Al-hammuri *et al.*
Visual Computing for Industry, Biomedicine, and Art (2023) 6:14
https://doi.org/10.1186/s42492-023-00140-9

Visual Computing for Industry, Biomedicine, and Art

REVIEW

Open Access

Vision transformer architecture and applications in digital health: a tutorial and survey

Khalid Al-hammuri[1*], Fayez Gebali[1], Awos Kanan[2] and Ilamparithi Thirumarai Chelvan[1]

Abstract

The vision transformer (ViT) is a state-of-the-art architecture for image recognition tasks that plays an important role in digital health applications. Medical images account for 90% of the data in digital medicine applications. This article discusses the core foundations of the ViT architecture and its digital health applications. These applications include image segmentation, classification, detection, prediction, reconstruction, synthesis, and telehealth such as report generation and security. This article also presents a roadmap for implementing the ViT in digital health systems and discusses its limitations and challenges.

Keywords Vision transformer, Digital health, Telehealth, Artificial intelligence, Medical imaging

Introduction

The coronavirus disease 2019 (COVID-19) pandemic demonstrated how artificial intelligence (AI) can help scale a system during emergencies with limited medical staff or existing safety concerns. AI algorithms are widely used in digital medicine solutions, mainly in image and text recognition tasks, to analyze medical data stored in clinical information systems and generate medical reports, and to assist in other technical operations such as robotic surgery. Among the various AI-assisted tools for analyzing medical images, the vision transformer (ViT) has emerged as a state-of-the-art algorithm that replaces or combines traditional techniques such as convolutional neural networks (CNNs). This article discusses the foundations and applications of the ViT in digital health.

*Correspondence:
Khalid Al-hammuri
khalidalhammuri@uvic.ca
[1] Electrical and Computer Engineering, University of Victoria, Victoria V8W 2Y2, Canada
[2] Computer Engineering, Princess Sumaya University for Technology, Amman 11941, Jordan

The ViT [1, 2] is a type of neural network for image processing in computer vision tasks [3]. The backbone of the ViT is a self-attention mechanism typically used in natural language processing (NLP). The ViT was introduced to deal with the image processing limitations of common machine learning architectures such as CNNs [4], recurrent neural networks (RNNs) [5], and even the traditional transformers for language models [1, 6]. The ViT provides a strong representation of image features and trains data using fewer computational resources compared with CNNs [1].

CNNs are widely used in the machine learning field and are suitable for feature extraction in specific local regions. However, they are unable to capture the contextual relationship between image features in the global context. In contrast, the ViT applies an attention mechanism to understand the global relationships among features.

RNNs are used to obtain inferences about sequence-to-sequence relationships and memorizes some past data. However, they require a large memory and are unsuitable for extracting image features compared with the ViT or CNNs. Bidirectional encoder representations from transformers (BERT) was developed by

Google to process language models [7] based on attention mechanisms [8]. BERT can efficiently process sequence-to-sequence models but requires a larger memory compared with an RNN or a long short-term memory (LSTM) [9].

BERT has limitations in processing imaging data and is effective only for flattened data in a sequential shape. To deal with this issue, the ViT splits images into patches then and flattens them for analysis as linear sequences [1] in a parallel processing mechanism.

The applications of the ViT in medical imaging include segmentation, classification, reconstruction, prognosis prediction, and telehealth (e.g., report generation and security).

The remainder of this paper is organized as follows. ViT architecture section describes the foundations of the ViT architecture. Applications of the ViT in digital health section presents an overview of the important applications of the ViT in medical imaging. Roadmap for implementing ViT section presents a roadmap for the end-to-end implementation of the ViT. Limitations and challenges of ViT in digital health section discusses the limitations and challenges of using the ViT, and Conclusions section concludes the paper.

ViT architecture

This section discusses the core principles and foundations of the ViT based on the attention mechanism. The ViT architecture consists of a hierarchy of different functional blocks, which will be explained in the following subsections. Figure 1 shows the typical transformer architecture proposed by ref. [8] based on the attention mechanism.

Researchers have proposed various modifications for typical transformer designs [8] (Fig. 18 for a typical transformer architecture in Appendix) for applications other than NLP tasks. The changes focus on the design framework of encoder-decoder blocks in the transformer architecture. In vision tasks, the transformer splits the image into patches and flattens them into sequential forms to be processed like time-series data, which is more suited to the nature of transformers (Fig. 19 in Appendix). To ensure that an image can be reconstructed without any data loss, positional encoding was utilized for the embedded features in a vector shape. The embedded features were fed into the encoder for image classification and then classified by multilayer perception [1]. However, in the segmentation task, the transformer is combined with the CNN either in the encoder stage, similar to the TransUNet architecture (Fig. 20 in Appendix) [10], or in both

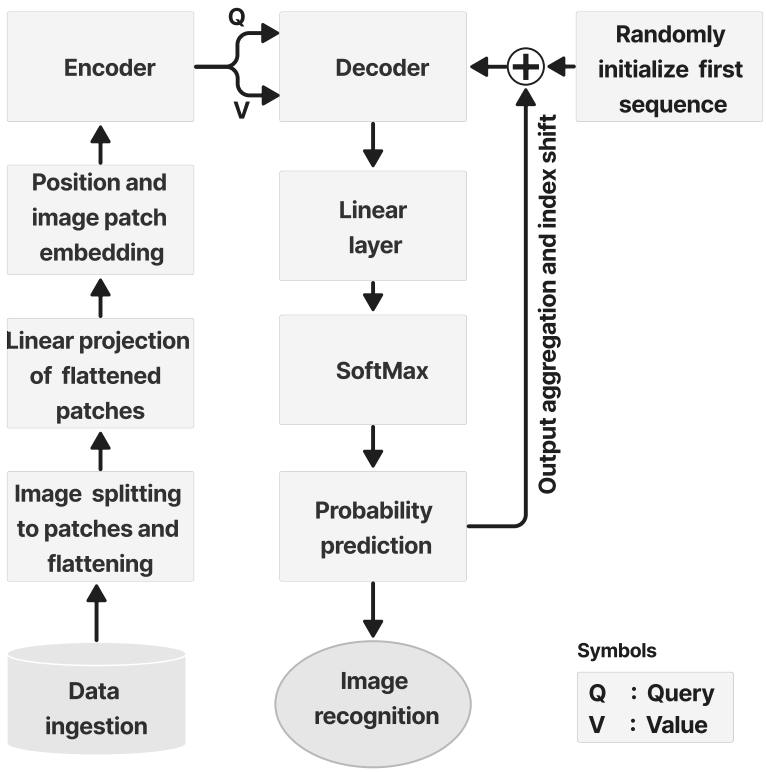


Fig. 1 Transformer architecture [1]

the encoder and decoder stages, such as the Ds-TransU-Net [11] (Fig. 21 in Appendix).

Encoder architecture

Figure 2 shows a typical encoder architecture [8] that consists of a stack of N identical layers, with each layer containing two sublayers. The first sublayer performs the multihead self-attention (MSA), while the second sublayer normalizes the output of the first sublayer and feeds it into the multilayer perceptron (MLP), which is a type of feedforward network. See Appendix A.1 for an example of the transformer architecture in ref. [1].

Image patches embedding

Owing to computer memory limitations, the simultaneous processing of an entire image is difficult. Therefore, the image is divided into different patches and processed sequentially. To conduct a detailed analysis of each image patch, each one was embedded into a set of feature values in the form of a vector.

The concept of image patch embedding in the ViT was inspired by the term 'embedding' in ref. [12]. The feature vectors were then graphically visualized in an embedding space. Visualizing the features in the embedding space is beneficial to identify the image patches with similar features [13]. The distance between each feature can be measured in the features map to determine the degree of similarity [14].

Figure 3 shows the feature embedding process, which begins by creating an embedding layer from the embedding vectors of each input feature. Random embedding values are initially assigned and updated during training inside the embedding layer. During training, similar features become closer to each other in the embedding or latent space. This is important to classify or extract similar features. However, not knowing the position of each feature makes it difficult to determine the relationship between them. In medical imaging applications, positional encoding and feature embedding enable accurate feature selection in a specific-use case.

Positional encoding

The transformer model has the advantage of simultaneously processing inputs in parallel, unlike the well-known LSTM algorithm [15, 16]. However, parallel processing is difficult because of the risk of information loss due to the inability to reconstruct the processed sequences in their original positions.

Figure 4 shows the positional encoding process for feature representation. Positional encoding was proposed to solve this problem and encode each feature vector to its accurate position [8, 17]. The feature vector and positional encoding values were added to form a new vector in the embedding space. In this study, sine and cosine functions were used as examples to derive the positional encoding values at different frequencies, expressed as Eqs.(1) and (2) [8], respectively.

$$P(x, 2i) = \sin\left(\frac{x}{10000^{\frac{2i}{d}}}\right) \tag{1}$$

$$P(x, 2i+1) = \cos\left(\frac{x}{10000^{\frac{2i}{d}}}\right) \tag{2}$$

where P is the positional encoding, d is the vector dimension, x is the position, and i is the index dimension. The sinusoidal function is beneficial for encoding the feature

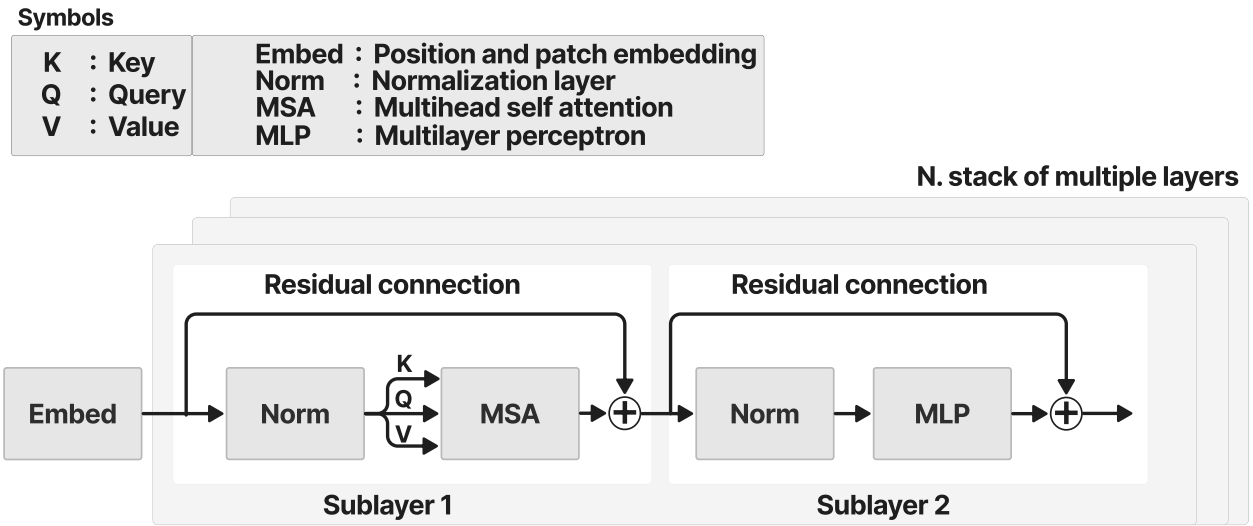


Fig. 2 Encoder block in the transformer architecture [1]

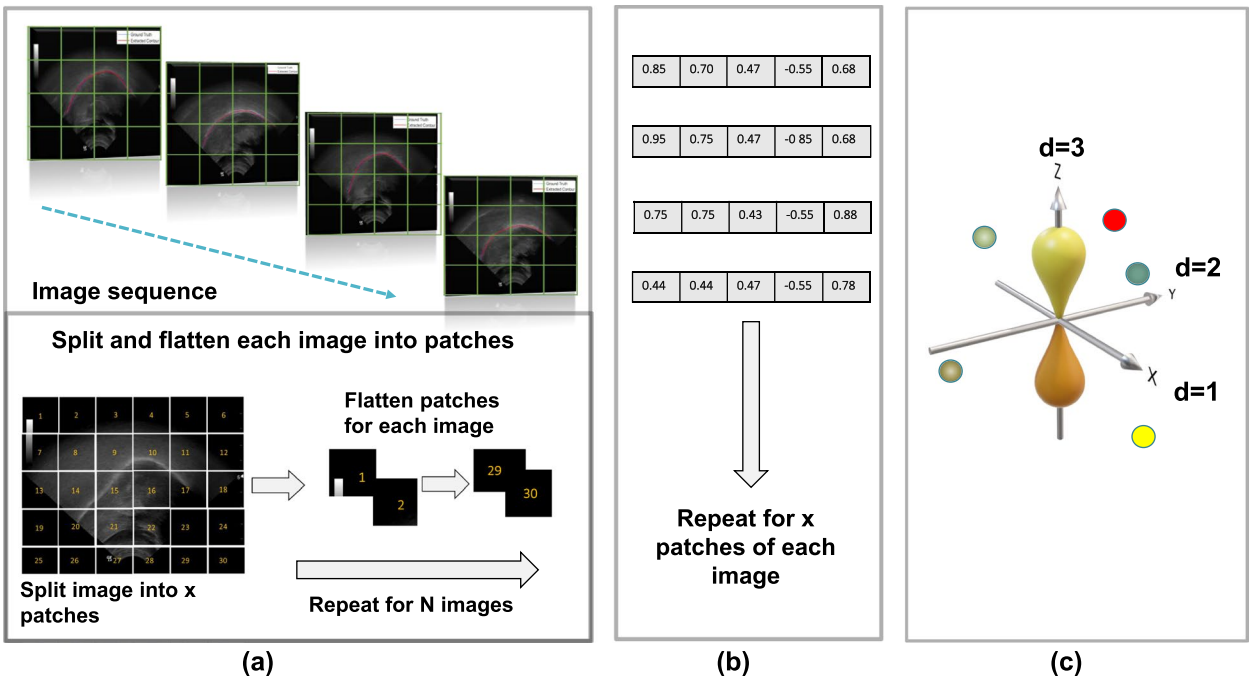


Fig. 3 **a** Illustration of splitting ultrasound images into patches and flattening them in a linear sequence; **b** Image patch vectorization and linear projection; **c** Patch embedding in multidimensional space

position in the embedding space using frequencies ranging from 2π to 10000. In Eqs. (1) and (2), the frequencies resembled the index dimension i [8].

MSA

Figure 5 shows the MSA process, which calculates the weighted average of feature representations based on the similarity scores between pairs of representations. Given the input sequence X of L tokens or entries with the dimension d, $X \in R^{L \times d}$ was projected using three matrices: $W_K \in R^{d \times dk}$, $W_Q \in R^{d \times dq}$, and $W_V \in R^{d \times dv}$ with the same dimensions to derive the representation of the features. Equation (3) presents the formulas used to derive the Key (K), Query (Q), and Value (V).

$$K = XW_K, Q = XW_Q, V = XW_V \tag{3}$$

The final embedding layer that includes the position encoding was copied into the three linear layers K, Q, and V. To derive the similarity between the input features, matrix multiplication between K and Q was performed using self-attention. The output was then scaled and normalized using SoftMax. The self-attention [3] process is explained in the following steps:

1. Calculate the score from the input of Q and K.

$$S = QK^T \tag{4}$$

2. Normalize the score to stabilize the training.

$$N_s = S^{\sqrt{d}} \tag{5}$$

3. Calculate the probabilities of the normalized score using *SoftMax*.

$$P = SoftMax(Ns) \tag{6}$$

4. Compute the self-attention filter by multiplying P and V.

$$Self - attention = PV \tag{7}$$

The multiplication of the outputs of K and Q were scaled by the square root of the input vector dimension, and then normalized by the SoftMax function to generate the probabilities. Equation (8) presents the SoftMax function, where x is the input data point. Equation (9) computes the attention filter.

$$SoftMax(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{8}$$

$$Self - attention(Q, K, V) = SoftMax\left(\frac{Q \cdot K^T}{\sqrt{d}}\right).V \tag{9}$$

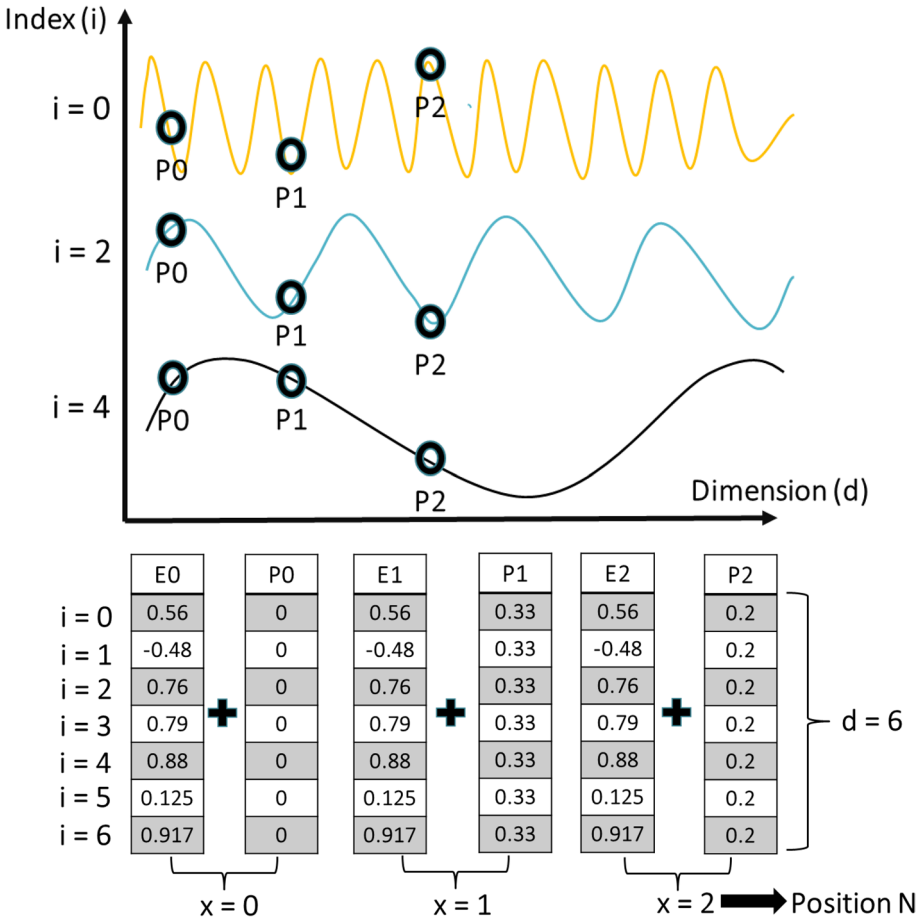


Fig. 4 Positional encoding for the feature representations. Top: Sinusoidal representation for the positional encoding (P0-P3) at different indices and dimensions. Bottom: Vector representation for the positional encoding and feature embedding; P is the position encoding and E is the embedding vector

The output probabilities from SoftMax and the value layer were multiplied to obtain the desired output with emphasis on the desired features to filter out unnecessary data. The principle behind a multihead is to concatenate the results of different attention filters, with each one focusing on the desired features. The self-attention process is repeated multiple times to form the MSA. The final output of the concatenated MSA was passed through a linear layer and resized to a single head. Equation (10) presents the *MSA* formula.

$$MSA(Q, K, V) = C(h_1, \ \dots \ , h_n)W_0 \tag{10}$$

where C is the concatenation of the multiheads; W_0 is the projection weight; Q, K and V denote the Query, Key and Value, respectively; and h resembles each head in the self-attention process and was replicated n times. The number of replications was dependent on the amount of attention or the desired features needed to extract the required information. Figure 5 shows the MSA process in the ViT architecture. Detailed information on the scaled dot products between K, Q, and V are also presented.

Layer normalization and residual connections

A residual connection is required to directly feed the output from the position encoding layer into the normalization layer by bypassing the MSA layer [18]. The residual connection is essential for knowledge preservation and to avoid vanishing gradient problems [19, 20]. The MSA layer is vital for extracting useful features from the input. However, it could also lead to the disregard of helpful information of lesser weight in the attention filter. Minimizing the value of the feature weight may cause a vanishing gradient during the model training stage. A vanishing gradient occurs when the gradient of the loss function is depleted and becomes almost or equal to zero while optimizing the weight in the backpropagation algorithm. The residual connection directly feeds information from the initial layers into the

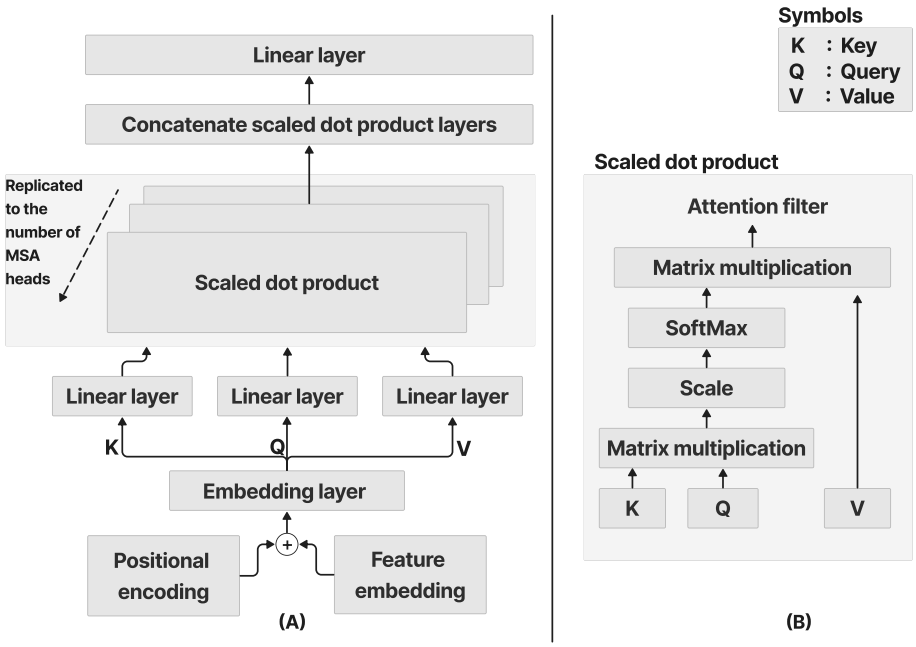


Fig. 5 MSA process. **a** MSA process with several attention layers in parallel; **b** Scaled dot product [8]. The diagram flows upwards from the bottom according to the direction of the arrow

layers at the end of the neural network to preserve features and retain important information.

The add and normalize layer [21] combine the input from the position encoding and MSA layers, and then normalize them. The normalization layer is essential during training to speed up and stabilize the loss convergence. Normalization can be achieved by standardizing the activation of neurons along the axis of the features. Equations (11) and (12) are the statistical components of layer normalization over all the hidden units in the same layer [21].

$$\mu^l = \frac{1}{H}\left(\sum_{i=1}^{H} a_i^l\right) \tag{11}$$

$$\sigma^l = \sqrt{\frac{1}{H}\sum_{i=1}^{H}\left(a_i^l - \mu^l\right)^2} \tag{12}$$

where a_i^l is the normalized value of the sum of the input features along the i^{th} hidden units in the l^{th} layers. H is the total number of hidden units in the layer. μ is the mean or average values of features along the axis in the normalization layer, σ is the standard deviation of the values of the features along the axis.

MLP

Figure 6 shows the MLP diagram, which is part of the ViT architecture. The MLP is a feedforward artificial neural network that combines a series of fully connected layers including the input, one or more hidden layers in the middle, and the output [22].

Fully connected layers are a type of layer in which the output of each neuron is connected to all the neurons in the next hidden layer. The diagram shows that each neuron from the layer in the feedforward neural network is connected to all the neurons in the next layer through an activation function. The residual connection preserves the knowledge from the initial layers and minimizes the vanishing gradient problem. Typical MLP layers include the input, output, and hidden layers.

Decoder and mask MSA

Figure 7 shows the decoder and mask MSA in the ViT architecture used to extract the final image. The decoder was stacked for N layers, the same as the number of encoder layers. The decoder includes the same sublayers as the encoder and mask MSA stacked on them. The mask MSA works similarly as the MSA, but focuses on the desired features in position i and ignores the undesirable features from the embedding layer by using the mask-only features before i. This is important to obtain an inference from the relationship between different features in the embedding space and a prediction from the features relevant to the desired position.

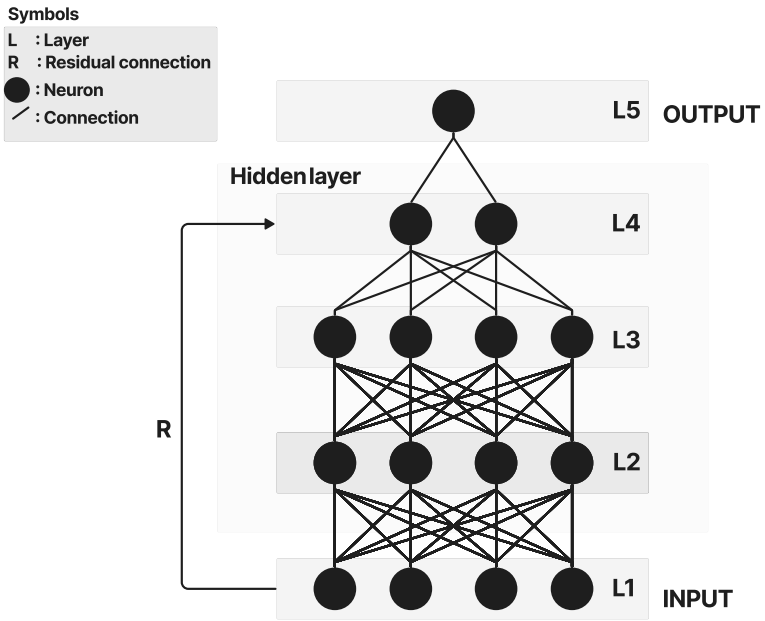


Fig. 6 MLP

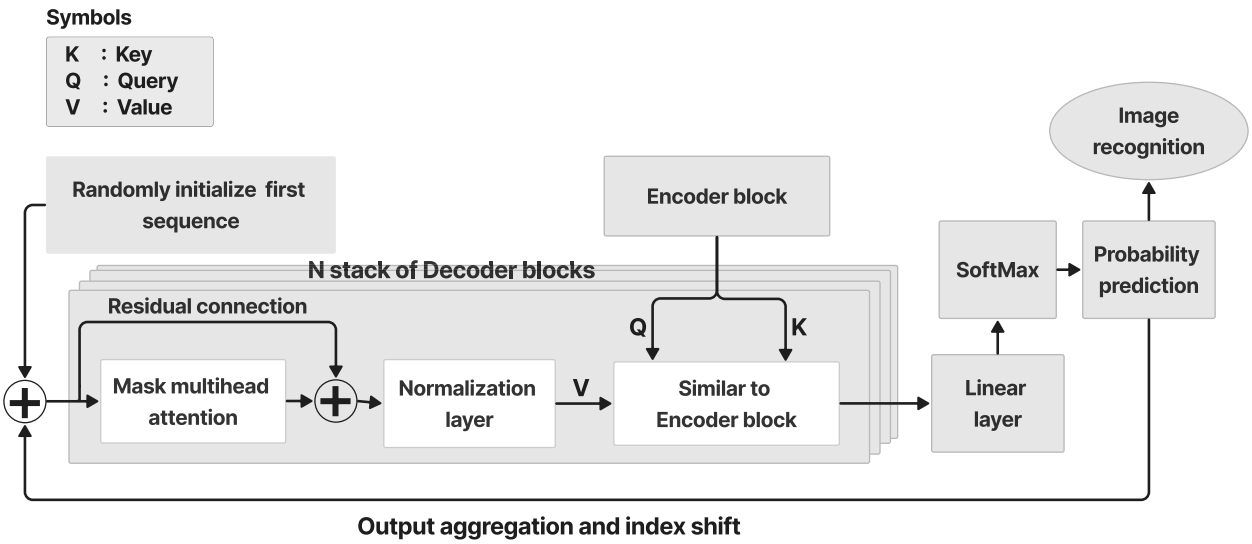


Fig. 7 Decoder and mask multihead attention block to produce the final image

The decoder obtains the V, Q, and K as inputs. The V was obtained from the previous embedding space, while Q and K were obtained from the encoder output. There are other MSA and normalization layers inside the decoder, which is common in ViT designs. Despite modifications to the decoder-encoder design, the core principle remains the same. The different architectures for different applications are explained in Applications of the ViT in digital health section.

In the image recognition task, the decoder output was flattened as a linear or dense layer. Then, SoftMax was used to derive the probability of the weight of each neuron in the dense layer. The final probability was used to classify or segment the features based on the training data to detect the final object or image.

Applications of the ViT in digital health

Computer vision and machine learning algorithms have been employed in recent medical studies on brain and breast tumors [23, 24], histopathology [25], speech recognition [26, 27], rheumatology [28], automatic captioning [29], endoscopy [30], fundus imaging [31], and telemedicine [32]. The ViT has emerged as the state-of-the-art in AI-based algorithms that use computer vision and machine learning for digital health solutions.

Figure 8 shows the distribution of ViT applications in the medical field. These include medical segmentation, detection, classification, report generation, registration, prognosis prediction, and telehealth.

Applications of ViT in medical image segmentation

TransUNet [10] is one of the earliest attempts to apply the ViT in medical imaging segmentation by combining it with the UNet [34] architecture. UNet is well known in the area of biomedical image segmentation. It is efficient in object segmentation tasks and can preserve the quality of fine image details after reconstruction. The UNet inherited the localization ability of a CNN for feature extraction. Although localization is essential in a segmentation task, it has limitations in processing sequence-to-sequence image frames or extracting global features within the same image outside a specific region. In contrast, the ViT has the advantages of processing sequence-to-sequence features and extracting the global relationships between them. However, the ViT has limitations in feature localization compared with CNNs. TransUNet proposed a robust architecture that combined the capabilities of the ViT and UNet in a single model.

TransUNet is a powerful tool for multiorgan segmentation. Segmenting different objects is essential to analyze complex structures in magnetic resonance imaging (MRI) and computed tomography (CT) images. Figure 9 shows an example of image segmentation of the abdomen in a CT scan using TransUNet, which was compared with ground truth (GT) images to validate the results.

To further improve the TransUNet architecture, a Dual-TransUNet was implemented in ref. [11]. The main difference is that the Dual-TransUNet used the transformer in the encoder to extract features and the decoder to reconstruct the desired image, while the TransUNet only used the transformer in the encoder stage. The Swin transformer [35] is another architecture for implementing the ViT in combination with Unet [34, 36] in medical imaging.

The ViT was also used in iSegFormer [37], which was proposed for the interactive segmentation of three-dimensional (3D) MRI images of the knee. The 3D UX-net [38] could segment brain tissues from the entire body in an MRI scan. UNesT [39] developed a hierarchical transformer using local spatial representation for brain, kidney, and abdominal multiorgan image segmentation. Similarly, the NestedFormer [40] was proposed to segment brain tumors in MRI images.

RECIST [41] used the ViT to automatically segment brain tumors to measure the size of the lesions in CT

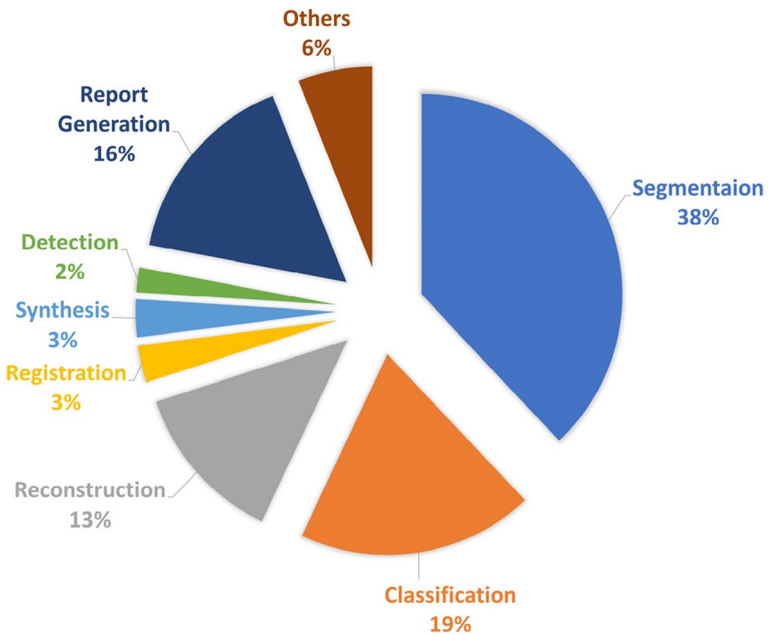


Fig. 8 Distribution of medical imaging applications of the ViT according to the survey [33]

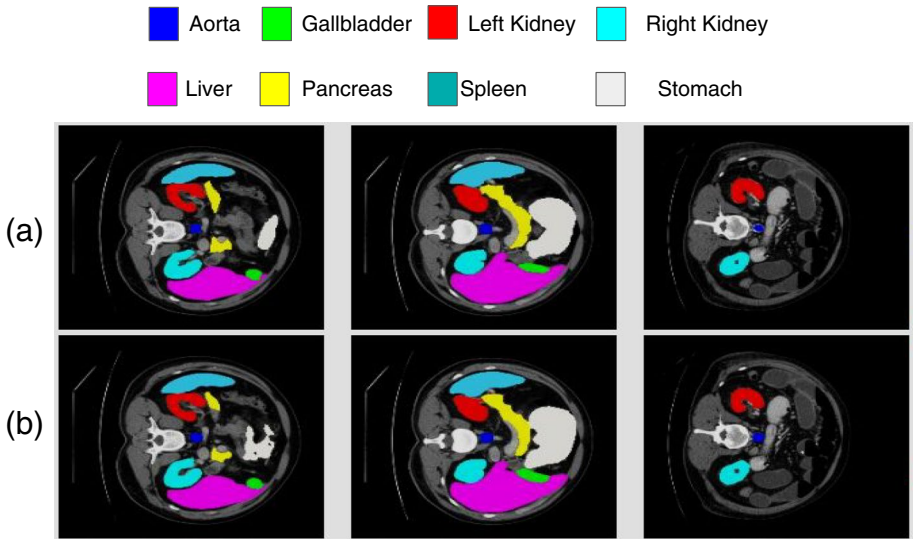


Fig. 9 Comparison of TransUNet and GT using output segmentation results of different organs: **a** GT (expert reference) and **b** TransUNet [10]

images. GT U-Net [42] was used for tooth therapy by segmenting the root canal in X-ray images. Colorectal cancer (CRC) images were segmented by the fully convolutional network (FCN) transformer [43] during a colonoscopy. The ViT was also used in the TraSeTR [44] to assist in robotic surgery by segmenting the image and generating instructions based on previous knowledge. Table 1 lists examples of ViT applications in medical image segmentation.

Applications of ViT in medical image detection

Image detection plays a key role in digital health and imaging analysis to identify objects in complex structures and share that information within the healthcare information system for further analysis. This is important to measure the cell size and count the number of suspicious objects or malignant tissues.

Object detection is essential in cancer screening when cell labeling or classification is difficult, and a careful analysis is required to identify cancers. The detection transformer (DETR) was proposed to detect lymphoproliferative diseases in MRI T2 images [48]. In MRI scans, the metastatic lymph nodes are small and difficult to identify. The application of the DETR can reduce false positives as well as improve the precision and sensitivity by 65.41% and 91.66%, respectively.

The convolutional transformer (COTR) [49] detects polyp lesions in colonoscopy images to diagnose CRC, which has the second highest cancer-related mortality risk worldwide. The COTR architecture employs a CNN

Table 1 Examples of ViT applications in medical image segmentation

Method	Category	Medical application
TransUnet [10]	MRI, CT	CT and MRI cardiac segmentation
Dual-TransUnet [11]	Microscopy	Skin lesion analysis [45]; gland segmentation in histology [46]; nuclei in divergent images [47]
Swin-Unet [35]	CT	Abdominal multiorgan segmentation
iSegFormer [37]	3D MRI	Knee image segmentation
3D UX-net [38]	3D MRI	Brain tissue segmentation
UNesT [39]	MRI, CT	Abdominal multiorgan segmentation + kidney segmentation + whole brain segmentation
NestedFormer [40]	MRI	Brain tumor segmentation
RECIST [41]	CT	Automatic tumor segmentation and diameter size prediction
GT U-Net [42]	X-ray	Tooth therapy: root canal segmentation
FCN-transformer [43]	Colonoscopy	CRC segmentation
TraSeTR [44]	Endoscopy	Robot-assisted surgery

Table 2 Examples of ViT applications in medical image detection

Method	Category	Medical application
DETR [48]	MRI	Lymphoproliferative diseases detection
COTR [49]	Colonoscopy	CRC detection
SATr [50]	CT	Universal lesion detection
UCLT [51]	CT	Lung nodule detection
IHD [52]	CT	Brain injury hemorrhage detection

for feature extraction and convergence acceleration. A transformer encoder is used to encode and recalibrate the features, a transformer decoder for object querying, and a feedforward network for object detection.

Global lesion detection in CT scans was performed using a slice attention transformer (SATr) [50]. The backbone of the SATr is a combination of convolution and transformer attention that detects log-distance feature dependencies while preserving the local features.

Lung nodule detection was investigated using an unsupervised contrastive learning-based transformer (UCLT) [51]. Lung nodules are small cancerous masses that are difficult to detect in complex lung structures because of their size. This study harnessed contrastive learning (CL) and the ViT to break down the volume of CT images into small patches of non-overlapping cubes, and extract the embedded features for processing using the transformer attention mechanism.

To predict the hemorrhage category of brain injuries in CT scans, a transformer-based architecture was used for intracranial hemorrhage detection (IHD) [52]. Table 2 lists examples of ViT applications in image classification.

Applications of ViT in medical image classification

Classification is an important digital health solution in medical imaging analysis that helps medical practitioners identify objects within a complex structure to immediately categorize medical cases. Utilizing AI while working in remote areas and using telehealth systems with limited medical resources ensures the accuracy of final clinical decisions. The importance of AI emerged during the pandemic when the pressure on healthcare systems exceeded the capacity of the healthcare infrastructure. The ViT has different applications in medical imaging classification.

TransMed [53] uses a combination of the ViT and a CNN to classify multimodal data for medical analysis. The classification system includes disease and lesion identification. Figure 10 shows an example of the application of TransMed in image classification.

Shoulder implant manufacturers [54] use a transformer in orthopedic applications to assist in shoulder replacement surgery with artificial implants and joints. Before surgery, shoulder X-ray images were used to detect and classify the shoulder implant manufacturer vendor to determine the required accessories. The GasHis-transformer [55] is a multiscale visual transformer for detecting and classifying gastric cancer images using histopathological images of hematoxylin and eosin obtained by a microscope. Table 3 lists examples of ViT applications in image classification.

A comparative analysis of cervical cancer classifications using various deep learning (DL) algorithms, including the ViT, was conducted using cytopathological images [56]. A transformer-based model was used in brain metastases classification [57] from an MRI of the brain. Brain metastases are among the main causes of malignant tumors in the central nervous system [61]. ScoreNet [58]

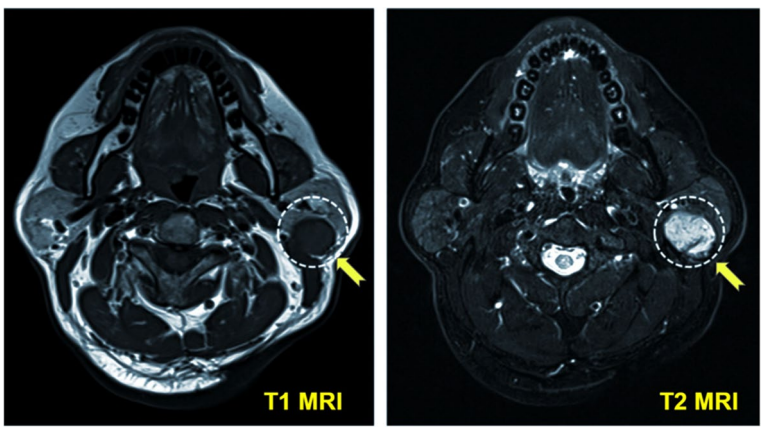


Fig. 10 Example of using the ViT for tumor classification in MRI images using TransMed [53]. The tumor is enclosed by the dashed circle indicated by the yellow arrow

Table 3 Examples of ViT applications in medical image classification

Method	Category	Medical application
TransMed [53]	MRI	Multi-modal classification: disease classification, lesion identification
Shoulder implant manufacture [54]	X-ray	Orthopedics: Shoulder implant manufacture classification
GasHis-transformer [55]	Histopathology microscopic images	Gastric cancer classification and detection
Multi-scale cytopathology [56]	Cytopathological images	Cervical cancer classification
Brain metastases classification [57]	MRI	Classification of the brain tumor of central nervous system
ScoreNet [58]	Histology Datasets of haematoxylin + eosin	Breast cancer classification
RadioTransformer [59]	X-ray	COVID-19 classification using chest X-ray images
TractoFormer [60]	Diffusion MRI	Nerve tracts modelling and 3D fiber representation

is a transformer-based model that classifies breast cancer using histopathology images. RadioTransformer [59] classifies COVID-19 cases based on chest X-rays. TractoFormer [60] classifies brain images based on tractography, which is a 3D model of the brain nerve tracts using diffusion MRI. TractoFormer discriminates between 3D fiber spatial relationships. It has proven to be accurate in classifying patients with schizophrenia vs controls.

Applications of ViT in medical imaging prognosis predication

The ability of the ViT to analyze time-series sequence data and obtain insights from previous data allows the prediction of future behaviors or patterns. In medical imaging, it is important to help healthcare practitioners predict the effects of diseases or cancers to treat them before they spread. Figure 11 shows the use of the transformer for surgical instructions, which are also implemented in Surgical Instruction Generation Transformer (SIGT) algorithm for surgical robots [62]. The algorithm used the ViT to analyze the visual scene during surgery and update the reinforcement learning (RL), reward, and status to predict the instructions for the robot.

The Sig-Former [63] can predict surgical instructions during an operation using the transformer attention mechanism to analyze the input image. The dataset includes images acquired during surgeries such as laparoscopic sleeve gastrectomy and laparoscopic ventral hernia repair.

The 3D Shuffle Mixer [64] analyzes 3D volumetric images from CT and MRI using context-aware dense predictions for different diseases, such as hemorrhagic stroke, abdominal CT images, and brain tumors.

Transformer Prediction:
Retract peritoneum incise with scissors and electrocautery.

Ground Truth:
While retracting peritoneum, identify correct plane by thin areolar.

Transformer Prediction:
Continue tying.

Ground Truth:
Tie knots in each suture with tails.

Transformer Prediction:
Process completed at first horizontal axis position.

Ground Truth:
Repeat process at horizontal axis positions.

Fig. 11 Examples of using ViT for surgical instruction prediction. Transformer prediction is based on the SIGT method [62]. GT is used as a reference for comparison and validation

Table 4 Examples of ViT applications in medical image prediction

Method	Category	Medical application
3D-SMx [64]	3D (MRI, CT)	Context-aware dense prediction for different diseases that includes hemorrhagic stroke, abdominal CT images, brain tumor
GBT [65]	Cancer genome (TCGA)	Computation pathology: genetic alteration
RTM [66]	Ultrasound	Fetal weigh at birth prediction
CLIMAT [67]	X-ray	Forecasts knee osteoarthritis trajectory
Sig-Former [63]	Laparoscopy	Surgical instructions prediction
SIGT [62]	Robot camera	Surgical instruction prediction and image captioning

Graph-based transformer models [65] predict genetic alteration. Ultrasound recordings are used for fetal weight prediction by the residual transformer model [66]. CLIMAT [67] forecasts the trajectory of knee osteoarthritis based on X-ray images from specialized radiologists. Table 4 lists examples of ViT applications in medical image prediction.

Applications of ViT in image reconstruction and synthesis

After acquiring data from medical imaging modalities such as MRI, CT, and digital X-ray, the images are stored as raw data in an unstructured format. To make this raw data readable, a reconstruction process is applied to retrieve images without any loss. However, this process is computationally expensive because of the size and complexity of reconstruction algorithms. The use of DL significantly improves the reconstruction performance by enhancing the preservation of fine image details within a reconstruction time of a few seconds. In contrast, traditional techniques such as image reconstruction using compressed sensing require more time [68]. Reconstructing magnetic resonance images is a challenge because of the size, complexity, and sparsity of the K-space matrix, in which the raw images are stored in the frequency domain.

SLATER [69] is a zero-shot adversarial transformer that performs the unsupervised reconstruction MRI images. SLATER maps the noise and latent representation to the MR coil-combined images. To maximize the consistency of the images, the operator input and maximum optimized prior information were combined using a zero-shot reconstruction algorithm. Figure 12 shows different methods for reconstructing fast MRI and the reconstruction error map using SLATER (ViT-based method) from T_1 weighted images. These were then compared with other techniques based on non-ViT methods.

The Task Transformer (T^2Net) [77] proposed an architecture to simultaneously reconstruct and enhance images using a super-resolution method for MRI. The T^2Net process can be divided into two parts. First, two CNN subtasks were used to extract

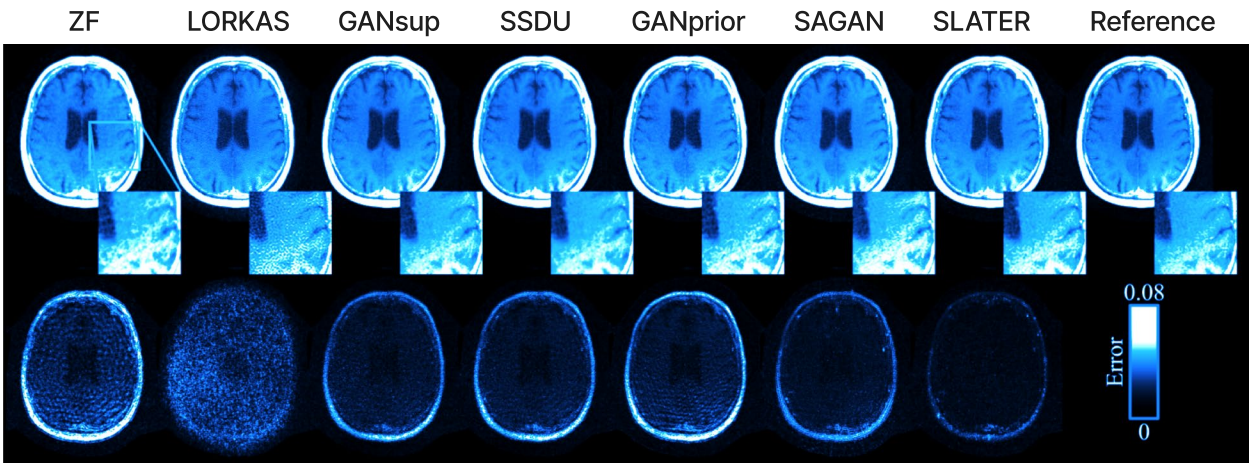


Fig. 12 Top: Different reconstruction methods from T_1 weighted acquisition of the fast MRI using different methods. ZF is a traditional Fourier method [70]. LORKAS [71, 72], GAN_{sub} [73], SSDU [74], GAN_{prior} [75], and SAGAN [76] are generative adversarial network (GAN) reconstruction-based methods. SLATER is a ViT-based method [69]. Bottom: Reconstruction error map [69]

domain-specific features. Second, T^2Net was embedded and the relationship between the two subtasks was synthesized. ReconFormer addresses the problem of under sampled K-space data by utilizing recurrent pyramid transformer layers to rapidly and efficiently retrieve the data [78]. Transformer-based methods for fast MRI reconstruction were evaluated in ref. [79]. The results showed that the combination of GANs and ViT achieved the best performance, i.e., a 30% improvement over standard methods such as the Swin transformer. Table 5 lists examples of ViT applications in image reconstruction.

A ViT-based (stereo transformer) was utilized in efficient dynamic surgical scene reconstruction [80] to reconstruct a robotic surgery scene acquired by an endoscope. This application is essential for surgical education, robotic guidance, and context-aware representation.

DuTrans adopted a Swin transformer as the core of their architectural design to reconstruct the sinograms of CT scans from the attenuation coefficient of the Hounsfield unit [81, 83]. The accurate reconstruction of CT scans is essential to obtain high-quality images, reduce radiation doses, and distinguish fine details to facilitate the early detection of cancers.

MIST-net proposed a multidomain transformer model to reconstruct CT scans [82]. MIST-net can reduce radiation doses without compromising image quality. MIST-net incorporates the Swin transformer architecture, residual features, and an edge enhancement filter to reconstruct the desired CT image.

Applications of ViT in telehealth

There is an increasing need for efficient techniques to process all medical information within the healthcare ecosystem. This is because of the complex nature of the unstructured format of medical data, such as images, clinical reports, and laboratory results. The ViT provides a comprehensive solution as it can process medical data in different formats and automatically generate reports or instructions. Figure 13 shows the main components of a telehealth ecosystem: the data source, ingestion, machine learning, and data analysis.

The hospital information system (HIS) and radiology information system register the patient and store data in electronic health records (EHRs) and picture archiving and communication systems (PACS) to be shared within the telehealth ecosystem. The HIS relies on standards such as Health Level 7 and Fast Healthcare Interoperability Resources for the exchange of patient metadata or EHRs [84, 85]. PACS is used to store and transfer medical images, mainly in the Digital Imaging [86] and Communications in Medicine [87] format, which are available to medical staff for further clinical analysis.

Patient data are shared in a cloud or server, either in real-time streaming or in batches from a data storage warehouse or data lake. The ViT or any other machine learning model is used to train the system on the ingested data. Once the model has been deployed, the ViT can be used to analyze medical data, approximately 90% of which are in an image format. Once the data have been analyzed, the results are sent to update patient records in the EHR or other storage systems.

Applications of ViT in report generation

The ViT provides a unified solution that processes text along with unstructured data, such as images. The advantage of using the ViT is that it can process and generate radiology reports, surgical instructions, and other clinical reports in a global context by retrieving huge amounts of information stored in health information systems.

Figure 14 shows the image capture, report consistency, completeness, and report generation by the Real Time Measurement, Instrumentation & Control (RTMIC) [88] and International Federation of Clinical Chemistry (IFCC) algorithms [89] from an input of medical images. The RTMIC is a ViT-based algorithm used for medical image captioning [88]. The GT is a manual reference written by an expert. Att2in is an attention-based method used for comparison [90]. The quality standards for health information systems state that the transferred data

Table 5 Examples of ViT applications in medical image reconstruction

Method	Category	Medical application
SLATER [69]	MRI	MRI unsupervised reconstruction
T^2Net [77]	MRI	Image reconstruction and super-resolution enhancement
ReconFormer[78], FastMRIRecon [79]	MRI	Accelerated MRI reconstruction
E-DSSR [80]	Endoscopy	Surgical robot scene reconstruction
DuTrans [81], MIST-net [82]	CT	CT sinograms reconstruction

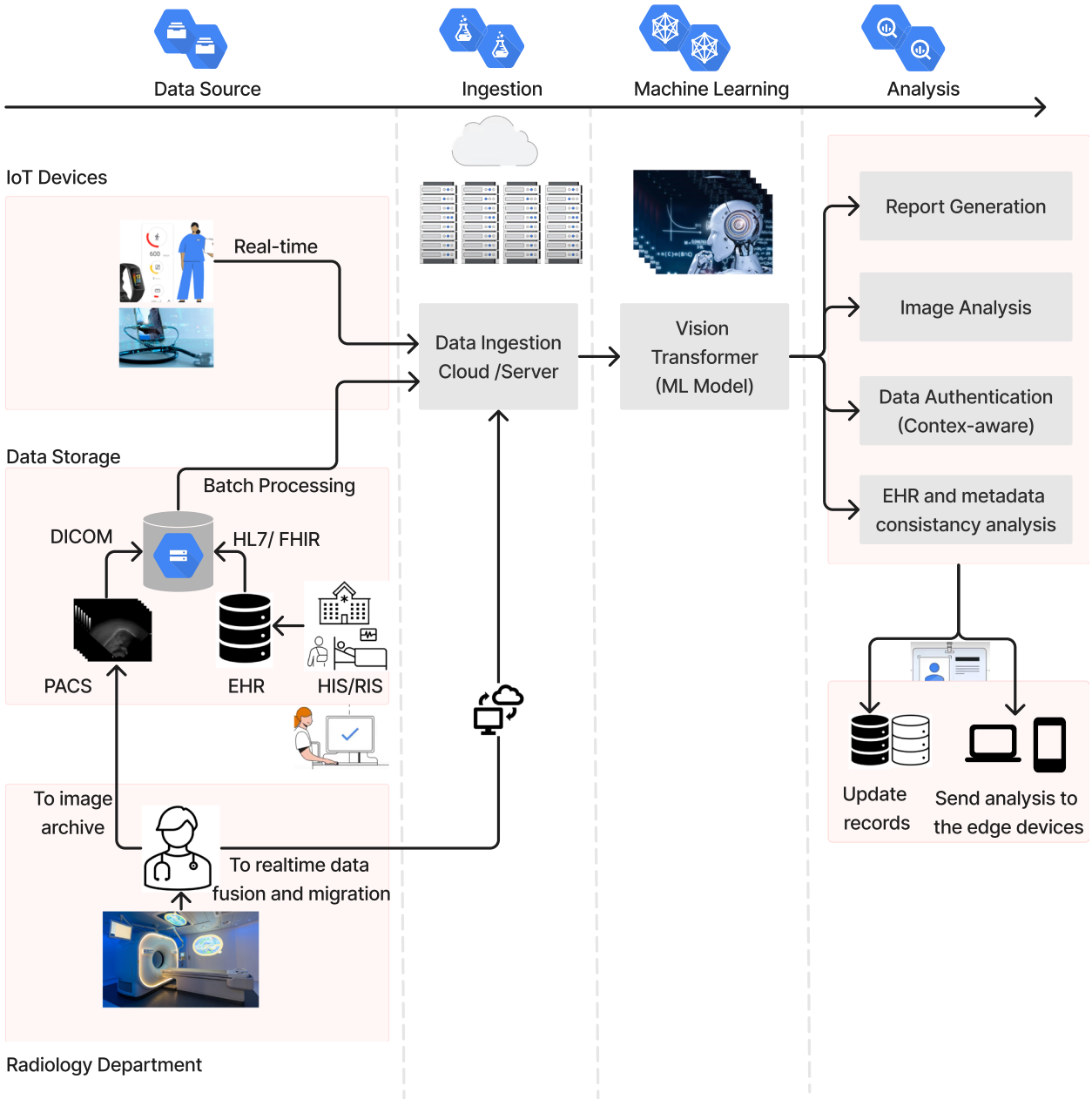


Fig. 13 Schematic of the components of the ViT in a telehealth ecosystem

should be consistent and complete. The IFCC algorithm [89] improves the factual completeness and consistency in image-to-text radiology report generation. The algorithm uses a combination of transformers to extract features and RL to optimize the results.

The transformer efficiently addresses the challenges of handling biased medical data and long and inconsistent paragraphs. The AlignTransformer can produce a long descriptive and coherent paragraph based on the analysis of medical images [91]. It mainly operates in two stages. First, it aligns the medical tags with the related medical images to extract the features. Second, the extracted features are used to generate a long report based on the training data for each medical tag.

The transformer is also used to generate surgical reports during robot-assisted surgery by learning domain adaptation in the Learning Domain Adaption Surgical Robot (LDASR) [92]. The LDASR uses a

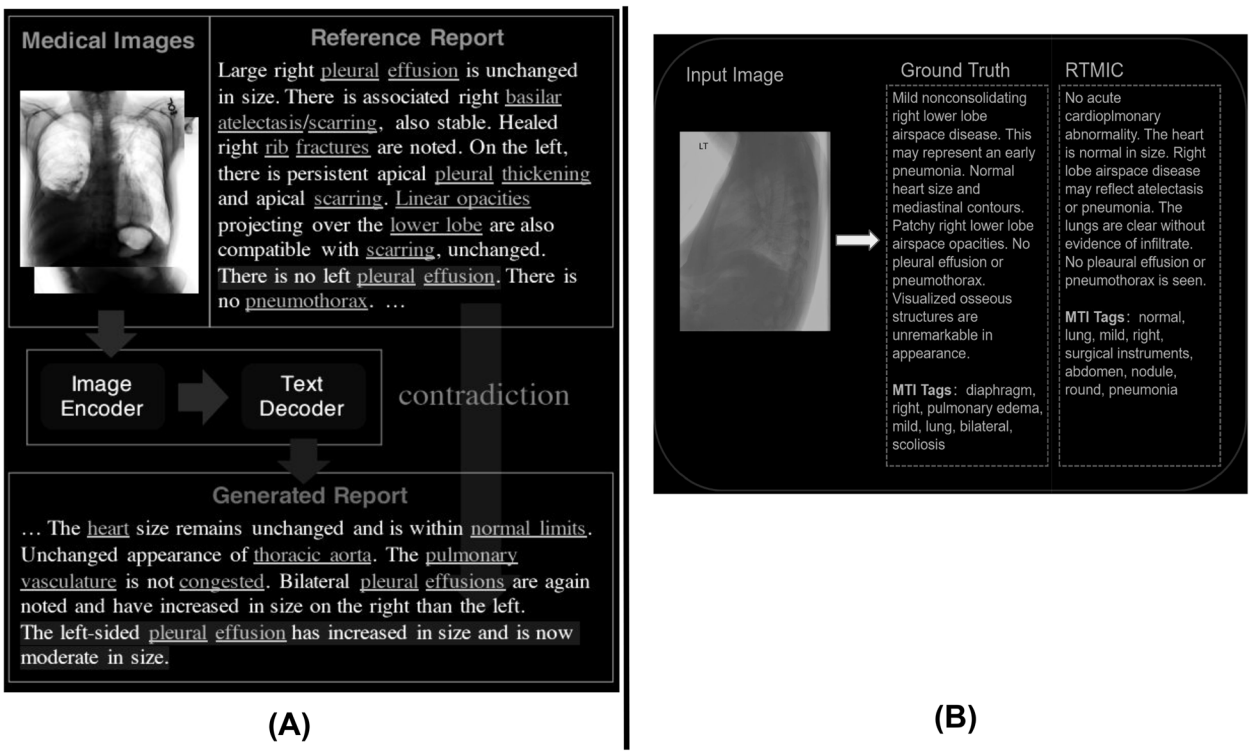


Fig. 14 Examples of report generation from the input image using the ViT. **a** Sample of results by the IFCC algorithm [89] for report completeness and consistency; **b** Example of report generation results by the RTMIC algorithm [88]

transformer to learn the relationships between the desired region of interest, surgical instruments, and images to generate image captions and reports during surgery. Table 6 lists examples of ViT applications in image generation.

Applications of ViT in telehealth security

Telehealth security is receiving significant attention from healthcare providers owing to the emerging risks associated with leveraging advanced technologies such as machine learning. In healthcare, there is a serious risk of misdiagnosing a patient with the wrong disease or even diagnosing a healthy person with a disease.

An adversarial attack refers to a malicious attack against the machine learning algorithm or data vulnerability. These attacks may include modifying the data or algorithm code, resulting in incorrect outputs [93, 94]. The accuracy of the algorithm may also be affected by the manipulation of the code or labeled data. Cybercriminals attempt to extort money from healthcare providers by threatening to publish patient information and encrypt the database. Figure 15 shows the effects of data poisoning by adversarial attacks on medical images that attempt to disrupt the behavior of the trained machine learning model.

Researchers have developed the following countermeasures against cybercrime:

Table 6 Examples of ViT applications in medical report generation

Method	Category	Medical application
RTMIC [88]	Medical images general	Report generation from medical images (e.g., MRI, CT, PET and X-ray)
IFCC [89]	Medical images general	Medical report completeness and consistency
AlignTransformer [91]	Medical images general	Long report generation from medical images tags
LDASR [92]	Surgical robot camera	Surgical report generation

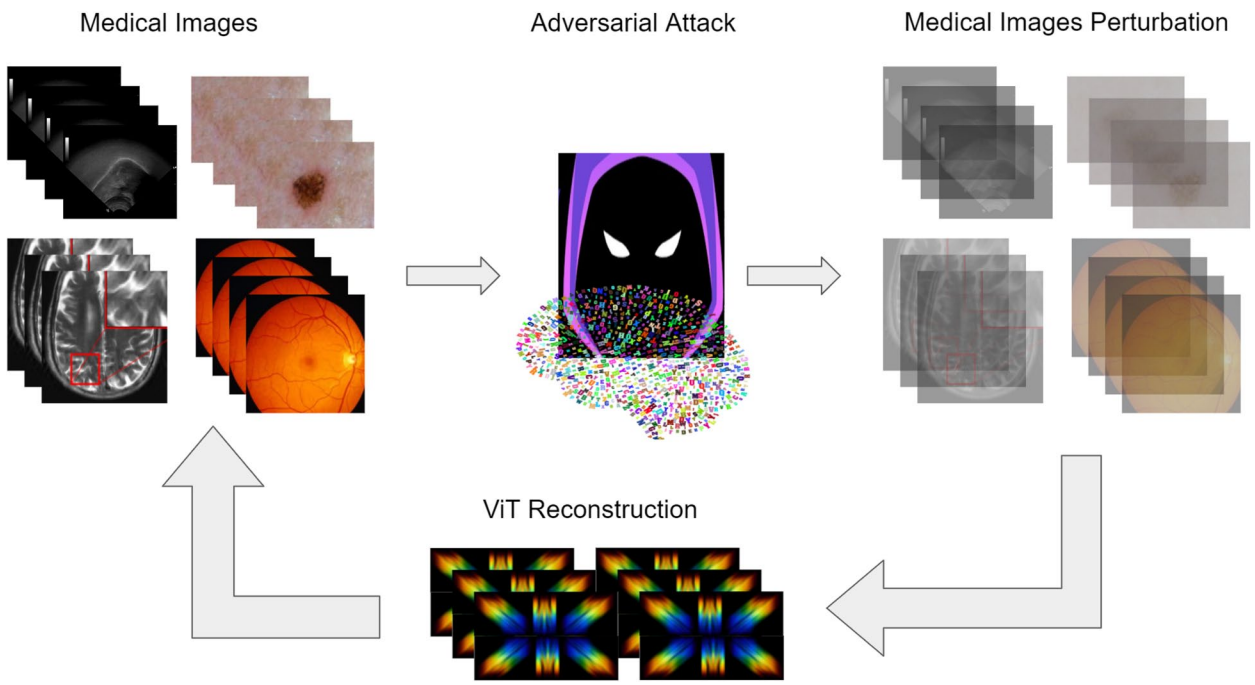


Fig. 15 Illustration of data poisoning by an adversarial attack that fools learning-based models trained on medical image datasets

1. Implement a context-aware system to ensure that the code is safe and not jeopardized.
2. Store data in an encrypted cloud environment and ensure that these are backed up.
3. Federated learning is another measure that uses a distributed computing engine to process data in geographically distributed environments that maintain data in different locations, making them difficult to hack.
4. Embrace a zero-trust policy when managing access control systems in digital health applications. This provides an additional authentication measure by considering different attributes before granting access instead of just relying on a role-based access system.

Unlike the ViT, traditional CNN-based algorithms are not robust against adversarial attacks because of the simplicity of their architecture [95]. The complexity of the ViT algorithm and its ability to extract features in a global context are solid grounds for detecting irregularities in data entry. The ViT has been used for data encryption [96], anomaly detection [97], network intrusion system detection [98], anti-spoofing [99], and patch processing [100]. Table 7 lists examples of the applications use of ViT in information system security.

Table 7 Examples of ViT applications in security

Method	Application
Jigsaw block-based encryption [96]	Data encryption
MFVT [97]	Anomaly detection
Image conversion from network data-flow [98]	Network intrusion system detection
Zero-shot face [99]	Anti-spoofing
Backdoor defender [100]	Patch processing

Roadmap for implementing ViT

Figure 16 shows the four stages in the end-to-end implementation of the ViT model pipeline. These are problem formulation, data processing; model implementation, training, and validation; and model deployment and quality assurance, respectively.

Problem formulation

Before implementing a machine learning model, the problem must be understood and formulated to fit the context of the desired product-use case.

Data preparation

Once the problem is understood, high-quality data must be prepared for the AI algorithm. The data must be relevant, accurate, statistically balanced, and sufficient for

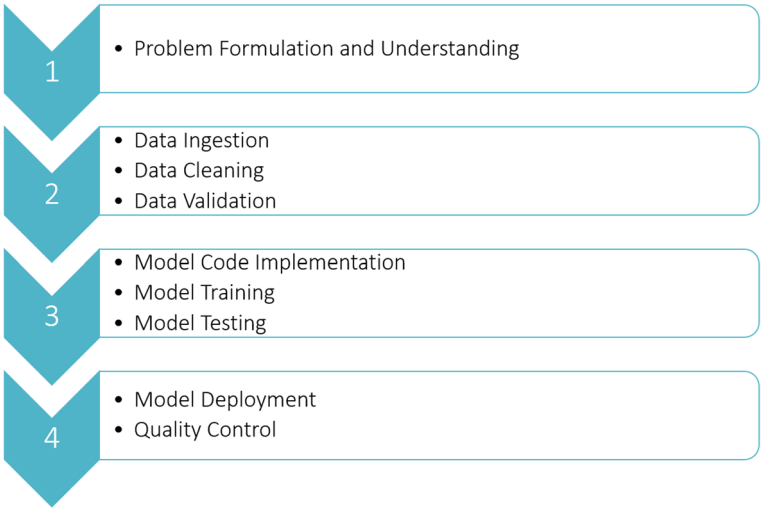


Fig. 16 Roadmap for ViT implementation

training. The data should also be verified by different qualitative and qualitative measures to ensure their validity. This helps stabilize the model during training and speeds up convergence to obtain the optimal solution.

Model and code implementation

There is no master algorithm that fits everything; each has its own advantages and disadvantages. The suitable ViT model or architecture is selected based on the available data and application to achieve the desired success metrics. The model hyperparameters are fine-tuned during the training stage to achieve the desired accuracy and prevent overfitting or underfitting. The model should also be validated and tested on datasets other than those used for training.

Model deployment and testing

Finally, once the model passes all the end-to-end testing and verification processes, it should be ready for deployment. Different environments can be used to deploy the final product in different cloud or on-premise applications. The recommended environment is a cloud-based system because it can automatically generate a model on a scale that fits the computational resources for different applications. The deployed model should undergo different quality assurance and monitoring processes to ensure that the target performance of the system is met during tests outside the laboratory or development environment. Any bugs found in the code should be fixed. If the performance of the trained model is insufficient, then a new dataset should be used for training.

Limitations and challenges of ViT in digital health

Transformer-based algorithms are emerging as the state-of-art in vision tasks to replace traditional standalone CNN architectures. However, transformer-based models have disadvantages in terms of technical or regulatory compliance requirements. These include data size and labeling, the need for a hybrid model, data bias and model fairness, and ethical and privacy challenges.

Dataset size and labeling challenges

Similar to other attention-based mechanisms, transformers inherently require a huge amount of data to train the model. The transformer achieved the best performance compared with the well-known ResNet architecture when trained on the JFT dataset [101], which contains 300 million images and 18000 classes. However, when trained on the ImageNet-21 k dataset [102], which contains approximately 14 million images and 21000 classes, the transformer performance did not surpass that of the ResNet architecture trained on the same dataset ImageNet-1 k [103, 104] with 1.28 million images and 1000 classes. Figure 17 shows the performance of the ViT and ResNet architectures with respect to the data size.

The results show that ResNet performed better when the dataset was small. ResNet and ViT exhibited almost the same performance when the trained on approximately 100 million samples. However, the ViT achieved superior performance compared with

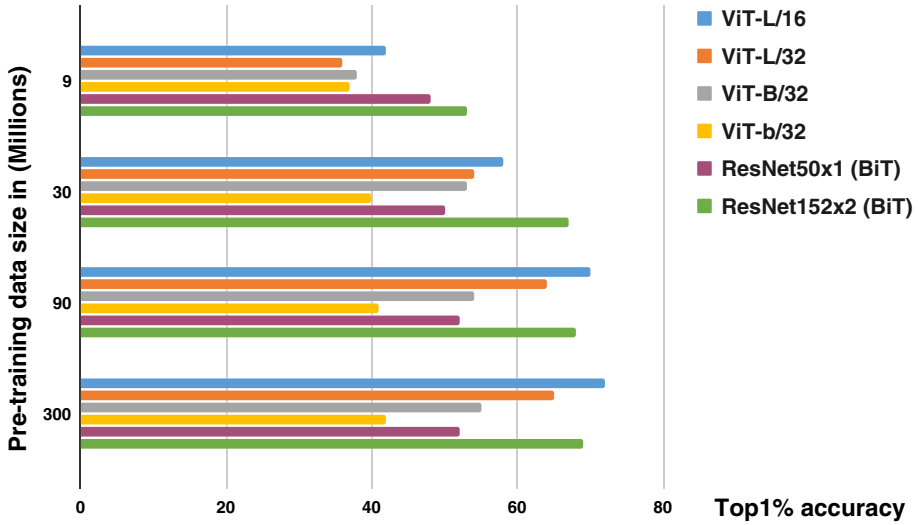


Fig. 17 Comparison between ViT and ResNet (BiT) architecture accuracies on different sizes of training data. The *y*-axis is the size of pretraining data in the ImageNet dataset. The *x*-axis is the accuracy selected from the top 1% of the selected five-shots of ImageNet. Results according to the study in ref. [1]

ResNet when the dataset size was larger than 100 million images [1].

The limited dataset size is challenging in medical applications because it is difficult to obtain a clean and high-quality dataset that is feasible for clinical application standards. Moreover, finding qualified specialists to annotate millions of images is difficult, expensive, and time-consuming.

Transfer learning, data augmentation, adversarial imaging synthesis, and automatic data labeling are among the best practices to deal with the problem of insufficient dataset size. The researchers in ref. [105] suggested that the ViT model outperformed ResNet when trained from scratch on the large ImageNet dataset without using data augmentation or a large pretrained model. Thus, there is a tradeoff between dataset size limitations and performance because having a large dataset but sufficient computational resources for training remains a challenge. The use of cloud-based data training could be a solution to limited resources. However, this is an expensive option for academia and more suitable for industrial applications. Similarly, ref. [106] proposed an effective weight initialization scheme to fine-tune the ViT using self-supervised inductive biases learned directly from small-scale datasets. This reduced the need for huge datasets for training, and hence required less computational resources.

CL is beneficial in medical image applications because it can minimize the difference between similar object representations in the latent space, while maximizing the difference between dissimilar objects [107]. CL has been used with ViT in medical histopathology to classify large images (in gigapixels) and obtain inferences to distinguish between multilabel cancer cells for classification [108].

The need for hybrid model with transformer

The transformer was initially designed to process language models in a sequential format. Since then, it has been modified to process vision tasks by splitting the image into small patches and processing them sequentially as a text-like model. The transformer can obtain inferences about the information in a global context to capture a wide range of dependencies between objects; however, it has a limited feature localization capacity. While the standalone transformer model is sufficient for most classification tasks, in the case of image segmentation for critical medical applications that require a high-quality image, the transformer performance is insufficient and must be combined with a hybrid model.

Unet or ResNet architectures are widely used as standard models for medical image segmentation that can preserve image details owing to the nature of the encoder-decoder architecture with residual connections. However, Unet and ResNet have inherited the limitation of CNNs in failing to capture a wide range of dependencies by having only local feature extraction capabilities. TransUNet was the first architecture proposed for medical imaging segmentation that combined the transformer

[10] and Unet architectures for local and global feature extraction.

The transformer was also combined with RL to generate instructions for surgical robots [62, 63]. The transformer can capture features to update the state-reward status in the RL to automate robot tasks. The RL-transformer combination has also been used in medical image captioning [88] to automatically generate medical reports within the hospital system.

Data bias and fairness

Training machine learning models using huge datasets (in millions or billions of examples) requires resources with sufficient computational power and storage. Therefore, many algorithms tend to apply dimensionality reduction to minimize model parameters, which reduces the extracted features. This allows model training with reduced computational and memory requirements. However, there is a possibility of losing information with less representation in the feature map or dataset. Consequently, the model may be biased toward labels or classes with the largest amount of training data. The bias in the results could be significant, particularly when label balancing was not performed before training. In medical applications, rare diseases and outliers could be disregarded from the model prediction.

In ref. [109], the fairness and interpretability of DL models were evaluated using the largest publicly available dataset, the Medical Information Mart for Intensive Care, version IV. The study found that some DL models lacked fairness when relying on demographics and ethnicity to predict mortality rates. In contrast, DL models that used proper and balanced critical features for training were not biased and tended to be fair. In many models, racial attributes were used unequally across subgroups. This resulted in inconsistent recommendations on the use of mechanical ventilators for treatments or in intensive care units when relying on demographic and racial categories such as gender, marital status, age, insurance type, and ethnicity. Figure 22 in Appendix A.5 shows examples of the global features importance scores used to predict mortality rates using different machine learning methods. The figure shows the bias of the importance score toward certain features when machine learning algorithms were changed.

Ethical and privacy challenges

Information-sharing in healthcare information systems is regulated, although privacy and ethical regulations may differ across jurisdictions. For example, the Health Insurance Portability and Accountability Act (HIPAA) of the United States regulates healthcare information systems to protect sensitive patient information. The HIPAA states that such information cannot be disclosed without patient consent. Patients also have the right to access their data, ask for modifications, and know who accesses them. While such regulations help preserve patient privacy, collecting health-related datasets or making them available to the public is a challenge. This is a critical issue in the case of the ViT as millions of examples are required to train the model and obtain accurate results. Using the ViT or any other machine learning model trained on a large dataset has a higher risk of errors, and the results are subject to ethical concerns. Many large datasets are obtained from the Internet; hence, the sources may be unknown or untrustworthy, and there is no previous consent to collect these data. Training the ViT from untrusted sources could generate false results, which could lead to errors or offensive content in generated patient reports. The consequences may be worse in the case of data breaches or cyberattacks on the healthcare information system as these could alter patient records, images, or the data streaming performance of the telehealth system. Although the ViT is more robust against adversarial attacks, there is no guarantee that the ViT-based model will not generate inappropriate content. This raises concerns regarding the need to regulate the current AI industry as well as applications in healthcare to ensure that the input and output of the systems are clean and valid for clinical applications. Federated learning from different healthcare facilities and edge devices or servers can help maintain a high level of data privacy. However, the research in ref. [110] reported vulnerabilities in retrieving original data from the shared model weights.

Conclusions

The ViT has emerged as the state-of-the-art in image recognition tasks, replacing traditional standalone machine learning algorithms such as CNN-based models. The ViT can extract information in a global context using an attention-based mechanism and analyze images, texts, patterns, and instructions.

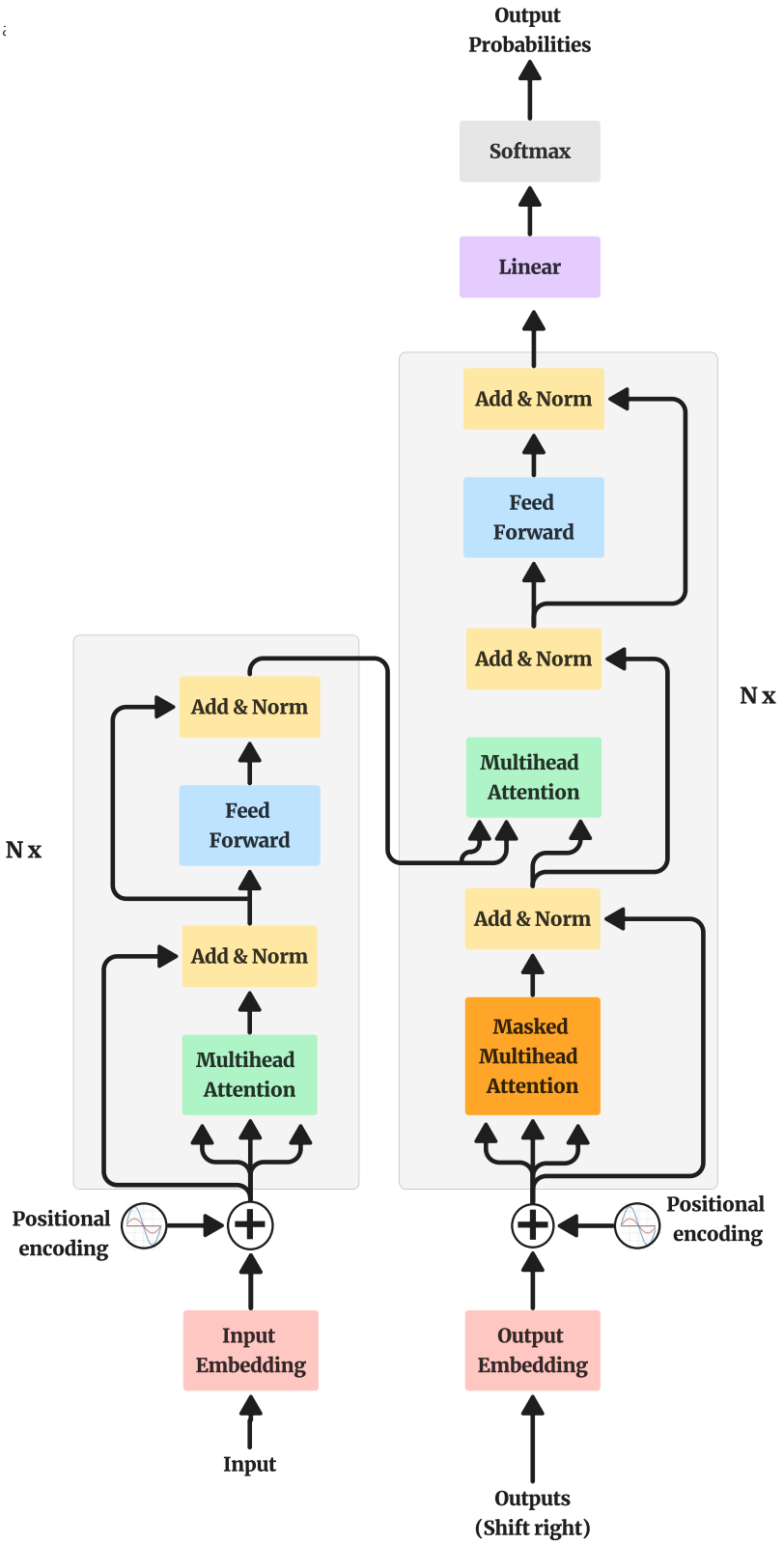
The superior performance of the ViT makes it practical for various digital medicine applications such as segmentation, classification, image reconstruction, image enhancement, data prognosis prediction, and telehealth security.

Appendix

ViT common architectures

A.1 Typical transformer architecture

Fig. 18 Transformer typical a



A.2 Architecture example of using transformer in image recognition

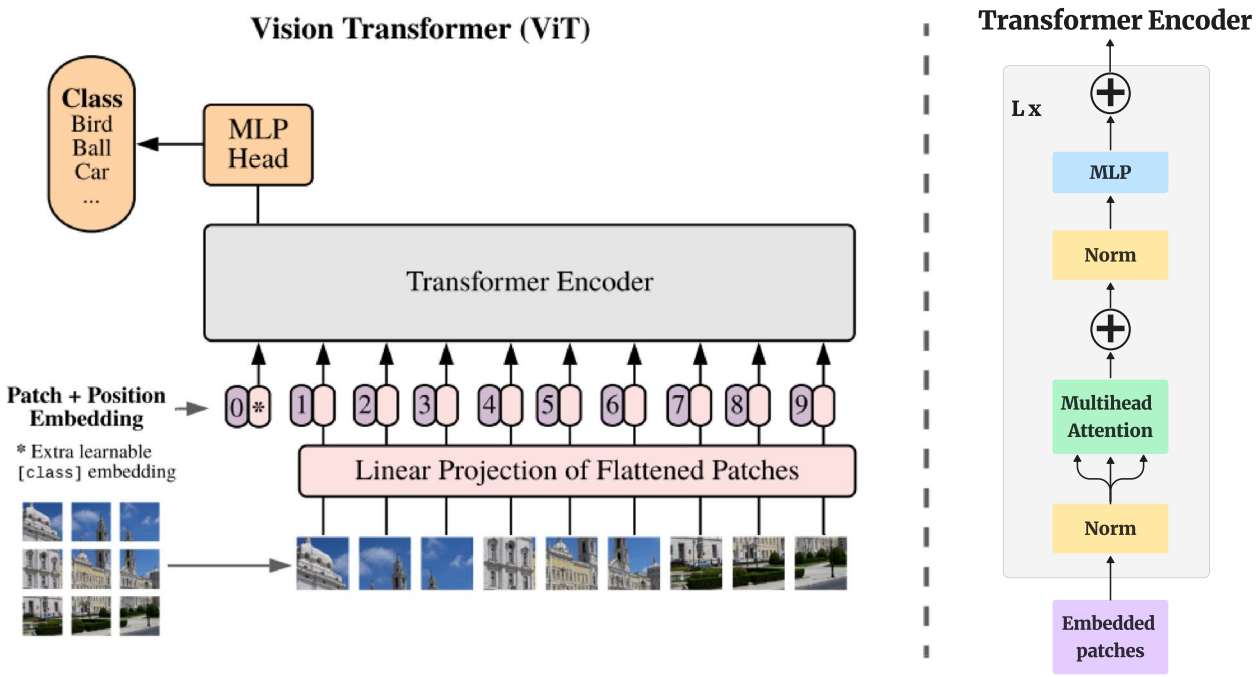


Fig. 19 Example of using Transformer architecture for image recognition [1]

A.3 TransUnet architecture diagram

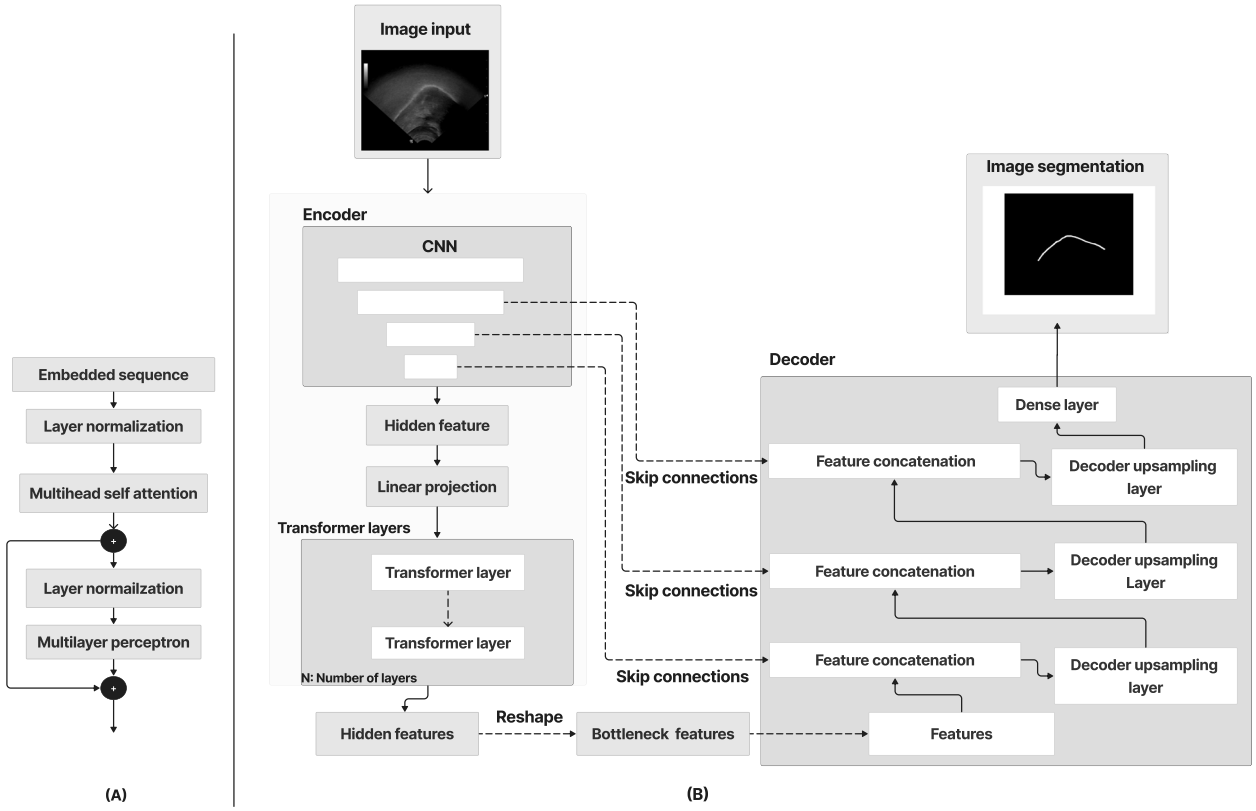
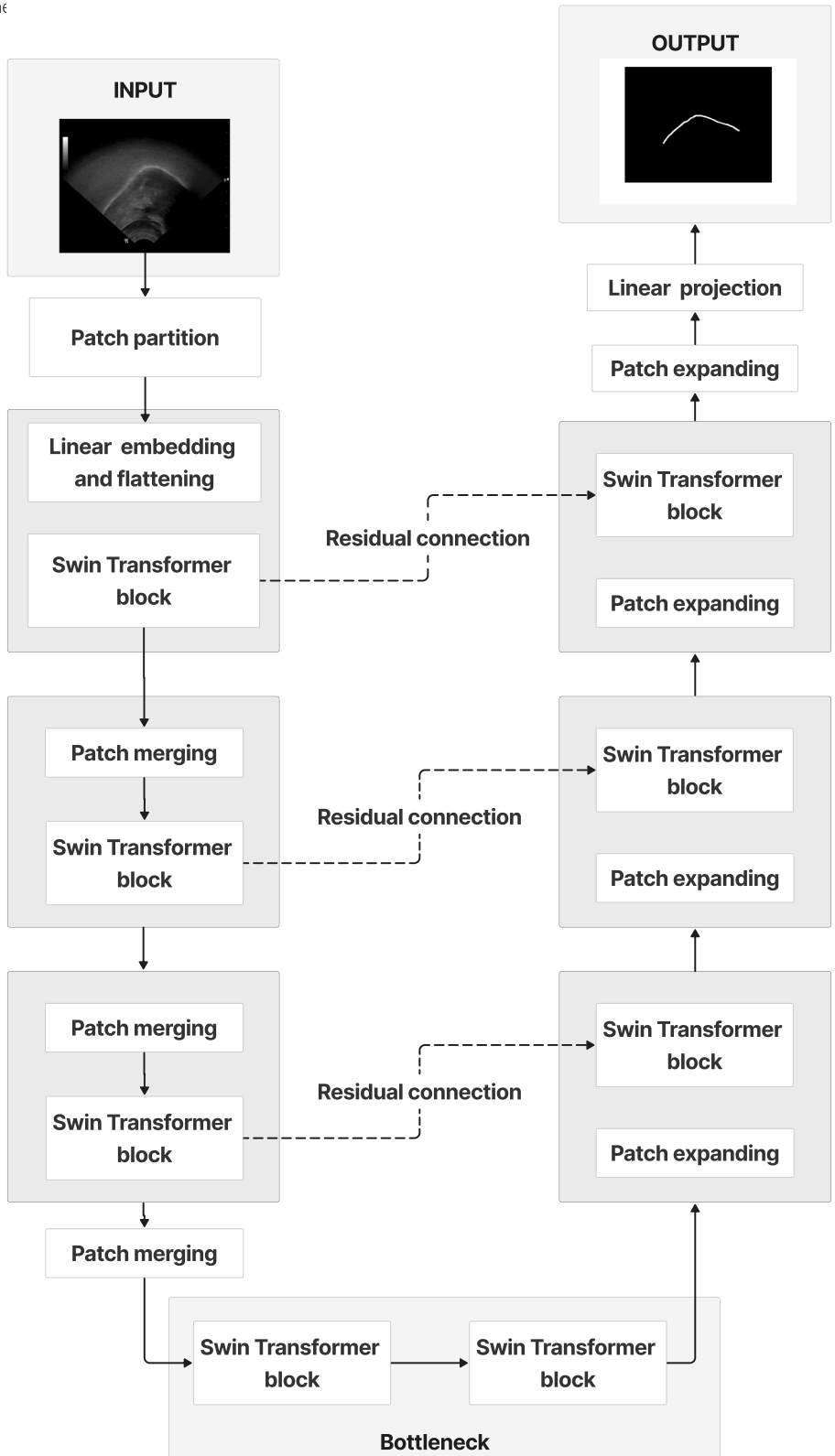


Fig. 20 **a** Transformer layer diagram; **b** TransUnet architecture [10]

A.4 Swin-transUnet architecture diagram

Fig. 21 Swin TransUne



A.5 Example of global features importance rank (Fig. 22)

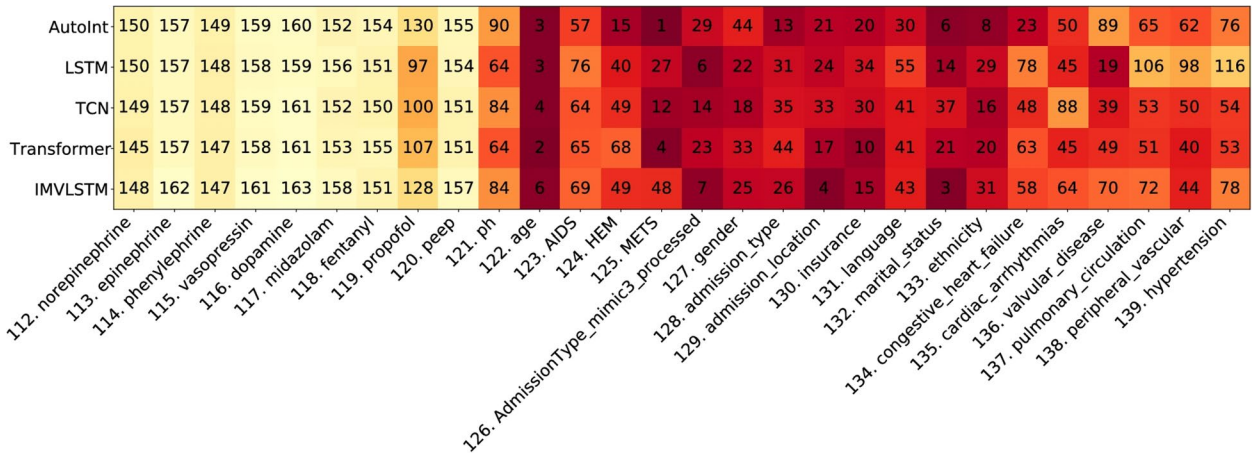


Fig. 22 Examples of global features that are used for mortality predictions are numbered from (112-139). The numbers in the table depicts the rank sore and each column represents a feature and its importance score by different methods on the horizontal line [109]. AutoInt [111], LSTM [112], TCN [113], Transformer [8], IMVLSTM [114] are the machine learning methodologies

Abbreviations

3D	Three-dimensional
AI	Artificial intelligence
BERT	Bidirectional encoder representations from transformers
CNN	Convolutional neural networks
COVID-19	Coronavirus disease 2019
CT	Computed tomography
CL	Contrastive learning
CRC	Colorectal cancer
COTR	Convolutional transformer
DL	Deep learning
EHR	Electronic health record
FCN	Fully convolutional network
GT	Ground truth
GAN	Generative adversarial network
HIS	Hospital information system
HIPAA	Health Insurance Portability and Accountability Act
K	Key
LSTM	Long short-term memory
MLP	Multilayer perceptron
MSA	Multihead self-attention
MRI	Magnetic resonance imaging
NLP	Natural language processing
PACS	Picture archiving and communication system
Q	Query
RL	Reinforcement learning
RNN	Recurrent neural network
V	Value
ViT	Vision transformer
DETR	Detection transformer
SATr	Slice attention transformer
UCLT	Unsupervised contrastive learning-based transformer
IHD	Intracranial hemorrhage detection
SIGT	Surgical Instruction Generation Transformer
RTMIC	Real Time Measurement, Instrumentation & Control
IFCC	International Federation of Clinical Chemistry
LDASR	Learning Domain Adaption Surgical Robot

Acknowledgements
Not applicable.

Authors' contributions
KA-h, FG and AK provided the conception; KA-h provided the methodology and investigation; KA-h and AK made the data analysis; KH prepared the original draft; FG, ITC, AK and KA-h reviewed and edited the manuscript; FG, AK and ITC provided the supervision; FG provided the funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding
This research was supported by a grant from the National Research Council of Canada through the Collaborative Research and Development Initiative.

Availability of data and materials
The data underlying this manuscript is based on existing publications and is available in the referenced literature or from the corresponding authors upon reasonable request.

Declarations

Competing interests
The authors declare no competing financial or non-financial interests.

Received: 21 March 2023 Accepted: 30 May 2023
Published online: 10 July 2023

References

1. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T et al (2021) An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the 9th international conference on learning representations, OpenReview.net, Vienna, 3-7 May 2021
2. Zhang QM, Xu YF, Zhang J, Tao DC (2023) ViTAEv2: vision transformer advanced by exploring inductive bias for image recognition and beyond. Int J Comput Vis 131(5):1141-1162. https://doi.org/10.1007/s11263-022-01739-w
3. Han K, Wang YH, Chen HT, Chen XH, Guo JY, Liu ZH et al (2023) A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 45(1):87-110. https://doi.org/10.1109/TPAMI.2022.3152247
4. Wang RS, Lei T, Cui RX, Zhang BT, Meng HY, Nandi AK (2022) Medical image segmentation using deep learning: a survey. IET Image Process 16(5):1243-1267. https://doi.org/10.1049/ipr2.12419
5. Bai WJ, Suzuki H, Qin C, Tarroni G, Oktay O, Matthews PM et al (2018) Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G (eds) Medical image computing and computer assisted intervention. 21st international conference, Granada, September 2018. Lecture notes in computer science (Image processing, computer vision, pattern recognition, and graphics), vol 11073. Springer, Cham, pp 586-594. https://doi.org/10.1007/978-3-030-00937-3_67
6. Wang YX, Xie HT, Fang SC, Xing MT, Wang J, Zhu SG et al (2022) PETR: rethinking the capability of transformer-based language model in scene text recognition. IEEE Trans Image Process 31:5585-5598. https://doi.org/10.1109/TIP.2022.3197981
7. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), Association for Computational Linguistics, Minneapolis, 2-7 June 2019
8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al (2017) Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems, Curran Associates Inc., Long Beach, 4-9 December 2017
9. Gao Y, Phillips JM, Zheng Y, Min RQ, Fletcher PT, Gerig G (2018) Fully convolutional structured LSTM networks for joint 4D medical image segmentation. In: Proceedings of the 15th international symposium on biomedical imaging, IEEE, Washington, 4-7 April 2018. https://doi.org/10.1109/ISBI.2018.8363764
10. Chen JN, Lu YY, Yu QH, Luo XD, Adeli E, Wang Y et al (2021) TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv: 2102.04306
11. Lin AL, Chen BZ, Xu JY, Zhang Z, Lu GM, Zhang D (2022) DS-TransUNet: dual Swin transformer U-Net for medical image segmentation. IEEE Trans Instrum Meas 71:4005615. https://doi.org/10.1109/TIM.2022.3178991
12. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Proceedings of the 1st international conference on learning representations, ICLR, Scottsdale, 2-4 May 2013
13. Maeda Y, Fukushima N, Matsuo H (2018) Taxonomy of vectorization patterns of programming for fir image filters using kernel subsampling and new one. Appl Sci 8(8):1235. https://doi.org/10.3390/app8081235
14. Jain P, Vijayanarasimhan S, Grauman K (2010) Hashing hyperplane queries to near points with applications to large-scale active learning. In: Proceedings of the 23rd international conference on neural information processing systems, Curran Associates Inc., Vancouver, 6-9 December 2010
15. Yu Y, Si XS, Hu CH, Zhang JX (2019) A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 31(7):1235-1270. https://doi.org/10.1162/neco_a_01199
16. Huang ZH, Xu W, Yu K (2015) Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv: 1508.01991
17. Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional sequence to sequence learning. In: Proceedings of the 34th international conference on machine learning, PMLR, Sydney, 6-11 August 2017
18. Takase S, Kiyono S, Kobayashi S, Suzuki J (2022) On layer normalizations and residual connections in transformers. arXiv preprint arXiv: 2206.00330
19. Topal MO, Bas A, van Heerden I (2021) Exploring transformers in natural language generation: GPT, BERT, and XLNet. arXiv preprint arXiv: 2102.08036
20. Wang SL, Liu F, Liu B (2021) Escaping the gradient vanishing: periodic alternatives of softmax in attention mechanism. IEEE Access 9:168749-168759. https://doi.org/10.1109/ACCESS.2021.3138201
21. Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. arXiv preprint arXiv: 1607.06450
22. Taud H, Mas JF (2018) Multilayer perceptron (MLP). In: Camacho Olmedo M, Paegelow M, Mas JF, Escobar F (eds) Geomatic approaches for modeling land change scenarios. Lecture notes in geoinformation and cartography. Springer, Cham, pp 451-455. https://doi.org/10.1007/978-3-319-60801-3_27
23. Akinyelu AA, Zaccagna F, Grist JT, Castelli M, Rundo L (2022) Brain tumor diagnosis using machine learning, convolutional neural networks, capsule neural networks and vision transformers, applied to MRI: a survey. J Imaging 8(8):205. https://doi.org/10.3390/jimaging8080205
24. Mahoro E, Akhloufi MA (2022) Breast cancer classification on thermograms using deep CNN and transformers. Quant Infrared Thermogr J. https://doi.org/10.1080/17686733.2022.2129135
25. Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN (2022) Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. Nat Cancer 3(9):1026-1038. https://doi.org/10.1038/s43018-022-00436-4
26. Al-Hammuri K, Gebali F, Thirumarai Chelvan I, Kanan A (2022) Tongue contour tracking and segmentation in lingual ultrasound for speech recognition: a review. Diagnostics 12(11):2811. https://doi.org/10.3390/diagnostics12112811
27. Al-Hammuri K (2019) Computer vision-based tracking and feature extraction for lingual ultrasound. Dissertation, University of Victoria
28. McMaster C, Bird A, Liew DFL, Buchanan RR, Owen CE, Chapman WW et al (2022) Artificial intelligence and deep learning for rheumatologists. Arthritis Rheumatol 74(12):1893-1905. https://doi.org/10.1002/art.42296
29. Beddiar DR, Oussalah M, Seppänen T (2023) Automatic captioning for medical imaging (MIC): a rapid review of literature. Artif Intell Rev 56(5):4019-4076. https://doi.org/10.1007/s10462-022-10270-w
30. Renna F, Martins M, Neto A, Cunha A, Libânio D, Dinis-Ribeiro M et al (2022) Artificial intelligence for upper gastrointestinal endoscopy: a roadmap from technology development to clinical practice. Diagnostics 12(5):1278. https://doi.org/10.3390/diagnostics12051278
31. Coan LJ, Williams BM, Adithya VK, Upadhyaya S, Alkafri A, Czanner S et al (2023) Automatic detection of glaucoma via fundus imaging and artificial intelligence: a review. Surv Ophthal 68(1):17-41. https://doi.org/10.1016/j.survophthal.2022.08.005
32. Chang A (2020) The role of artificial intelligence in digital health. In: Wulfovich S, Meyers A (eds) Digital health entrepreneurship. Health informatics. Springer, Cham, pp 71-81. https://doi.org/10.1007/978-3-030-12719-0_7
33. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS et al (2022) Transformers in medical imaging: a survey. arXiv preprint arXiv: 2201.09873. https://doi.org/10.1016/j.media.2023.102802
34. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A (eds) Medical image computing and computer-assisted intervention. 18th international conference, Munich, October 2015. Lecture notes in computer science (Image processing, computer vision, pattern recognition, and graphics), vol 9351. Springer, Cham, pp 234-241. https://doi.org/10.1007/978-3-319-24574-4_28
35. Cao H, Wang YY, Chen J, Jiang DS, Zhang XP, Tian Q et al (2023) Swin-Unet: unet-like pure transformer for medical image segmentation. In: Karlinsky L, Michaeli T, Nishino K (eds) Computer vision. Tel Aviv, October 2022. Lecture notes in computer science, vol 13803. Springer, Cham, 205-218. https://doi.org/10.1007/978-3-031-25066-8_9
36. Dong H, Yang G, Liu FD, Mo YH, Guo YK (2017) Automatic brain tumor detection and segmentation using U-Net based fully convolutional

networks. In: Valdés Hernández M, González-Castro V (eds) Medical image understanding and analysis. 21st annual conference, Edinburgh, July 2017. Communications in computer and information science, vol 723. Springer, Cham, pp 506-517. https://doi.org/10.1007/978-3-319-60964-5_44
37. Liu Q, Xu ZL, Jiao YN, Niethammer M (2022) iSegFormer: interactive segmentation via transformers with application to 3D knee MR images. In: Wang LW, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer-assisted intervention. 25th international conference, Singapore, September 2022. Lecture notes in computer science, vol 13435. Springer, Cham, pp 464-474. https://doi.org/10.1007/978-3-031-16443-9_45
38. Lee HH, Bao SX, Huo YK, Landman BA (2022) 3D UX-Net: a large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv: 2209.15076
39. Yu X, Yang Q, Zhou YC, Cai LY, Gao RQ, Lee HH et al (2022) UNesT: local spatial representation learning with hierarchical transformer for efficient medical segmentation. arXiv preprint arXiv: 2209.14378
40. Xing ZH, Yu LQ, Wan L, Han T, Zhu L (2022) NestedFormer: nested modality-aware transformer for brain tumor segmentation. In: Wang LW, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer-assisted intervention. 25th international conference, Singapore, September 2022. Lecture notes in computer science, vol 13435. Springer, Cham, pp 140-150. https://doi.org/10.1007/978-3-031-16443-9_14
41. Tang YB, Zhang N, Wang YR, He SH, Han M, Xiao J et al (2022) Accurate and robust lesion RECIST diameter prediction and segmentation with transformers. In: Wang LW, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer assisted intervention. 25th international conference, Singapore, September 2022. Lecture notes in computer science, vol 13434. Springer, Cham, pp 535-544. https://doi.org/10.1007/978-3-031-16440-8_51
42. Li YX, Wang S, Wang J, Zeng GD, Liu WJ, Zhang QN et al (2021) GT U-Net: a U-Net like group transformer network for tooth root segmentation. In: Lian CF, Cao XH, Rekik I, Xu XN, Yan PK (eds) Machine learning in medical imaging. 12th international workshop, Strasbourg, September 2021. Lecture notes in computer science (Image processing, computer vision, pattern recognition, and graphics), vol 12966. Springer, Cham, pp 386-395. https://doi.org/10.1007/978-3-030-87589-3_40
43. Sanderson E, Matuszewski BJ (2022) FCN-transformer feature fusion for polyp segmentation. In: Yang G, Aviles-Rivero A, Roberts M, Schönlieb CB (eds) Medical image understanding and analysis. 26th annual conference, Cambridge, July 2022. Lecture notes in computer science, vol 13413. Springer, Cham, pp 892-907. https://doi.org/10.1007/978-3-031-12053-4_65
44. Zhao ZX, Jin YM, Heng PA (2022) TraSeTR: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In: Proceedings of the 2022 international conference on robotics and automation, IEEE, Philadelphia, 23-27 May 2022. https://doi.org/10.1109/ICRA46639.2022.9811873
45. Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D et al (2019) Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv: 1902.03368
46. Valanarasu JMJ, Sindagi VA, Hacihaliloglu I, Patel VM (2020) KiU-Net: towards accurate segmentation of biomedical images using over-complete representations. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK et al (eds) Medical image computing and computer-assisted intervention. 23rd international conference, Lima, October 2020. Lecture notes in computer science (Image processing, computer vision, pattern recognition, and graphics), vol 12264. Springer, Cham, pp 363-373. https://doi.org/10.1007/978-3-030-59719-1_36
47. Caicedo JC, Goodman A, Karhohs KW, Cimini BA, Ackerman J, Haghighi M et al (2019) Nucleus segmentation across imaging experiments: the 2018 data science bowl. Nat Methods 16(12):1247-1253. https://doi.org/10.1038/s41592-019-0612-7
48. Mathai TS, Lee S, Elton DC, Shen TC, Peng YF, Lu ZY et al (2022) Lymph node detection in T2 MRI with transformers. In: Proceedings of the SPIE 12033, Medical imaging 2022: computer-aided diagnosis, SPIE, San Diego, 20 February-28 March 2022. https://doi.org/10.1117/12.2613273
49. Shen ZQ, Fu RD, Lin CN, Zheng SH (2021) COTR: convolution in transformer network for end to end polyp detection. In: Proceedings of the 7th international conference on computer and communications, IEEE, Chengdu, 10-13 December 2021. https://doi.org/10.1109/ICCC54389.2021.9674267
50. Li H, Chen L, Han H, Zhou SK (2022) SATr: slice attention with transformer for universal lesion detection. In: Wang LW, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer assisted intervention. 25th international conference, Singapore, September 2022. Lecture notes in computer science, vol 13433. Springer, Cham, pp 163-174. https://doi.org/10.1007/978-3-031-16437-8_16
51. Niu C, Wang G (2022) Unsupervised contrastive learning based transformer for lung nodule detection. Phys Med Biol 67(20):204001. https://doi.org/10.1088/1361-6560/ac92ba
52. Shang FX, Wang SQ, Wang XR, Yang YH (2022) An effective transformer-based solution for RSNA intracranial hemorrhage detection competition. arXiv preprint arXiv: 2205.07556
53. Dai Y, Gao YF, Liu FY (2021) TransMed: transformers advance multi-modal medical image classification. Diagnostics 11(8):1384. https://doi.org/10.3390/diagnostics11081384
54. Zhou M, Mo SL (2021) Shoulder implant X-ray manufacturer classification: exploring with vision transformer. arXiv preprint arXiv: 2104.07667
55. Chen HY, Li C, Wang G, Li XY, Rahaman M, Sun HZ et al (2022) GasHis-transformer: a multi-scale visual transformer approach for gastric histopathological image detection. Pattern Recognit 130:108827.https://doi.org/10.1016/j.patcog.2022.108827
56. Liu WL, Li C, Rahaman MM, Jiang T, Sun HZ, Wu XC et al (2022) Is the aspect ratio of cells important in deep learning? A robust comparison of deep learning methods for multi-scale cytopathology cell image classification: from convolutional neural networks to visual transformers. Comput Biol Med 141:105026.https://doi.org/10.1016/j.compbiomed.2021.105026
57. Lyu Q, Namjoshi SV, McTyre E, Topaloglu U, Barcus R, Chan MD et al (2022) A transformer-based deep-learning approach for classifying brain metastases into primary organ sites using clinical whole-brain MRI images. Patterns 3(11):100613.https://doi.org/10.1016/j.patter.2022.100613
58. Stegmüller T, Bozorgtabar B, Spahr A, Thiran JP (2023) ScoreNet: learning non-uniform attention and augmentation for transformer-based histopathological image classification. In: Proceedings of the 2023 IEEE/CVF winter conference on applications of computer vision, IEEE, Waikoloa, 2-7 January 2023. https://doi.org/10.1109/WACV56688.2023.00611
59. Bhattacharya M, Jain S, Prasanna P (2022) RadioTransformer: a cascaded global-focal transformer for visual attention-guided disease classification. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T (eds) Computer vision. 17th European conference, Tel Aviv, October 2022. Lecture notes in computer science, vol 13681. Springer, Cham, pp 679-698. https://doi.org/10.1007/978-3-031-19803-8_40
60. Zhang F, Xue TF, Cai WD, Rathi Y, Westin CF, O'Donnell LJ (2022) TractoFormer: a novel fiber-level whole brain tractography analysis framework using spectral embedding and vision transformers. In: Wang LW, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer assisted intervention. 25th international conference, Singapore, September 2022. Lecture notes in computer science, vol 13431. Springer, Cham, pp 196-206. https://doi.org/10.1007/978-3-031-16431-6_19
61. Bertolini F, Spallanzani A, Fontana A, Depenni R, Luppi G (2015) Brain metastases: an overview. CNS Oncol 4(1):37-46. https://doi.org/10.2217/cns.14.51
62. Zhang JL, Nie YY, Chang J, Zhang JJ (2021) Surgical instruction generation with transformers. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng YF et al (eds) Medical image computing and computer assisted intervention. 24th international conference, Strasbourg, September 2021. Lecture notes in computer science (Image processing, computer vision, pattern recognition, and graphics), vol 12904. Springer, Cham, pp 290-299. https://doi.org/10.1007/978-3-030-87202-1_28
63. Zhang JL, Nie YY, Chang J, Zhang JJ (2022) SIG-Former: monocular surgical instruction generation with transformers. Int J Comput Assisted Radiol Surg 17(12):2203-2210. https://doi.org/10.1007/s11548-022-02718-9

64. Pang JY, Jiang C, Chen YH, Chang JB, Feng M, Wang RZ et al (2022) 3D shuffle-mixer: an efficient context-aware vision learner of transformer-MLP paradigm for dense prediction in medical volume. IEEE Trans Med Imaging. https://doi.org/10.1109/TMI.2022.3191974
65. Reisenbüchler D, Wagner SJ, Boxberg M, Peng TY (2022) Local attention graph-based transformer for multi-target genetic alteration prediction. In: Wang LW, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer assisted intervention. 25th international conference, Singapore, September 2022. Lecture notes in computer science, vol 13432. Springer, Cham, pp 377-386. https://doi.org/10.1007/978-3-031-16434-7_37
66. Płotka S, Grzeszczyk MK, Brawura-Biskupski-Samaha R, Gutaj P, Lipa M, Trzciński T et al (2022) BabyNet: residual transformer module for birth weight prediction on fetal ultrasound video. In: Wang LW, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer-assisted intervention. 25th international conference, Singapore, September 2022. Lecture notes in computer science, vol 13434. Springer, Cham, pp 350-359. https://doi.org/10.1007/978-3-031-16440-8_34
67. Nguyen HH, Saarakkala S, Blaschko MB, Tiulpin A (2021) CLIMAT: clinically-inspired multi-agent transformers for knee osteoarthritis trajectory forecasting. arXiv preprint arXiv: 2104.03642. https://doi.org/10.1109/ISBI52829.2022.9761545
68. Xie YT, Li QZ (2022) A review of deep learning methods for compressed sensing image reconstruction and its medical applications. Electronics 11(4):586. https://doi.org/10.3390/electronics11040586
69. Korkmaz Y, Dar SUH, Yurt M, Özbey M, Çukur T (2022) Unsupervised MRI reconstruction via zero-shot learned adversarial transformers. IEEE Trans Med Imaging 41(7):1747-1763. https://doi.org/10.1109/TMI.2022.3147426
70. Huang W, Hand P, Heckel R, Voroninski V (2021) A provably convergent scheme for compressive sensing under random generative priors. J Fourier Anal Appl 27(2):19. https://doi.org/10.1007/s00041-021-09830-5
71. Haldar JP, Zhuo JW (2016) P-LORAKS: low-rank modeling of local k-space neighborhoods with parallel imaging data. Magn Reson Med 75(4):1499-1514. https://doi.org/10.1002/mrm.25717
72. Haldar JP (2015) Low-rank modeling of local k-space neighborhoods: from phase and support constraints to structured sparsity. In: Proceedings of the SPIE Optical Engineering + Applications, SPIE, San Diego, 2 September 2015. https://doi.org/10.1117/12.2186705
73. Dar SUH, Yurt M, Shahdloo M, Ildız ME, Tınaz B, Çukur T (2020) Prior-guided image reconstruction for accelerated multi-contrast MRI via generative adversarial networks. IEEE J Sel Top Signal Process 14(6):1072-1087. https://doi.org/10.1109/JSTSP.2020.3001737
74. Yaman B, Hosseini SAH, Moeller S, Ellermann J, Uğurbil K, Akçakaya M (2020) Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. Magn Reson Med 84(6):3172-3191. https://doi.org/10.1002/mrm.28378
75. Narnhofer D, Hammernik K, Knoll F, Pock T (2019) Inverse GANs for accelerated MRI reconstruction. In: Proceedings of the SPIE 11138, wavelets and sparsity XVIII, SPIE, San Diego, 11-15 August 2019. https://doi.org/10.1117/12.2527753
76. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, IEEE, Seattle, 13-19 June 2020. https://doi.org/10.1109/CVPR42600.2020.00813
77. Feng CM, Yan YL, Fu HZ, Chen L, Xu Y (2021) Task transformer network for joint MRI reconstruction and super-resolution. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng YF et al (eds) Medical image computing and computer-assisted intervention. 24th international conference, Strasbourg, September 2021. Lecture notes in computer science, (Image processing, computer vision, pattern recognition, and graphics), vol. 12906. Springer, Cham, pp 307-317. https://doi.org/10.1007/978-3-030-87231-1_30
78. Guo PF, Mei YQ, Zhou JY, Jiang SS, Patel VM (2022) ReconFormer: accelerated MRI reconstruction using recurrent transformer. arXiv preprint arXiv: 2201.09376
79. Huang JH, Wu YZ, Wu HJ, Yang G (2022) Fast MRI reconstruction: how powerful transformers are? In: Proceedings of the 44th annual international conference of the IEEE engineering in medicine & biology society, IEEE, Glasgow, 11-15 July 2022. https://doi.org/10.1109/EMBC48229.2022.9871475
80. Long YH, Li ZS, Yee CH, Ng CF, Taylor RH, Unberath M et al (2021) E-DSSR: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng YF et al (eds) Medical image computing and computer assisted intervention. 24th international conference, Strasbourg, September, 2021. Lecture notes in computer science, (Image processing, computer vision, pattern recognition, and graphics), vol 12904. Springer, Cham, pp 415-425. https://doi.org/10.1007/978-3-030-87202-1_40
81. Wang C, Shang K, Zhang HM, Li Q, Hui Y, Zhou SK (2021) DuDoTrans: dual-domain transformer provides more attention for sinogram restoration in sparse-view CT reconstruction. arXiv preprint arXiv: 2111.10790
82. Pan JY, Zhang HY, Wu WF, Gao ZF, Wu WW (2022) Multi-domain integrative Swin transformer network for sparse-view tomographic reconstruction. Patterns 3(6):100498.https://doi.org/10.1016/j.patter.2022.100498
83. Razi T, Niknami M, Ghazani FA (2014) Relationship between Hounsfield unit in CT scan and gray scale in CBCT. J Dent Res Dent Clin Dent Prospects 8(2):107-110
84. Duda SN, Kennedy N, Conway D, Cheng AC, Nguyen V, Zayas-Cabán T et al (2022) HL7 FHIR-based tools and initiatives to support clinical research: a scoping review. J Am Med Inf Assoc 29(9):1642-1653. https://doi.org/10.1093/jamia/ocac105
85. Auer F, Abdykalykova Z, Müller D, Kramer F (2022) Adaptation of HL7 FHIR for the Exchange of Patients' Gene Expression Profiles. Stud Health Technol Inform 295:332-335. https://doi.org/10.1101/2022.02.11.22270850
86. Carter C, Veale B (2022) Digital radiography and PACS, 4th edn. Elsevier, Amsterdam
87. Twa MD, Johnson CA (2022) Digital imaging and communication standards. Optom Vis Sci 99(5):423. https://doi.org/10.1097/OPX.0000000000001909
88. Xiong YX, Du B, Yan PK (2019) Reinforced transformer for medical image captioning. In: Suk HI, Liu M, Yan P, Lian C (eds) Machine learning in medical imaging. 10th international workshop, Shenzhen, October 2019. Lecture notes in computer science (Image processing, computer vision, pattern recognition, and graphics), vol 11861. Springer, Cham, pp 673-680. https://doi.org/10.1007/978-3-030-32692-0_77
89. Miura Y, Zhang YH, Tsai E, Langlotz C, Jurafsky D (2021) Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, Association for Computational Linguistics, Online, 6-11 June 2021. https://doi.org/10.18653/v1/2021.naacl-main.416
90. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V (2017) Self-critical sequence training for image captioning. In: Proceedings of the 2017 IEEE conference on computer vision and pattern recognition, IEEE, Honolulu, 21-26 July 2017. https://doi.org/10.1109/CVPR.2017.131
91. You D, Liu FL, Ge S, Xie XX, Zhang J, Wu X (2021) AlignTransformer: hierarchical alignment of visual regions and disease tags for medical report generation. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng YF et al (eds) Medical image computing and computer assisted intervention. 24th international conference, Strasbourg, September 2021. Lecture notes in computer science, (Image processing, computer vision, pattern recognition, and graphics), vol 12903. Springer, Cham, pp 72-82. https://doi.org/10.1007/978-3-030-87199-4_7
92. Xu MY, Islam M, Lim CM, Ren HL (2021) Learning domain adaptation with model calibration for surgical report generation in robotic surgery. In: Proceedings of the 2021 IEEE international conference on robotics and automation, IEEE, Xi'an, 30 May-5 June 2021. https://doi.org/10.1109/ICRA48506.2021.9561569

93. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS (2019) Adversarial attacks on medical machine learning. Science 363(6433):1287-1289. https://doi.org/10.1126/science.aaw4399
94. Papangelou K, Sechidis K, Weatherall J, Brown G (2019) Toward an understanding of adversarial examples in clinical trials. In: Berlingerio M, Bonchi F, Gärtner T, Hurley N, Ifrim G (eds) Machine learning and knowledge discovery in databases. European conference, Dublin, September 2018. Lecture notes in computer science (Lecture notes in artificial intelligence), vol 11051. Springer, Cham, pp 35-51. https://doi.org/10.1007/978-3-030-10925-7_3
95. Benz P, Ham S, Zhang CN, Karjauv A, Kweon IS (2021) Adversarial robustness comparison of vision transformer and MLP-mixer to CNNs. In: Proceedings of the 32nd british machine vision conference 2021, BMVA Press, Online, 22-25 November 2021
96. Chuman T, Kiya H (2022) Security evaluation of block-based image encryption for vision transformer against jigsaw puzzle solver attack. In: Proceedings of the 4th global conference on life sciences and technologies (LifeTech), IEEE, Osaka, 7-9 March 2022. https://doi.org/10.1109/LifeTech53646.2022.9754937
97. Li M, Han DZ, Li D, Liu H, Chang CC (2022) MFVT: an anomaly traffic detection method merging feature fusion network and vision transformer architecture. EURASIP J Wirel Commun Netw 2022(1):39. https://doi.org/10.1186/s13638-022-02103-9
98. Ho CMK, Yow KC, Zhu ZW, Aravamuthan S (2022) Network intrusion detection via flow-to-image conversion and vision transformer classification. IEEE Access 10:97780-97793. https://doi.org/10.1109/ACCESS.2022.3200034
99. George A, Marcel S (2021) On the effectiveness of vision transformers for zero-shot face anti-spoofing. In: Proceedings of the 2021 IEEE international joint conference on biometrics, IEEE, Shenzhen, 4-7 August 2021. https://doi.org/10.1109/IJCB52358.2021.9484333
100. Doan KD, Lao YJ, Yang P, Li P (2022) Defending backdoor attacks on vision transformer via patch processing. arXiv preprint arXiv: 2206.12381
101. Riquelme C, Puigcerver J, Mustafa B, Neumann M, Jenatton R, Susano Pinto A et al (2021) Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems 34: 8583-8595
102. Ridnik T, Ben-Baruch E, Noy A, Zelnik-Manor L (2021) ImageNet-21K pretraining for the masses. arXiv preprint arXiv: 2104.10972
103. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE conference on computer vision and pattern recognition, IEEE, Miami, 20-25 June 2009. https://doi.org/10.1109/CVPR.2009.5206848
104. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma SA et al (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211-252. https://doi.org/10.1007/s11263-015-0816-y
105. Chen XN, Hsieh CJ, Gong BQ (2022) When vision transformers outperform ResNets without pre-training or strong data augmentations. In: Proceedings of the 10th international conference on learning representations, OpenReview.net, 25-29 April 2022
106. Gani H, Naseer M, Yaqub M (2022) How to train vision transformer on small-scale datasets? arXiv preprint arXiv: 2210.07240
107. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th international conference on machine learning, PMLR, Online, 13-18 July 2020
108. Wang XY, Yang S, Zhang J, Wang MH, Zhang J, Yang W et al (2022) Transformer-based unsupervised contrastive learning for histopathological image classification. Med Image Anal 81:102559.https://doi.org/10.1016/j.media.2022.102559
109. Meng CZ, Trinh L, Xu N, Liu Y (2021) MIMIC-IF: interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. https://doi.org/10.21203/rs.3.rs-402058/v1
110. Lu JH, Zhang XS, Zhao TL, He XY, Cheng J (2022) APRIL: finding the Achilles' heel on privacy for vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, IEEE, New Orleans, 18-24 June 2022. https://doi.org/10.1109/CVPR52688.2022.00981
111. Song WP, Shi CC, Xiao ZP, Duan ZJ, Xu YW, Zhang M et al (2019) AutoInt: automatic feature interaction learning via self-attentive neural networks. In: Proceedings of the 28th ACM international conference on information and knowledge management, ACM, Beijing, 3-7 November 2019. https://doi.org/10.1145/3357384.3357925
112. Yu K, Zhang MD, Cui TY, Hauskrecht M (2019) Monitoring ICU mortality risk with a long short-term memory recurrent neural network. In: Proceedings of the pacific symposium on Biocomputing 2020, World Scientific, Kohala Coast, 3-7 January 2020. https://doi.org/10.1142/9789811215636_0010
113. Bai SJ, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv: 1803.01271
114. Guo T, Lin T, Antulov-Fantulin N (2019) Exploring interpretable LSTM neural networks over multi-variable data. In: Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, 9-15 June 2019

Publisher's Note