Review

Developments in Deep Learning Artificial Neural Network Techniques for Medical Image Analysis and Interpretation

Olamilekan Shobayo [1,2,*] and **Reza Saatchi** [1]

1   School of Engineering and Built Environment, Sheffield Hallam University, Pond Street, Sheffield S1 1WB, UK; r.saatchi@shu.ac.uk
2   School of Computing and Digital Technologies, Sheffield Hallam University, 151 Arundel Street, Sheffield S1 2NU, UK
*   Correspondence: o.shobayo@shu.ac.uk

Abstract: Deep learning has revolutionised medical image analysis, offering the possibility of automated, efficient, and highly accurate diagnostic solutions. This article explores recent developments in deep learning techniques applied to medical imaging, including convolutional neural networks (CNNs) for classification and segmentation, recurrent neural networks (RNNs) for temporal analysis, autoencoders for feature extraction, and generative adversarial networks (GANs) for image synthesis and augmentation. Additionally, U-Net models for segmentation, vision transformers (ViTs) for global feature extraction, and hybrid models integrating multiple architectures are explored. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) process were used, and searches on PubMed, Google Scholar, and Scopus databases were conducted. The findings highlight key challenges such as data availability, interpretability, overfitting, and computational requirements. While deep learning has demonstrated significant potential in enhancing diagnostic accuracy across multiple medical imaging modalities—including MRI, CT, US, and X-ray—factors such as model trust, data privacy, and ethical considerations remain ongoing concerns. The study underscores the importance of integrating multimodal data, improving computational efficiency, and advancing explainability to facilitate broader clinical adoption. Future research directions emphasize optimising deep learning models for real-time applications, enhancing interpretability, and integrating deep learning with existing healthcare frameworks for improved patient outcomes.

Keywords: artificial intelligence; artificial neural networks; medical image analysis; deep learning; image classification and pattern recognition

1. Introduction

The use of imaging for diagnosis in healthcare is substantial, amounting to about USD 100 billion globally per year [1]. Mounting pressures on healthcare facilities and the market for imaging diagnosis have led to increasing demands for diagnostic excellence in the clinical setting due to the rising number of clinical images, greater image complexity, and faster results demanded by clinicians. As a result, the need for new technologies is centred on providing solutions that will increase the effectiveness of the clinical process, improving healthcare systems, and provide accurate diagnosis for patients while improving care quality. Therefore, there has been high demand for technologies that can aid in the automation of workflows associated with the use of medical imaging for diagnosis, leading to advances in the use of artificial intelligence (AI) methods such as deep learning to assist radiologists in analysing complex image datasets [2].

1.1. Deep Learning Overview

Deep learning is a subfield of machine learning that leverages artificial neural network (ANN) architecture to acquire knowledge from large datasets and perform intricate operations. One of the main advantages of deep learning techniques is their ability to mimic the information processing complexity of the human brain [2,3]. The field has been studied since the 1980s, but gained prominence in recent years. This is because of access to large datasets for model training, improved algorithm development, and increased processing power of microprocessors. The structure of an ANN is an interconnection of nodes (sometimes referred to as processing elements or neurons) that can span up to multiple layers, depending on the intricacy of the tasks and capabilities of hardware resources. Each node gathers information from the previous layer and then transmits to the subsequent layer based on the configured characteristics and set parameters. The values of the parameters used are normally initially randomised, but then iteratively updated during training based on a set learning rate [4]. A deep learning network extracts deeper and intricate information as its number of architectural layers increases, resulting in optimised performance with large datasets and training iterations. This in turn ensures precise recognition of patterns in the data [3].

1.2. Deep Learning in Medical Imaging, Classification, and Segmentation

Deep learning has seen extensive applications in medical image analysis. It has been used for different medical imaging modalities, including X-ray radiographs, computed tomography (CT), ultrasound (US), and magnetic resonance imaging (MRI) scans, to provide predictive diagnosis and treatment. Subtle and intricate patterns presented by these medical images are effectively identified by the adapted deep learning approach, thereby providing a means to automate the feature extraction process. Image feature extraction or selection is a process that can be performed manually by a qualified specialist, but this can be time-consuming and subjective. When deep learning models are properly trained, they have an ability to accurately identify lesion or tumours, examine membranes and tissues for differences, and undertake diverse medicine-related tasks, thereby providing accelerated diagnostic outcomes. Therefore, deep learning is emerging in the realm of medical image analysis assisting with diagnosis. A summary of deep learning applications is shown in Figure 1.
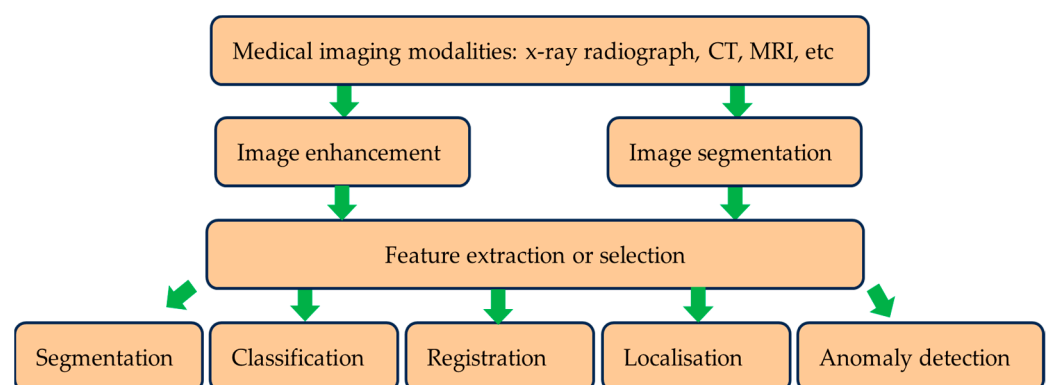
Figure 1. A summary of deep learning usage with medical images.

1.3. Challenges in Utilising Deep Learning in the Medical Field

Despite the significant advancements and capabilities of deep learning techniques in medical image analysis, there are certain limitations and challenges in their implementation and acceptance. For instance, most deep learning algorithms often lack explainability, i.e., they typically operate as black boxes [5]. In the medical field, where decision-making

processes are required for diagnosis and provision of treatment, the deep learning method of convolutional neural networks (CNNs), for instance, does not provide an insight into the manner by which it came up with a decision. Other challenges that applications of deep learning techniques can face include the following

- Overfitting during training. Overfitting causes poor accuracy when training deep artificial neural networks in recognising images not included during their training dataset (i.e., unseen or test images), even though the images in the training set could have been correctly recognised [6–10]. It has several possible causes that include insufficient training examples to facilitate adequate generalisation and excessive model parameters for the architecture. However, there are techniques that could be valuable to deal with overfitting. One technique is known as dropout [6,11–14], whereby some nodes in the architecture are temporarily left out during training. Another approach is to artificially extend the number of examples in the training dataset through a process known as data augmentation [15–17]. A comparison of different image data augmentation methods was reported in [18]. The technique has been applied to mammograms [19] and CT images [20]. Augmentation can be performed by manipulating the images through processes such as kernel filters, geometric transformation, random erasing, and mixing images. It can also be carried out through deep learning approaches such as adversarial training, neural style transformers, and generative adversarial networks [7]. An issue with performing data augmentation is selecting the best approach for a given set of images [21]. An exploration of the influence of different data augmentation techniques on the explainability of deep learning methods was reported in [22].
- Image annotation. Many deep learning algorithms are supervised, i.e., they require labelled images indicating their categories during their training phase. The labelling requires annotation of images by qualified medical practitioners. Because deep learning requires large datasets for training, this process can be time-consuming. However, there have been reports of automated and interactive image annotation methods that can assist with this operation, e.g., [23–28].
- Noisy images. Medical images can be noisy. Noise distorts the quality of images being used to train deep learning networks [29] and can reduce their ability to learn effectively.
- Interpretability. This is an area of great research interest to render decision-making by deep learning artificial neural networks more transparent, i.e., moving away from so-called black box behaviour to more interpretable decision-making. The issue of interpretability has been explored in many studies, e.g., [30–36]. Interpretability can be considered from multiple perspectives, e.g., user orientation for explanations provided, visualisation through graphs, charts rules etc, user comprehensibility through comprehensive reasoning, simplicity of explanations, local interpretation of a single datum and global interpretation of overall data, consistency in explanations, transparency in decision-making, and ethics and fairness in revealing bias and discrimination [31].
- Data sharing complexities and small datasets. Medical data gathered by a single institution may be insufficient to allow effective training of deep learning algorithms, and thus sharing across many institutions would be required. This, however, could be challenging due to regulatory, technical, and privacy concerns [37] and financial and time constraints, and limited availability of patients can limit the dataset. However, valuable techniques have been devised to address small-dataset problems in deep learning [38,39].

- Ethical issues. There are numerous ethical issues requiring careful consideration, e.g., patient anonymity, consent/assent, data handling and protection, data sharing, and vulnerable participants [40–43].
- Trust. There is an ongoing issue of relying on critical medical diagnostic results generated when the manner of their generation is not sufficiently transparent [4].
- Computational requirements and environmental issues. Training deep learning algorithms typically requires high computational capabilities and long durations. Many general-purpose computers do not have the means of delivering the required computational resources, and there is also the issue of the environmental aspects of using so much electricity to perform the required deep learning training [44,45].

This article explores deep learning techniques for medical image analysis, highlighting their roles in assisting diagnostic tasks. It examines CNNs for classification and segmentation, recurrent neural networks (RNNs) for sequential data, autoencoders for feature extraction, and generative adversarial networks (GANs) for data augmentation. It also discusses U-Net models for segmentation, vision transformers (ViTs) for long-range dependencies, and hybrid models that integrate multiple architectures. Different algorithms and their architectures are discussed. Associated mathematical models are presented to demonstrate the manner in which different imaging modalities operate. The study emphasises the transformative impact of deep learning in medical diagnostics and suggests future improvements in efficiency and interpretability. In the following sections, the materials and methods are explained and the results are discussed.

2. Materials and Methods

A systematic review was undertaken to explore the latest developments in deep learning ANNs for analysing medical images. As this field is growing, with many related articles published daily, a rapid review approach was undertaken [6]. The necessary constraints and criteria for inclusion were used when searching for literature that provided information about recent deep learning techniques used for medical image analysis. The following keywords were combined as part of the literature search: "deep learning" and "medical image analysis". We also included imaging modalities and used "OR" and "AND" operators:

"deep learning" AND "medical image analysis".
"deep learning" AND "medical image analysis" OR "CT".
"deep learning" AND "medical image analysis" OR "MRI".
"deep learning" AND "medical image analysis" OR "X-ray".
"deep learning" AND "medical image analysis" OR "infrared thermal image".
"deep learning" AND "medical image analysis" OR "ultrasound".

The search was performed on different scientific databases, including Scopus, PubMed, and Google Scholar, and the inclusion and exclusion criteria are shown in Figure 2. The imaging modalities are shown in Figure 3.

We examined each article with a principal focus on the deep learning techniques deployed and the imaging modalities adopted. Articles that did not report either a deep learning algorithm or a medical imaging modality were excluded. The remaining articles were then filtered to include the deep learning techniques in whole or in part, as many works used a combination of deep learning techniques. As this is a wider review of deep learning methodologies, we have not looked critically into the clinical aspects of diseases to be diagnosed. The categories included were CNNs, RNNs, autoencoders, GANs, transfer learning (TL), ViTs, and hybrid models. We also considered studies that combined multiple deep learning algorithms, i.e., hybrid systems further refined our search to include the terms "hybrid". For completeness, a summary of each method is included as part of the review.

As CNN architectures were widely deployed, mostly with other deep learning techniques, we further narrowed the search terms to include only publications that specifically used CNNs by updating the search terms to include ("convolutional neural network" AND "medical image analysis") OR ("CNN" AND "medical image analysis").
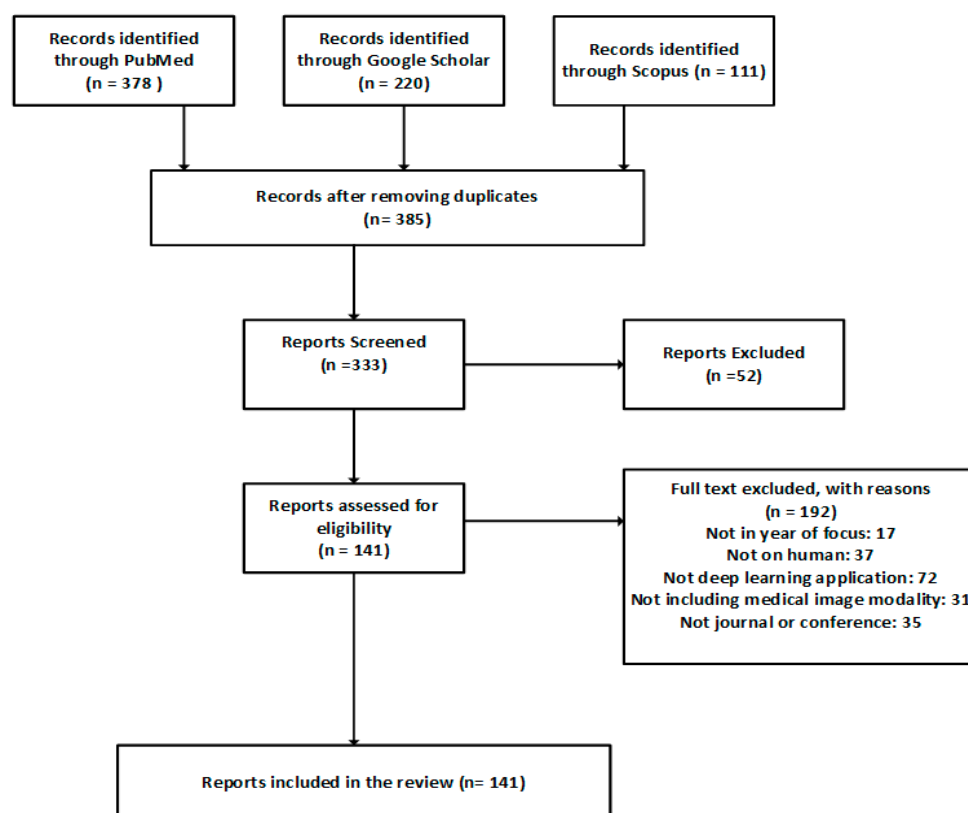

Figure 2. Inclusion and exclusion criteria of the systematic literature review using the PRISMA framework.
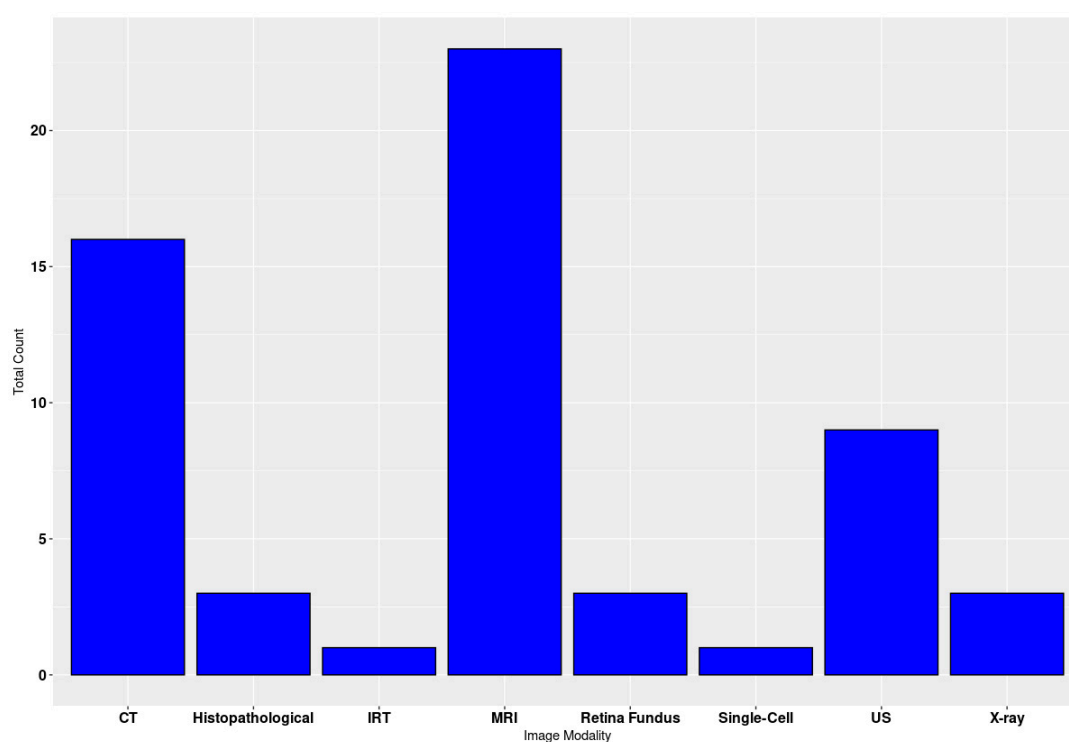

Figure 3. Count of imaging modalities on a section of articles included in the review.

Over 500 different studies were initially identified across the databases, with some duplicated across the databases, especially on PubMed and Google Scholar. After applying the exclusion criteria, this was refined to 141 articles being eventually reviewed in this study. The 141 articles were distributed over all the identified categories. We included a broad range of imaging modalities. The objective was to identify the deep learning techniques most suitable for medical imaging tasks, the processes employed to improve the monitoring and diagnosis performance of these techniques, and the optimisation performed.

3. Results

This section contains an overview of deep learning neural network concepts.

3.1. Convolutional Neural Networks

CNNs were the dominant technique for analysing medical images for diagnostic purposes [46]. It has gained immense popularity since 2012, when high-performance computing (HPC) became more accessible. This led to the ImageNet competition for different combinations of deep CNNs to achieve better diagnostic results when compared to human experts. CNNs are effective for several tasks, such as image segmentation, detection, registration, localisation, and classification [5]. They consist of numerous layers of convolutional filters with nonlinear activation functions combined with pooling layers, dropout layers, and fully connected layers. Their ability to extract complex spatial relationships and patterns in images has seen them used in various medical imaging modalities, such as magnetic resonance imaging (MRI), computed tomography (CT), X-rays, ultrasound (US), and histopathology, and more recently in infrared thermal (IRT) imaging [47,48]. Images associated with multiple conditions, e.g., bone fractures, cancer, liver diseases, pneumonia, and COVID-19, were segmented, classified, registered, and interpreted. The CNN architecture applied to the discrete Fourier-transformed infrared thermal images is shown in Figure 4. Although the study was a pilot, the model was effective in screening wrist fracture in children.
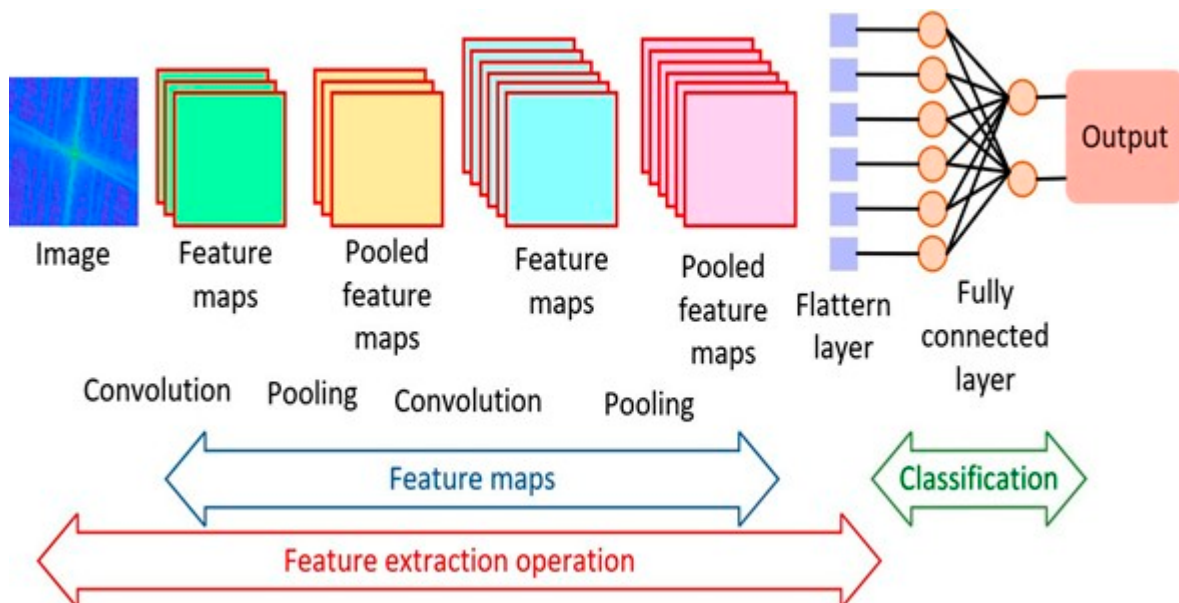

Figure 4. An example of a CNN used for the analysis of infrared thermal images [48].

The concept of convolution is simply a mathematical operation where a filter, also known as a kernel, is applied on an input image for feature extraction. The convolution

process is described mathematically next. Let the image input to the CNN be represented by a 3D matrix such that:

$$X \in \mathbb{R}^{H \times W \times D} \tag{1}$$

where *H* is the height of the image, *W* is the width of the image, and *D* is the depth of the image (representing the number of channels, which is usually three for an RGB image). When applied to a set of convolutional filters for each input, the output of the convolutional filter can be represented by:

$$Z_{i,j,k} = \sum_{h=0}^{H_k} \sum_{w=0}^{W_k} \sum_{d=0}^{D} X_{i+h,j+w,d} \cdot K_{h,w,d,k} \tag{2}$$

where $Z_{i,j,k}$ is the output feature map for the filter *k* at position (i, j), *K* is the convolutional filter (kernel), and H_k and W_k are the height and width of the kernel, which are usually a sample of the size of the input. This process helps to identify local patterns such as the edges, shapes, and textures from the input image.

When the CNN algorithm is used for classification purposes, it is combined with other layers such as the activation layers, pooling layer, and the fully connected layer. The output from the convolutional filter is passed through an activation function $A_{i,j,k}$. If the rectified linear unit (ReLU) activation function is applied, the output from the convolutional layer becomes:

$$A_{i,j,k} = max\left(0,\ Z_{i,j,k}\right) \tag{3}$$

This operation helps to add nonlinearity by ensuring only positive values are retained, thereby helping to learn complex patterns from the image.

The pooling layer is then applied to the output of the activation layers. The effect of the pooling operation is to reduce spatial dimensions from the convolution operation and thus help the network to capture small translations in the image. This process also helps to reduce the computational complexity of the network. Assuming the max (maximum) pooling function is applied to the activation layer with a $p \times p$ window size, the output from the pooling layer is represented by:

$$A_{i,j,k} = \max_{h,w \in [1,p]}\ A_{i+h,j+w,k} \tag{4}$$

The fully connected layer is connected to the output of a flattened max pooling layer output. This layer takes a vector value and is a standard neural network, with each input connected to all the neurons in the next layer. The output of the fully connected (FC) layer is given by:

$$Z_l = W_l A_{l-1} + b_l \tag{5}$$

where A_{l-1} is the input into the FC layer and W_l and b_l represent the weights and biases of the FC layer. The final or output layer, which is also a vector, is mostly passed through a softmax function for a classification task. This function helps to convert the output classes into probabilities with the following expression:

$$\hat{y}_i = \frac{exp\ (Z_i)}{\sum_{j=1}^{C} exp\ (Z_j)} \tag{6}$$

where $\hat{y}_i$ represents the probability of predicting class *i*, and *C* represents the total number of classes being differentiated. The full CNN undergoes training to correctly learn the features of the image input. This training uses a loss function, usually a cross-entropy loss, when considering a classification task. The cross-entropy loss is given by:

$$L = -\sum_{i=1}^{C} y_i log\ (\hat{y}_i) \tag{7}$$

where $\hat{y}_i$ and y_i are the predicted probabilities for class i and the target label, respectively.

The gradient of the loss function is calculated with respect to the weights through a back-propagation formula:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial Z}.\frac{\partial Z}{\partial W} \tag{8}$$

The weights are updated using the gradient descent, i.e.,

$$W_{new} = W_{old} - \eta\frac{\partial L}{\partial W} \tag{9}$$

where η $(0 < \eta \leq 1)$ represents the learning rate controlling the convergence rate.

Literature Review Findings for CNNs

Most studies have used a CNN individually, though there have been studies that applied a CNN in combination with another algorithm such as UNET and GANs for medical image diagnostics. However, in this section, the focus is on a CNN application on its own (other CCN-based algorithms, such as transfer learning, are discussed in Section 3.6). A lightweight CNN algorithm was used to detect COVID-19 from chest X-ray radiograph images [49]. In that study, their proposed CNN was inspired by the ResNet model [50], in which no layers were connected in sequence, creating a skip connection whereby neurons in a particular layer can be connected to another neuron further ahead. This arrangement helped to create a lightweight model that was effective in edge detection applications. Their study was compared with other CNN models such as CVDNet and deep GRU-CNN, and showed similar performance to models with reduced computational complexity. A deeper CNN in their work was proposed that used CT scans and X-ray radiographs for the detection of several pulmonary diseases, including COVID-19 and viral pneumonia [51]. They used varied datasets consisting of different imaging modalities for their work, which provided another dimension to the efficacy of CNNs in diagnostic imaging. They proposed a 26-layer deep CNN which was inspired by a wide residual network (WRN) [52]. It provided a faster training time when considering the deep nature of the architecture. This was achievable based on the sophistication of their hardware. Their model was effective in terms of the accuracy in detecting the different pulmonary diseases when compared to traditional methods. A CNN model was used to classify ultrasonic images of fatty liver [53]. A pertinent problem in that research was the similarity in the pathological ultrasonic images used for training the CNN algorithm. This sort of challenge might pose a problem for the CNN architecture, as it may struggle to extract distinct features for different pathological images. Therefore, there will be a need for deep convolutional layers, leading to computational complexities. However, they used pixel-level feature extraction as a preprocessing step and then proposed a CNN architecture comprising two convolutional layers, a pooling layer and a fully connected layer. They also tested the proposed method with a skip connection and improved its accuracy when compared to other algorithms such as VGGNet. The authors in [54] developed a CNN model to distinguish between benign and malignant tumours using radial endobronchial ultrasound (rEBUS) images. A total of 769 images were collected from hospitals in Taiwan for model training and internal validation, with an additional 300 and 92 images from two other hospitals used for external validation. The model was trained with image augmentation techniques to improve its robustness. Internal validation showed strong performance, with an AUC of 0.88, sensitivity of 0.80, and specificity of 0.75. For external validation, the AUC was 0.76 and 0.72 at the two external sites, with varying sensitivity and specificity.

After fine-tuning, performance improved, reaching AUC values of 0.78 and 0.82. The CNN also showed potential in identifying lung cancer subtypes, with moderate success for adenocarcinoma and squamous cell carcinoma, but limited accuracy for small cell lung cancer. Overall, the model proved promising for aiding rEBUS-based lung cancer diagnosis. The authors in [55] developed a deep convolutional neural network (DCNN) for the classification of ovarian cancer in the pelvic area using ultrasound images. They used data collected from ten hospitals in China between 2003 and 2019, including adult patients with adnexal (ovary-related) lesions and healthy controls. Validation of the proposed model was performed using three datasets: one internal and two external, all containing images from patients with either ovarian cancer or benign lesions. The DCNN showed high diagnostic performance, with area under the curve (AUC) scores of 0.911, 0.870, and 0.831 across the three datasets, respectively. They benchmarked their model against the diagnostic results of 35 radiologists, and the proposed DCNN model outperformed the 35 radiologists in accuracy for detecting ovarian cancer.

3.2. Recurrent Neural Networks

Recurrent neural networks (RNNs) are another class of neural network that have gained significant research consideration for modelling of sequential data. Their variability in length of their input and output also makes them suitable for natural language processing tasks [1]. An RNN has an internal memory unlike feed-forward networks, which helps to keep memory of the hidden states across different time stamps. In RNNs, there is a feedback loop between the outputs of the hidden layers [56]. This arrangement allows the RNNs to learn sequential patterns, making them suitable for tasks such as time-series prediction and video analysis [57]. RNN architecture normally suffers from the issue of vanishing gradients [53]. Their use transcends sequential and textual data. They can also be applied to imaging modalities, such as dynamic imaging in functional MRIs, with time-series characteristic information to monitor disease progress in a patient and to check how they respond to treatments. This functionality has been made popular by a variant of the network known as the long short-term memory network (LSTM). The LSTM works by introducing self-loops to allow for the flow of gradients for long durations [58]. The recurrent structure of the neural network is always enforced by the LSTM by introducing gating functions on the neurons in the hidden layer. The architecture of an RNN is shown in Figure 5.
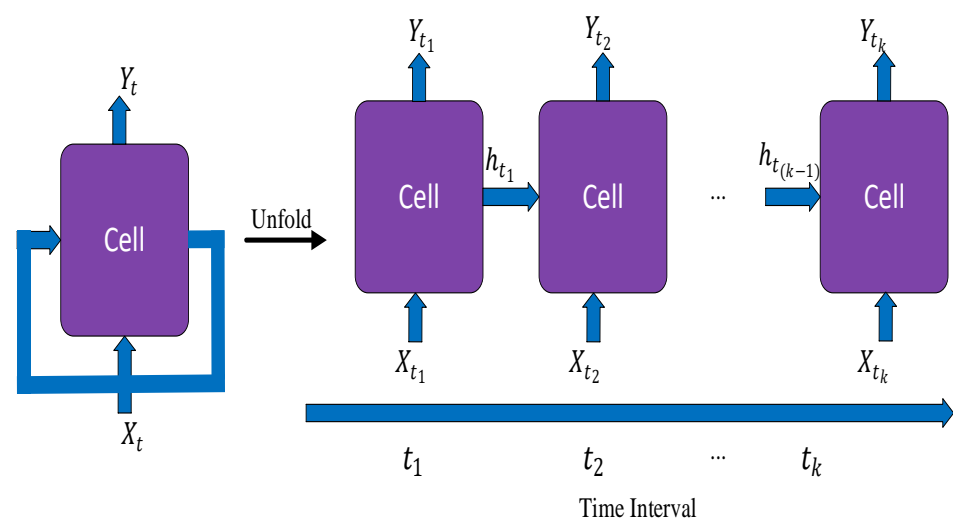

Figure 5. Structure of an RNN [53].

The mathematical representation of an RNN is described next. Let the sequential input to the RNN at time step t be x_t where $x_t \in \mathbb{R}^n$ is a vector, hence the input sequence becomes:

$$x_1, x_2, \ldots, x_T \tag{10}$$

where T represents the number of time steps in the sequence. For every time step t, a hidden state (h_t) is maintained. Each hidden state is updated as:

$$h_t = f(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \tag{11}$$

where W_{hh} represents the hidden states or recurrent weights, W_{xh} is the weight of the connection between the input layer and the hidden layer, and b_h is the bias for the hidden state. The function f represents the activation function, typically the ReLU or the hyperbolic tangent. To obtain the output of the RNN at each time step, the output from each hidden state is multiplied by its associated weights and added to a chosen bias vector. This is represented as:

$$y_t = g\left(W_{hy}h_t + b_y\right) \tag{12}$$

where W_{hy} represents the weight matrix between the hidden state and the output and b_y represents the output bias vector. The function g is the activation function for the output, which could be task-dependent. The softmax activation function is mostly employed for classification tasks.

Literature Review Findings for RNNs

RNNs are widely used for machine learning or deep learning tasks that are time-related. For image diagnostics, they can be used to monitor disease progression and patient response to treatments based on different time stamps of when the images were taken. Most studies have based their classification tasks on an RNN, typically combining it with a CNN for the feature extraction stage, as shown in Table 1.

Table 1. Recurrent neural network applications.

Article	Imaging Modality	Task	CNN Feature Extraction	Disease/Body Part	Variant Used
[58]	MRI	Classification	Y	Alzheimer's	BGRU
[59]	Histopathological images	Classification	N	Breast cancer	None
[60]	MRI	Classification/segmentation	N	Brain tumour	LSTM
[57]	MRI	Segmentation	Y	Aorta	LSTM
[61]	MRI	Classification/localisation	Y	Knee ligament	LSTM
[62]	IRT	Classification	Y	Diabetes mellitus	LSTM
[56]	CT	Image denoising	N	Lungs	LSTM
[63]	MRI	Registration	Y	Brain cancer	LSTM

In one study, MR images were used for classification of Alzheimer's disease by analysing the longitudinal sequence of the MR images taken at different time steps to measure disease progression [58]. They combined a CNN for feature extraction and RNN for classification. For the RNN architecture, they cascaded three bidirectional gated recurrent units (BGRU) with the inputs from the CNN at multiple points, providing a longitudinal analysis. RNN architectures can suffer from vanishing gradients when the sequence of the images becomes too long, as they do not contain memory units to store sequences [64], hence the reason for developing variants such as LSTM and GRU. RNNs can also be used on their own without the addition of gated memory units. An RNN with slight modification was used in another study for the classification of breast cancer with

microscopic histopathological imaging [58]. The RNN architecture was gain-modulated using the honey badger algorithm (HBA) by updating the weights of the RNN during training. The capacity in which RNN was used in their study was for classification of different stages of breast cancer and not disease progression, hence why gated memory approaches were not used. RNNs can also be used for image segmentation tasks, especially in cases where time-series images of a part of the body are taken, requiring sequencing of the time frame for identification of anomalies. In another study, an RNN-based LSTM model was proposed together with U-Net to label aortic MR images [60]. This sort of task uses the capacity of the RNN architecture to segment temporal images, as most annotated images normally used for classification are static. The U-Net algorithm was combined with a CNN for feature extraction, with the sequencing part achieved with the RNN algorithm. The ability of a neural network to visualise several distinct images when compared to a human makes RNNs suitable for segmenting image labels for disease identification. RNNs can also been used to denoise diagnostic images. Images used for diagnosis can be susceptible to such interference as white and salt-and-pepper noise. An LSTM model was combined with particle swarm optimisation (PSO) algorithms to optimise the batch normalisation process of the RNN training to remove noise from CT images of the lungs [59]. The technique provided improved peak signal-to-noise ratio (PSNR) when compared to traditional noise removal techniques such as filtering-based and diffusion-based techniques.

3.3. Autoencoders

Autoencoders are unsupervised learning models used for dimensionality reduction and feature extraction with minimal distortion when their input is compared to their output [65]. They play an important role in the deep learning paradigm for medical image analysis [66]. They can help denoise or compress medical images and are useful in anomaly detection, where unusual patterns in images indicate potential medical issues. They can also be used as a semi-supervised deep learning model to produce annotated data in situations where there is a lack of substantial amount of annotated data available for training any deep learning network for tasks such as classification or segmentation [67]. Their architecture consists of an encoder and decoder structure with a latent space to store the value of the compressed data. Both the encoder and the decoder comprise a fully connected feed-forward neural network. The encoder converts the input image into a low-dimension compressed version, which is referred to as latent space or the encoder. The latent space contains only essential features of the input from the encoder and is kept as shallow as possible in terms of the number of neurons used to retain the compressed version of the input and computational efficiency. The encoder in turn transforms the latent space to a reconstruction of the input. A loss function is generally used during training to compare the input with its reconstruction [68]. The architecture of an autoencoder with an MR image as its input is shown in Figure 6.

The mathematical representation of the dimensionality reduction function of the autoencoder's encoder and decoder is described next. Let the input image to the encoder be represented by:

$$x \, \epsilon \, \mathbb{R}^n \tag{13}$$

The encoder in turn transforms the input x into a lower-dimension latent representation:

$$z \, \epsilon \, \mathbb{R}^d \tag{14}$$

where $d < n$.

The transformation function that produced the latent information is given as:

$$z = f_{enc}(x) = \sigma(W_{enc}x + b_{enc}) \tag{15}$$

where $W_{enc} \in \mathbb{R}^{d \times n}$ is the encoder weight matrix, $b_{enc} \in \mathbb{R}^{d \times n}$ represents the encoder's bias vector, and σ represents the selected activation function of the encoder, which is commonly ReLU or a sigmoid function. The latent representation z of the input x provides compressed information by only capturing the essential features of the input image, a feature referred to as bottleneck, in which the dimension of the latent feature is smaller than the input vector. This facilitates the autoencoder to learn a better way to efficiently represent the input vectors. The decoder network tries to transform the latent space z back to the reconstructed input $\hat{x}$ to match the input vector x. The decoder function can be represented thus:

$$\hat{x} = f_{dec}(z) = \sigma'(W_{dec}z + b_{dec}) \tag{16}$$

The loss function is used to minimize the difference between the input x and the reconstructed input $\hat{x}$. This difference is normally quantified by binary cross-entropy error (BCE) when the dataset is binary or mean squared error (MSE) for multi-class data.
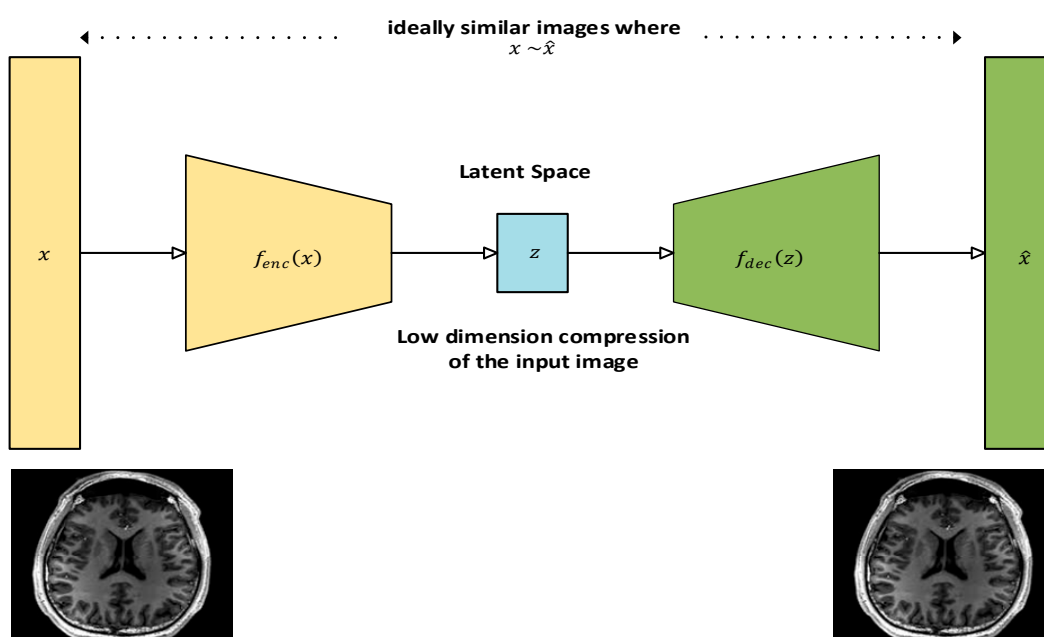

Figure 6. Autoencoder architecture reproducing an MR image.

The MSE loss is computed with the following function:

$$L(x, \hat{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{17}$$

and the BCE loss is computed as follows:

$$L(x, \hat{x}) = -\sum_{i=1}^{n}[x_i log\,(\hat{x}_i) + (1 - x_i)log\,(1 - \hat{x}_i)] \tag{18}$$

Training of the autoencoder requires finding parameters for the weights and biases of the encoder and decoder such that the reconstruction loss is minimised over the training data. Gradient-based optimisation techniques such as stochastic gradient descent (SGD) or Adam can be used for this purpose. Table 2 shows the summary of autoencoder based techniques.

Table 2. Autoencoder-based techniques.

Article	Image Modality	Task	Disease/Body Part
[67]	MRI	Augmentation/segmentation	Brain
[69]	MRI	Denoising	Prostate
[70]	CT + others	Classification	Face
[71]	CT	Augmentation	Various
[72]	MRI and CT	Classification	Intracerebral haemorrhage
[73]	X-ray/digital histopathology	Anomaly detection	Various
[74]	Single-cell images	Classification	Myeloid leukaemia
[75]	None	Anomaly detection	None
[76]	CT	Classification	COVID-19
[77]	MRI	Denoising/classification	Autism/brain

Literature Review Findings for Autoencoders

Autoencoders are associated with both unsupervised and semi-supervised deep learning tasks. These tasks are normally preferred in the absence of the ample annotated datasets required for deep learning activities, especially in imaging analytics. Autoencoders are used to augment small annotated datasets, often before a segmentation task. For example, a method known as GenSeg was proposed that combines the generative aspects of autoencoders to generate a latent representation of tumour cells from a labelled health image and uses U-Net architecture to obtain the unique information of tumours present in the MR images. There is also other related research [69]. Noise reduction is another task that has benefited from the generative features of the autoencoder framework. The fusion of Bayes shrinkage fused wavelet transform (BSbFWT) was proposed for noise removal and an autoencoder block for generating a noiseless variant of an MR image of prostate cancer [72]. MR images can be prone to Gaussian and Rician noise, which is introduced during image capture by the MRI device and the imaging environment. The effectiveness of noise-reduced images was measured using several parameters, such as peak signal-to-noise ratio (PSNR), mean squared error (MSE), structural similarity index metric (SSIM), and mean absolute error (MAE), which outperformed numerous traditional filtering approaches presented in their work. Autoencoders can also be used in classification tasks [70,72,74,76]. Autoencoders were used to mesh a fully convoluted network for a classification task [73]. The authors proposed a convolutional mesh autoencoder (CMA) framework for the classification of syndromic craniosynostosis (SC) of three known SC variants—Muenke, Crouzon, and Apert disease—in infants and adults using 3D computed tomography (CT) images and others with very good percentages on evaluation metrics such as sensitivity, specificity, and accuracy when compared to normal subjects. These tasks are very critical, as late and inaccurate diagnosis might prove irreversible, causing permanent damage to the brain. As indicated, an autoencoder which is infused with four convolutional layers for encoding and decoding, was used for the construction of face models from the CT images. Autoencoders were also used to detect complex anomalies present in medical imaging [73,75,77]. In another study, autoencoders were applied to chest X-ray and digital pathology images [76]. Abnormalities that were barely visible, such as metastases in lymph nodes, always proved difficult to detect as they resemble normal images in pathological slides. The authors proposed a deep perpetual autoencoder that learnt the shared patterns of normal images and content similarities to abnormal ones and restored them correctly. For evaluation of their mode, they used the receiver operating characteristic (ROC), as it

integrates the classification performance of the normal and the abnormal class. Their model was also evaluated on non-medical images and performed well in comparison, indicating the suitability of their autoencoder for medical imaging purposes.

3.4. Generative Adversarial Networks

Generative adversarial networks (GANs) are groups of deep learning artificial neural networks that can be used for generating synthetic medical images, data augmentation, and improving image resolution. They are valuable for enhancing small datasets, which are common in medical imaging, and can also be used to create better training data for improving model performance. A GAN uses an unsupervised learning algorithm and can be used for mostly semi-supervised and unsupervised learning [78,79]. A GAN consists of two main parts—the generator (G) and the discriminator (D). The generator, which comprises a multilayer perceptron (LP), learns the data distribution of the input image and produces a similar image to the input, also known as "fake data". The job of the discriminator, which is also an MLP, is to discriminate between the generated image from the generator network and the input image. The result of this discrimination, which constitutes an error between the original input image and the generated image, is fed back to the generator input to make the generated image more realistic, i.e., closer to the original image. During training, the weights of the generator and discriminator are alternately updated, and the weight updates of the generator come from the discrimination error. Both networks are engaged in a competing optimisation process. This process continues until there is an equilibrium between the generator and the discriminator networks [1,80,81]. The architecture of a GAN is shown in Figure 7.
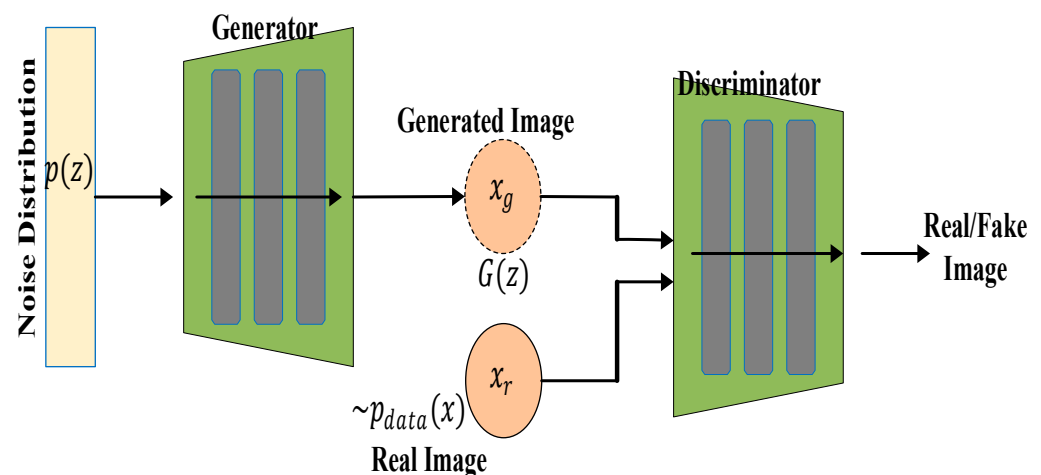

Figure 7. Structure of a GAN [79].

The generator (G) consists of an input noise vector z such that $z \sim p_z(z)$, where p_z is a prior distribution that could be Gaussian. The output of the generator, which could be an image, is represented by $G(z)$, and the objective is to maximize $\log (D(G(z)))$. The input to the discriminator (D) is the image data sample x and the output of the generator network. The output to the discriminator is the probability that x is real or fake. This is represented by $D(x) \epsilon [0, 1]$, where 0 represents fake and 1 represents real. The goal of the discriminator network is to maximize $\log(D(x) + log(1 - D(G(z))))$. The discriminator loss $\mathcal{L}_D$ is given by:

$$\mathcal{L}_D = -\mathbb{E}_{x \sim p_{data}}[\log D(x)] - \mathbb{E}_{x \sim p_z}[\log(1 - (G(z)))] \tag{19}$$

where $-\mathbb{E}_{x \sim p_{data}}[\log D(x)]$ maximises the probability of correctly classifying the real data and $\mathbb{E}_{x \sim p_z}[\log(1 - D(G(z)))]$ minimises the probability of misclassifying the generated data. When both generator and discriminator are at a stable equilibrium, the combined minimax game function is defined as:

$$\underset{G}{\text{Min}} \, \underset{D}{\max} \, \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{x \sim p_z}[\log(1 - (G(z)))] \tag{20}$$

Table 3 shows the summary of GAN-based techniques.

Table 3. GAN-based techniques.

Article	Imaging Modality	Task	Disease/Body Part	Variant Used
[79]	MRI/retina fundus	Image synthesis	-	-
[82]	MRI	Image resolution	Brain	Cycle-GAN
[83]	CT	Image synthesis	COVID-19	Enhanced vanilla
[84]	X-ray/CT	Image Denoising	Chest/thorax	CGAN
[85]	Various	Image resolution	Various	Enhanced vanilla
[86]	-	Image synthesis	Skin cancer	DCGAN
[87]	MRI/CT	Image synthesis	Head/neck	Vanilla GAN
[88]	MRI/CT	Image resolution	Bladder cancer	Enhanced vanilla
[89]	Retina fundus/MRI	Image resolution	Various	Vanilla GAN
[90]	CT/MRI	Translation	Thorax/brain	CGAN

Literature Review Findings for GANs

Deep learning technique such as GANs can be used in medical imaging in two different ways. The generative network in GANs has been used for image synthesis purposes to generate synthetic datasets for tasks where there are limited annotated datasets and using the discriminating networks of GANs for anomaly detection [79]. The quality of the generated synthetic images can be measured by the use of qualitative metrics such as Fréchet inception distance (FID), which is a measure of similarities between the representations of the generated and the real input images, structural similarity index measure (SSIM), which indicates the similarities of the structures (usually image contrast and brightness), and peak signal-to-noise ratio (PSNR), which is used to analyse the sensitivity of the generated image. PSNR is the most important metric when dealing with medical images. Quantitative metrics used for generated image quality include number of parameters (NoPs) to represent the total number of trainable parameters in the GAN and floating points of operations (FLoPs), which measure the cost of computation of the network [78]. Most studies that have used GANs have used them for image synthesis, image resolution, image translation, and image denoising. For example, a lightweight GAN (LEGAN) was proposed to generate high-fidelity images from MRIs and retina fundus images [79]. That technique boasts fewer parameters used in the training process and lower FID when compared to other variants of GAN such as CGAN, DCGAN, Pix2Pix, and so on. To achieve this, they used a two-stage GAN to create a coarse-to-fine paradigm, which is necessary in generating images with high sensitivity to the fine patterns of the original image. To lower the NoPs, redundancy in the convolutional kernel was eliminated by using the principal components of a normally fully ranked convolutional kernel for feature extraction. The resolution of MR images of the brain was improved by increasing the contrast using a GAN variant known as Cycle-GAN to aid in the segmentation of the images [85]. To achieve this, they used the image-to-image translation technique to create a high tissue contrast (HTC) of the real image. The attention block of the Cycle-GAN used in that study helped in focusing on a single tissue type and increasing the contrast within the tissue. Other studies have used GANs for medical image resolution, including [85]. For

denoising tasks, a GAN was used for denoising X-ray images using the CGAN variant of the GAN architecture [87]. The authors customised it to deal with spatially varying noise, which is often overlooked when dealing with medical images. To achieve this, the gradients of original images were merged with noisy images to obtain conditional information for the CGAN, thereby enhancing the contrast. The convolutional layers of the generator were used in full for better feature extraction. For improved consistency between the real and fake images, the reconstruction loss was combined with WGAN loss to create an objective loss for the network, obtaining remarkable PSNR and SSIM performance when compared to other state-of-the-art (SOTA) GAN architectures.

3.5. U-Net

U-Net architecture combines the best of CNN and encoder–decoder models, specifically for the purposes of segmenting medical images [91,92]. They have found application in major medical imaging modalities such as CT, X-ray, and MRI. U-Net's ability to exploit small, annotated data samples (based on its fully connected layers) by leveraging data augmentation and improved feature extraction made it a valuable technique for medical image segmentation [92]. The U-shaped architecture with skip connections help to delineate objects in images, making it highly effective in medical image analysis, particularly in tasks like tumour detection, organ delineation, and segmentation of medical images from various modalities, and variants have been widely embraced among the many different deep learning networks [93].

U-Net architecture is mainly composed of two paths. The first path is referred to as the contracting or encoder path. It uses a downsampling module that consists of several repeating convolutional blocks for semantic and contextual feature extraction. Each convolutional block has two successive 3×3 convolutions, ReLU activation functions, and a pooling layer. The pooling layer serves to increase the receptive field of the convolutional network with no extra burden of computing resources that might be introduced by an additional convolutional block. The second path of U-Net is the expansive or decoder path. It is saddled with the task of upsampling spatial resolutions of the feature maps from the contracting path, usually by a factor of two. During this operation, the dimensions of the features are reduced and a pixel-wise classification/resolution score is produced. The expansive path is made up of a 2×2 transposed convolutional layer (reversing the operation in the contracting path), which is followed by a 3×3 convolutional layer and a ReLU activation function. There is also a bottleneck layer that serves as a connection between the two paths. It is also comprised of two blocks of 3×3 convolutional layers and a ReLU activation function. The embedded skip connections in the bottleneck copy the output of each stage of the paths, helping to learn contextual and semantic representations in the deep and shallow layers, respectively [92,94]. The U-Net architecture is shown in Figure 8.

The encoders/contracting paths shown on the left-hand side of Figure 8 comprise several layers. Each layer *l* in the encoder is represented by the function:

$$f^l = \sigma\left(W^l * f^{l-1} + b^l\right) \tag{21}$$

where f^{l-1} represents input features from a previous layer, W^l, b^l is the convolutional weights and biases, and σ represents the activation function, which is ReLU in most cases.

The output of the encoder is normally max-pooled and can be represented as:

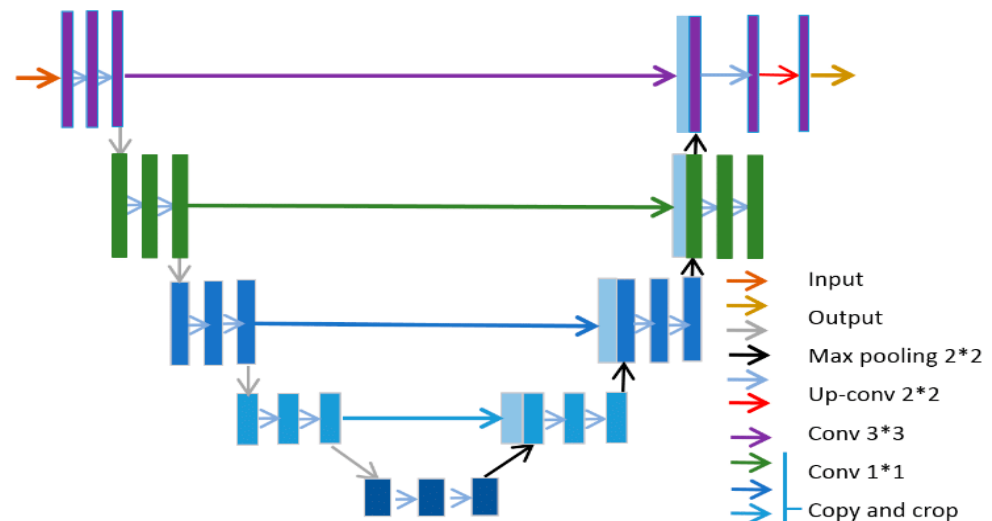$$f^l_{pooled} = MaxPool\left(f^l\right) \tag{22}$$

Figure 8. The architecture of U-Net [95].

The output is then passed through the bottleneck layer, which is another convolution function, represented by:

$$f_{bottleneck} = \sigma\left(W_b * f^L + b_b\right) \tag{23}$$

with L representing the number of layers in the encoder/contracting path. The upsampling taking place in the decoder is performed via a skip connection at every layer on the encoder. As such, each layer in the decoder upsamples the feature map, concatenates the corresponding encoder features, and applies convolution with the given function as:

$$f^l_{upsampled} = ConvTranspose\left(f^l_{bottleneck}\right) \tag{24}$$

$$f^l_{concat} = Concat\left(f^l_{upsampled}, f^{L-1}_{encoder}\right) \tag{25}$$

$$f^l = \sigma\left(W^l_{decoder} * f^l_{concat} + b^l_{decoder}\right) \tag{26}$$

At the output layer, there is a 1×1 convolutional layer that uses softmax activation to map the desired output segments:

$$y = Softmax\left(W_{out} * f^{L_{decoder}} + b_{out}\right) \tag{27}$$

During training of a U-Net, the loss function used is usually the cross-entropy loss, and it is normally applied pixel-wise.

Table 4 provides a summary of works that have implemented the U-Net segmentation.

Table 4. U-Net segmentation techniques.

Article	Imaging Modality	Disease/Body Part	Variant Used
[95]	MRI	Brain tumour	Enhanced U-Net
[96]	Various	Various	Enhanced U-Net
[97]	CT	Hepatocellular carcinoma	Enhanced U-Net
[98]	CT	Liver	Enhanced U-Net
[99]	MRI	Brain tumour	None
[100]	Colour fundus	Diabetic retinopathy	Enhanced U-Net
[101]	MRI	Various/musculoskeletal	Enhanced U-Net
[102]	MRI	Lower limb muscle	Attention U-Net/SCU-Net
[103]	MRI	Musculoskeletal	Various
[104]	Various	Various	Enhanced U-Net (U-Net++)
[105]	Ultrasound	Breast cancer	Enhanced U-Net (attention gate)

Literature Review Findings for U-Net

U-Net is mostly used for segmentation tasks, especially for segmentation of cancer for various imaging modalities. Over time, there have been a lot of modifications to the vanilla U-Net model as different researchers have tried to enhance different facets, ranging from the skip connections to modifying the convolutional layers with attention networks in a bid to increase segmentation quality or reduce computational burden. A multi-level feature assembly MFLA U-Net was reported that was integrated with a multi-scale information attention and pixel-vanishing attention mechanism [106]. This enhanced U-Net model was designed to boost segmentation performance. The model was tested on different medical imaging datasets with different modalities such as colonoscopy and dermoscopic images. Dice index coefficients were used as a metric to evaluate the effectiveness of the model in segmenting these images. It outperformed many state-of-the-art U-Net models on the datasets used for testing. A lightweight U-Net architecture was applied to a publicly available brain tumour dataset (BraTs) to segment brain tumours [99]. The focus of the study was mainly developing a low-resource U-Net framework that had a multimodal CNN encoder–decoder. Augmentation was excluded to reduce the computational demand on the network. The model achieved remarkable performance, with Dice coefficient values of up to 0.93 for specific segmented classes when compared to other U-Net models. An enhanced U-Net model with minimal parameters was reported [97]. The authors achieved this by developing a framework known as stack multi-connection simple reducing net, otherwise known as SRNet. This network used fewer convolution operation in the downsampling and upsampling processes, which in turn helped to reduce the total parameters of a vanilla U-Net algorithm by 20%. They also modified the original architecture by ensuring the convolutional layers were not stacked, helping to reduce information loss. Their model was also tested using the BraTs dataset. They obtained matching results with popular variants of the U-Net model, also using Dice coefficients as an evaluation tool. Others modified a vanilla U-Net for the purpose of improving the accuracy of segmentation [99]. They explored the weakness of U-Net models that only focus on contextual information and neglect other useful features of the channel. The developed HDA-ResUNet combined the best of attention mechanisms, U-Net, and dilated convolution. They evaluated their model on ISI and LiTS segmentation datasets and achieved than a conventional U-Net.

3.6. Transfer Learning

Deep learning techniques are known to be computationally intensive, owing to the large number of trainable parameters available in their networks. These parameters increase substantially as the network deepens. Also, the availability of large annotated medical image datasets is scarce, and this is a very important consideration in the use of deep learning for medical image analysis [107,108]. Transfer learning algorithms were developed to help solve such problems by providing a means to reparametrize an already-trained large deep learning network. These networks are CNN-based and trained on millions of images with different classes. The resulting learnt parameters are saved to be reused on other datasets. Some aspects of these networks are modified to suit new datasets. Transfer learning involves using pretrained deep learning models (e.g., ResNet, VGG, DenseNet, GogleNet, XceptionNet, AlexNet, Inception V3, and SqueezeNet) and fine-tuning them on medical images. Developing models when using transfer learning is performed in two stages: initialising the weights and fine-tuning. During initialisation of the weights, the weights of a previously trained model with a different dataset, i.e., AlexNet trained on ImageNet, as shown in Figure 9, are copied. When a new training dataset, such as medical images in this case, is used to train this model, the weights are updated. In the fine-tuning process, some of the CNN layers are frozen and thereby the

weights are not being updated. Another method of fine-tuning is freezing all the CNN layers barring the classification layer that is adjusted according to the requirements of the medical image data. This technique is highly effective when the available dataset is small or specific to a particular medical condition. It reduces training time and improves diagnostic accuracy [109]. Transfer learning can be categorised into inductive, transductive, and unsupervised learning based on data labels. These can also be categorised as homogeneous and heterogeneous based on how consistent the dataset features and labels are between the source and target domain [110,111].
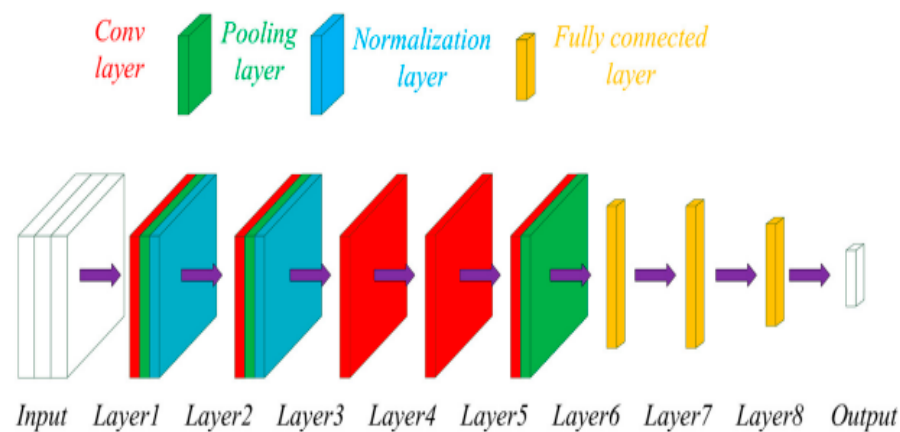

Figure 9. Structure of a transfer learning model—AlexNet [108].

The concept of transfer learning can be modelled mathematically as described next. The source domain be represented as:

$$D_s = \left\{ x_s^{(i)}, y_s^{(i)} \right\} \text{ for } {}_{i=1}^{N_s}, \tag{28}$$

where $x_s^{(i)} \in X_s$ represents the input data from the source domain, $y_s^{(i)} \in Y_s$ represents the corresponding labels, and N_s is the number of samples. The target domain is represented as:

$$D_T = \left\{ x_T^{(i)}, y_T^{(i)} \right\} \text{ for } {}_{i=1}^{N_T}, \tag{29}$$

where $x_T^{(i)} \in X_T$ represents the input data from the target domain and $y_s^{(i)} \in Y_s$ represents the corresponding labels. N_T is the number of samples, with the assumption that $X_s \neq X_T$ or $Y_s \neq Y_T$. The objective of the transfer learning function is to find a model $f_T(x_T; \varnothing_T)$ that minimises the loss in the target domain. If we represent the loss function by $\mathcal{L}_T$, then $\mathcal{L}_T$ can be defined as:

$$L_T = \frac{1}{N_T} \sum_{i=1}^{N_T} l(f_T(x_T^{(i)}; \varnothing_T), \ y_T^{(i)}) \tag{30}$$

where l is the specific loss algorithm, which can be cross-entropy or mean squared error and $\varnothing_T$ represents the target domain model parameters. The transfer learning process between the source and target domain takes two steps. We pretrain the model on the source domain using the objective loss function given by the equation:

$$L_s = \frac{1}{N_s} \sum_{i=1}^{N_s} l(f_s(x_s^{(i)}; \varnothing_s), \ y_s^{(i)}) \tag{31}$$

The learned parameters $\varnothing_s$ from the source domain are transferred to the target domain by freezing the base layers, where $\varnothing_s^{base}$ is fixed or by fine-tuning specific layers i.e., $\varnothing_T^{new}$.

The target domain overall model becomes:

$$f_T\left(x_T^{(i)};\varnothing_T\right) = f_s^{base}\left(x_T;\varnothing_s^{base}\right) + f_T^{new}(x_T;\varnothing_T^{new}) \tag{32}$$

The learning process is then optimised by regularising the losses. Therefore, a combined loss function L is given by:

$$L = L_T + \lambda \cdot L_{regularisation} \tag{33}$$

where λ is the regularisation weight and $L_{regularisation}$ can be any term used for smooth transfer, e.g., L_2-norm.

Findings for Transfer Learning

Transfer learning (TL) techniques are mostly applied in the classification of medical images, as summarised in Table 5.

Table 5. Summary of transfer learning techniques.

Article	Imaging Modality	Disease/Body Part	TL Variant/Best Model
[112]	Histopathological images	Breast cancer	ResNet 50
[113]	MRI	Brain tumour	Improved ResNet 50
[114]	MRI	Alzheimer's	Various (EffiecientNet)
[115]	CT	Pulmonary nodules	Various (DenseNet)
[116]	X-ray/CT	COVID-19	Various (VGG 16)
[117]	MRI	Alzheimer's	Modified ResNET 18
[118]	Ultrasound	Thyroid	VGG-16

This is because the networks they are learning from have been pretrained for classification as well. Several transfer learning models, such as VGG16, DenseNet 121, and ResNet 50, were used for the binary classification tasks of X-ray radiographs and CT images of COVID-19 patients [118]. Most of the model parameters were frozen and the weights were not initialised. However, they were able to obtain very good results in terms of classification accuracy, with VGG16 performing best, with accuracy of 99%. Transfer learning models like those in [118] were compared and a custom CNN built from scratch to compare the efficacy of the TL process. The models were trained on different publicly available datasets with varying modalities, ranging from X-rays to CT scans of different diseases, including lung cancer and brain tumour. The initial convolutional layers of the TL networks used in that work was frozen and the weights of the top layers were updated. The models were trained on all the datasets, and improved accuracy was obtained with the TL models, with ResNet 50 showing the highest accuracy of 90% for the histopathological images. In another study, transfer learning techniques were trained on the ImageNet datasets [119]. The authors pretrained a novel hybrid DCNN model that combined convolutional layers with global average pooling for each layer with a skip connection in each of the convolutional layers. The 200,000 augmented unlabelled data were sourced from different repositories of biopsy breast cancer image datasets and were used to train the DCNN model. The pretrained DCNN model was then used to classify annotated skin cancer lesions as benign or malignant based on the weights from the novel pretrained model. They obtained improved accuracy when compared to other models tested on similar datasets. A transfer learning model (ResNet 18) was used for the diagnosis of different stages of Alzheimer's disease (AD) [120]. The model was fine-tuned by unfreezing all the layers, allowing it to update all it weights based on the publicly available DICOM MRI datasets used. High accuracy was obtained for the three different classes (in the range of 99%), and their model outperformed

other related works that classified AD. Qing Guan et al. [121] used an Inception-v3 TL model for the classification of papillary thyroid carcinomas (PTCs) and benign thyroid nodules using ultrasound images. A total of 2836 images from 2235 patients were used. The model was trained to crop nodule margins and make diagnostic predictions. The best performance was achieved using a 50-pixel margin and 384×384 image size. In the test group, Inception-v3 achieved sensitivity of 93.3% and specificity of 87.4%. Their proposed model was most accurate for nodules sized 0.5–1.0 cm.

3.7. Vision Transformers

Despite the significant successes recorded in enhancing the diagnostic accuracy of deep learning models such as CNN, RNN, and U-Net in the classification and segmentation of medical images, there remain some limitations. Their reliance on localised feature extraction leading to inductive bias and sequential operation makes them fall short when the medical imaging tasks requiring long-range dependency and global feature extraction [122]. Although initially designed for natural language processing tasks such as sentiment analysis [123], machine translation, and text summarisation [124], their ability to capture long range dependencies in image pixels helps to build a more robust segmentation–classification model. The vision transformer is a relatively new deep learning architecture that is increasingly being applied to medical imaging. Developed by Google in 2020, it performs segmentation or classification using the transformer architecture. The ViT creates a partition of the input images into multiple patches of 16×16 pixels and linearly embeds them. For the pixels to be suitable for the transformer architecture, they must be transformed into fixed-length vectors [125]. The self-attention mechanism represents the main feature of the ViT architecture, as this forms the basis of the interactions among the pixel patches. It also uses positional encoding to represent the spatial location of the image patches. Feed-forward layers placed after the self-attention layers are generally used by the model for final decisions [126]. Several large language models have been developed for medical image analysis [127]; however, we have focused on the legacy transformer model in our analysis. The original architecture of the vision transformer is shown in Figure 10.
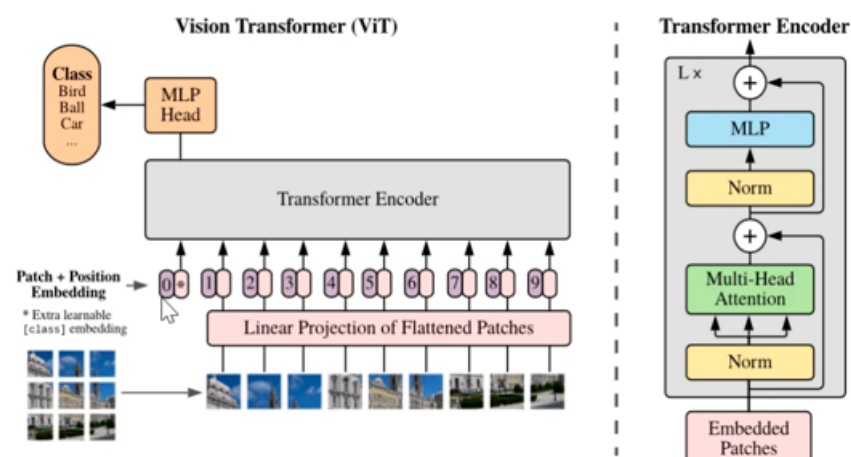

Figure 10. Vision transformer architecture [128].

If we represent the input to the ViT model by an image X, then X can be defined such that:

$$X \in \mathbb{R}^{H \times W \times C} \tag{34}$$

where H and W represent the height and width of the image, respectively, and C is the number of channels, which is usually three for RGB images. The input image is divided into non-overlapping patches of size P such that:

$$X_{patches} = \{x_1, x_2, \ldots, x_{N,}\} \tag{35}$$

where $N = \frac{H \times W}{P^2}$ is the total number of patches and each patch is flattened as:

$$x_1 \in \mathbb{R}^{P^2.C} \tag{36}$$

The patches are sent through the transformer encoder layers consisting of the multi-head attention layer, which computes the relationships between patches using the K, V, Q (keys, values, and queries) matrices, which are computed as:

$$K = Z^{(l-1)}W_K, \; V = Z^{(l-1)}W_V, \; Q = Z^{(l-1)}W_Q \tag{37}$$

where $Z^{(l-1)}$ is the positional patch embedding for each patch and W_K, W_V, and W_Q are the learnable weight projection matrices, which have the same size as embedding space. This is then passed to the feed-forward neural network (FFNN) and a classification head for a classification task [128].

Literature Review Findings for Vision Transformers

Several studies have used ViTs and their variants, including TNT, Swin, DeiT, and PVT [129], for classification, registration, and segmentation of medical images. Just like the convolution-based transfer learning models, learnable parameters from transformer models such as those can also be used for specific DL tasks. For example, a pretrained Swin transformer was used for the classification of breast cancer using publicly available breast X-ray images [130]. The dataset was resized to fit the Swin transformer input size and augmentation was also performed to improve generalisation, achieving very high classification rates based on the selected metrics for evaluation, with accuracy of 99.9% and precision of 99.8%. The results were compared with convolution-based TL algorithms (ResNet50 and VGG16), and the Swin transformer had superior performance. MR images of ischemic strokes were classified using ViT by [131]. The vanilla ViT or ViT base used in the study also had limited fine-tuning, and hyperparameters were adjusted to fit the datasets to be classified. They also augmented and resized the MR images as preprocessing steps before deploying the ViT model. They achieved an impressive accuracy score of 97.59% when compared to the VGG16 model used in a similar study, demonstrating the superiority of transformer models for classification tasks. Transformers were used for image registration tasks [132]. In that study, the authors developed the TransMorph algorithm, which is a hybrid transformer and convolutional network. The network leveraged the encoder–decoder architecture of transformers, but instead of the attention mechanism at the decoder, this was replaced with a convolutional network. For the transformer network, they used the Swin variant due to its ability to extract feature maps at different resolutions by merging patch layers, making it suitable for the image registration task. The algorithm was tested on different image-pair datasets comprising mainly MRI and CT modalities for registration purposes. They obtained very competitive results based on Dice score evaluations when compared to both traditional and other DL methods used for similar tasks. Chuantao Wang et al. [133] proposed a ViT model with a deep neural network called ConvTrans-Net for the segmentation of lymph node tumours using ultrasound images from Kaggle. The model works by concatenating the different feature vectors of the ultrasound images that are passed on to a multilayer perceptron, and the output of the multiple attention mechanism is passed on to a feed-forward layer for segmentation purposes. Their model was evaluated using the Jaccard similarity coefficient, precision, and recall, and was benchmarked with the NCP and WS-2017 models. They obtained a

Jaccard coefficient 85.21%, recall of 89.65%, and precision of 85.17 when compared to the benchmark models.

3.8. Hybrid Models

Most deployed deep learning methods use CNNs, as described in the preceding sections, with variants such as VGG, ResNet, LSTM, RNN, GAN, and GRU all being used for different medical image analysis. However, different authors have tried to combine the strengths of some of the models to deal with the weaknesses of the others by combining them, and this forms the basis of hybrid models [3]. Some studies have also combined different deep learning methodologies by focusing on the strengths of the methodologies. For example, a convolutional network's strong local feature extraction has been combined with the long-range dependencies of transformers when performing medical analysis.

3.8.1. Convolution-Based Hybrid Models

A hybrid model of convolution algorithms was developed comprising SegNet, MultResUNet, and the krill herd optimisation algorithm (KHO) to improve the segmentation of CT scans of liver lesions and RNA genome sequencing [134]. The SegNet framework provided the segmentation capacity of their model, utilising pixel-wise classification through the softmax layer. A CNN-based architecture was employed with MultiResUNet to handle the lesion segmentation, together with the SegNet framework. The hyperparameters of the models, when optimised through the KHO algorithm, helped to improve the segmentation process. The algorithm was tested with a publicly available LiTs dataset using evaluation metrics such as Dice coefficient, F1 score, and accuracy, and obtained better results, with F1 scores comparatively higher than the models used for comparison. A hybrid model of convolutional methods involving RestNet and U-Net (ResUNet) was developed for the segmentation of liver and tumours using CT images [135]. The model focused on improving the available models by providing improved image contrast and segmenting irregular tumour shapes and small tumours. The combination used the best of RestNet's residual connection and U-Net's encoder and decoder structure to enhance feature learning, segmentation precision, and efficiency. Various augmentation techniques such as rotation and reflection were also implemented to increase the variability in the dataset, and accuracy of 99.6% and a Dice coefficient of 99.2% were achieved. Yangyang Zhu et al. [136] developed a hybrid model for the classification of unexplained cervical lymphadenopathy (CLA) using ultrasound images of patients in an underdeveloped area of China. The CLA-HDM model was made up of three smaller models, each designed to handle a specific diagnosis task related to unexplained lymph node abnormalities (CLA). Model 1 checked whether the issue was benign (non-cancerous) or malignant (cancerous). Model 2 looked deeper into benign cases to decide between tuberculosis or a reactive condition. Model 3 examined malignant cases to determine if they were due to metastasis (spread from another cancer) or lymphoma. Each model had two input branches, one for greyscale ultrasound (BUS) images and another for colour Doppler (CDFI) images. The CDFI images were processed to emphasize important colour details. Then, both images were analysed by a deep learning model (ResNet-50). The efficacy of their model was evaluated using the area under the curve, which was above the 0.8 benchmark for each model. The authors in [137] developed an ensemble model of a custom-built convolutional neural network (5-CNN) and a transfer learning model using the pretrained VGG-19 architecture for the classification of thyroid disorders using ultrasound images of four categories: autoimmune, nodular, micro-nodular, and normal. The combined CNN-VGG method showed superior performance, achieving test accuracy of 97.35%, specificity of 98.43%, and sensitivity of 95.75%. It also demonstrated

strong predictive capabilities, with high positive and negative predictive values, and an area under the ROC curve of 0.96.

3.8.2. Convolution–Transformer-Based Hybrid Models

A hybrid model called TBConvl-NET that combines CNN, LSTM, and ViT for the segmentation of several diseases was developed using publicly available datasets of different modalities such as ultrasound and MRI [138]. The hybrid model targeted some well-known challenges in segmenting medical images, such as scale, texture, and shape of pathology. Due to the high computational resources required, they used a depth-wise separable convolution, thereby reducing computational overhead. Swin transformer blocks were used in the skip connections to help deal with the varying scales of the data and help preserve semantic information. They used the Dice index, accuracy, and Jaccard index to evaluate their model, compared the developed model with other hybrid segmentation models, and obtained improved results. A hybrid classification model that combines the transformer and the convolution model to improve the classification of skin lesion was developed using publicly available datasets [139]. It used the Swin-Unet architecture to perform image segmentation, leveraging the self-attention of the Swin transformer and robust hierarchical analysis of the UNet. This was combined with the Xception and ResNet 18 models for feature extraction to further improve on the image analysis. For hyperparameter tuning, a hybrid salp swam algorithm (HSSA) was used to obtain the optimal parameters, hence avoiding local minima during training. A gated recurrent unit (GRU) network was used for the eventual classification. Accuracy of 94.51% and 95.38% was achieved on the two datasets used in the study. The model performed better when compared to TL models like AlexNet and ResNet18. Laifa Yan et al. [140], proposed a framework focusing on the three-vessel view (3VV) that identifies three major heart vessels—the pulmonary artery, aorta, and superior vena cava—using ultrasound images. In the first step, a YOLOv5 (based on CNN architecture) model was used to detect these vessels and earmark the region of interest. In the second step, they used a modified DeepLabv3 model with a new attentional multi-scale feature fusion (AMFF) module to perform segmentation. Using a dataset of 511 images, the model achieved high accuracy, with Dice scores of over 85% for the pulmonary artery and aorta.

4. Discussion

In this article, we have provided a broad view of various deep learning methodologies used for medical imaging analysis. These included convolutional neural networks (CNNs), recurrent neural networks (RNNs), autoencoders, generative adversarial networks (GANs), U-Net architectures, vision transformers, and hybrid models. For each of these methodologies, some of the key findings were presented.

CNNs are extensively used for medical imaging tasks like disease detection, classification, and segmentation, and the most important contribution of the algorithms is the unrivalled feature extraction through convolutional layers, making them effective for tasks requiring spatial hierarchy. CNN models can be lightweight in terms of computing resources, as inspired by ResNet, while also balancing model performance. Deeper CNN architecture can be computationally expensive, but can be very useful in multi-disease detection, e.g., in the classification and detection of COVID-19 and viral pneumonia, achieving high accuracy due to advanced hardware. It can also be effective when the task includes pixel-level feature extraction. Examples can be seen in its use in the classification of ultrasound image classification of fatty liver. RNNs, particularly their variants like long short-term memory (LSTM) and gated recurrent units (GRUs), are suitable for sequential data tasks. Their ability to retain temporal dependencies is utilised in dynamic imaging

and disease progression monitoring. RNNs, which usually have convolutional layers, are effective for classification tasks for different imaging modalities, with MRI dominating. They have been applied in the detection of several diseases, such as Alzheimer's and breast cancer. LSTM variants of the RNN algorithm handle tasks like segmentation of temporal MR images and noise removal in diagnostic images. Some modifications have also been applied to RNN algorithm training, such as using the honey badger algorithm (HBA) for optimisation of the model training parameters.

Autoencoders are unsupervised models used for dimensionality reduction, anomaly detection, and data augmentation. Their encoder–decoder structure helps compress and reconstruct medical images with minimal distortion. Autoencoders are employed for image analysis tasks like segmentation and denoising and have been used with several imaging modalities for the analysis of different diseases. Enhancing autoencoders by fusion with wavelet transform can also help improve noise removal from medical images. They can also be combined with convolutional networks to improve disease classification. GANs are used for generating synthetic medical images, data augmentation, and improving image resolution or quality. The main components of a GAN comprise a generator for creating synthetic images and a discriminator for distinguishing between real and generated images. GANs can be very computationally intensive due to the dual deep convolutional networks. Research in this area is focused on reducing models' parameters and maintaining performance. Different variants of GANs have employed different image analysis tasks, with each variant tuned to fit the specific tasks.

U-Nets are specialised for medical image segmentation and have been widely adopted for tasks like tumour detection and organ delineation. The architecture combines encoder–decoder pathways with skip connections for improved feature representation. Different variants of the U-Net architecture also exist to enhance the vanilla model, making them suitable for specific imaging modalities while performing the main task of segmentation. Also, enhancements have been created to reduce model operating costs while retaining the segmentation process. More recent studies have been adding transformers to U-Nets to further enhance segmentation accuracy. Transfer learning leverages pretrained models (e.g., ResNet, VGG, DenseNet) on large datasets and fine-tunes them for medical imaging tasks. They are mostly used for classification tasks. Models like VGG16 and ResNet50 perform well for COVID-19 classification, achieving very high accuracy on different imaging modalities. Some transfer learning models are quite deep, with millions of parameters, making them very computationally intensive. Custom hybrid models pretrained on unlabelled medical datasets have also been used to improve classification accuracy for skin lesions and breast cancer images, mimicking the transfer learning ideology. They provide an option in terms of computing resources and domain-specific training and learning. ViTs are a relatively newer deep learning architecture, adapted for medical imaging tasks. They require global feature extraction and long-range dependencies. They divide images into patches and process them using self-attention mechanisms. ViT models can also be pretrained like most CNN models and learning transferred for newer task. Several variants have also been developed depending on the task. With larger training datasets, ViTs have been shown to outperform CNN models when used for tasks with fewer annotated data.

Hybrid models combine the strengths of multiple architectures (e.g., CNNs with RNNs or CNNs with transformers) to improve medical image analysis. CNN-based hybrids like ResUNet combine ResNet's feature extraction with U-Net's segmentation capabilities. They are used widely for cancer related image segmentation. Convo-transformer hybrids like TBConvl-NET integrate CNN, LSTM, and ViT for improved segmentation and classification of diseases. Hybrid models can also be optimised. Algorithms such as the hybrid salp

swarm algorithm (HSSA) could be used to obtain optimal parameters, thereby increasing model performance.

5. Conclusions

This review explored the transformative impact of deep learning algorithms on medical image analysis. Convolutional neural networks (CNNs) are widely used for feature extraction and disease classification across medical imaging modalities, achieving high accuracy with optimised computational costs. Recurrent neural networks (RNNs), including LSTM and GRU, enhance temporal analysis for disease progression monitoring. Autoencoders and generative adversarial networks (GANs) assist in data augmentation, denoising, and synthetic image generation. U-Net architectures improve segmentation for tumour detection and organ delineation. Vision transformers (ViTs) leverage attention mechanisms for superior classification and registration. Hybrid models combining CNNs, transformers, and optimisation techniques enhance performance, while transfer learning mitigates data scarcity, ensuring robust results across imaging applications. Together, these advancements underscore the versatility and efficiency of deep learning in medical diagnostics, paving the way for improved clinical outcomes and personalised healthcare solutions. Future advancements may focus on computational efficiency, integrating multimodal data, and enhancing interpretability for clinical adoption.

Author Contributions: Conceptualisation, O.S. and R.S.; methodology, O.S. and R.S.; investigation, O.S. and R.S.; resources, O.S. and R.S.; data curation, O.S. and R.S.; writing—original draft preparation, O.S. and R.S.; writing—review and editing, O.S. and R.S. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jeong, J.J.; Tariq, A.; Adejumo, T.; Trivedi, H.; Gichoya, J.W.; Banerjee, I. Systematic Review of Generative Adversarial Networks (GANs) for Medical Image Classification and Segmentation. *J. Digit. Imaging* **2022**, *35*, 137–152. [CrossRef] [PubMed]
2. Yu-Jen Chen, Y.; Hua, K.; Hsu, C.; Cheng, W.; Hidayati, S.C. Computer-aided Classification of Lung Nodules on Computed Tomography Images via deep Learning Technique. *OncoTargets Ther.* **2015**, *8*, 2015–2022. [CrossRef] [PubMed]
3. Zhu, Z. Advancements in Automated Classification of Chronic Obstructive Pulmonary Disease based on Computed Tomography Imaging Features Through Deep Learning Approaches. *Respir. Med.* **2024**, *234*, 107809. [CrossRef] [PubMed]
4. Sarmadi, A.; Razavi, Z.S.; Van Wijnen, A.J.; Soltani, M. Comparative Analysis of Vision Transformers and Convolutional Neural Networks in Osteoporosis Detection from X-ray Images. *Sci. Rep.* **2024**, *14*, 18007. [CrossRef]
5. Takahashi, S.; Sakaguchi, Y.; Kouno, N.; Takasawa, K.; Ishizu, K.; Akagi, Y.; Aoyama, R.; Teraya, N.; Bolatkan, A.; Shinkai, N.; et al. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. *J. Med. Syst.* **2024**, *48*, 84. [CrossRef]
6. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
7. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
8. Tian, Y.; Zhang, Y. A Comprehensive Survey on Regularization Strategies in Machine Learning. *Inf. Fusion* **2022**, *80*, 146–166. [CrossRef]
9. Xiao, M.; Wu, Y.; Zuo, G.; Fan, S.; Yu, H.; Shaikh, Z.A.; Wen, Z. Addressing Overfitting Problem in Deep Learning-Based Solutions for Next Generation Data-Driven Networks. *Hindawi Wirel. Commun. Mob. Comput.* **2021**, *2021*, 8493795. [CrossRef]
10. Ying, X. An Overview of Overfitting and its Solutions. *IOP Conf. Ser. J. Phys. Conf. Ser.* **2019**, *1168*, 022022. [CrossRef]

11. Garbin, C.; Zhu, X.; Marques, O. Dropout vs. Batch Normalization: An Empirical Study of Their Impact to Deep Learning. *Multimed. Tools Appl.* **2020**, *79*, 12777–12815. [CrossRef]

12. Helmbold, D.P. Surprising Properties of Dropout in Deep Networks. *J. Mach. Learn. Res.* **2018**, *18*, 1–28.

13. Piotrowskia, A.P.; Napiorkowskia, J.J.; Piotrowska, A.E. Impact of Deep Learning-Based Dropout on Shallow Neural Networks Applied to Stream Temperature Modelling. *Earth Sci. Rev.* **2020**, *201*, 103076. [CrossRef]

14. Salehin, I.; Kang, D.-K. A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain. *Electronics* **2023**, *12*, 3106. [CrossRef]

15. Xu, M.; Yoon, S.; Fuentes, A.; Park, D.S. A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recognit.* **2023**, *137*, 109347. [CrossRef]

16. Kumar, T.; Brennan, R.; Mileo, A.; Bendechache, M. Image Data Augmentation Approaches: A Comprehensive Survey and Future Directions. *IEEE Access* **2024**, *12*, 187536–187571. [CrossRef]

17. Yang, Z.; Sinnott, R.O.; Bailey, J.; Ke, Q. A Survey of Automated Data Augmentation Algorithms for Deep Learning-Based Image Classification Tasks. *Knowl. Inf. Syst.* **2023**, *65*, 2805–2861. [CrossRef]

18. Nanni, L.; Paci, M.; Brahnam, S.; Lumini, A. Comparison of Different Image Data Augmentation Approaches. *J. Imaging* **2021**, *7*, 254. [CrossRef]

19. Oza, P.; Sharma, P.; Patel, S.; Adedoyin, F.; Bruno, A. Image Augmentation Techniques for Mammogram Analysis. *J. Imaging* **2022**, *8*, 141. [CrossRef]

20. Sandfort, V.; Yan, K.; Pickhardt, P.J.; Summers, R.M. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Sci. Rep.* **2019**, *9*, 16884. [CrossRef]

21. Wei, J.; Wang, Q.; Song, X.; Zhao, Z. The Status and Challenges of Image Data Augmentation Algorithms. *J. Phys. Conf. Ser.* **2023**, *2456*, 012041. [CrossRef]

22. Liu, X.; Karagoz, G.; Meratnia, N. Analyzing the Impact of Data Augmentation on the Explainability of Deep Learning-Based Medical Image Classification. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 1. [CrossRef]

23. Mamat, N.; Othman, M.F.; Abdulghafor, R.; Alwan, A.A.; Gulzar, Y. Enhancing Image Annotation Technique of Fruit Classification Using a Deep Learning Approach. *Sustainability* **2023**, *15*, 901. [CrossRef]

24. Jayaraj, R.; Lokesh, S. Automatic Image Annotation Using Adaptive Convolutional Deep Learning. *Intell. Autom. Soft Comput.* **2023**, *36*, 481–497. [CrossRef]

25. Zhang, D.; Islam, M.M.; Lu, G. A Review on Automatic Image Annotation Techniques. *Pattern Recognit.* **2012**, *45*, 346–362. [CrossRef]

26. Galbusera, F.; Cina, A. Image Annotation and Curation in Radiology: An Overview for Machine Learning practitioners. *Eur. Radiol. Exp.* **2024**, *8*, 11. [CrossRef]

27. Yin, S.; Bi, J. Medical Image Annotation Based on Deep Transfer Learning. *J. Appl. Sci. Eng.* **2019**, *22*, 385–390.

28. Zhang, Y.; Chen, J.; Xiangxun, M.; Wang, G.; Bhatti, U.A.; Huang, M. Interactive Medical Image Annotation Using Improved Attention U-net With Compound Geodesic Distance. *Expert Syst. Appl.* **2024**, *237 Pt A*, 121282. [CrossRef]

29. Sriwong, K.; Kerdprasop, K.; Kerdprasop, N. The Study of Noise Effect on CNN-Based Deep Learning from Medical Images. *Int. J. Mach. Learn. Comput.* **2021**, *11*, 202–207. [CrossRef]

30. Huang, Z.; Li, F.; Wang, Z.; Wang, Z. Interpretability of Deep Learning. *Int. J. Future Comput. Commun.* **2022**, *11*, 34–39. [CrossRef]

31. Xu, B.; Yang, G. Interpretability Research of Deep Learning: A Literature Survey. *Inf. Fusion* **2025**, *115*, 102721. [CrossRef]

32. Antamis, T.; Drosou, A.; Vafeiadis, T.; Nizamis, A.; Ioannidis, D.; Tzovaras, D. Interpretability of Deep Neural Networks: A Review of Methods, Classification and Hardware. *Neurocomputing* **2024**, *601*, 128204. [CrossRef]

33. Marcinkeviĉs, R.; Vogt, J.E. Interpretable and Explainable Machine Learning: A Methods-Centric Overview with Concrete Examples. *WIREs Data Min. Knowl. Discov.* **2023**, *13*, e1493. [CrossRef]

34. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef]

35. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* **2019**, *8*, 832. [CrossRef]

36. Jiang, L.; Xu, C.; Bai, Y.; Liu, A.; Gong, Y.; Wang, Y.-P.; Deng, H.-W. Autosurv: Interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics datawas partially supported by previous studies. *Npj Precis. Oncol.* **2024**, *8*, 4. [CrossRef]

37. Balachandar, N.; Chang, K.; Kalpathy-Cramer, J.; Rubin, D.L. Accounting for Data Variability in Multi-Institutional Distributed Deep Learning for Medical Imaging. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 700–708. [CrossRef]

38. Safonova, A.; Ghazaryan, G.; Stiller, S.; Main-Knorn, M.; Nendel, C.; Ryo, M. Ten Deep Learning Techniques to Address Small Data Problems with Remote Sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *125*, 103569. [CrossRef]

39. Piffer, S.; Ubaldi, L.; Tangaro, S.; Retico, A.; Talamonti, A. Tackling the Small Data Problem in Medical Image Classification with Artificial Intelligence: A Systematic Review. *Prog. Biomed. Eng.* **2024**, *6*, 032001. [CrossRef]

40. D'Antonoli, T.A. Ethical Considerations for Artificial Intelligence: An Overview of the Current Radiology Landscape. *Diagn. Interv. Radiol.* **2020**, *26*, 504–511.

41. Safdar, N.M.; Banja, J.D.; Meltzer, C.C. Ethical Considerations in Artificial Intelligence. *Eur. J. Radiol.* **2020**, *122*, 108768. [CrossRef] [PubMed]

42. Chen, I.Y.; Pierson, E.; Rose, S.; Joshi, S.; Ferryman, K.; Ghassemi, M. Ethical Machine Learning in Healthcare. *Annu. Rev. Biomed. Data Sci.* **2021**, *4*, 123–144. [CrossRef] [PubMed]

43. Chau, M. Ethical, Legal, and Regulatory Landscape of Artificial Intelligence in Australian Healthcare and Ethical Integration in Radiography: A Narrative Review. *J. Med. Imaging Radiat. Sci.* **2024**, *55*, 101733. [CrossRef] [PubMed]

44. Chen, C.; Zhang, P.; Zhang, H.; Dai, J.; Yi, Y.; Zhang, H.; Zhang, Y. Deep Learning on Computational-Resource-Limited Platforms: A Survey. *Mob. Inf. Syst.* **2020**, *1*, 8454327. [CrossRef]

45. Pietrołaj, M.; Blok, M. Resource Constrained Neural Network Training. *Sci. Rep.* **2024**, *14*, 2421. [CrossRef]

46. Chlap, P.; Min, H.; Vandenberg, N.; Dowling, J.; Holloway, L.; Haworth, A. A review of Medical Image Data Augmentation Techniques for Deep Learning Applications. *J. Med. Imaging Radiat. Oncol.* **2021**, *65*, 545–563. [CrossRef]

47. Masumoto, R.; Eguchi, Y.; Takeuchi, H.; Inage, K.; Narita, M.; Shiga, Y.; Inoue, M.; Toshi, N.; Tokeshi, S.; Okuyama, K.; et al. Automatic Generation of Diffusion Tensor imaging for the Lumbar Nerve using Convolutional Neural Networks. *Magn. Reson. Imaging* **2024**, *114*, 110237. [CrossRef]

48. Shobayo, O.; Saatchi, R.; Ramlakhan, S. Convolutional Neural Network to Classify Infrared Thermal Images of Fractured Wrists in Pediatrics. *Healthcare* **2024**, *12*, 994. [CrossRef]

49. Chauhan, S.; Edla, D.R.; Boddu, V.; Rao, M.J.; Cheruku, R.; Nayak, S.R.; Martha, S.; Lavanya, K.; Nigat, T.D. Detection of COVID-19 Using Edge Devices by a Light-Weight Convolutional Neural Network from Chest X-ray Images. *BMC Med. Imaging* **2024**, *24*, 1. [CrossRef]

50. Wu, Z.; Shen, C.; Van Den Hengel, A. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. *Pattern Recognit.* **2019**, *90*, 119–133. [CrossRef]

51. Zhu, H.; Liu, Y.; Gao, X.; Zhang, L. Combined CNN and Pixel Feature Image for Fatty Liver Ultrasound Image Classification. *Comput. Math. Methods Med.* **2022**, *2022*, 9385734. [CrossRef] [PubMed]

52. Yu, K.; Tseng, Y.; Yang, H.; Liu, C.; Kuo, P.; Lee, M.; Huang, C.; Kuo, L.; Wang, J.; Ho, C.; et al. Deep learning with test-time augmentation for radial endobronchial ultrasound image differentiation: A multicentre verification study. *BMJ Open Respir. Res.* **2023**, *10*, e001602. [CrossRef] [PubMed]

53. Gao, Y.; Zeng, S.; Xu, X.; Li, H.; Yao, S.; Song, K.; Li, X.; Chen, L.; Tang, J.; Xing, H.; et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: A retrospective, multicentre, diagnostic study. *Lancet Digit. Health* **2022**, *4*, e179–e187. [CrossRef]

54. Abdulahi, A.T.; Ogundokun, R.O.; Adenike, A.R.; Shah, M.A.; Ahmed, Y.K. PulmoNet: A Novel Deep learning based Pulmonary Diseases Detection Model. *BMC Med. Imaging* **2024**, *24*, 51. [CrossRef] [PubMed]

55. Zhang, K.; Sun, M.; Han, T.X.; Yuan, X.; Guo, L.; Liu, T. Residual networks of residual networks: Multilevel residual networks. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 1303–1314. [CrossRef]

56. Kim, J.; Hong, J.; Park, H. Prospects of Deep Learning for Medical Imaging. *Precis. Future Med.* **2018**, *2*, 37–52. [CrossRef]

57. Zhang, H.; Qie, Y. Applying Deep Learning to Medical Imaging: A Review. *Appl. Sci.* **2023**, *13*, 10521. [CrossRef]

58. Rajeev, R.; Samath, J.A.; Karthikeyan, N.K. An Intelligent Recurrent Neural Network with Long Short-Term Memory (LSTM) BASED Batch Normalization for Medical Image Denoising. *J. Med. Syst.* **2019**, *43*, 234. [CrossRef]

59. Yao, W.; Bai, J.; Liao, W.; Chen, Y.; Liu, M.; Xie, Y. From CNN to Transformer: A Review of Medical Image Segmentation Models. *J. Digit. Imaging Inform. Med.* **2024**, *37*, 1529–1547. [CrossRef]

60. Cui, R.; Liu, M. RNN-based Longitudinal Analysis for Diagnosis of Alzheimer's Disease. *Comput. Med. Imaging Graph.* **2019**, *73*, 1–10. [CrossRef]

61. Anbalagan, V.; Balasubramanian, V. HBO-GMRNN: Honey Badger Optimization Based Gain Modulated Recurrent Neural Network for Classification of Breast Cancer. *Biomed. Signal Process. Control.* **2024**, *91*, 105910. [CrossRef]

62. Amarneni, S.; Valarmathi, D.R.S. Diagnosing the MRI Brain Tumour Images Through RNN-LSTM. *E Prime Adv. Electr. Eng. Electron. Energy* **2024**, *9*, 100723. [CrossRef]

63. Zhu, K.; Chen, Y.; Ouyang, X.; White, G.; Agam, G. Fully RNN for Knee Ligament Tear Classification and Localization in MRI scans. *Electron. Imaging* **2022**, *34*, 227-1–227-6. [CrossRef]

64. Gulshan; Arora, A.S. Automated Prediction of Diabetes Mellitus Using Infrared Thermal Foot Images: Recurrent Neural Network Approach. *Biomed. Phys. Eng. Express* **2024**, *10*, 025025. [CrossRef]

65. Ayub, S.; Kannan, R.J.; Alsini, R.; Hasanin, T.; Sasidhar, C. LSTM-Based RNN Framework to Remove Motion Artifacts in Dynamic Multicontrast MR Images with Registration Model. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 5906877. [CrossRef]

66. Das, T.; Saha, G. Addressing Big Data Issues Using RNN Based Techniques. *J. Inf. Optim. Sci.* **2020**, *40*, 1773–1785. [CrossRef]

67. Zhang, Y. A Better Autoencoder for Image: Convolutional Autoencoder. In Proceedings of the ICONIP17-DCEC, Guangzhou, China, 14–18 October 2017; pp. 11–21. Available online: http://users.cecs.anu.edu.au/Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58.pdf (accessed on 17 April 2025).

68. Baldi, P. Autoencoders, unsupervised learning, and deep architectures. In Proceedings of the ICML Workshop on Unsupervised and Transfer Learning, Edinburgh, UK, 26 June–1 July 2012; JMLR Workshop and Conference Proceedings. pp. 37–49.

69. Vorontsov, E.; Molchanov, P.; Gazda, M.; Beckham, C.; Kautz, J.; Kadoury, S. Towards Annotation-Efficient Segmentation Via Image-to-Image Translation. *Med. Image Anal.* **2022**, *82*, 102624. [CrossRef]

70. Wang, W.; Huang, Y.; Wang, Y.; Wang, L. Generalized autoencoder: A neural network framework for dimensionality reduction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 490–497.

71. Juneja, M.; Kaur Saini, S.; Kaul, S.; Acharjee, R.; Thakur, N.; Jindal, P. Denoising of Magnetic Resonance Imaging Using Bayes Shrinkage Based Fused Wavelet Transform and Autoencoder Based Deep Learning Approach. *Biomed. Signal Process. Control.* **2021**, *69*, 102844. [CrossRef]

72. O' Sullivan, E.; Van De Lande, L.S.; Papaioannou, A.; Breakey, R.W.F.; Jeelani, N.O.; Ponniah, A.; Duncan, C.; Schievano, S.; Khonsari, R.H.; Zafeiriou, S.; et al. Convolutional Mesh Autoencoders for the 3-Dimensional Identification of FGFR-Related Craniosynostosis. *Sci. Rep.* **2022**, *12*, 2230. [CrossRef]

73. Wolf, D.; Payer, T.; Lisson, C.S.; Lisson, C.G.; Beer, M.; Götz, M.; Ropinski, T. Self-supervised Pre-training with Contrastive and Masked Autoencoder Methods for Dealing with Small Datasets in Deep Learning for Medical Imaging. *Sci. Rep.* **2023**, *13*, 20260. [CrossRef]

74. Chen, R.; Song, Y.; Huang, J.; Wang, J.; Sun, H.; Wang, H. Rapid Diagnosis and Continuous Monitoring of Intracerebral Hemorrhage with Magnetic Induction Tomography Based on Stacked Autoencoder. *Rev. Sci. Instrum.* **2021**, *92*, 084707. [CrossRef] [PubMed]

75. Shvetsova, N.; Bakker, B.; Fedulova, I.; Schulz, H.; Dylov, D.V. Anomaly Detection in Medical Imaging with Deep Perceptual Autoencoders. *IEEE Access* **2021**, *9*, 118571–118583. [CrossRef]

76. Elhassan, T.A.; Mohd Rahim, M.S.; Siti Zaiton, M.H.; Swee, T.T.; Alhaj, T.A.; Ali, A.; Aljurf, M. Classification of Atypical White Blood Cells in Acute Myeloid Leukemia Using a Two-Stage Hybrid Model Based on Deep Convolutional Autoencoder and Deep Convolutional Neural Network. *Diagnostics* **2023**, *13*, 196. [CrossRef] [PubMed]

77. Zhang, H.; Guo, W.; Zhang, S.; Lu, H.; Zhao, X. Unsupervised Deep Anomaly Detection for Medical Images Using an Improved Adversarial Autoencoder. *J. Digit. Imaging* **2022**, *35*, 153–161. [CrossRef]

78. Li, D.; Fu, Z.; Xu, J. Stacked-Autoencoder-Based Model for COVID-19 Diagnosis on CT Images. *Appl. Intell.* **2020**, *51*, 2805–2817. [CrossRef]

79. Zhang, H.; Chen, J.; Liao, B.; Wu, F.; Bi, X. Deep Canonical Correlation Fusion Algorithm Based on Denoising Autoencoder for ASD Diagnosis and Pathogenic Brain Region Identification. *Interdiscip. Sci. Comput. Life Sci.* **2024**, *16*, 455–468. [CrossRef]

80. Gao, J.; Zhao, W.; Li, P.; Huang, W.; Chen, Z. LEGAN: A Light and Effective Generative Adversarial Network for medical image synthesis. *Comput. Biol. Med.* **2022**, *148*, 105878. [CrossRef]

81. Singh, N.K.; Raza, K. *Medical Image Generation Using Generative Adversarial Networks: A Review*; Springer: Singapore, 2021; pp. 77–96.

82. Xun, S.; Li, D.; Zhu, H.; Chen, M.; Wang, J.; Li, J.; Chen, M.; Wu, B.; Zhang, H.; Chai, X.; et al. Generative adversarial networks in medical image segmentation: A review. *Comput. Biol. Med.* **2021**, *140*, 105063. [CrossRef]

83. Yi, X.; Walia, E.; Babyn, P. Generative adversarial network in medical imaging: A review. *Med. Image Anal.* **2019**, *58*, 101552. [CrossRef]

84. Hamghalam, M.; Wang, T.; Lei, B. High tissue contrast image synthesis via multistage attention-GAN: Application to segmenting brain MR scans. *Neural Netw.* **2020**, *132*, 43–52. [CrossRef]

85. Zhu, Q.; Ye, H.; Sun, L.; Li, Z.; Wang, R.; Shi, F.; Shen, D.; Zhang, D. GACDN: Generative adversarial feature completion and diagnosis network for COVID-19. *BMC Med. Imaging* **2021**, *21*, 154. [CrossRef] [PubMed]

86. Li, Y.; Zhang, K.; Shi, W.; Miao, Y.; Jiang, Z. A Novel Medical Image Denoising Method Based on Conditional Generative Adversarial Network. *Comput. Math. Methods Med.* **2021**, *2021*, 9974017. [CrossRef]

87. Ahmad, W.; Ali, H.; Shah, Z.; Azmat, S. A new generative adversarial network for medical images super resolution. *Sci. Rep.* **2022**, *12*, 9533. [CrossRef] [PubMed]

88. Mutepfe, F.; Kalejahi, B.K.; Meshgini, S.; Danishvar, S. Generative Adversarial Network Image Synthesis Method for Skin Lesion Generation and Classification. *J. Med. Signals Sens.* **2021**, *11*, 237–252. [CrossRef]

89. Touati, R.; Le, W.T.; Kadoury, S. A feature invariant generative adversarial network for head and neck MRI/CT image synthesis. *Phys. Med. Biol.* **2021**, *66*, 095001. [CrossRef] [PubMed]

90. Xiao, Y.; Chen, C.; Wang, L.; Yu, J.; Fu, X.; Zou, Y.; Lin, Z.; Wang, K. A novel hybrid generative adversarial network for CT and MRI super-resolution reconstruction. *Phys. Med. Biol.* **2023**, *68*, 135007. [CrossRef]

91. Mahapatra, D.; Bozorgtabar, B.; Garnavi, R. Image super-resolution using progressive generative adversarial networks for medical image analysis. *Comput. Med. Imaging Graph.* **2019**, *71*, 30–39. [CrossRef]

92. Uzunova, H.; Ehrhardt, J.; Handels, H. Memory-efficient GAN-based domain translation of high resolution 3D medical images. *Comput. Med. Imaging Graph.* **2020**, *86*, 101801. [CrossRef]

93. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Lecture Notes in Computer Science*; Springer Science + Business Media: Berlin, Germany, 2015; pp. 234–241. [CrossRef]

94. Azad, R.; Aghdam, K.; Rauland, A.; Jia, Y.; Avval, H.; Bozorgpour, A.; Karimijafarbigloo, S.; Cohen, J.P.; Adeli, E.; Merhof, D. Medical Image Segmentation Review: The Success of U-Net. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10076–10095. [CrossRef]

95. Krithika Alias Anbudevi, M.; Suganthi, K. Review of Semantic Segmentation of Medical Images Using Modified Architectures of UNET. *Diagnostics* **2022**, *12*, 3064. [CrossRef]

96. Ehab, W.; Li, Y. Performance Analysis of UNet and Variants for Medical Image Segmentation. *arXiv* **2023**, arXiv:2309.13013.

97. Ding, Y.; Chen, F.; Zhao, Y.; Wu, Z.; Zhang, C.; Wu, D. A Stacked Multi-Connection Simple Reducing Net for Brain Tumor Segmentation. *IEEE Access* **2019**, *7*, 104011–104024. [CrossRef]

98. Wang, Z.; Zou, Y.; Liu, P.X. Hybrid dilation and attention residual U-Net for Medical Image Segmentation. *Comput. Biol. Med.* **2021**, *134*, 104449. [CrossRef]

99. Khan, R.A.; Luo, Y.; Wu, F. RMS-UNet: Residual Multi-Scale UNet for Liver and Lesion Segmentation. *Artif. Intell. Med.* **2022**, *124*, 102231. [CrossRef]

100. Kong, Z.; Zhang, M.; Zhu, W.; Yi, Y.; Wang, T.; Zhang, B. Data enhancement based on M2-Unet for liver segmentation in Computed Tomography. *Biomed. Signal Process. Control* **2022**, *79*, 104032. [CrossRef]

101. Chetty, G.; Yamin, M.; White, M. A low resource 3D U-Net based deep learning model for medical image analysis. *Int. J. Inf. Technol.* **2022**, *14*, 95–103. [CrossRef]

102. Huang, K.; Yang, Y.; Huang, Z.; Liu, Y.; Lee, S. Retinal Vascular Image Segmentation Using Improved UNet Based on Residual Module. *Bioengineering* **2023**, *10*, 722. [CrossRef]

103. Lin, Z.; Dall'ara, E.; Guo, L. A novel mean shape based post-processing method for enhancing deep learning lower-limb muscle segmentation accuracy. *PLoS ONE* **2024**, *19*, e0308664. [CrossRef]

104. Ding, T.; Shi, K.; Pan, Z.; Ding, C. AI-based automated breast cancer segmentation in ultrasound imaging based on Attention Gated Multi ResU-Net. *PeerJ Comput. Sci.* **2024**, *10*, e2226. [CrossRef]

105. Henson, W.H.; Li, X.; Lin, Z.; Guo, L.; Mazzá, C.; Dall'ara, E. Automatic segmentation of lower limb muscles from MR images of post-menopausal women based on deep learning and data augmentation. *PLoS ONE* **2024**, *19*, e0299099. [CrossRef]

106. Lin, Z.; Henson, W.H.; Dowling, L.; Walsh, J.; Dall'ara, E.; Guo, L. Automatic segmentation of skeletal muscles from MR images using modified U-Net and a novel data augmentation approach. *Front. Bioeng. Biotechnol.* **2024**, *12*, 1355735. [CrossRef] [PubMed]

107. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. *UNet++: A Nested U-Net Architecture for Medical Image Segmentation*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; p. 3.

108. Garbaz, A.; Oukdach, Y.; Charfi, S.; El Ansari, M.; Koutti, L.; Salihoun, M. MLFA-UNet: A multi-level feature assembly UNet for medical image segmentation. *Methods* **2024**, *232*, 52–64. [CrossRef]

109. Kora, P.; Ooi, C.P.; Faust, O.; Raghavendra, U.; Gudigar, A.; Chan, W.Y.; Meenakshi, K.; Swaraja, K.; Plawiak, P.; Rajendra Acharya, U. Transfer learning techniques for medical image analysis: A review. *Biocybern. Biomed. Eng.* **2021**, *42*, 79–107. [CrossRef]

110. Yu, X.; Wang, J.; Hong, Q.; Teku, R.; Wang, S.; Zhang, Y. Transfer learning for medical images analyses: A survey. *Neurocomputing* **2022**, *489*, 230–254. [CrossRef]

111. Ayana, G.; Dese, K.; Abagaro, A.M.; Jeong, K.C.; Yoon, S.; Choe, S. Multistage Transfer Learning for Medical Images. *Artif. Intell. Rev.* **2024**, *57*, 232. [CrossRef]

112. Atasever, S.; Azginoglu, N.; Terzi, D.S.; Terzi, R. A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning. *Clin. Imaging* **2022**, *94*, 18–41. [CrossRef]

113. Kim, H.E.; Cosa-Linan, A.; Santhanam, N.; Jannesari, M.; Maros, M.E.; Ganslandt, T. Transfer learning for medical image classification: A literature review. *BMC Med. Imaging* **2022**, *22*, 69. [CrossRef]

114. De Santana, M.A.; Pereira, J.M.S.; Silva, F.L.D.; de Lima, N.M.; de Sousa, F.N.; de Arruda, G.M.S.; de Lima, R.d.C.F.; da Silva, W.W.A.; Santos, W.P.d. Breast cancer diagnosis based on mammary thermography and extreme learning machines. *Res. Biomed. Eng.* **2018**, *34*, 45–53. [CrossRef]

115. Kumar, S.; Choudhary, S.; Jain, A.; Singh, K.; Ahmadian, A.; Bajuri, M.Y. Brain Tumor Classification Using Deep Neural Network and Transfer Learning. *Brain Topogr.* **2023**, *36*, 305–318. [CrossRef]

116. Archana, B.; Kalirajan, K. An Intelligent Magnetic Resonance Imagining-Based Multistage Alzheimer's Disease Classification using Swish-Convolutional Neural Networks. *Med. Biol. Eng. Comput.* **2024**, *63*, 885–899. [CrossRef]

117. Zhu, Y.; Jin, P.; Bao, J.; Jiang, Q.; Wang, X. Thyroid ultrasound image classification using a convolutional neural network. *Ann. Transl. Med.* **2021**, *9*, 1526. [CrossRef] [PubMed]

118. Kalusivalingam, A.K.; Sharma, A.; Patel, N.; Singh, V. Enhancing Diagnostic Accuracy in Medical Imaging through Convolutional Neural Networks and Transfer Learning Algorithms. *Int. J. Artif. Intell. Mach. Learn.* **2021**, *2*, 1–24.

119. Alzubaidi, L.; Al-Amidie, M.; Al-Asadi, A.; Humaidi, A.J.; Al-Shamma, O.; Fadhel, M.A.; Zhang, J.; Santamaría, J.; Duan, Y. Novel Transfer Learning Approach for Medical Imaging with Limited Labeled Data. *Cancers* **2021**, *13*, 1590. [CrossRef] [PubMed]

120. Odusami, M.; Maskeliūnas, R.; Damaševičius, R.; Krilavičius, T. Analysis of Features of Alzheimer's Disease: Detection of Early Stage from Functional Brain Changes in Magnetic Resonance Images Using a Finetuned ResNet18 Network. *Diagnostics* **2021**, *11*, 1071. [CrossRef]

121. Guan, Q.; Wang, Y.; Du, J.; Qin, Y.; Lu, H.; Xiang, J.; Wang, F. Deep learning based classification of ultrasound images for thyroid nodules: A large scale of pilot study. *Ann. Transl. Med.* **2019**, *7*, 137. [CrossRef]

122. Tian, D.; Jiang, S.; Zhang, L.; Lu, X.; Xu, Y. The role of large language models in medical image processing: A narrative review. *Quant. Imaging Med. Surg.* **2023**, *14*, 1108–1121. [CrossRef]

123. Ogunleye, B.; Sharma, H.; Shobayo, O. Sentiment Informed Sentence BERT-Ensemble Algorithm for Depression Detection. *BDCC* **2024**, *8*, 112. [CrossRef]

124. Bora, A.; Cuayáhuitl, H. Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. *Make* **2024**, *6*, 2355–2374. [CrossRef]

125. Pu, Q.; Xi, Z.; Yin, S.; Zhao, Z.; Zhao, L. Advantages of transformer and its application for medical image segmentation: A survey. *Biomed. Eng. OnLine* **2024**, *23*, 14. [CrossRef]

126. Berroukham, A.; Housni, K.; Lahraichi, M. Vision Transformers: A Review of Architecture, Applications, and Future Directions. In Proceedings of the 2023 7th IEEE Congress on Information Science and Technology (CiSt), Agadir/Essaouira, Morocco, 16–22 December 2023; pp. 205–210. [CrossRef]

127. Van, M.; Verma, P.; Wu, X. On large visual language models for medical imaging analysis: An empirical study. In Proceedings of the 2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), Wilmington, DE, USA, 19–21 June 2024; pp. 172–176.

128. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.

129. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110. [CrossRef] [PubMed]

130. Tanimola, O.; Shobayo, O.; Popoola, O.; Okoyeigbo, O. Breast Cancer Classification Using Fine-Tuned SWIN Transformer Model on Mammographic Images. *Analytics* **2024**, *3*, 461–475. [CrossRef]

131. Abbaoui, W.; Retal, S.; Ziti, S.; El Bhiri, B. Automated Ischemic Stroke Classification from MRI Scans: Using a Vision Transformer Approach. *J. Clin. Med.* **2024**, *13*, 2323. [CrossRef] [PubMed]

132. Chen, J.; Frey, E.C.; He, Y.; Segars, W.P.; Li, Y.; Du, Y. TransMorph: Transformer for unsupervised medical image registration. *Med. Image Anal.* **2022**, *82*, 102615. [CrossRef]

133. Wang, C.; Zhang, J.; Liu, S. Medical ultrasound image segmentation with deep learning models. *IEEE Access* **2022**, *11*, 10158–10168. [CrossRef]

134. Ramamurthy, M.; Krishnamurthi, I.; Vimal, S.; Robinson, Y.H. Deep Learning Based Genome Analysis and NGS-RNA LL Identification with a Novel Hybrid Model. *Biosystems* **2020**, *197*, 104211. [CrossRef]

135. Rahman, H.; Bukht, T.F.N.; Imran, A.; Tariq, J.; Tu, S.; Alzahrani, A. A Deep Learning Approach for Liver and Tumor Segmentation in CT Images Using ResUNet. *Bioengineering* **2022**, *9*, 368. [CrossRef]

136. Zhu, Y.; Meng, Z.; Fan, X.; Duan, Y.; Jia, Y.; Dong, T.; Wang, Y.; Song, J.; Tian, J.; Wang, K.; et al. Deep learning radiomics of dual-modality ultrasound images for hierarchical diagnosis of unexplained cervical lymphadenopathy. *BMC Med.* **2022**, *20*, 269. [CrossRef]

137. Vasile, C.M.; Udriștoiu, A.L.; Ghenea, A.E.; Popescu, M.; Gheonea, C.; Niculescu, C.E.; Ungureanu, A.M.; Udriștoiu, Ș.; Drocaş, A.I.; Gruionu, L.G. Intelligent diagnosis of thyroid ultrasound imaging using an ensemble of deep learning methods. *Medicina* **2021**, *57*, 395. [CrossRef]

138. Iqbal, S.; Khan, T.M.; Naqvi, S.S.; Naveed, A.; Meijering, E. TBConvL-Net: A hybrid deep learning architecture for robust medical image segmentation. *Pattern Recognit.* **2025**, *158*, 111028. [CrossRef]

139. Obayya, M.; Saeed, M.K.; Alruwais, N.; Alotaibi, S.S.; Assiri, M.; Salama, A.S. Hybrid Metaheuristics with Deep Learning-Based Fusion Model for Biomedical Image Analysis. *IEEE Access* **2023**, *11*, 117149–117158. [CrossRef]

140. Yan, L.; Ling, S.; Mao, R.; Xi, H.; Wang, F. A deep learning framework for identifying and segmenting three vessels in fetal heart ultrasound images. *Biomed. Eng. OnLine* **2024**, *23*, 39. [CrossRef] [PubMed]