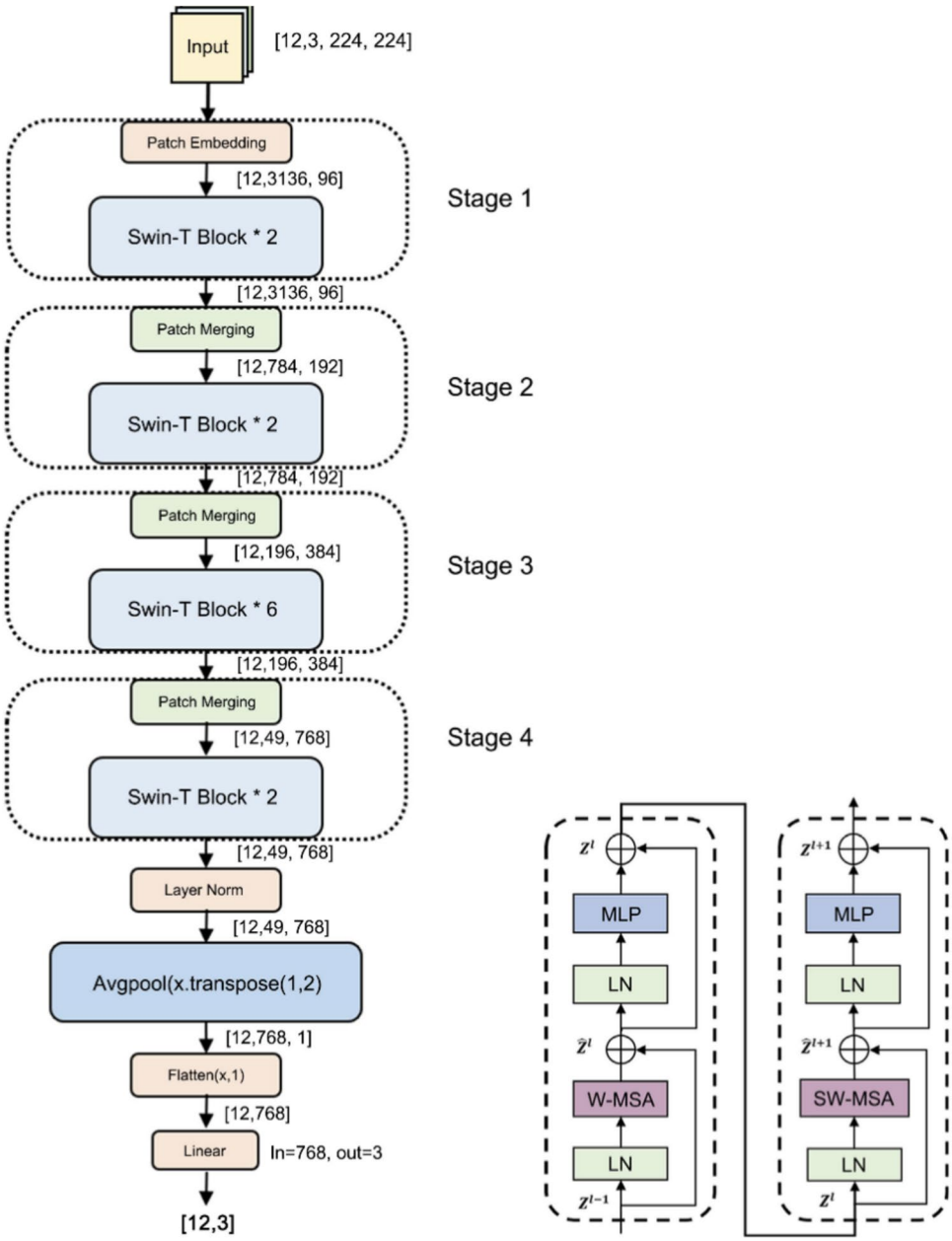


Fig. 12 The model processes input lung images through hierarchical Swin Transformer blocks, employing patch embeddings, merging layers, and self-attention mechanisms to extract multi-scale features [47]

the lab's own dataset, outperforming other models such as AlexNet and ResNet50. The pre-trained ViT model also outperforms the ViT model without pre-training and has comparable accuracy to other methods for the public dataset. The results show that the ViT model can accurately classify emphysema subtypes, making it useful for computer-aided diagnosis and other medical applications. Future work needed to improve accuracy and generalization to new datasets.

Bharati et al. [50] propose a new hybrid deep learning framework called VDSNet which combines VGG (Visual Geometry Group), data augmentation, and spatial transformer network with CNN. The implementation tools used are Jupyter Notebook, Tensorflow, and Keras. The model is tested on the NIH chest X-ray image dataset from Kaggle and outperforms existing methods in terms of precision, recall, F0.5 score, and validation accuracy. On the full dataset, VDSNet has a validation accuracy of 73% compared to other models. Figure 14 shows the proposed approach.

The study of Shamrat et al. [51] aims to determine the most efficient transfer learning method for classification by evaluating eight pre-trained models, including AlexNet, GoogLeNet, InceptionV3, ResNet50, MobileNetV2, EfficientNetB7, DenseNet121, and VGG16. The highest accuracy was achieved by adding several layers to the VGG16 model, creating a customized fine-tuned transfer learning

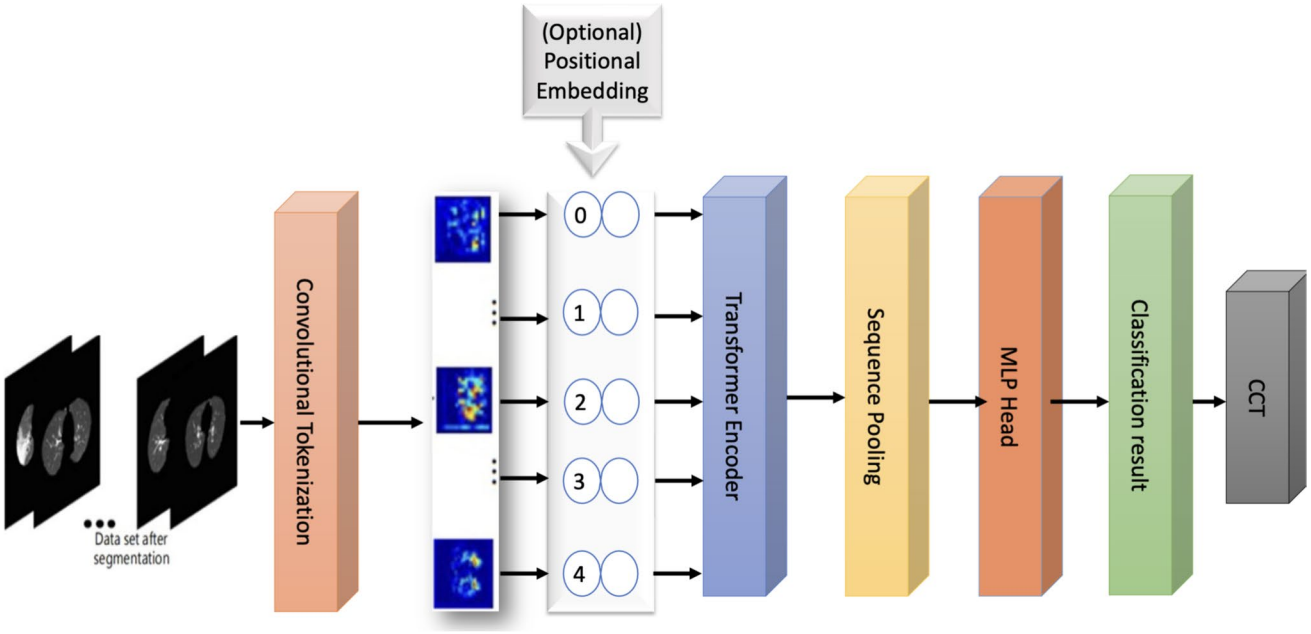


Fig. 13 The approach includes data input after segmentation, convolutional tokenization, an optional positional embedding, a transformer encoder, sequence pooling, an MLP head, and classification output, utilizing Compact Convolutional Transformers (CCT) for enhanced feature extraction [48]

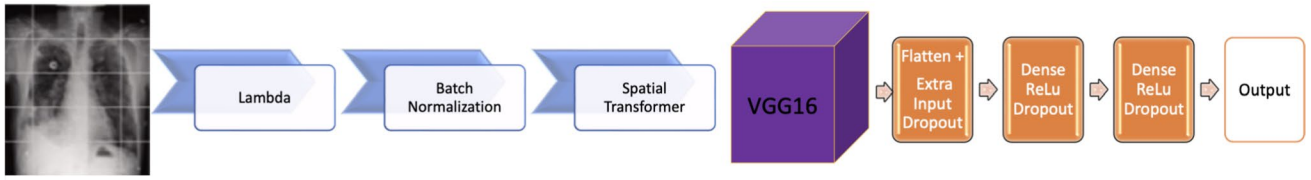


Fig. 14 VDSNet framework for lung disease classification, integrating VGG16, spatial transformer networks to enhance classification performance on chest X-ray images [50]

method. The various pre-trained models have different architectures and are pre-trained with various techniques, such as using max pooling and convolutions in AlexNet, using inception modules in GoogLeNet, using factorized convolutions in InceptionV3, using a mixture of convolutional filters in ResNet50, and using an inverting residual structure in MobileNetV2. Figure 15 shows the proposed approach.

Ma and Lv [52] compare the performance of a model for pneumonia identification in chest X-rays based on a transformer backbone network with models based on traditional convolutional neural backbone networks. The experiments were done on two publicly available chest X-ray datasets, Chest X-ray dataset from National Institutes of health and CXR images (pneumonia) data set from Guangzhou Women and Children's Medical Center. The model used in the experiment was based on the Swin Transformer backbone network and had steps for image enhancement, parameter optimization, feature extraction, classification, and lesion area output. The performance of the model was compared with models based on AlexNet, GoogLeNet, VGGNet16, ResNet50, DenseNet, and DenseNet121. The decision factors were also extracted using Grad-CAM to perform discriminative output of the lesion.

Jiang et al. [53] propose a new variant of the pyramid vision transformer for multi-label chest X-ray image classification (MXT) to assist in the diagnosis of lung diseases during the COVID-19 outbreak. MXT captures both short and long-range visual information, reduces resource consumption and encodes position information for effective multi-label classification. The proposed method was evaluated on two datasets and yielded a high mean AUC score of 83.0% and 94.6%. Ablation studies showed that the proposed MLOP embedding and dynamic position feed forward perform better than the traditional methods. The comprehensive experiments demonstrate the effectiveness of the proposed MXT in assisting radiologists in diagnosing lung diseases and reducing the work pressure of medical staff. Limited to binary and multi-label classification, future work could

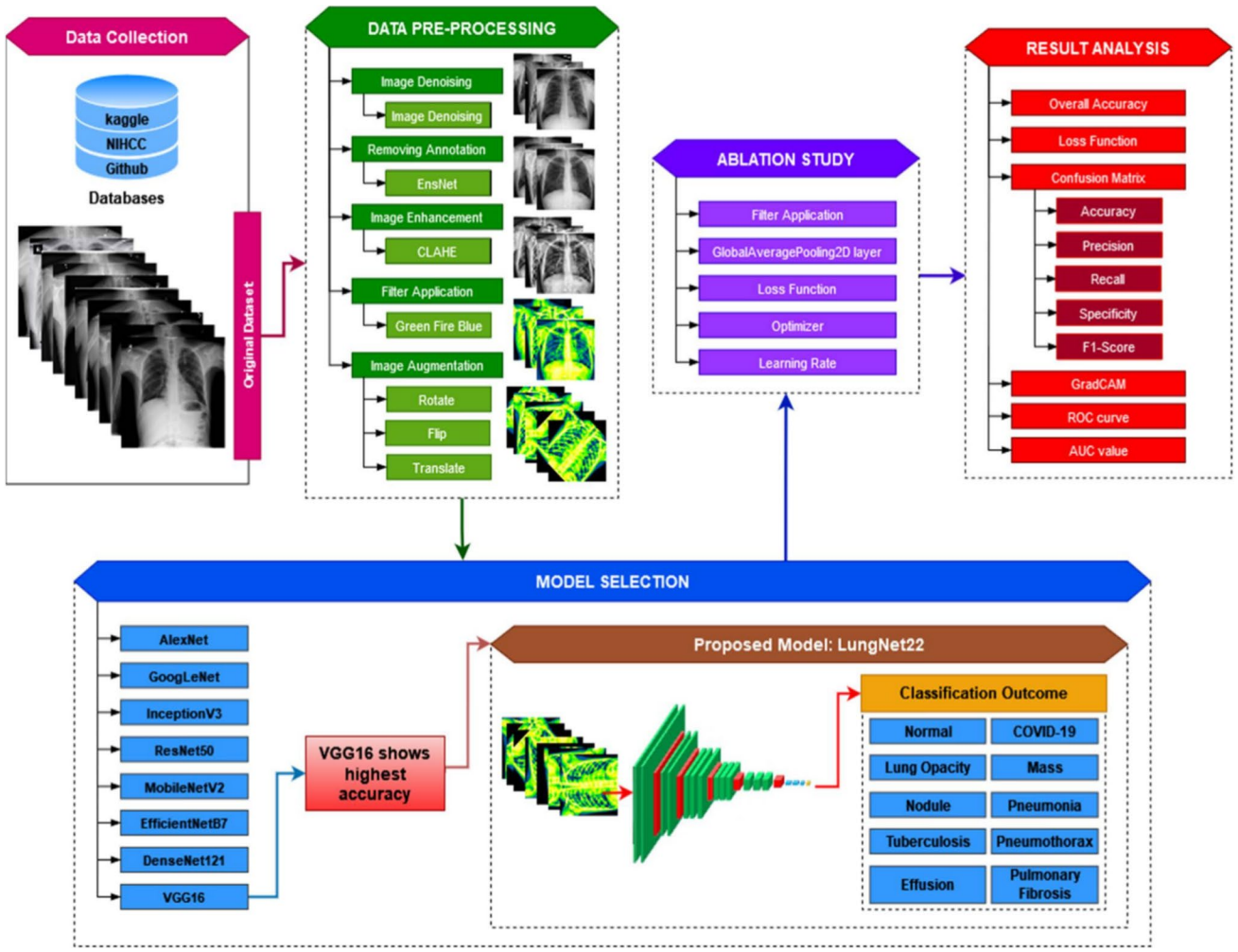


Fig. 15 The process includes data collection from multiple databases, data preprocessing with enhancement and augmentation techniques, model selection where VGG16 showed the highest accuracy, and the development of a proposed model, LungNet22. The model undergoes ablation studies to refine performance, followed by result analysis using multiple evaluation metrics such as accuracy, precision, recall, and GradCAM visualization [51]

investigate its performance on other medical image classification tasks.

Imran et al. [54] proposed a Transformer-Based Hierarchical Model for detecting and classifying Non-Small Cell Lung Cancer (NSCLC) using histopathological images. Their approach integrates CNNs and ViTs to capture both local and long-range dependencies in medical images. The model classifies NSCLC into three categories: normal, adenocarcinoma, and squamous cell carcinoma. The study utilized the LC25000 dataset, a benchmark dataset for NSCLC histopathology classification. The model outperformed state-of-the-art methods, achieving an accuracy of 0.988, F1 score of 0.980, specificity of 0.991, recall of 0.982, and precision of 0.980. Additionally, it demonstrated a low inference time of 1.816 ms, making it suitable for real-world applications.

Table 4 shows a summary of the above approaches in Lung Cancer classification.

Retinal and Eye Analysis

Playout et al. [55] evaluate the performance of ViTs for retinal disease classification compared to traditional CNN models, taking into account multi-modality imaging and generalization to external data. The study introduces a mechanism called Focused Attention to produce high-resolution interpretable predictions via attribution maps, and through a survey with retinal specialists, confirms the superior interpretability of ViTs and the relevance of Focused Attention as a lesion detector. Future work includes exploring the effectiveness of the mechanism on other image classification tasks.

Rodriguez et al. [56] used a data preprocessing step to extract useful information from the images in a dataset of retinal diseases, the authors perform background removal to reduce the size of the images and remove redundant information, the classification model used is the C-Tran architecture,

Table 4 Comparative overview of vision transformer studies in lung cancer classification

	Sun and Pang [45]	Khademi et al. [46]	Chen et al. [47]	Sun et al. [48]	Wu et al. [49]	Bharati et al. [50]	Shamrat et al. [51]	Ma and Lv [52]	Jiang et al. [53]	Imran et al. [54]
Specific body organs the method is applied to	Lungs	Lungs	Lung	Lung	Lungs	Chest/Lungs	Lungs	Chest	Chest	Lung
Modality specificity or medical image analysis in general	CT images	CT images	Microscopic	CT	CT images	X-ray images	X-ray images	Chest X-ray	X-ray	Histopathological images
Dimensionality applicability	2D	3D + time	2D	2D	3D	2D	2D	2D	2D	2D

which is a transformer encoder that takes both visual features from a CNN and masked labels as input. The C-Tran model consists of three parts: generating features, labels, and state embeddings, modeling feature and label interactions using a transformer encoder, and prediction using a fully connected layer. The model also uses state embeddings to handle partially labeled data; the experiments show that the model achieved an AUC score of 0.962 compared to other approaches.

Shen et al. [57] propose Structure-Oriented Transformer (SoT) framework for retinal disease grading, which aims to improve the accuracy of vision loss prediction by better capturing the relationship between the local lesion and the whole retina, the SoT framework uses structure guidance as a model-oriented filter, a pre-trained vision transformer, and a Token vote classifier to obtain the final grading results, the SoT framework is shown to outperform existing methods for neovascular Age-related Macular Degeneration (nAMD) grading and to have good generality for different retinal diseases. Future work includes expanding the method to other diseases and datasets and exploring other transformer models.

Z. Ma et al. [58] used two retinal optical coherence tomography (OCT) datasets, the OCT2017 and the Srinivasan2014, obtained from 4686 and 45 patients, respectively, the datasets consist of various retinal images divided into different classes such as normal, dry form of AMD, wet form of AMD, and diabetic macular edema, the proposed method in the study, HCTNet, improves retinal OCT classification by combining the advantages of vision transformer and CNN. The HCTNet takes a retinal OCT image as input and uses a low-level feature extraction module to generate low-level features, followed by two branches for global and local analysis using the Transformer and ConvNet respectively, the output of these two branches is then fused and a final retinal disease classification is performed using a fully connected layer. The proposed method achieves an accuracy of 91.56%. Figure 16 shows the proposed approach.

Jiang et al. [59] used a dataset of OCT fundus images collected by Duke University, which includes 1407 images of normal retinal, 723 images of AMD, and 1101 images of DME. The images are preprocessed by resizing and normalization before being input into a vision transformer, which is an image classification model based on the transformer structure used in NLP. The vision transformer consists of an encoder component made up of six identical encoders, each consisting of a multi-head attention layer and a feed-forward layer. The images are divided into patches and processed through linear embedding, positional embedding, class embedding, and an MLP head structure before being classified. The loss function used in this study is the symmetric cross-entropy loss function, which reduces the impact of noise in the dataset on training and prevents overfitting. The model achieved a classification accuracy of 99.69%.

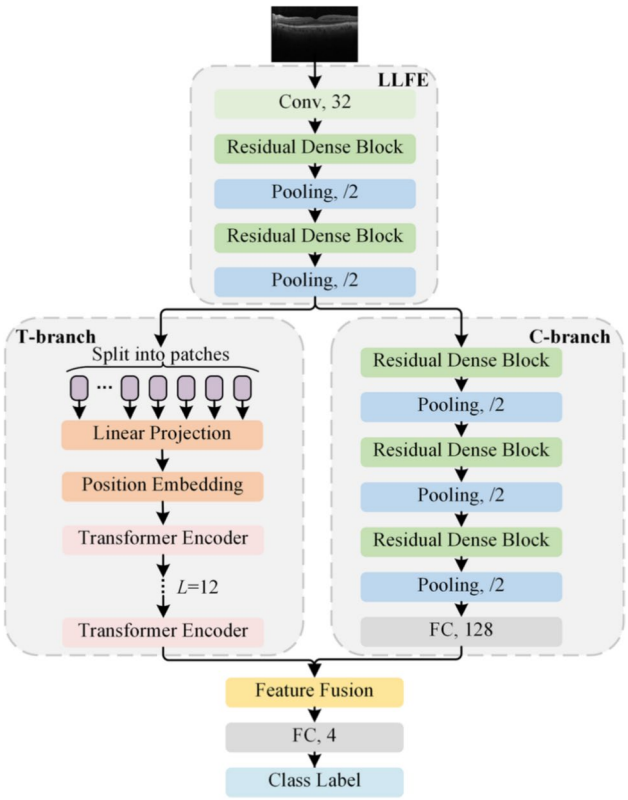


Fig. 16 ViT-based model for Retinal and Eye Analysis, integrating convolutional feature extraction (LLFE) with dual-branch processing: a Transformer-based T-branch for spatial feature learning and a convolutional C-branch for hierarchical feature extraction [58]

Wang et al. [60] present MBSaNet, a custom neural network for multilabel fundus disease recognition, which combines the generalization ability of CNNs and the global feature modeling ability of Transformer models. The model consists of two parts: a feature extractor with five stages and a multilabel classifier. The evaluation was done on a dataset of 3,500 patients with fundus images, and the results were evaluated using AUC, Kappa, and F1 score, achieving 0.89 0.44 0.88. Future directions include exploring the use of larger datasets and incorporating more advanced deep learning techniques. Figure 17 shows the proposed approach.

Torki et al. [61] discuss fundus imaging in the medical field, specifically for diagnosing eye diseases such as glaucoma. The article outlines various publicly available and private datasets of fundus images, including the LAG database, ODIR-5 K, ORIGA, REFUGE, DRISHTI-GS1, HRF, ACRIMA, and RIM-one DL. The authors merged some of these datasets to create a merged dataset for their research, dividing the dataset into training, testing, and validation splits while preserving the ratio of glaucoma to non-glaucoma images. The authors also briefly discuss the use of CNNs and ViTs in computer vision tasks and their benefits and limitations. The model achieves 92.57% in sensitivity, 96.94% in specificity, and 97.9% in AUC. Future directions include exploring the use of larger datasets and investigating the effectiveness of other vision transformer models.

Wang et al. [62] proposed a novel ViT-based architecture, retinal ViT, for multi-label classification of retinal diseases. The study aims to demonstrate that

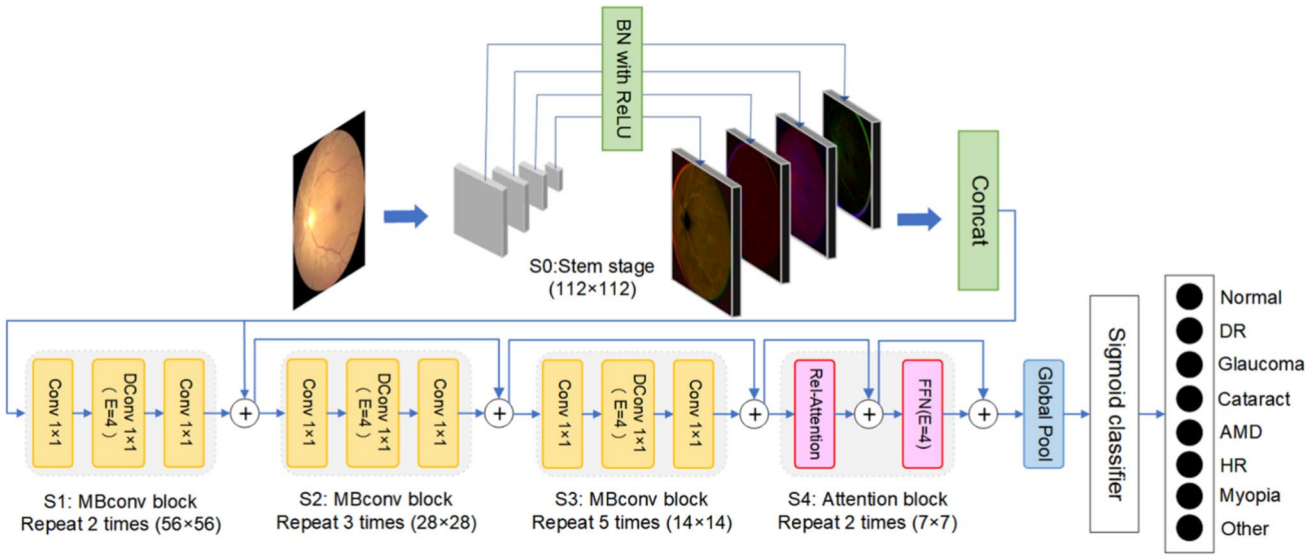


Fig. 17 The model processes retinal fundus images through a series of MobileNetV2 (MBConv) blocks, attention mechanisms, and feature extraction stages to classify eye conditions such as diabetic retinopathy (DR), glaucoma, cataract, age-related macular degeneration (AMD), high myopia (HR), and others [60]

transformer-based models can achieve performance comparable to or better than CNN-based models by removing convolutional modules entirely. The retinal ViT model employs self-attention mechanisms and a feed-forward multi-label classifier with a sigmoid activation function. The proposed approach was tested on ODIR-2019, a publicly available retinal fundus image dataset, which contains images labeled with multiple retinal diseases, including diabetic retinopathy, glaucoma, myopia, and cataract. The model's performance was compared with state-of-the-art CNN architectures such as ResNet, VGG, DenseNet, and MobileNet. Results demonstrated that the retinal ViT outperformed CNN-based models in key evaluation metrics, including Kappa coefficient 0.645, F1 score 0.919, and AUC 0.938, highlighting the effectiveness of self-attention in retinal disease classification.

He et al. [63] proposed an interpretable Swin-Poly Transformer network for automatic classification of retinal diseases using optical coherence tomography (OCT) images. The model leverages a Swin Transformer architecture to extract multi-scale hierarchical features and integrates PolyLoss to refine cross-entropy loss, optimizing classification performance. Additionally, it employs Score-CAM visualization to generate confidence score maps, enhancing interpretability by highlighting regions critical for decision-making. The proposed approach was evaluated on OCT2017 and OCT-C8 datasets. It outperformed both CNNs and conventional vision transformer models, achieving 99.80% accuracy and an AUC of 99.99% on OCT2017.

Table 5 shows a summary of the above approaches in retinal disease classification.

Coronavirus 2019 Classification

Fan et al. [64] propose a classification model for COVID-19 CT images based on the combination of the Transformer module and CNN module. The COVID-19 CT image first passes through the Transformer branch module to extract global features, then through the CNN branch module to extract local features. The two branches are then combined using a bi-directional feature fusion structure to improve the classification accuracy. The Transformer module is composed of an encoder and uses the Self-Attention module, which has a global receptive field. The CNN module uses multiple convolution kernels to extract local features. The final classification vectors from the two branches are fused, and the model is adjusted through inverse gradient calculation. Future directions include investigating the effectiveness of the method on different datasets and exploring the use of more advanced deep learning techniques.

This study of Ascencio-Cabral and Reyes-Aldasoro [65] evaluates the performance of 5 deep learning architectures in classifying COVID-19, the classifiers were built on ResNet-50, ResNet-50r, DenseNet-121, MobileNet-v3, and CaiT. 20 experiments were conducted and the metrics of accuracy, balanced accuracy, F1, F2, Matthews correlation coefficient, sensitivity, and specificity were obtained, the results show that less complex architectures such as ResNet-50, ResNet-50r, and DenseNet-121 performed better with higher rankings for the Matthews correlation coefficient, while MobileNet-v3 and CaiT had lower rankings. Future directions include exploring the use of more advanced deep learning techniques and incorporating other modalities such as X-ray and ultrasound.

Table 5 Comparative overview of vision transformer studies in retinal disease classification

	Playout et al. [55]	Rodriguez et al. [56]	Shen et al. [57]	Z. Ma et al. [58]	Z. Jiang et al. [59]	Wang et al. [60]	Torki et al. [61]	Wang et al. [62]	He et al. [63]
Specific body organs the method is applied to	Retinal images	Eye	Retina	Retina	Retinopathy in the eye	Retinal images	Eyes	Retinal images	Retinal images
Level of application	Tissues	Organs	Organ	Tissues	Tissues	Organs	Organs	Tissues	Tissues
Types of diseases included	Disease	Disease	Retinal diseases (multi-label classification)	Retinal diseases	Retinopathy	Diabetic retinopathy, glaucoma, and	Glaucoma	Disease	Disease
Modality specificity or medical image analysis in general	Retinal images	Fundus Images	Medical image analysis of retinal fundus images	OCT	Optical Coherence Tomography (OCT)	Medical image analysis in general	Medical image analysis in general	Retinal fundus image	OCT

Mehboob et al. [66] propose a method for COVID-19 diagnosis using ViTs, a computationally efficient vision model that is more domain agnostic compared to traditional CNNs, the method works by dividing an input image into fixed patches, flattening them, adding positional embeddings, and applying a transformer encoder to produce a contextualized encoding, the model employs self-attention mechanism, multi-layer perceptron, and layer normalization with residual connections. Input image resolution, patch size, number of layers, and number of epochs are varied to optimize the model, data augmentation using image flipping, resizing, and rotation is also performed, a multi-layer perceptron head is used as the classification module to output the number of classes. The accuracy of the model was 99.7% on the multi-class HUST-19 dataset and 98% on the binary-class SARS-CoV-2 dataset. When evaluated across corpora by training on the HUST-19 dataset and testing on the Brazilian COVID dataset, the accuracy was found to be 93%. Limitations include the need for larger datasets and further investigation into the effectiveness of the method on different modalities and disease types, future directions include exploring the use of more advanced deep learning techniques and investigating the interpretability of the model's predictions. Figure 18 shows the proposed approach.

Peng et al. [67] propose a DeepDSR model to classify COVID-19 CT images using three available datasets from Kaggle and Github. The datasets contain 7,398 pulmonary CT images of COVID-19 infected, normal, and other pneumonia patients. To boost the accuracy of DeepDSR, the authors use an ensemble model that integrates three state-of-the-art network architectures (DenseNet, Swin Transformer, and RegNet). The weights of these networks are pre-trained on the ImageNet dataset and then trained on the integrated larger dataset of CT images. The final classification results are obtained by integrating the results from these three models and using a soft voting approach. The authors explain the pipeline of DeepDSR and the architecture of the three models used in the ensemble model. The results indicate that DeepDSR performed best in binary and three-class classification problems. It achieved the highest precision of 0.98, recall of 0.98, accuracy of 0.98, F1 score of 0.98, AUC of 0.99, and AUPR of 0.99 in binary classification. In three-classifications, DeepDSR had the best precision of 0.97, recall of 0.96, accuracy of 0.97, and F1 score of 0.96. Limitations include lack of diversity in the dataset and the need for more rigorous validation, future directions include incorporating multi-modal data and applying the method to other types of pneumonia. Figure 19 shows the proposed approach.

Cao et al. [68] propose a network for automatic detection of pneumonia in X-ray images has two main stages: data processing and image classification. Data processing includes data transformation, data augmentation, and adaptive normalization. The data is transformed to a tensor of the same resolution, and data augmentation involves image translation and rotation to balance the dataset. Normalization is done

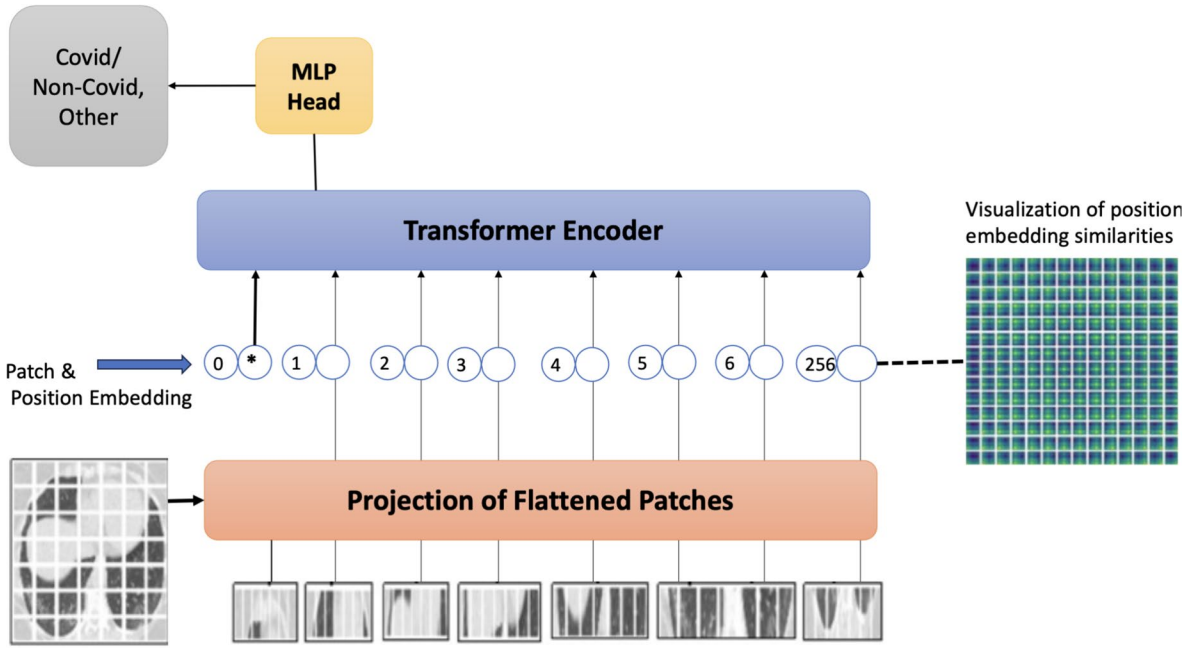


Fig. 18 The model involves the projection of flattened patches and position embeddings, followed by a transformer encoder, which processes the input. The output is then passed through a multi-layer perceptron (MLP) head for classification, distinguishing between COVID and non-COVID cases [66]

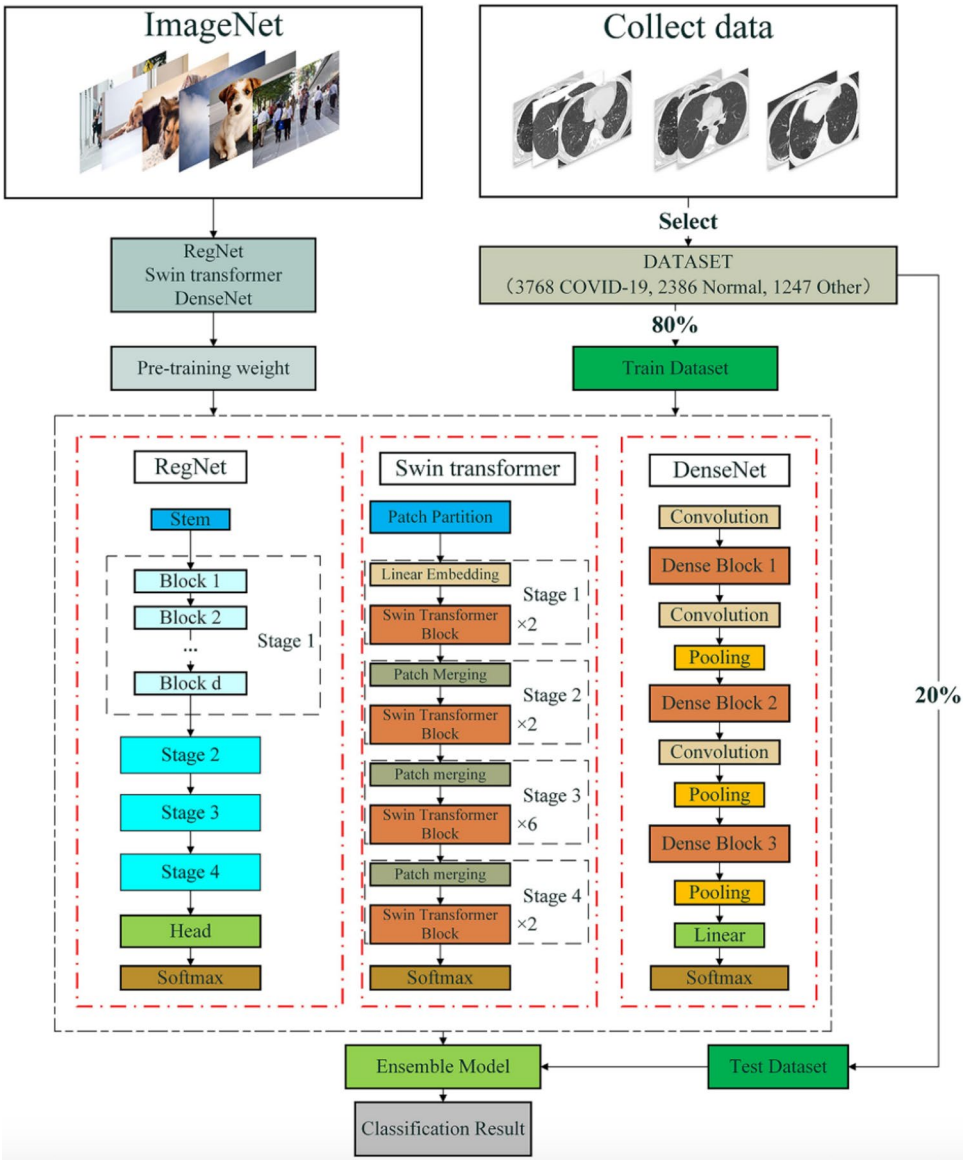


Fig. 19 The approach utilizes a pre-trained weight initialization from ImageNet and a collected dataset. The dataset is split into an 80% training set and a 20% test set. Feature extraction is performed using RegNet, Swin Transformer, and DenseNet models, which are subsequently combined into an ensemble model for final classification [67]

to adjust the dynamic range of medical images. The image classification stage involves using CNN and a Transformer network. CNN uses convolutional kernels to extract local features of images, while Transformer uses self-attention and MLP to extract global features of the image. The results show that the accuracy, precision, recall, and F1 score of the network implemented on benchmark datasets were 97.09%, 97.16%, 96.93%, and 97.04% respectively. Limitations include limited dataset size and future directions include incorporating multi-modal data and exploring different model architectures..

Zhang and Wen [69] present a vision transformer classification network for COVID-19 diagnosis in chest CTs. The network consists of two stages: lung segmentation as pre-processing followed by the image classification using Swin transformer. The Swin vision transformer is used as the backbone for feature learning. It is a hierarchical transformer built by replacing the standard Multi-head Self Attention with the shifted window in the transformer block. The Swin transformer provides a better speed-accuracy trade-off compared to the standard ViTs. The authors have compared the performance of the Transformer architecture with a CNN-based architecture. The method with a Swin transformer backbone achieved the best F1 score of 0.935 on the validation dataset, while the CNN-based EfficientNetV2 backbone had the best precision of 93.7%. The final prediction model with a Swin transformer achieved an F1 score of 0.84 on the test dataset. Limitations include dataset bias and future directions include exploring multi-task learning and incorporating clinical data.

Shome et al. [70] present a COVID-19 detection pipeline using a ViT model for transfer learning. The ViT model resembles the Transformer architecture and is trained on a

dataset with a custom MLP block added. The input image is reshaped into patches and added with positional embeddings to form a sequence. The transformer encoder used has multi-headed self-attention layers and multiple MLP blocks. The authors used the ViT L-16 model with pre-trained weights from ImageNet and fine-tuned it by removing the pre-trained MLP prediction block and adding a custom MLP block consisting of batch normalization and dense layers. The model was trained with a NovoGrad optimizer and a categorical or binary cross-entropy loss function with label smoothing. Multiple metrics including accuracy and AUC score were monitored during the training process. The proposed transformer model has a high accuracy of 98% in differentiating COVID-19 from normal chest X-rays, with an AUC score of 99% in binary classification. In multi-class classification of COVID-19, normal and pneumonia patient's X-rays, the accuracy is 92% with an AUC score of 98%. Limitations include lack of diversity in the dataset and future directions include incorporating more clinical features and exploring multi-task learning. Figure 20 shows the proposed approach.

Hongyan et al. [71] propose a Transformer (COVT) for COVID-19 diagnosis from chest-X-ray images. The network consists of a CNN block and a Transformer block. The CNN block increases the receptive field of the input image, extracts multi-scale and rich context information through atrous convolution, and inputs the information into the Transformer block. The Transformer block is based on vision transformer and has an improved multi-layer perceptron (MLP) module. The improved MLP uses both local information from the original MLP and global information from average pooling, which makes up for the original MLP's insufficient attention to global information while having a small computational cost. Limitations include lack of diverse data and future directions include exploring other modalities and incorporating clinical information. Figure 21 shows the proposed approach.

Liu and Yin [72] used two datasets for COVID-19 X-ray images, one with 3616 COVID-19 X-ray images and 3600 normal X-ray images (Dataset-1) and the other with 340 COVID-19 X-ray images and 340 normal X-ray images (Dataset-2). The study uses the Vision Outlooker (VOLO) model, which is based on a transformer and has a new attention mechanism. The study uses a pre-trained version of VOLO on ImageNet-1 K. The study uses VOLO-D3 because it outperforms D4 and D5 on the dataset used. The images are scaled to 224×224 before being fed into the model and the training is done on 8 Tesla V100 GPU with 32 GB memory each, using the AdamW optimizer and a cosine learning rate scheduler with an initial learning rate of 1.6e-3 and a decay epoch of 30. The total number of epochs is 300. The model achieved a classification accuracy of 99.7%. Limitations include the need for more diverse and balanced datasets and future directions include exploring multi-modal data and interpretability.

Heart Disease Classification

Randazzo et al. [73] proposed a vision transformer model for predicting fatal arrhythmic events in patients with Brugada Syndrome. The study utilized 12-lead electrocardiogram (ECG) images as input data, training the ViT on a dataset of 278 ECGs from 210 diagnosed Brugada Syndrome patients. The dataset was divided into two classes: "event" (94 ECGs with documented ventricular tachycardia, ventricular fibrillation, or sudden cardiac death) and "no event" (184 ECGs from patients without these conditions). The ViT model, initially trained on a balanced dataset, achieved strong performance with 89% accuracy, 94% specificity, 84% sensitivity, and an F1 score of 89%. Further optimization using an

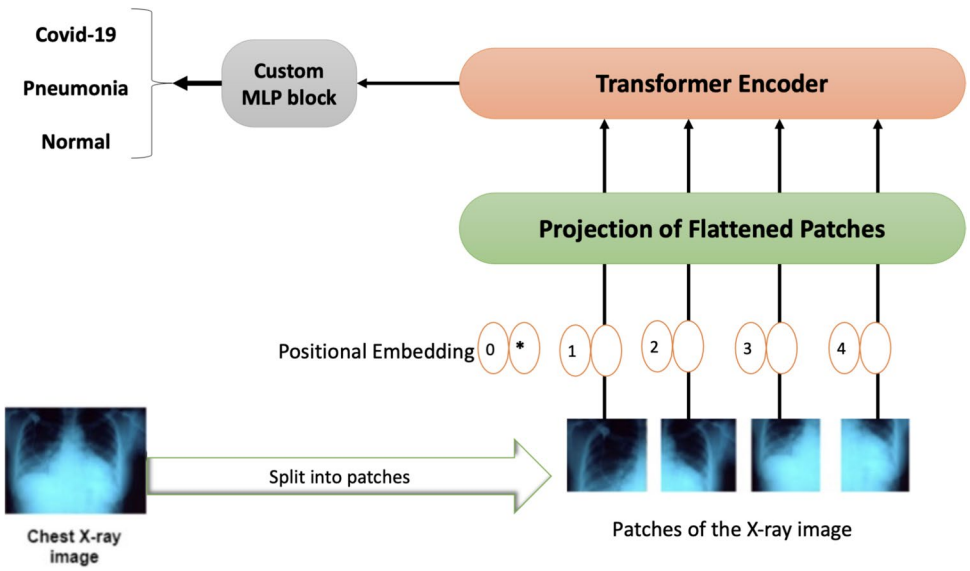


Fig. 20 An overview of the proposed ViT-based approach for COVID-19. The image is split into patches, which are projected into a flattened representation with positional embeddings. These embeddings are then processed by a transformer encoder, followed by a custom MLP block to classify the input as COVID-19, pneumonia, or normal [70]

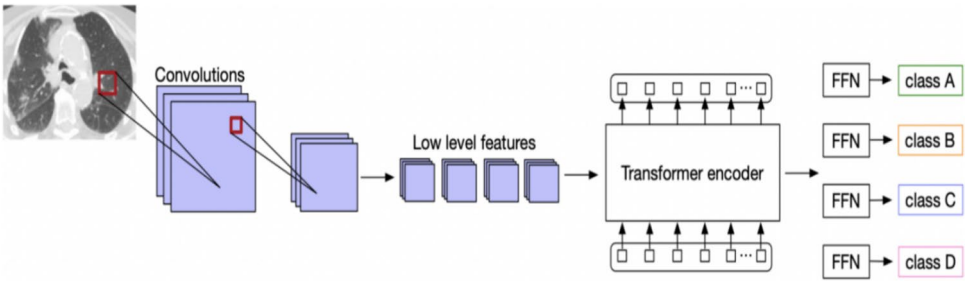


Fig. 21 The model extracts features from chest CT scan images through convolutional layers, generating low-level feature representations. These features are then processed by a transformer encoder, which captures spatial dependencies and high-level representations. Finally, a feedforward network (FFN) classifies the image into one of four categories [71]

adjusted classification threshold improved prediction on an unbalanced dataset, reaching 74% accuracy, 95% negative predictive value, and 90% sensitivity.

Che et al. [74] propose a model for classifying ECG signals into 9 categories (one normal type and eight abnormal types) based on their characteristics. The ECG data used in the paper was collected from 11 hospitals and contains 6877 records from both men and women with a frequency of 500 Hz. Pre-processing was done to improve the quality of the signals and remove noise, including using difference method and wavelet transform. The ECG signals were then split into fixed-length segments and fed into a combination of CNN and transformer networks to capture deep features and output embedding vectors, which were finally fed into a classification layer to obtain the class probability. The model also used link constraints to improve the quality of embedding features for downstream tasks. The results show that the model achieved an F1 score of 78.6%. Future work includes extending the proposed method to other arrhythmia classification tasks and evaluating the method on large-scale datasets. Figure 22 shows the proposed approach.

Reynaud et al. [75] propose a new transformer model to interpret US videos. The model is end-to-end trainable and comprises three modules: a dimensionality reduction encoder, a BERT-based spatio-temporal reasoning module, and two regressors for labeling end-systole and end-diastole frames and estimating the left ventricular ejection fraction (LVEF). The encoder uses ResNetAE to distill US frames into a smaller dimensional embedding and the BERT encoder acts as a spatiotemporal information extractor. The resulting features are passed through two regressors to predict the ES, ED frames, and the LVEF, respectively. The LVEF prediction is done by averaging the predictions over the frames to output a single prediction per video. Figure 23 shows the proposed approach.

Rafi and Woong Ko [76] propose a new deep learning method called HeartNet for automatic ECG classification in the medical field to assist medical professionals in detecting arrhythmias, which is a leading cause of mortality. The method combines a multi-head attention-based convolutional neural network and adopts a generative adversarial network to synthesize data and overcome the issue of data-label deficiency and unbalanced class distribution. The results showed an improvement in accuracy by 5–10% and a high accuracy of $99.67\% \pm 0.11$ and precision of $99.27\% \pm 0.17$ compared to other recent models. The robustness and effectiveness of the proposed method were validated by extensive experiments and comparisons with other datasets. Although the proposed method shows promising results, further research with new ML (machine learning) techniques is recommended.

Jumphoo et al. [77] propose a transfer learning approach for valvular heart disease (VHD) detection using the Data-Efficient Image Transformer (DeiT) model, pre-trained on image datasets. To enhance classification performance, they introduce a hybrid Convolution-DeiT (Conv-DeiT) model, which integrates a convolutional block with a Squeeze-and-Excitation (SE) attention mechanism. This combination refines channel and spatial information before feeding it into the DeiT model.

The study uses the Heart Sound Murmur (HSM) database, containing phonocardiogram (PCG) signals, as the input modality. The extracted Mel-Frequency Cepstral Coefficients (MFCCs) and their derivatives are transformed into an image-like representation to be processed by the transformer model.

Experimental results demonstrate that the DeiT-based transfer learning method achieves an accuracy of 97.44%, while the proposed Conv-DeiT model further improves classification performance, achieving a 99.44% accuracy.

Table 6 shows a summary of the above approaches in heart disease classification.

Colon Cancer Classification

Zeid et al. [78] applied Visito classify colorectal cancer histology images into eight categories. The dataset consisted of 5000 images, the results showed high accuracy, with the

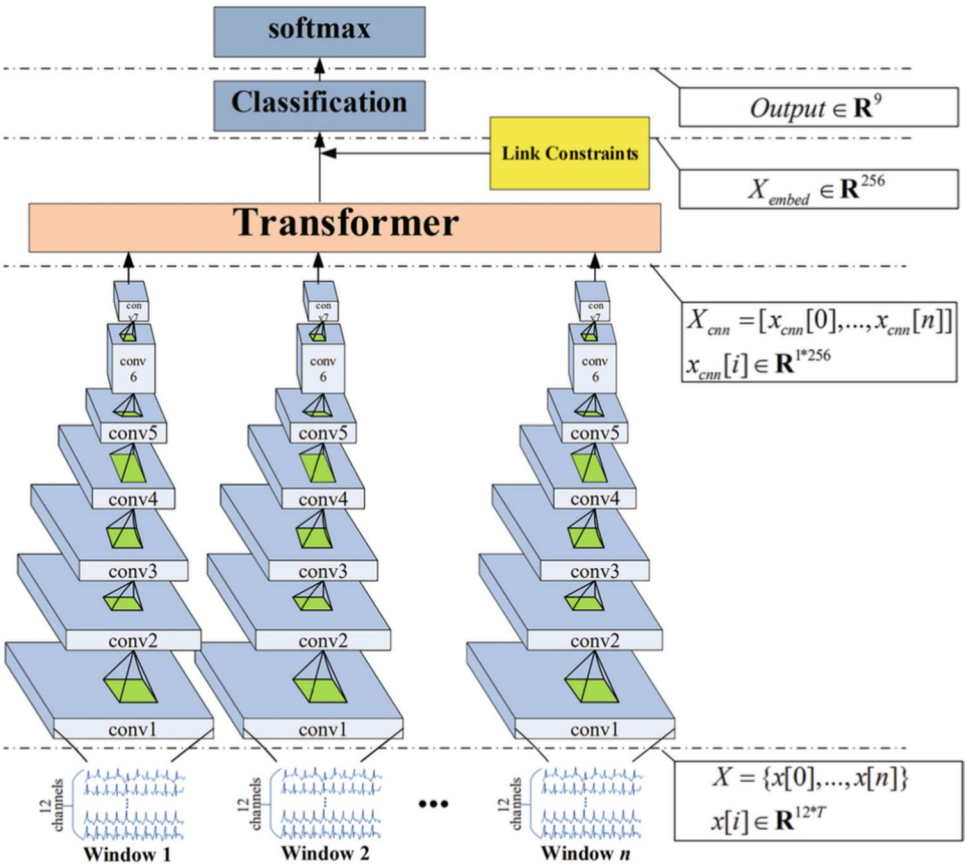


Fig. 22 The model processes input signals through multiple convolutional layers, extracting hierarchical features from segmented time windows. These feature embeddings are then passed to a Transformer module, incorporating link constraints to enhance learning. The final classification layer, followed by a softmax function, predicts heart disease conditions based on the extracted features [74]

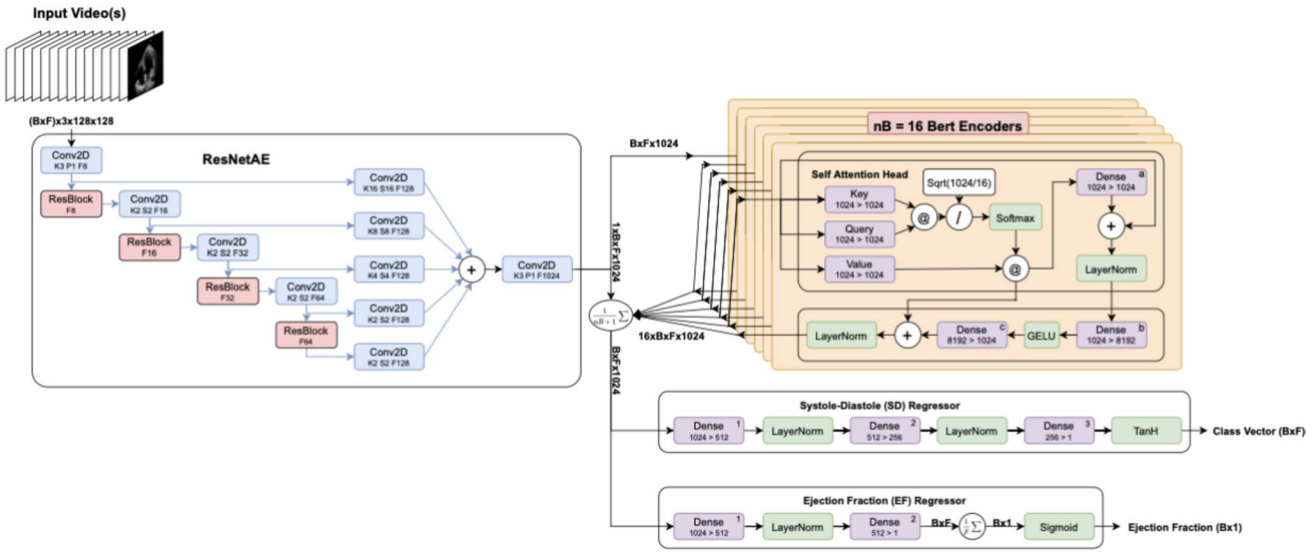


Fig. 23 The framework integrates a ResNetAE-based feature extractor for video input processing, followed by multiple BERT encoders for feature transformation. The model further branches into two regressors for systole-diastole ratio (SDR) and ejection fraction (EF) prediction, facilitating comprehensive heart disease classification [75]

vision transformer achieving 93.3% and the Compact Convolutional Transformer achieving 95%. These results outperformed previous work on the same dataset and demonstrated the potential for using Transformers in the histopathological image domain, further investigation of interpretability and visualization methods required.

Table 6 Comparative overview of vision transformer studies in heart disease classification

	Randazzo et al. [73]	Che et al. [74]	Reynaud et al. [75]	Rafi and Woong Ko [76]	Jumphoo et al. [77]
Level of application	Organs	Cells/tissues	Organs	Organs	Organs
Types of diseases included	Arrhythmia	Arrhythmia	Arrhythmia	Arrhythmia	Valvular heart disease
Modality specificity or medical image analysis in general	ECG	ECG signal processing	Ultrasound	ECG	Phonocardiogram
Dimensionality applicability	1D	1D	Video	1D	1D
Training approaches discussed	Transfer learning	Transfer learning	Transfer learning	Adversarial data synthesis	Transfer learning
Full architectures versus lightweight models	Pre-trained ViT mode	Constrained Transformer Network	Ultrasound Video Transformers	HeartNet	Full DeiT architectures with a hybrid Conv-DeiT model

Cai et al. [79] present a new approach for classifying colorectal adenoma whole slide images (WSIs) for early screening of colorectal cancer, the approach, called MIST, uses a Swin Transformer-based network to extract features and a dual-stream multiple instance learning network to predict the class of slides without manual annotation. The study found that MIST's accuracy was superior to existing methods and comparable to the accuracy of local pathologists, MIST is a low-burden, interpretable, and effective approach that may assist clinicians in the decision-making process and potentially reduce the mortality of CRC patients, further investigation of generalizability to other types of adenomas required.

Sui et al. [80] present a novel dataset for colorectal cancer MRI image analysis consisting of 375 cases with 289 negatives and 86 positives; the dataset is divided into training, testing, and validation sets. The authors propose a multitask framework for tumor region detection and segmentation the framework consists of two pipelines, the tumor region detector, and the tumor segmentation pipeline. The detection pipeline uses an encoder-decoder model, and the segmentation pipeline uses image patches and a transformer. The loss function for the tumor detection part is optimized through the Hungarian algorithm. The final output of the framework is a predicted tumor bounding box and tumor/nontumor class labels. Figure 24 shows the proposed approach.

Chen et al. [81] proposed a framework called IL-MCAM that consists of 5 steps: (1) dividing original images into training, validation, and test sets, (2) augmenting the training set images, (3) training the MCAM model using three parallel channels with different attention mechanisms, (4) iteratively retraining the model with human–machine interaction on misclassified images from the validation set, and (5) testing the final model with the test set images. The MCAM model uses three parallel channels with different attention mechanisms to extract global and local information from the images. The training process uses a transfer learning approach, and the final model parameters are obtained through several iterations of human–machine interaction on misclassified images from the validation set. The framework achieved an accuracy of 98.98% on HE-CRC-DS dataset and 99.77% on HE-NCT-CRC-100 K dataset. The authors discuss limitations and future directions related to the proposed multi-channel attention mechanism and weakly supervised learning approach. Figure 25 shows the proposed approach.

Lv et al. [82] present a new approach for colorectal cancer survival analysis using a transformer-based model called "TransSurv." It integrates multiple sources of data, including histopathological images, genomic data, and clinical information, by using a multi-scale histopathological features fusion transformer and a cross-modal fusion transformer based on cross attention. The results of the study on the Cancer Genome Atlas (TCGA) colorectal cancer cohort show that TransSurv outperforms existing methods and improves prognosis prediction.

Khalid et al. [83] propose an efficient colorectal cancer detection network (CCDNet) that integrates atrous convolution with a coordinate attention transformer to enhance histopathological image classification. Their approach employs a Wiener-based Midpoint Weighted Non-Local Means filter for denoising and preserving image features, followed by a multiscale atrous convolution (MAConv) and a Cross-shaped Window (CrSWin) transformer to capture both local and global features. The model was evaluated on colorectal histological images and the NCT-CRC-HE-100 K dataset, achieving impressive accuracy rates of 98.61% and 98.96%, respectively.

Table 7 shows a summary of the above approaches in colon cancer classification.

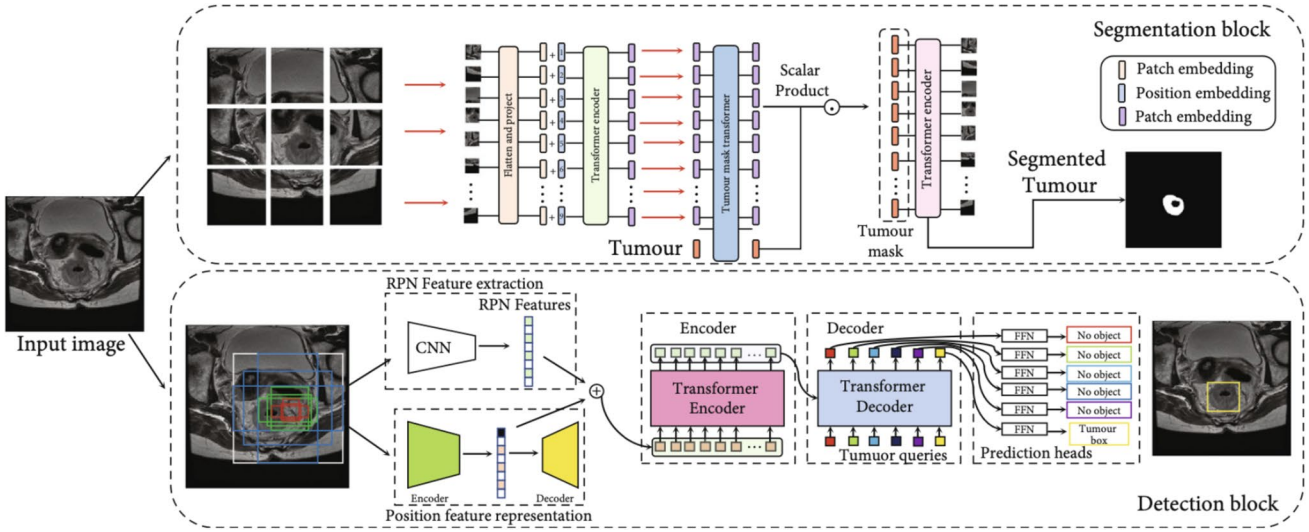


Fig. 24 The approach consists of two main blocks: segmentation block and detection block. The segmentation block extracts tumour regions from the input image using patch embeddings and position embeddings, producing a tumour mask. The detection block utilizes a CNN for RPN feature extraction, a transformer encoder-decoder architecture for tumour queries, and prediction heads for classification [80]

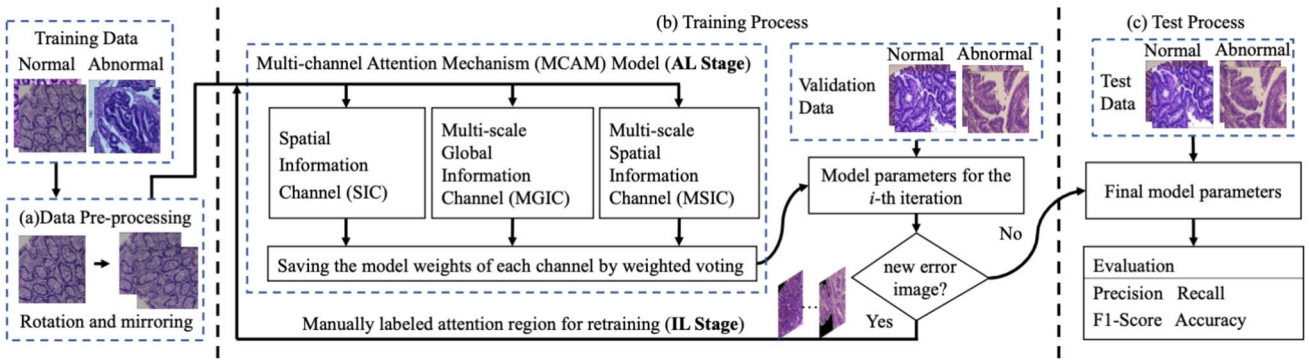


Fig. 25 The approach consists of three main stages: **a** data pre-processing, where training images undergo augmentation techniques such as rotation and mirroring. **b** Training process, which involves feeding the pre-processed images into the MCAM model, integrating spatial and multi-scale information channels, followed by iterative model updates and validation. **c** Test process, where the final trained model is evaluated on test data using metrics like precision, recall, F1 score, and accuracy [81]

Neurological Disease Diagnosis with ViTs

The study of Lyu et al. [84] focuses on cross-domain transfer learning using a ViT as a backbone. The ViT was initially trained on natural images using ImageNet-21 K dataset, and then transferred to the brain imaging domain with limited samples. The brain imaging data was collected from 505 subjects, but 375 were used after quality control. The T1-weighted MRI data was pre-processed and resized into slices that were used to train the model. The architecture of ViT consists of a stack of transformer encoder layers, each with multi-head attention and multi-layer perceptron sub-layers. The pipeline of ViT is formulated by equations. The pretrained ViT was adapted to brain imaging data by convolutional patch-wise embedding. The model was trained for AD/CN classification tasks. The approach is limited to Alzheimer's disease (AD) classification and needs further validation on other datasets. Future work may include incorporating multimodal data and explainability. Figure 26 shows the proposed approach.

Li et al. [85] propose a new architecture called TransResNet, which combines the strengths of convolutional neural networks and transformers, for brain disease classification from MRI data. This architecture outperforms state-of-the-art CNN-based methods in Alzheimer's disease prediction by utilizing the attention mechanism of Transformers and being pre-trained on a large-scale dataset for brain age

Table 7 Comparative overview of vision transformer studies in colon cancer classification

	Zeid et al. [78]	Cai et al. [79]	Sui et al. [80]	Chen et al. [81]	Lv et al.[82]	Khalid et al. [83]
Types of diseases included	Colorectal cancer	Colorectal adenomas	Malignant	Malignant	Colorectal cancer	Colorectal cancer
Modality specificity or medical image analysis in general	Histopathology	Whole slide imaging	Histopathology	Histopathology	Histopathology, Genomic data	Histopathology
Training approaches discussed	Transfer learning	Multiple instance learning	Transfer learning	Transfer learning	Transfer learning	Full architecture

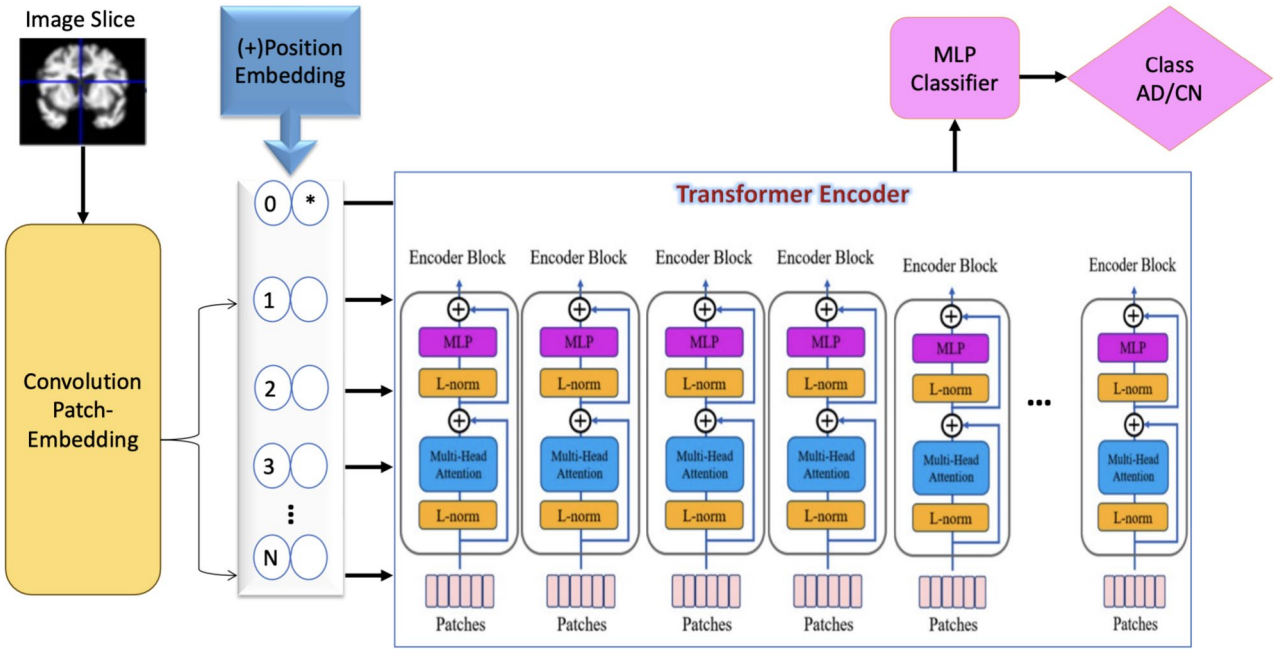


Fig. 26 The approach involves extracting image slices, applying convolutional patch embedding, and incorporating positional embeddings before feeding the sequence into a transformer encoder. The encoder consists of multiple blocks with multi-head self-attention, layer normalization, and MLP layers [84]

estimation, the approach needs further validation on other datasets and may benefit from incorporating multimodal data and interpretability.

Sarraf et al. [86] propose an architecture for computer vision tasks called optimized vision transformer (OViTAD); this architecture is based on the transformer architecture for natural language processing and uses multi-head self-attention (MHSA) layers as its core; the input dimension of the transformer is set based on the data for the task of multiple-class Alzheimer's prediction; the number of heads in MHSA is reduced to improve efficiency. A positional embedding is used to feed arrays to the transformer and a multi-perceptron layer is used to translate the features to a format usable for the cross-entropy loss function. The paper also uses two other ViTs, DeepViT and class-attention in image transformers (CaIT), as baselines for comparison; the 4D fMRI data is decomposed into 2D images and preprocessed before being split into training, validation, and testing sets; the main objective is to perform multiclass prediction and binary classifications. The model achieved an F1 score of 97% with a ± 0.0 margin of error and 99.55% with a ± 0.39 margin of error in the triple-class prediction experiments using rs-fMRI and sMRI modalities, respectively. Additionally, it accomplished these results with 30% fewer parameters compared to a basic transformer. The approach needs further validation on other datasets, and future work may include incorporating multimodal data and developing a more explainable model.

Hu et al. [87] proposed a deep learning model called VGG-TSwinformer for the short-term longitudinal study of mild cognitive impairment (MCI) which is a transition state between normal aging and Alzheimer's disease. The model is based on CNN and transformer and uses VGG-16-based CNN to extract low-level spatial features of longitudinal

sMRI images, sliding-window attention to fine-tune feature representations and temporal attention to measure the evolution of features. The model was tested on the ADNI dataset and outperformed other cross-sectional studies in terms of accuracy, sensitivity, and AUC for the classification task of stable MCI vs progressive MCI.

Schizophrenia (SZ) is a complex psychiatric disorder requiring early and accurate diagnosis to improve patient outcomes. Liu et al. [88] utilized vision transformers to extract meaningful facial feature representations directly from 921 facial diagnostic images, establishing a benchmark with six comparison methods and four evaluation metrics. The ViT model demonstrated a 3–10% increase in accuracy over traditional SZ detection approaches, emphasizing its ability to capture key facial regions relevant to the disorder. The study further employed attention rollout visualization, highlighting that the eyes, mouth, forehead, cheeks, and nose contribute to SZ diagnosis in descending order of importance—findings that align with TCM diagnostic observations.

By integrating self-attention mechanisms for deep semantic feature learning, this method offers an efficient, interpretable, and non-invasive alternative to conventional SZ detection techniques. The results suggest that ViTs can facilitate a clinically meaningful, automated approach for psychiatric diagnosis, broadening the scope of AI-driven mental health assessments.

Bedel et al. [89] proposed BolT (Blood-Oxygen-Level-Dependent Transformer), a vision transformer-based model designed for functional MRI (fMRI) time-series analysis, addressing limitations in traditional deep learning models that struggle with integrating contextual representations across diverse time scales. BolT introduces a fused window attention mechanism that hierarchically encodes local-to-global representations by processing overlapping temporal windows, allowing for efficient and scalable fMRI data analysis. The model was evaluated on large-scale public datasets, including the Human Connectome Project (HCP) and the Autism Brain Imaging Data Exchange (ABIDE), where it demonstrated superior classification performance in tasks such as gender detection, cognitive task classification, and Autism Spectrum Disorder (ASD) diagnosis. By incorporating cross-window attention and cross-window regularization, BolT enhances feature alignment across time-series windows, leading to improved classification accuracy while maintaining computational efficiency. Furthermore, its gradient-weighted attention visualization technique enables the identification of critical time points and brain regions contributing most significantly to predictions, corroborating established neuroscientific findings.

Alharthi and Alzahrani [90] proposed a multi-slice generation approach for Autism Spectrum Disorder (ASD) diagnosis using both structural MRI (sMRI) and functional MRI (fMRI) modalities. Their study leveraged four vision transformers (ConvNeXt, MobileNet, Swin, and ViT) for sMRI analysis and a 3D-CNN model for both sMRI and fMRI classification. To improve ASD detection accuracy, they explored various methods of extracting slices from 3D sMRI and 4D fMRI along axial, coronal, and sagittal planes. The ConvNeXt model achieved the highest performance among ViT-based methods, particularly when using 50-middle slices from sMRI, while the 3D-CNN model on fMRI data outperformed other approaches, obtaining a maximum accuracy of 87.1% and an F1 score of 82.6%. Their findings demonstrate the effectiveness of vision transformers and 3D-CNN architectures in ASD diagnosis, particularly when utilizing optimal slice selection techniques to enhance classification accuracy.

Table 8 shows a summary of the above approaches in neurological disease diagnosis.

Diabetic Retinopathy

Gu et al. [91] propose a network which is a classification model that distinguishes 5 classes of diabetic retinopathy, the model consists of two key components: the feature extraction block (FEB) and the grading prediction block (GPB), the FEB is used to extract features from the image by reshaping the image into a sequence of flattened 2D patches and passing it through a transformer encoder. The GPB uses a class-specific residual attention algorithm to generate class-specific features for the input image extracted by the FEB, and then uses average pooling to fuse the features, the model outputs 5 probability scores corresponding to the 5 classes of diabetic retinopathy. Figure 27 shows the proposed approach.

Adak et al. [92] propose a solution architecture for predicting the severity stage of diabetic retinopathy (DR) in a fundus photograph; the task is formulated as a multi-class classification problem where features are extracted from the input image and fed to a classifier; the solution uses an ensemble of image transformers (ViT, BEiT, CaiT, and DeiT) and preprocesses the raw fundus images with techniques such as resizing, data augmentation, and CLAHE. The individual image transformers are adapted for the task, and the ViT model converts the input image into a series of patches and extracts features through transformer layers.

Wu et al. [93] present a vision transformer-based method for recognizing diabetic retinopathy using fundus images, the images are divided into patches, converted into sequences, and input into multi-head attention layers to generate a final representation, which is then fed into a softmax classification layer for recognition. The proposed method is compared against current CNNs and extreme learning machines and achieves an accuracy of 91.4% with high specificity, precision, sensitivity, quadratic weighted kappa score, and area under the curve.

Table 8 Comparative overview of vision transformer studies in neurological disease diagnosis

	Lyu et al. [84]	Li et al. [85]	Sarraf et al.[86]	Hu et al. [87]	Liu et al.[88]	Bedel et al. [89]	Alharthi and Alzahrani [90]
Modality specificity or medical image analysis in general	MRI	MRI	Resting-state fMRI and structural MRI	MRI	Face diagnosis image	fMRI	sMRI, fMRI
Dimensionality applicability	3D	2D	3D	2D	2D	4D	3D, 4D
Full architectures versus lightweight models	Vision transformer	Trans-ResNet: integrating transformers and CNNs	Optimized vision transformer (OViTAD)	VGG-TSwinformer: Transformer-based deep learning model	Vision transformer	Blood-oxygen-level-dependent transformer	ConvNeXt, MobileNet, Swin, and ViT

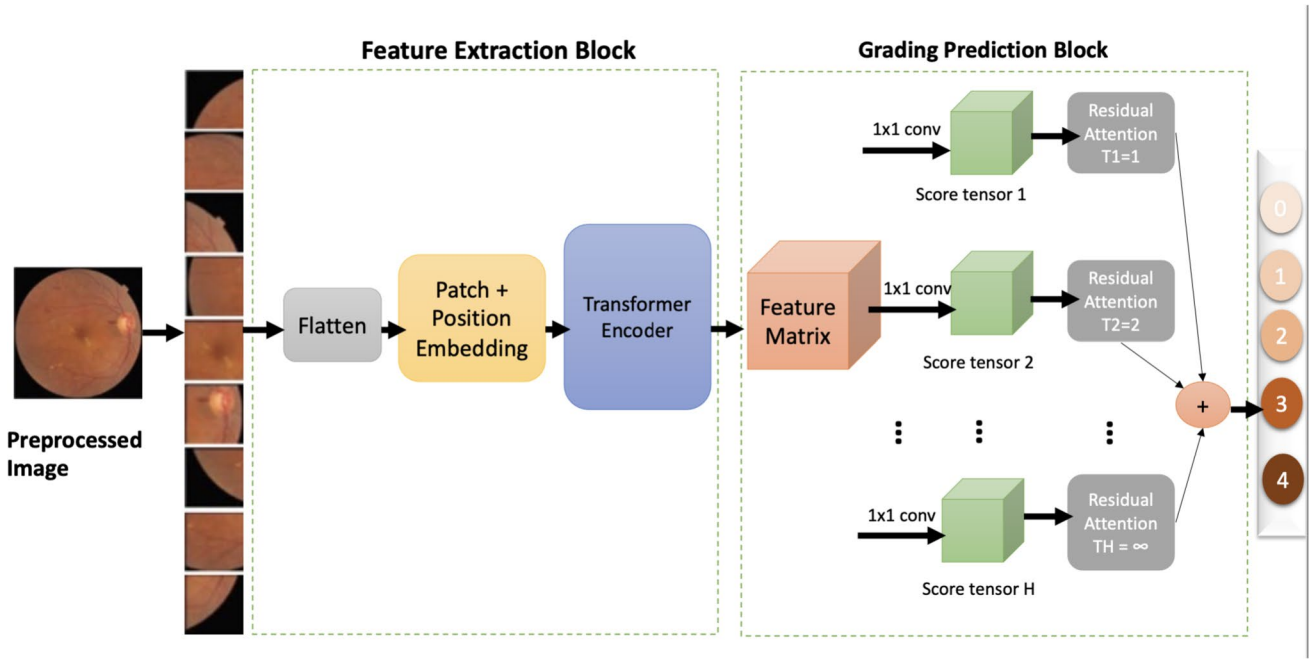


Fig. 27 The model comprises a feature extraction block utilizing a transformer encoder and a grading prediction block leveraging residual attention mechanisms for severity classification [91]

Oulhadj et al. [94] proposed a novel deep learning-based approach for diabetic retinopathy (DR) severity level classification using a hybrid model combining a fine-tuned vision transformer and a modified Capsule Network. Their methodology includes an enhanced preprocessing pipeline that applies Power Law Transformation (PLT) and Contrast-Limited Adaptive Histogram Equalization (CLAHE) to improve the quality of retinal fundus images. The model was evaluated on four public datasets—APTOS, EyePACS, Messidor-2, and DDR—achieving test accuracies of 88.18%, 87.78%, 80.36%, and 78.64%, respectively.

The papers present different vision transformer models for recognizing diabetic retinopathy in fundus images, all four models divide the input image into patches, convert them into sequences, and feed them into a transformer-based architecture for feature extraction, the difference lies in the architecture of the feature extraction block and the grading prediction block. Gu et al. propose a feature extraction block (FEB) that reshapes the image into a sequence of flattened 2D patches, followed by a transformer encoder, and a grading prediction block (GPB) that uses a class-specific residual attention algorithm. Mohan et al. use a single ViT for representation, while Adak et al. use an ensemble of image

transformers (ViT, BEiT, CaiT, and DeiT) and preprocess the raw fundus images with techniques such as resizing and data augmentation. Wu et al. use a single ViT for recognition and achieve an accuracy of 91.4%. All four models aim to recognize diabetic retinopathy in fundus images, with slightly different approaches to feature extraction and classification.

Skin Disease Classification

Cai et al. [95] present a model for multimodal fusion of image and metadata features for skin disease classification. The model consists of two encoders (Image Encoder and Metadata Encoder) and a decoder. The Image Encoder uses a ViT (NesT or ViT-L) model as the backbone for transfer learning to extract image deep features. The Metadata Encoder uses a Soft Label Encoder to embed metadata into soft labels and extract metadata features. The decoder fuses the multimodal features using a proposed Mutual Attention block and outputs the result through a FFN and SoftMax layer. The choice of ViT model as the backbone in Image Encoder was based on comparison of different ViT and CNN models on private and benchmark datasets (ISIC 2018). The metadata consists of limited textual descriptions of the patient's clinical information and is encoded as a 13-dimensional or 39-dimensional vector based on the dataset used. The experiments show that the proposed model achieves an accuracy of 0.9381 and an AUC of 0.99 on ISIC 2018 dataset. Figure 28 shows the proposed approach.

Xin et al. [96] propose an end-to-end architecture for skin cancer classification consisting of five steps: (1) acquisition of dermoscopic skin cancer images using two datasets, (2) data normalization, augmentation, and balanced sampling to deal with insufficient data and imbalanced data, (3) feature extraction using a multi-scale vision transformer, (4) contrastive learning to encode similar skin cancer data similarly, and (5) validation of results. The authors used z-score normalization to normalize the data, data augmentation to increase the amount of data, and balanced sampling to alleviate the problem of data imbalance. The architecture is designed to improve the accuracy of the model in clinical data. The proposed model has shown accuracy of 94.3% on the HAM10000 dataset and 94.1% accuracy on the datasets.

Wu et al. [97] provide a comprehensive overview of the latest deep learning algorithms for skin cancer classification. It starts by introducing three types of dermatological images used in diagnosis and lists publicly available datasets. Then, it reviews the successful applications of typical convolutional neural networks for skin cancer classification. The paper also highlights several challenges in the skin cancer classification task, such as data imbalance, data limitation, and model efficiency, and provides corresponding solutions. The general development direction of these approaches is found to be structured, lightweight, and multimodal. The paper concludes that while there are still many challenges to overcome in the future, the growing popularity of deep learning offers many opportunities.

Huang et al. [98] propose LSATrans, a ViT-based framework tailored for severity scoring of skin diseases. Unlike standard self-attention mechanisms, they introduce Lesion-aware Self-Attention (LSA), which captures both lesion and normal surrounding skin features while modeling their relationship. Additionally, they employ contrastive learning to enhance the robustness of the learned representations. The model is evaluated on three skin disease severity scoring tasks—scar (VSS score), atopic dermatitis (EASI score), and psoriasis (PASI score)—achieving mean absolute errors (MAEs) of 0.5895, 0.5614, and 0.5416, respectively, outperforming existing CNN and ViT-based methods. LSATrans is further validated on skin disease classification tasks, achieving AUCs of 0.9774 and 0.9801 for general and fine-grained skin disease classification.

These papers discuss the use of deep learning algorithms for skin disease classification. The first paper introduces a

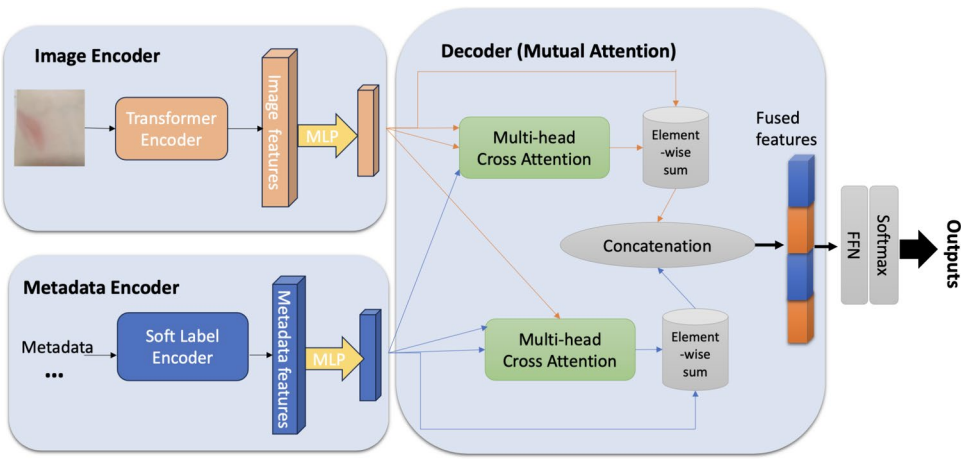


Fig. 28 The model consists of an Image Encoder using a transformer encoder to extract image features and a Metadata Encoder that processes auxiliary metadata using a Soft Label Encoder. A Decoder with Mutual Attention Mechanism integrates these features using Multi-head Cross Attention layers. The fused features undergo concatenation and element-wise summation before being passed through an MLP classifier with Softmax activation to generate the final skin disease classification output [95]

multimodal fusion model using a ViT model as the backbone of the image encoder and a Soft Label Encoder for metadata features. The second paper proposes the Swin-SimAM network for melanoma detection using the Swin Transformer for feature extraction and the SimAM attention module for focusing on important parts of skin lesions. The third paper presents an end-to-end architecture for skin cancer classification, including data normalization, augmentation, feature extraction using a multi-scale vision transformer, and contrastive learning. The fourth paper is a comprehensive overview of deep learning algorithms for skin cancer classification, including a review of CNNs and an analysis of challenges and solutions in the field.

Kidney Disease Classification

Islam et al. [99] collected a dataset of kidney images for the diagnosis of kidney stone, tumor, normal, and cyst. The data was preprocessed using techniques such as data augmentation and normalization and split into training and testing sets. Six models were trained on the dataset including three Visual Transformer variants, Inception v3, Vgg16, and Resnet 50. The models were trained on Google Colab Pro Edition with GPU and evaluated using previously unseen data. The authors made the dataset publicly available on Kaggle. The results show that the different diagnoses have partially overlapping distributions, making it difficult to classify using only statistical features.

Zhang et al. [100] proposed a ViT model for classifying renal pathology using whole-slide histopathology images stained with HE, MASSON, PAS, and PASM. The study involved 1861 WSIs from 635 patients, with images tiled into 384×384 pixels and classified into 14 pathological types. The ViT model, built using the Timm library, was compared against ResNet50, demonstrating an accuracy (0.96–0.99) across all staining methods. The study also employed visualization techniques, including Deep Taylor Decomposition and a random mask method, to interpret the ViT's focus on pathological features, revealing its ability to capture significant renal structures better than CNNs.

Lymph Node Disease Classification

Liu et al. [101] propose a new solution for cervical ultrasound lymph-node-level classification, which is important for accurate diagnosis and surgical treatment of diseases. Currently, this classification is subject to human error as it relies on physician judgment. The proposed solution, Depth-wise Separable Convolutional Swin Transformer, combines deep-wise separable convolution and self-attention mechanisms to capture global and local features, and a new loss function is proposed to address data imbalance. The model was tested on 1146 cases with an average accuracy of 80.65% and showed consistency with expert observations.

Jiang et al. [102] propose CSMViT, a lightweight hybrid Transformer-CNN model for diagnosing lymph node metastases in pathological images. The model enhances MobileViT by integrating an improved Ghost module to reduce parameters and improve efficiency, a Cross Scale Attention (CSA) module for multi-scale feature fusion, and a modified network backbone for improved feature extraction.

The study focuses on lymph node pathological images from papillary thyroid carcinoma (PTC) patients. Experimental results show CSMViT achieving 99.42% accuracy, 99.4% F1 score, and 99.6% specificity, with an 8% improvement in detection accuracy over the original MobileViT. Additionally, the model achieves 72.19% fewer FLOPs, significantly reducing computational complexity while maintaining high classification performance.

Bone Analysis

Tanzi et al. [103] proposed a retrospective study of patients with proximal femur fractures in a Level-I trauma center. The study was approved by the Ethics Review Board and included patients over 18 years old who were treated in the Emergency Department between 2013 and 2020. Radiographs and demographic data were collected and analyzed using a computerized dataset. The images were labeled and reviewed by a senior trauma surgeon and then divided into different fracture types. The final dataset consisted of 4207 manually annotated images. The study used ViT (vision transformer) architecture to classify the fractures, but due to a limited number of images, a Compact Convolutional Transformer was applied instead. The performance of the proposed ViT architecture in the paper is good, with the ViT correctly predicting 83% of the test images and having a precision of 0.77, recall of 0.76, and F1 score of 0.77.

Wu et al. [104] proposed SVTNet, a ViT-based approach for automatic bone age assessment (BAA) using Tanner-Whitehouse 3 (TW3) methodology. The method processes left-hand X-ray images and extracts 20 key regions (13 from the radius-ulna-short (RUS) bones and 7 from the carpal (C) bones). SVTNet first segments the hand region using a pre-trained YOLO network, then localizes key points with a Spatial Configuration Network (SCN), and finally classifies the extracted patches using multiple ViT models. The study utilized 3,871 pediatric hand X-rays from a tertiary hospital, and results showed an average absolute error of 0.50 years for RUS bones and 0.47 years for C bones, demonstrating an accuracy comparable to expert orthopedic surgeons.

Table 9 shows a summary of the above approaches in kidney disease classification, lymph node disease classification, and bone analysis.

Limitations of Vision Transformers in Medical Imaging

Vision transformers have significantly advanced medical image analysis by enabling long-range contextual understanding and achieving state-of-the-art results in disease classification, segmentation, and reconstruction. However, despite their impressive performance, ViTs exhibit several inherent limitations that impact their applicability in medical imaging. These challenges primarily involve high computational costs, data efficiency concerns, local spatial precision, and model interpretability. Consequently, researchers are exploring alternative architectures such as State-Space Models (SSMs), [105, 106], as complementary or alternative approaches to ViTs in certain tasks. This section outlines the fundamental challenges of ViTs and their impact on medical imaging applications.

Computational Complexity and Resource Constraints

A major drawback of ViTs is their computational overhead, primarily due to the self-attention mechanism. Unlike convolutional neural networks, which process images with localized filters, ViTs utilize global self-attention, where each image patch interacts with every other patch. This results in quadratic complexity ($O(N^2)$) concerning image resolution, making ViTs inefficient for high-dimensional medical imaging modalities such as MRI, CT, and histopathological slides.

- Memory Bottleneck: ViTs demand significant GPU memory, which can be a limiting factor for deployment in resource-constrained medical imaging environments.
- Slow Training and Inference: ViTs typically have hundreds of millions of parameters, leading to prolonged training times and slow inference speeds, which may hinder real-time clinical applications.
- Scalability Issues: Processing high-resolution 3D medical images (e.g., volumetric MRI or CT scans) is particularly challenging due to the exponential increase in memory and computation requirements compared to CNN-based approaches.

Data Requirements and Annotation Challenges

ViTs are data-hungry models, requiring extensive labeled datasets for effective training. Unlike CNNs, which leverage inductive biases (e.g., spatial hierarchies), ViTs learn feature representations from scratch, making them more dependent on large-scale data to generalize effectively. However, obtaining well-annotated medical datasets is challenging due to:

- Limited Access to Labeled Data: Medical datasets are often small due to patient privacy regulations and ethical concerns regarding data sharing.
- High Annotation Cost: Expert radiologists and pathologists are required for precise annotation, which increases the difficulty of obtaining large-scale labeled datasets.
- Class Imbalance in Medical Imaging: Many medical conditions are rare, leading to severe class imbalance, which can degrade ViT performance unless specific training techniques such as self-supervised learning (SSL) and knowledge distillation are employed.

Local Feature Representation and Spatial Precision

Medical images contain fine-grained anatomical structures crucial for disease diagnosis and segmentation. Unlike CNNs, which use localized receptive fields to maintain

Table 9 Comparative overview of vision transformer studies in kidney disease classification, lymph node disease classification, and bone analysis

	Islam et al. [99]	Zhang et al. [94]	Liu et al. [101]	Jiang et al. [102]	Tanzi et al. [103]	Wu et al. [104]
Specific body organs the method is applied to	Kidney	Kidney	Neck	Neck	Lower extremities	Hand bones
Level of application	Organs	Organs	Organs	Organs	Tissues	Organs
Types of diseases included	Tumor	Tumor	Tumor	Tumor	Fracture	Growth disorders
Modality specificity or medical image analysis in general	CT-radiography	Histopathological	Ultrasound	Pathological	Medical image analysis in general	X-ray analysis
Dimensionality applicability	3D	2D	2D	2D	2D	2D
Training strategies covered	Supervised	Supervised	Supervised	Supervised	Supervised	Supervised
Training approaches discussed	Transfer learning	Transfer learning	Transfer learning	Lightweight architecture	Transfer learning	Full architecture

spatial hierarchies, ViTs tokenize images into patches and apply self-attention globally. While this allows for excellent global context modeling, it may fail to capture fine spatial details, leading to:

- Blurry segmentation boundaries, which impact tumor detection and organ segmentation.
- Loss of microstructural details in high-resolution pathology images.
- Reduced edge detection accuracy, critical for tasks like retinal vessel segmentation or bone fracture detection.

Hybrid CNN-ViT architectures have been explored to combine local and global feature extraction, but these models introduce additional complexity and require extensive fine-tuning.

Model Interpretability and Clinical Trustworthiness

AI models in healthcare must be interpretable and transparent, particularly for high-stakes applications such as cancer detection, stroke diagnosis, and radiology. ViTs, being complex architectures, pose challenges in explainability:

- Lack of intrinsic interpretability: Unlike CNNs, which allow visualization through activation maps, ViTs do not naturally produce saliency maps or feature importance scores.
- Regulatory and Ethical Constraints: Many healthcare regulations require AI models to provide explainable justifications for medical decisions, a challenge for ViTs due to their black-box nature.
- Potential for Bias and Spurious Correlations: Without explicit local constraints, ViTs might learn non-medically relevant biases, potentially leading to erroneous diagnoses.

Research efforts, including attention heatmaps and gradient-based explainability methods, aim to improve ViT interpretability but remain an open challenge.

Emerging Alternatives: State-Space Models

Given these limitations, State-Space Models are being investigated as an alternative or complementary approach to ViTs in medical imaging. SSMs, such as Mamba-based models (I2I-Mamba, MambaRoll), have demonstrated promising results in tasks like medical image synthesis and reconstruction.

While SSMs are not direct replacements for ViTs, they offer a promising pathway to address computational and spatial limitations, particularly in multi-modal medical image synthesis and reconstruction.

Future Research Directions for ViTs in Medical Imaging

Despite these challenges, ViTs remain a crucial component of medical AI research. To enhance their utility, future efforts should focus on:

Hybrid ViT-SSM Architectures: Combining ViTs with SSM modules to optimize computational efficiency and feature representation.
Self-Supervised and Transfer Learning: Reducing data dependency by leveraging SSL and domain adaptation techniques.
Efficient ViT Variants: Developing lightweight ViTs with reduced parameters tailored for medical imaging applications.
Explainable ViTs: Enhancing model interpretability using improved attention visualization techniques and post-hoc explainability methods.

ViTs have played a transformative role in medical image analysis, achieving state-of-the-art results in disease classification, segmentation, and reconstruction. However, their high computational costs, data inefficiency, and weak local feature modeling present significant challenges. Emerging research suggests that State-Space Models may provide more computationally efficient and spatially precise alternatives, particularly for tasks like medical image synthesis and reconstruction. Moving forward, a hybrid approach integrating ViTs and SSMs may offer the best balance between global context modeling, computational efficiency, and spatial precision in medical imaging applications.

Discussion and Future Directions

Vision transformers have demonstrated significant potential in medical imaging tasks, including classification, segmentation, and disease detection. Their self-attention mechanism allows for the effective capture of both local and global image features, making them a compelling alternative to convolutional neural networks (CNNs). However, despite their advantages, ViTs still face several challenges and limitations that must be addressed for their wider adoption in clinical practice. This section discusses the primary issues associated with ViTs in medical imaging, including difficulties in training, computational complexity, deployment in resource-constrained settings, generalization capabilities, and model interpretability [107–109].

Challenges in Training and Data Requirements

One of the most pressing challenges with ViTs is their reliance on large amounts of labeled training data. Unlike

CNNs, which are often effective with moderate-sized datasets, ViTs require extensive training datasets to learn meaningful representations. This requirement is particularly problematic in medical imaging, where obtaining well-annotated datasets is difficult due to privacy concerns, high annotation costs, and the need for expert radiologists or pathologists for accurate labeling.

A major consequence of limited labeled data is overfitting, where ViTs perform well on training data but fail to generalize to unseen medical images. Additionally, variations in imaging modalities, acquisition protocols, and disease presentations introduce dataset heterogeneity, further complicating model training. Addressing these issues requires advanced techniques such as self-supervised learning, data augmentation, transfer learning, and domain adaptation. Developing large-scale, publicly available medical imaging datasets tailored for ViT training could also be a crucial step toward overcoming data limitations.

Computational Complexity and Resource Constraints

ViTs are computationally intensive models that demand significant processing power and memory resources, limiting their accessibility, particularly in low-resource clinical settings. Unlike CNNs, which leverage spatial locality and weight sharing for computational efficiency, ViTs process entire images as sequences of patches, requiring complex matrix operations and self-attention computations that scale quadratically with the input size.

The high computational cost makes it challenging to train ViTs on standard hardware, necessitating specialized infrastructure such as high-end GPUs or TPUs. Additionally, deploying ViTs in real-time or mobile healthcare applications remains a challenge due to their high memory footprint and latency issues. To mitigate these limitations, researchers are exploring techniques such as model pruning, quantization, knowledge distillation, and efficient transformer architectures like Swin Transformers, which introduce hierarchical processing and localized self-attention to improve efficiency.

Generalization Across Medical Imaging Domains

Pre-training ViTs on large-scale datasets such as ImageNet has been a common approach to leverage transfer learning. However, medical images are vastly different from natural images, leading to domain shift issues when applying pre-trained ViTs to clinical imaging tasks. This domain gap can result in suboptimal performance and reduced generalization capabilities, especially when applied to new imaging modalities or unseen disease conditions.

To address this, researchers are investigating domain-specific pre-training strategies, including the use of large-scale medical imaging datasets for self-supervised learning. Additionally, fine-tuning strategies that incorporate medical-specific priors, multi-modal learning approaches, and contrastive learning techniques could help ViTs adapt better to the nuances of medical imaging data.

Challenges in Model Interpretability and Clinical Trust

One of the critical barriers to the clinical adoption of ViTs is their lack of interpretability and explainability. In high-stakes medical applications, clinicians need to understand and trust AI-generated predictions before integrating them into decision-making workflows. However, ViTs operate as complex black-box models, making it difficult to determine the rationale behind their decisions.

Enhancing interpretability requires the development of attention visualization techniques, feature attribution methods, and explainable AI (XAI) frameworks that can highlight key image regions contributing to model predictions. Integrating human feedback mechanisms, interactive AI systems, and multimodal interpretability approaches (e.g., combining textual reports with image classification) could further bridge the gap between AI models and clinical decision-making.

Future Research Directions

To fully realize the potential of ViTs in medical imaging, future research should focus on several key areas:

- Data-Efficient Training: Investigate semi-supervised, weakly supervised, and self-supervised learning methods to reduce dependency on large labeled datasets.
- Efficient Architectures: Develop lightweight and optimized ViT architectures for deployment in real-world medical applications, particularly in low-resource settings.
- Domain-Specific Adaptations: Explore domain adaptation techniques and create large-scale medical-specific pre-training datasets to bridge the gap between natural and medical image domains.
- Interpretable AI: Integrate explainability techniques to ensure transparency, accountability, and clinical trust in AI-assisted diagnostics.
- Multi-Modal Learning: Combine ViTs with other data sources, such as electronic health records, genomic data, and patient history, to improve diagnostic accuracy and robustness.

Despite their potential, vision transformers face several challenges in medical imaging, ranging from computational demands to data limitations and interpretability issues. Addressing these challenges requires interdisciplinary efforts in machine learning, medical imaging, and clinical practice. By advancing data-efficient training techniques, optimizing computational efficiency, enhancing model interpretability, and improving generalization capabilities, ViTs could become transformative tools in AI-powered healthcare diagnostics and decision support systems. Continued research in this domain is essential to unlocking the full potential of ViTs in improving patient outcomes and advancing medical image analysis.

Conclusion

This comprehensive review has synthesized findings from research papers that explored the application of vision transformer models in medical image classification across a broad spectrum of medical fields, through our systematic analysis, we have demonstrated that ViT offers promising advancements in the domain of medical image analysis, frequently outperforming traditional convolutional neural networks. The surveyed studies underscore the flexibility and power of ViT in handling diverse medical imaging tasks, from breast cancer classification and skin lesion analysis to COVID-19 classification and bone analysis. Notably, the success of ViT is facilitated by careful data preprocessing, strategic model architecture modifications, and effective training methodologies, including transfer learning techniques.

However, the interpretability of ViT remains a crucial challenge, echoing the broader issue of model explainability in the field of artificial intelligence, further research is needed to enhance the transparency of these models, ensuring their decisions are understandable to clinicians and foster trust in their implementation.

In summary, this review affirms the potential of ViT in revolutionizing medical image classification, the breadth and depth of research in this rapidly evolving field anticipates a future where the integration of AI and healthcare is not just desirable, but essential. While challenges persist, the prospects for future innovation and refinement in this space are substantial, promising significant improvements in diagnostic accuracy and patient care.

Author Contribution All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Sanad Aburass, Osama Dorgham, Jamil Al-Shaqsi, Maha Abu Rumman, and Omar Al-Kadi. The first draft of the manuscript was written by Sanad Aburass and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors declare no competing interests.

References

1. K. Han *et al*., "A Survey on Vision Transformer," Dec. 2020, https://doi.org/10.1109/TPAMI.2022.3152247.
2. A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif Intell Rev*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, https://doi.org/10.1007/s10462-020-09825-6.
3. Aburass, "Cubixel: a novel paradigm in image processing using three-dimensional pixel representation," *Multimed Tools Appl*, Sep. 2024, https://doi.org/10.1007/s11042-024-20081-6.
4. X. Yao, X. Wang, S. H. Wang, and Y. D. Zhang, "A comprehensive survey on convolutional neural network in medical image analysis," *Multimed Tools Appl*, vol. 81, no. 29, pp. 41361–41405, Dec. 2022, https://doi.org/10.1007/s11042-020-09634-7.
5. S.S. Aburass, A. Huneiti, and M. B. Al-Zoubi, "Classification of Transformed and Geometrically Distorted Images using Convolutional Neural Network," *Journal of Computer Science*, vol. 18, no. 8, pp. 757–769, 2022, https://doi.org/10.3844/jcssp.2022.757.769.
6. Zewen Li, Wenjie Yang, Shouheng Peng, and Fan Liu, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, https://doi.org/10.1109/TNNLS.2021.3084827.
7. J. Al Shaqsi, O. Drogham, and S. Aburass, "Advanced machine learning based exploration for predicting pandemic fatality: Oman dataset," *Inform Med Unlocked*, vol. 43, p. 101393, 2023, https://doi.org/10.1016/j.imu.2023.101393.
8. S. AbuRass, A. Huneiti, and M. B. Al-Zoubi, "Enhancing Convolutional Neural Network using Hu's Moments," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 130–137, Dec. 2020, https://doi.org/10.14569/IJACSA.2020.0111216.
9. R. Azad *et al*., "Advances in medical image analysis with vision Transformers: A comprehensive review," *Med Image Anal*, vol. 91, p. 103000, Jan. 2024, https://doi.org/10.1016/j.media.2023.103000.
10. S. S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput Surv*, vol. 54, no. 10s, pp. 1–41, Jan. 2021, https://doi.org/10.1145/3505244.
11. O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," *Comput Biol Med*, vol. 157, p. 106791, May 2023, https://doi.org/10.1016/j.compbiomed.2023.106791.
12. P. Komorowski, H. Baniecki, and P. Biecek, "Towards Evaluating Explanations of Vision Transformers for Medical Imaging," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2023, pp. 3726–3732. https://doi.org/10.1109/CVPRW59228.2023.00383.

13. S. Aburass, "Quantifying Overfitting: Introducing the Overfitting Index," 2023. Accessed: Nov. 10, 2023. [Online]. Available: https://arxiv.org/abs/2308.08682
14. S. Aburass, O. Dorgham, and J. Al Shaqsi, "A Hybrid Machine Learning Model for Classifying Gene Mutations in Cancer using LSTM, BiLSTM, CNN, GRU, and GloVe," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.14361
15. A. Dosovitskiy *et al*., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929
16. K. Islam, "Recent Advances in Vision Transformer: A Survey and Outlook of Recent Work," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.01536
17. F. Almalik, M. Yaqub, and K. Nandakumar, "Self-Ensembling Vision Transformer (SEViT) for Robust Medical Image Classification," 2022, pp. 376–386. https://doi.org/10.1007/978-3-031-16437-8_36.
18. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, Sep. 2021, pp. 10347–10357. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html
19. Z. Liu *et al*., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.14030
20. S. S. D'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, Sep. 2021, pp. 2286–2296. [Online]. Available: https://proceedings.mlr.press/v139/d-ascoli21a.html
21. B. Graham et al., "Levit: a vision transformer in convnet's clothing for faster inference," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12259–12269.
22. B. Gheflati and H. Rivaz, "Vision Transformer for Classification of Breast Ultrasound Images," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.14731
23. S. Tummala, J. Kim, and S. Kadry, "BreaST-Net: Multi-Class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers," *Mathematics*, vol. 10, no. 21, p. 4109, Nov. 2022, https://doi.org/10.3390/math10214109.
24. M. L. Abimouloud, K. Bensid, M. Elleuch, M. Ben Ammar, and M. Kherallah, "Advancing breast cancer diagnosis: token vision transformers for faster and accurate classification of histopathology images," *Vis Comput Ind Biomed Art*, vol. 8, no. 1, p. 1, Jan. 2025, https://doi.org/10.1186/s42492-024-00181-8.
25. W. Wang, R. Jiang, N. Cui, Q. Li, F. Yuan, and Z. Xiao, "Semi-supervised vision transformer with adaptive token sampling for breast cancer classification," *Front Pharmacol*, vol. 13, Jul. 2022, https://doi.org/10.3389/fphar.2022.929755.
26. X. Chen *et al*., "Transformers Improve Breast Cancer Diagnosis from Unregistered Multi-View Mammograms," *Diagnostics*, vol. 12, no. 7, Jul. 2022, https://doi.org/10.3390/diagnostics12071549.
27. Z. He *et al*., "Deconv-transformer (DecT): A histopathological image classification model for breast cancer based on color deconvolution and transformer architecture," *Inf Sci (N Y)*, vol. 608, pp. 1093–1112, 2022, https://doi.org/10.1016/j.ins.2022.06.091.
28. S. Mehta *et al*., "End-to-End diagnosis of breast biopsy images with transformers," *Med Image Anal*, vol. 79, Jul. 2022, https://doi.org/10.1016/j.media.2022.102466.
29. E. Kaczmarek, A. Jamzad, T. Imtiaz, J. Nanayakkara, N. Renwick, and P. Mousavi, "Multi-Omic Graph Transformers for Cancer Classification and Interpretation," 2021. [Online]. Available: https://www.worldscientific.com
30. C. He *et al*., "A Vision Transformer Network With Wavelet-Based Features for Breast Ultrasound Classification," *Image Analysis and Stereology*, vol. 43, no. 2, pp. 185–194, Jun. 2024, https://doi.org/10.5566/ias.3116.
31. G. L. Baroni, L. Rasotto, K. Roitero, A. Tulisso, C. Di Loreto, and V. Della Mea, "Optimizing Vision Transformers for Histopathology: Pretraining and Normalization in Breast Cancer Classification," *J Imaging*, vol. 10, no. 5, p. 108, Apr. 2024, https://doi.org/10.3390/jimaging10050108.
32. C. F. and R. J. and E. E. Sarker Md Mostafa Kamal and Moreno-García, "TransSLC: Skin Lesion Classification in Dermatoscopic Images Using Transformers," in *Medical Image Understanding and Analysis*, A. and R. M. and S. C.-B. Yang Guang and Aviles-Rivero, Ed., Cham: Springer International Publishing, 2022, pp. 651–660.
33. Y. Nie, P. Sommella, M. Carratù, M. O'Nils, and J. Lundgren, "A Deep CNN Transformer Hybrid Model for Skin Lesion Classification of Dermoscopic Images Using Focal Loss," *Diagnostics*, vol. 13, no. 1, Jan. 2023, https://doi.org/10.3390/diagnostics13010072.
34. X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Med Image Anal*, vol. 77, p. 102357, 2022, https://doi.org/10.1016/j.media.2022.102357.
35. L. M. de Lima and R. A. Krohling, "Exploring Advances in Transformers and CNN for Skin Lesion Diagnosis on Small Datasets," 2022, pp. 282–296. https://doi.org/10.1007/978-3-031-21689-3_21.
36. W. Wu *et al*., "Scale-Aware Transformers for Diagnosing Melanocytic Lesions," *IEEE Access*, vol. 9, pp. 163526–163541, 2021, https://doi.org/10.1109/ACCESS.2021.3132958.
37. Z. Zhao, "Skin Cancer Classification Based on Convolutional Neural Networks and Vision Transformers," in *Journal of Physics: Conference Series*, Institute of Physics, 2022. https://doi.org/10.1088/1742-6596/2405/1/012037.
38. L. Zhou and Y. Luo, "Deep Features Fusion with Mutual Attention Transformer for Skin Lesion Diagnosis," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 3797–3801. https://doi.org/10.1109/ICIP42928.2021.9506211.
39. G. M. S. Himel, Md. M. Islam, Kh. A. Al-Aff, S. I. Karim, and Md. K. U. Sikder, "Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-Based Noninvasive Digital System," *Int J Biomed Imaging*, vol. 2024, pp. 1–18, Feb. 2024, https://doi.org/10.1155/2024/3022192.
40. C. Xin *et al*., "An improved transformer network for skin cancer classification," *Comput Biol Med*, vol. 149, p. 105939, Oct. 2022, https://doi.org/10.1016/j.compbiomed.2022.105939.
41. S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling," *Current Oncology*, vol. 29, no. 10, pp. 7498–7511, Oct. 2022, https://doi.org/10.3390/curroncol29100590.
42. Q. Lyu *et al*., "A transformer-based deep-learning approach for classifying brain metastases into primary organ sites using clinical whole-brain MRI images," *Patterns*, vol. 3, no. 11, p. 100613, Nov. 2022, https://doi.org/10.1016/j.patter.2022.100613.
43. C. K. K. Reddy *et al*., "A fine-tuned vision transformer based enhanced multi-class brain tumor classification using MRI scan imagery," *Front Oncol*, vol. 14, Jul. 2024, https://doi.org/10.3389/fonc.2024.1400341.

44. M. Zhang *et al*., "Augmented Transformer network for MRI brain tumor segmentation," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 1, p. 101917, Jan. 2024, https://doi.org/10.1016/j.jksuci.2024.101917.
45. R. Sun, Y. Pang, and W. Li, "Efficient Lung Cancer Image Classification and Segmentation Algorithm Based on an Improved Swin Transformer," *Electronics (Basel)*, vol. 12, no. 4, p. 1024, Feb. 2023, https://doi.org/10.3390/electronics12041024.
46. S. Khademi *et al*., "Spatio-Temporal Hybrid Fusion of CAE and SWIn Transformers for Lung Cancer Malignancy Prediction," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.15297
47. Y. Chen, J. Feng, J. Liu, B. Pang, D. Cao, and C. Li, "Detection and Classification of Lung Cancer Cells Using Swin Transformer," *J Cancer Ther*, vol. 13, no. 07, pp. 464–475, 2022, https://doi.org/10.4236/jct.2022.137041.
48. W. Sun, Y. Pang, and G. Zhang, "CCT: Lightweight compact convolutional transformer for lung disease CT image classification," *Front Physiol*, vol. 13, Nov. 2022, https://doi.org/10.3389/fphys.2022.1066999.
49. Y. Wu, S. Qi, Y. Sun, S. Xia, Y. Yao, and W. Qian, "A vision transformer for emphysema classification using CT images," *Phys Med Biol*, vol. 66, no. 24, p. 245016, Dec. 2021, https://doi.org/10.1088/1361-6560/ac3dc8.
50. S. S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid deep learning for detecting lung diseases from X-ray images," *Inform Med Unlocked*, vol. 20, Jan. 2020, https://doi.org/10.1016/j.imu.2020.100391.
51. F. M. Javed Mehedi Shamrat *et al*., "LungNet22: A Fine-Tuned Model for Multiclass Classification and Prediction of Lung Disease Using X-ray Images," *J Pers Med*, vol. 12, no. 5, May 2022, https://doi.org/10.3390/jpm12050680.
52. Y. Ma and W. Lv, "Identification of Pneumonia in Chest X-Ray Image Based on Transformer," *Int J Antennas Propag*, vol. 2022, 2022, https://doi.org/10.1155/2022/5072666.
53. X. Jiang, Y. Zhu, G. Cai, B. Zheng, and D. Yang, "MXT: A New Variant of Pyramid Vision Transformer for Multi-label Chest X-ray Image Classification," *Cognit Comput*, vol. 14, no. 4, pp. 1362–1377, 2022, https://doi.org/10.1007/s12559-022-10032-4.
54. M. Imran, B. Haq, E. Elbasi, A. E. Topcu, and W. Shao, "Transformer-Based Hierarchical Model for Non-Small Cell Lung Cancer Detection and Classification," *IEEE Access*, vol. 12, pp. 145920–145933, 2024, https://doi.org/10.1109/ACCESS.2024.3449230.
55. C. Playout, R. Duval, M. C. Boucher, and F. Cheriet, "Focused Attention in Transformers for interpretable classification of retinal images," *Med Image Anal*, vol. 82, p. 102608, 2022, https://doi.org/10.1016/j.media.2022.102608.
56. M. A. Rodríguez, H. AlMarzouqi, and P. Liatsis, "Multi-Label Retinal Disease Classification Using Transformers," *IEEE J Biomed Health Inform*, vol. 27, no. 6, pp. 2739–2750, Jun. 2023, https://doi.org/10.1109/JBHI.2022.3214086.
57. J. Shen, Y. Hu, X. Zhang, Y. Gong, R. Kawasaki, and J. Liu, "Structure-Oriented Transformer for retinal diseases grading from OCT images," *Comput Biol Med*, vol. 152, p. 106445, 2023, https://doi.org/10.1016/j.compbiomed.2022.106445.
58. Z. Ma, Q. Xie, P. Xie, F. Fan, X. Gao, and J. Zhu, "HCTNet: A Hybrid ConvNet-Transformer Network for Retinal Optical Coherence Tomography Image Classification," *Biosensors (Basel)*, vol. 12, no. 7, Jul. 2022, https://doi.org/10.3390/bios12070542.
59. Z. Jiang *et al*., "Computer-aided diagnosis of retinopathy based on vision transformer," *J Innov Opt Health Sci*, vol. 15, no. 2, Mar. 2022, https://doi.org/10.1142/S1793545822500092.
60. K. Wang, C. Xu, G. Li, Y. Zhang, Y. Zheng, and C. Sun, "Combining convolutional neural networks and self-attention for fundus diseases identification," *Sci Rep*, vol. 13, no. 1, Dec. 2023, https://doi.org/10.1038/s41598-022-27358-6.
61. M. Wassel, A. M. Hamdi, N. Adly, and M. Torki, "Vision Transformers Based Classification for Glaucomatous Eye Condition," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, Aug. 2022, pp. 5082–5088. https://doi.org/10.1109/ICPR56361.2022.9956086.
62. D. Wang, J. Lian, and W. Jiao, "Multi-label classification of retinal disease via a novel vision transformer model," *Front Neurosci*, vol. 17, Jan. 2024, https://doi.org/10.3389/fnins.2023.1290803.
63. J. He, J. Wang, Z. Han, J. Ma, C. Wang, and M. Qi, "An interpretable transformer network for the retinal disease classification using optical coherence tomography," *Sci Rep*, vol. 13, no. 1, p. 3637, Mar. 2023, https://doi.org/10.1038/s41598-023-30853-z.
64. X. Fan, X. Feng, Y. Dong, and H. Hou, "COVID-19 CT image recognition algorithm based on transformer and CNN," *Displays*, vol. 72, p. 102150, Apr. 2022, https://doi.org/10.1016/j.displa.2022.102150.
65. A. Ascencio-Cabral and C. C. Reyes-Aldasoro, "Comparison of Convolutional Neural Networks and Transformers for the Classification of Images of COVID-19, Pneumonia and Healthy Individuals as Observed with Computed Tomography," *J Imaging*, vol. 8, no. 9, Sep. 2022, https://doi.org/10.3390/jimaging8090237.
66. F. Mehboob *et al*., "Towards robust diagnosis of COVID-19 using vision self-attention transformer," *Sci Rep*, vol. 12, no. 1, Dec. 2022, https://doi.org/10.1038/s41598-022-13039-x.
67. L. Peng *et al*., "Analysis of CT scan images for COVID-19 pneumonia based on a deep ensemble framework with DenseNet, Swin transformer, and RegNet," *Front Microbiol*, vol. 13, Sep. 2022, https://doi.org/10.3389/fmicb.2022.995323.
68. K. Cao, T. Deng, C. Zhang, L. Lu, and L. Li, "A CNN-transformer fusion network for COVID-19 CXR image classification," *PLoS One*, vol. 17, no. 10 October, Oct. 2022, https://doi.org/10.1371/journal.pone.0276758.
69. L. Zhang and Y. Wen, "A transformer-based framework for automatic COVID19 diagnosis in chest CTs."
70. D. Shome *et al*., "Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare," *Int J Environ Res Public Health*, vol. 18, no. 21, Nov. 2021, https://doi.org/10.3390/ijerph182111086.
71. Hongyan Xu, Xiu Su, and Dadong Wang, "CNN-based Local Vision Transformer for COVID-19 Diagnosis," *Proc ACM Hum Comput Interact*, vol. 5, no. CSCW1, Apr. 2021, https://doi.org/10.1145/1122445.1122456.
72. C. Liu and Q. Yin, "Automatic diagnosis of COVID-19 using a tailored transformer-like network," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Sep. 2021. https://doi.org/10.1088/1742-6596/2010/1/012175.
73. V. Randazzo *et al*., "A Vision Transformer Model for the Prediction of Fatal Arrhythmic Events in Patients with Brugada Syndrome," *Sensors*, vol. 25, no. 3, p. 824, Jan. 2025, https://doi.org/10.3390/s25030824.
74. C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin, "Constrained transformer network for ECG signal processing and arrhythmia classification," *BMC Med Inform Decis Mak*, vol. 21, no. 1, Dec. 2021, https://doi.org/10.1186/s12911-021-01546-2.
75. H. Reynaud, A. Vlontzos, B. Hou, A. Beqiri, P. Leeson, and B. Kainz, "Ultrasound Video Transformers for Cardiac Ejection

Fraction Estimation," 2021, pp. 495–505. https://doi.org/10.1007/978-3-030-87231-1_48.
76. T. H. Rafi and Y. Woong Ko, "HeartNet: Self Multihead Attention Mechanism via Convolutional Network With Adversarial Data Synthesis for ECG-Based Arrhythmia Classification," *IEEE Access*, vol. 10, pp. 100501–100512, 2022, https://doi.org/10.1109/ACCESS.2022.3206431.
77. T. Jumphoo, K. Phapatanaburi, W. Pathonsuwan, P. Anchuen, M. Uthansakul, and P. Uthansakul, "Exploiting Data-Efficient Image Transformer-Based Transfer Learning for Valvular Heart Diseases Detection," *IEEE Access*, vol. 12, pp. 15845–15855, 2024, https://doi.org/10.1109/ACCESS.2024.3357946.
78. M. A.-E. Zeid, K. El-Bahnasy, and S. E. Abo-Youssef, "Multi-class Colorectal Cancer Histology Images Classification Using Vision Transformers," in *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 2021, pp. 224–230. https://doi.org/10.1109/ICICIS52592.2021.9694125.
79. H. Cai *et al*., "MIST multiple instance learning network based on Swin Transformer for whole slide image classification of colorectal adenomas," *J Pathol*, vol. 259, no. 2, pp. 125–135, Feb. 2023, https://doi.org/10.1002/path.6027.
80. D. Sui, K. Zhang, W. Liu, J. Chen, X. Ma, and Z. Tian, "CST: A Multitask Learning Framework for Colorectal Cancer Region Mining Based on Transformer," *Biomed Res Int*, vol. 2021, 2021, https://doi.org/10.1155/2021/6207964.
81. H. Chen *et al*., "IL-MCAM: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach," *Comput Biol Med*, vol. 143, p. 105265, Apr. 2022, https://doi.org/10.1016/j.compbiomed.2022.105265.
82. Z. Lv, Y. Lin, R. Yan, Y. Wang, and F. Zhang, "TransSurv: Transformer-based Survival Analysis Model Integrating Histopathological Images and Genomic Data for Colorectal Cancer," *IEEE/ACM Trans Comput Biol Bioinform*, pp. 1–10, 2022, https://doi.org/10.1109/TCBB.2022.3199244.
83. M. Khalid, S. Deivasigamani, S. V, and S. Rajendran, "An efficient colorectal cancer detection network using atrous convolution with coordinate attention transformer and histopathological images," *Sci Rep*, vol. 14, no. 1, p. 19109, Aug. 2024, https://doi.org/10.1038/s41598-024-70117-y.
84. Y. Lyu, X. Yu, D. Zhu, and L. Zhang, "Classification of Alzheimer's Disease via Vision Transformer: Classification of Alzheimer's Disease via Vision Transformer," in *ACM International Conference Proceeding Series,* Association for Computing Machinery, Jun. 2022, pp. 463–468. https://doi.org/10.1145/3529190.3534754.
85. C. Li et al., "Trans-ResNet: Integrating Transformers and CNNs for Alzheimer's disease classification," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–5. https://doi.org/10.1109/ISBI52829.2022.9761549.
86. S. Sarraf, A. Sarraf, D. D. DeSouza, J. A. E. Anderson, and M. Kabia, "OViTAD: Optimized Vision Transformer to Predict Various Stages of Alzheimer's Disease Using Resting-State fMRI and Structural MRI Data," *Brain Sci*, vol. 13, no. 2, p. 260, Feb. 2023, https://doi.org/10.3390/brainsci13020260.
87. Z. Hu, Z. Wang, Y. Jin, and W. Hou, "VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer's disease prediction," *Comput Methods Programs Biomed*, vol. 229, p. 107291, 2023, https://doi.org/10.1016/j.cmpb.2022.107291.
88. H. Liu *et al*., "ViT-Based Face Diagnosis Images Analysis for Schizophrenia Detection," *Brain Sci*, vol. 15, no. 1, p. 30, Dec. 2024, https://doi.org/10.3390/brainsci15010030.
89. H. A. Bedel, I. Sivgin, O. Dalmaz, S. U. H. Dar, and T. Çukur, "BolT: Fused window transformers for fMRI time series analysis," *Med Image Anal*, vol. 88, p. 102841, Aug. 2023, https://doi.org/10.1016/j.media.2023.102841.
90. A. G. Alharthi and S. M. Alzahrani, "Multi-Slice Generation sMRI and fMRI for Autism Spectrum Disorder Diagnosis Using 3D-CNN and Vision Transformers," *Brain Sci*, vol. 13, no. 11, p. 1578, Nov. 2023, https://doi.org/10.3390/brainsci13111578.
91. Z. Gu, Y. Li, Z. Wang, J. Kan, J. Shu, and Q. Wang, "Classification of Diabetic Retinopathy Severity in Fundus Images Using the Vision Transformer and Residual Attention," *Comput Intell Neurosci*, vol. 2023, pp. 1–12, Jan. 2023, https://doi.org/10.1155/2023/1305583.
92. C. Adak, T. Karkera, S. Chattopadhyay, and M. Saqib, "Detecting Severity of Diabetic Retinopathy from Fundus Images using Ensembled Transformers," Jan. 2023, [Online]. Available: http://arxiv.org/abs/2301.00973
93. J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, "Vision Transformer-based recognition of diabetic retinopathy grade," *Med Phys*, vol. 48, no. 12, pp. 7850–7863, Dec. 2021, https://doi.org/10.1002/mp.15312.
94. M. Oulhadj *et al*., "Diabetic retinopathy prediction based on vision transformer and modified capsule network," *Comput Biol Med*, vol. 175, p. 108523, Jun. 2024, https://doi.org/10.1016/j.compbiomed.2024.108523.
95. G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," *Visual Computer*, 2022, https://doi.org/10.1007/s00371-022-02492-4.
96. C. Xin *et al*., "An improved transformer network for skin cancer classification," *Comput Biol Med*, vol. 149, Oct. 2022, https://doi.org/10.1016/j.compbiomed.2022.105939.
97. Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin Cancer Classification With Deep Learning: A Systematic Review," Jul. 13, 2022, *Frontiers Media S.A*. https://doi.org/10.3389/fonc.2022.893972.
98. K. Huang *et al*., "Intelligent strategy for severity scoring of skin diseases based on clinical decision-making thinking with lesion-aware transformer," *Artif Intell Rev*, vol. 58, no. 4, p. 95, Jan. 2025, https://doi.org/10.1007/s10462-024-11083-9.
99. M. N. Islam, M. Hasan, M. K. Hossain, M. G. R. Alam, M. Z. Uddin, and A. Soylu, "Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography," *Sci Rep*, vol. 12, no. 1, Dec. 2022, https://doi.org/10.1038/s41598-022-15634-4.
100. J. Zhang *et al.*, "Vision transformer introduces a new vitality to the classification of renal pathology," *BMC Nephrol*, vol. 25, no. 1, p. 337, Oct. 2024, https://doi.org/10.1186/s12882-024-03800-x.
101. Y. Liu, J. Zhao, Q. Luo, C. Shen, R. Wang, and X. Ding, "Automated classification of cervical lymph-node-level from ultrasound using Depthwise Separable Convolutional Swin Transformer," *Comput Biol Med*, vol. 148, p. 105821, 2022, https://doi.org/10.1016/j.compbiomed.2022.105821.
102. P. Jiang, Y. Xu, C. Wang, W. Zhang, and N. Lu, "CSMViT: A Lightweight Transformer and CNN fusion Network for Lymph Node Pathological Images Diagnosis," *IEEE Access*, vol. 12, pp. 155365–155378, 2024, https://doi.org/10.1109/ACCESS.2024.3483769.
103. L. Tanzi, A. Audisio, G. Cirrincione, A. Aprato, and E. Vezzetti, "Vision Transformer for femur fracture classification," *Injury*, vol. 53, no. 7, pp. 2625–2634, Jul. 2022, https://doi.org/10.1016/j.injury.2022.04.013.
104. J. Wu, Q. Mi, Y. Zhang, and T. Wu, "SVTNet : Automatic bone age assessment network based on TW3 method and vision

transformer," *Int J Imaging Syst Technol*, vol. 34, no. 2, Mar. 2024, https://doi.org/10.1002/ima.22990.
105. B. Kabas, F. Arslan, V. A. Nezhad, S. Ozturk, E. U. Saritas, and T. Çukur, "Physics-Driven Autoregressive State Space Models for Medical Image Reconstruction," Dec. 2024.
106. O. F. Atli *et al*., "I2I-Mamba: Multi-modal medical image synthesis via selective state space modeling," May 2024.
107. K. He *et al*., "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, Feb. 2023, https://doi.org/10.1016/j.imed.2022.07.002.
108. F. Shamshad *et al*., "Transformers in medical imaging: A survey," *Med Image Anal*, vol. 88, p. 102802, Aug. 2023, https://doi.org/10.1016/j.media.2023.102802.
109. A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision Transformers in medical computer vision—A contemplative retrospection," *Eng Appl Artif Intell*, vol. 122, p. 106126, Jun. 2023, https://doi.org/10.1016/j.engappai.2023.106126.