UNIVERSIDAD SAN FRANCISCO DE QUITO
USFQ

Colegio de Ciencias e Ingenierías

Cardiac lesions: a comparative study based on Convolutional Neural Networks and Visual Transformers in the segmentation of MR cardiac images.

William Sebastián Granizo López

Paula Milena Iñiguez Vaca

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero Industrial

Quito, 06 de diciembre de 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO
USFQ

Colegio de Ciencias e Ingenierías

Cardiac lesions: a comparative study based on Convolutional Neural Networks and Visual Transformers in the segmentation of MR cardiac images.

William Sebastián Granizo López

Paula Milena Iñiguez Vaca

Nombre del profesor, Título académico **María Gabriela Baldeón Calisto, PhD.**

Quito, 06 de diciembre de 2023

© DERECHOS DE AUTOR

Nombres y Apellidos: William Sebastián Granizo López
Código: 00215684
Cédula de Identidad: 1804837770

Nombres y Apellidos: Paula Milena Iñiguez Vaca
Código: 00212884
Cédula de Identidad: 1750324228

Lugar y Fecha: Quito, 6 de diciembre de 2023

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

La segmentación de imágenes cumple un rol importante para el diagnóstico, pronto diagnóstico y monitoreo correcto de enfermedades cardiovasculares. Esta tarea manual se realiza por radiólogos, los cuales pueden inducir error por fatiga visual. Para ello, se han desarrollado algoritmos de inteligencia artificial, específicamente de Deep Learning, los cuales han ayudado a incrementar la precisión en los diagnósticos por parte de los médicos. Tradicionalmente, los Convolutional Neural Networs (CNN) se han optado como los mejores modelos para la segmentación de imágenes; sin embargo, debido a sus limitaciones, se han propuestos alternativas con los modelos de Visual Transformers (ViT).

En esta investigación, realizamos una comparación estadística entre estos dos modelos para la segmentación de imágenes cardiacas MR del dataset ACDC. Se escogieron 5 modelos por cada tipo, se entrenaron y se compararon las métricas de Dice Coefficient y ASSD de forma estadística. Se encontró que los CNNs son más robustos en la segmentación de imágenes cardiacas sin la utilización de *data augmentation*, mientras que se observaron grandes oportunidades con los ViTs para tareas más complejas dado a su mecanismo de *self-attetion*.

Palabras Clave: estudio comparativo, segmentación de imágenes, enfermedades cardiovasculares, CNNs, ViT, Deep Learning.

ABSTRACT

Image segmentation plays a crucial role in the accurate and timely diagnosis and monitoring of cardiovascular diseases. This task is typically performed by radiologists, who may introduce errors due to visual fatigue. To address this, artificial intelligence algorithms, specifically based on Deep Learning, have been developed, aiding in increased diagnostic precision for medical professionals. Conventionally, Convolutional Neural Networks (CNNs) have been preferred for image segmentation; however, due to their limitations, alternatives such as Visual Transformers (ViTs) have been proposed.

In this research, we conducted a statistical comparison between these two models for the segmentation of cardiac MR images from the ACDC dataset. We selected five models for each type, trained them, and statistically compared the Dice Coefficient and Average Symmetric Surface Distance (ASSD) metrics. The findings revealed that CNNs exhibit greater robustness in cardiac image segmentation without the use of data augmentation. On the other hand, ViTs showed significant potential for more complex tasks, given their self-attention mechanism.

Keywords: comparative study, image segmentation, cardiovascular diseases, CNNs, ViT, Deep Learning.

TABLE OF CONTENTS

TABLES INDEX

FIGURE INDEX

INTRODUCTION

Cardiovascular diseases (CVDs) are the number one cause of deaths in the world; in 2021 20,5 million of people died because of them. According to the World Heart Federation (WHF), CVD was the leading cause of death worldwide in 2021 [1]. Nowadays, the quantity of heart diseases that have been identified by doctors is huge and every one of them affects in a unique way a patient's cardiovascular health involving the heart, the arteries, even the circulatory system [2]. Some of these pathologies include hypertrophic cardiomyopathy, a pathology involving the heart muscle; myocardial infarction, a disease caused by the obstruction of one of the coronary arteries that supplies blood to the heart; dilated cardiomyopathy, a pathology of the heart muscle that results in the enlargement of the ventricles causing the heart to function inadequately and not pumping enough blood to the rest of the body; abnormal right ventricle, a condition in which the right ventricle of the heart experiences an abnormality, either in its structure or its performance. All these diseases are the leading causes of sudden death among young individuals and can result in heart failure and stroke-related functional limitations, a huge portion of those affected go undetected [3], [4], [5], [6]. Historically, the statistics show the quantity, but it does not show the pain and the difficulties that people who have these diseases face every day. Therefore, the tools to detect it on time should be prioritized and investigated.

Over the years, healthcare professionals have created ways to diagnose these diseases and search for an accurate treatment for the patients. Imaging techniques such as magnetic resonance image (MRI), chest X-ray, computed tomography, and ultrasound are widely used to assess cardiac structures and their functions, and aid in disease diagnosis, monitoring, and treatment planning [7], [8], [9], [10]. All these methods are used for identifying anatomical structures and diagnosing pathologies.

Recognizing and segmenting specific anatomical structures in medical images such as MRI, doctors can better understand the location and extent of diseases, which can lead to timely treatment. Cardiac image segmentation is a crucial step in the diagnosis of diseases and can have a significant impact on patient outcomes [11].

Image segmentation is a technique where the input image is split into regions or segments with similar characteristics. Hence, segmentation is a precise way to identify anatomical structures, to facilitate and automate certain tasks in the field of radiology [12]. Accurate segmentation of cardiac images holds particular significance when it comes to the preparation and supervision of cardiac intervention procedures, as well as evaluating cardiac performance and identifying cardiac conditions [13]. To evaluate cardiac anatomy in a medical image, it is necessary to identify some parts of the heart like for example: the left ventricular (LV) endocardium, the left ventricular epicardium (or myocardium - MYO) and the right ventricular (RV) endocardium. Additionally, evaluation of the function of the left and right ventricles should be performed [14]. Another important part of the heart to analyze is the left atrium [15].

The analysis of parts of the heart in an MRI is performed by a radiologist so the results are exposed to human error. The segmentation work of a radiologist nowadays is mostly done manually. Therefore, it is not only inefficient but also tedious to complete due to the demand that exists in reading medical images. Furthermore, it depends on the perspective of the specialist, and it can be subject to errors such as exhaustion or distractions on the part of the radiologist. For this reason, the task consumes time and is prone to intra- and inter-observer variability [16], [17], [18]. This task continues to be semi-automatic because of the lack of accuracy of fully automatic cardiac segmentation methods. These methods have limitations in the cardiac medical field, such as over-segmentation and high sensitivity to noise, additionally, the segmentation process entails

longer processing time and some of the features involved become redundant, resulting in a lack of precision in the results obtained [17].

In the last decade, deep learning techniques have been used to improve the accuracy and efficiency of medical image segmentation. Moreover, within deep learning, Convolutional Neural Networks (CNNs) have become the preferred methods due to their outstanding performance in image analysis tasks. CNNs have the advantage of learning about the characteristics and the complex patterns in an automatic manner. Also, they can process a huge amount of data and identify patterns regardless of their location in the image [19].

In the last two years, Vision Transformers (Vit) have sparked the interest of researchers for image analysis tasks. ViT are a variation of the Transformers models, but their main difference is that Transformers were originally designed for processing sequences of words, whereas ViT are tailored for image data. These models use an attention structure called "multi-head attention," enabling them to focus on distinct parts of the input image simultaneously, which is a significant advantage [20]. One advantage of ViT models is their capability to generalize effectively to new input images, making them versatile for a wide range of image processing applications. Additionally, ViT are flexible in terms of input image size and resolution.

CNNs and ViT have become the most widely used approaches for medical image segmentation [21], obtaining state-of-the-art performance in various benchmarks [22], [17], [16], [23]. CNNs are a popular choice due to their ability to learn relevant features from images and their ability to process enormous amounts of data. Vision Transformers, on the other hand, are selected due to their capacity to capture large-scale context information and their ability to process high-resolution images. In medical image segmentation some CNN and ViT models include Grid Net [24], MA-Net [25],

FCN [26], U-Net [14], U-Net ++ [27], DeepLabV3 [28], DeepLabV3+ [29], LinkNet [30]. Lvit [31], TransFuse [32], TransBTS [33], TransBridge [34], TransUNet [35], MissFormer [36], ScaleFormer [37], SegFormer [38], SwinUnet [39].

Both, Vision Transformers and Convolutional Neural Networks have a neural network architecture. Although, there are some differences, for example, the mechanism of attention. ViT uses multiple attention or "multi-head" while, CNN focuses its attention on local characteristics using convolutional layers in specific segments of the image [40], [41]. Another interesting difference is flexibility in applications, CNN could be re-used for more than one project just changing or adjusting the last layers (Full connected layers) to the specific purpose. Meanwhile, ViT uses a lot of resources to be applied so in a different context it would be difficult for it to work effectively [42]. Additionally, ViT models have something called "weak inductive bias" which is related to the amount of data that the model needs to be trained. So, while ViT needs a great amount of data for the model to be well trained, CNNs does not need much data to be well trained. This is due to pooling layers, and its capacity to recognize visual patterns regardless of the location in the image [43].

Recently, some of the comparative studies between CNNs and ViTs have shown some improvement in terms of choosing the best model and seeing the differences in application [40], [43], [44], [45]. Selecting the optimal tool makes it easier to satisfy clinical needs so, this research not only could help in the advances of the medical field, but also promotes the reduction of costs and times; in this way, quality of patient care is improved by the automation and optimization of clinical processes, which leads to more precise and timely diagnoses, reducing mortality [46].

The demand for image segmentation tasks is massive, it is an essential and fundamental component of computer vision and machine learning applied to medicine

[47]. Since this is a demanded task, new opportunities are emerging to develop artificial intelligence tools that can directly support healthcare professionals in diagnosing CVDs [48]. With technological evolution, new models are continually being developed to address specific goals in medical image segmentation so, a comparative study between these models could develop an accurate tool for the job and in an unbiased manner determine which model is statistically better than others.

In this paper, we perform a comparative study between Transformers and CNNs for cardiac image segmentation. After a thorough literature review, 5 CNNs architectures and 5 ViT architectures were selected for comparison. Namely, the U-Net, U-Net ++, DeepLabV3, DeepLabV3+, LinkNet, MissFormer, TransUNet, ScaleFormer, SegFormer and SwinUNet. The architectures are evaluated on the Automated Cardiac Diagnosis Challenge (ACDC) dataset [16]. Two widely used metrics are used for evaluation, the Dice Coefficient and Average Symmetric Surface Distance (ASSD) [49], [50]. To obtain statistically significant conclusions, we perform a one-way ANOVA followed by a Tukey test. Our results demonstrate that CNNs models are statistically better than Transformers in terms of Dice Coefficient and ASSD, in addition to having lesser trainable parameters. Furthermore, the model that achieved the best performance was Linknet with a 0,90 mean dice coefficient and 0,30 ASSD mean.

DEVELOPMENT

Related work

Image segmentation is one of the earliest challenges in the field of computer vision. The initial endeavors in this field date back to as early as 1970-72 when researchers started exploring traditional methods like region growing and optimization techniques. The first one is a segmentation method based on the approximation of regions. It sets the starting points in the image (seeds) and then grows the region around

these seeds, adding pixels that have similar characteristics to those of the seed. Optimization techniques for image segmentation involve tuning neural network models to minimize a loss function adjusting the model parameters [51], [52], [53]. Before the year 2000, a lot of methods were applied in the field of digital image processing, some of these methods are: Clustering, Threshold, Region, and Edge Segmentation [54]. Sanchez-Ortiz et al. [55] defines the Clustering algorithm and mentions that it uses the intensity distribution of the image to group pixels into regions and assigns them a degree of membership. On the other hand, Santiago, C et al. [56] defines the Edge segmentation method saying that it uses the information of the edges in the vicinity of the shape model to detect the structure of interest. Liu, T et al. [57] provides a definition of the Threshold method, which involves choosing a threshold value and classifying each pixel as part of the object or the background according to whether its intensity is greater or less than that threshold. Pairs of thresholds ensure the pixel points from the region of interest are not excluded. Meanwhile, Galea, R et al. [58] deals with the Region segmentation method and says that this method is based on detecting the area of interest using a localization procedure, which is then fed to the segmentation network.These historic advances in image segmentation laid the foundation for the evolution of neural networks models used today for cardiac image segmentation. Then, the convolutional neural networks were benefit from technological advances and knowledge accumulated over the years to achieve outstanding accuracy in the segmentation of cardiac structures in medical images, thereby improving the diagnosis and treatment of cardiac disorders. There are many titles where it specifies modifications of a base CNN structure [16], [47], [59]. Since most algorithms applied in medical image segmentation fall under the category of Convolutional Neural Networks or its variations, this architecture has been proof as one of the most used and common in

the field of image segmentation, especially when its medical images [60], [61], [62]. Zotti et al. [14] presented a U-Net based architecture with a multi-resolution grid architecture; the network learns high- and low-level features useful for recording and segmenting cardiac anatomy. The model was tested in the ACDC dataset and achieved a Dice Score of 0,91. El-Taraboulsi, T et al. [63] introduced a V-Net model which employs a reversible mechanism and asymmetrical convolutions maintaining image size and quality so, V-Net can train high-quality images on a single GPU. The model was tested in MICCAI dataset and has a Dice Score of 0,92. Holger R. Roth et al. [64] bring in a Fully Convolutional Network model that is trained in an end-to-end, allowing it to learn features and make predictions at the pixel level. It has a cascaded approach, where a second-stage FCN is employed to focus more on boundary regions, improving the accuracy of segmentation results. This model has shown promising results in achieving state-of-the-art segmentation performance in medical imaging applications having a Dice Score of 0,82. Zhanwei Xu et al. [65] show a model that combines Faster R-CNN and U-net Network for efficient segmentation. It uses a region proposal Network, a 3D U-net Network, and an Edge-loss head to achieve competitive segmentation performance with reduced computational cost and inference time. The model was tested in MM-WHS2017 dataset and has a Dice Score of 0,86. Mahendra Khened et al. [26] propose a model that uses multiple scales and dense residual connections for cardiac segmentation and automated diagnosis. The network uses a dual loss function that combines the advantages of cross entropy loss and Dice loss. Furthermore, an expert classifier is proposed to improve the classification accuracy of patients with specific pathologies. The model was tested in ACDC dataset and has a mean Dice Score of 0,91. Ange Lou et al. [66] introduce a model called DC-UNet which uses a dual-channel CNN block to provide more effective features with fewer parameters and replaces the

skip connection between encoder and decoder with a residual module. DC-UNet has been evaluated on three datasets with tough cases and has shown a relative improvement in performance compared to the classical U-Net model. This model was tested in two datasets: ISBI-2012 Electron Microscopy dataset and the CVC-ClinicDB dataset getting an accuracy of 92.71%.

These historic advances in image segmentation laid the foundation for the evolution of neural networks models used today for cardiac image segmentation. Then, the convolutional neural networks were benefit from technological advances and knowledge accumulated over the years to achieve outstanding accuracy in the segmentation of cardiac structures in medical images, thereby improving the diagnosis and treatment of cardiac disorders. There are many titles where it specifies modifications of a base CNN structure [16], [47], [59]. Since most algorithms applied in medical image segmentation fall under the category of Convolutional Neural Networks or its variations, this architecture has been proof as one of the most used and common in the field of image segmentation, especially when its medical images [60], [61], [62]. Zotti et al. [14] presented a U-Net based architecture with a multi-resolution grid architecture; the network learns high- and low-level features useful for recording and segmenting cardiac anatomy. The model was tested in the ACDC dataset and achieved a Dice Score of 0,91. El-Taraboulsi, T et al. [63] introduced a V-Net model which employs a reversible mechanism and asymmetrical convolutions maintaining image size and quality so, V-Net can train high-quality images on a single GPU. The model was tested in MICCAI dataset and has a Dice Score of 0,92. Holger R. Roth et al. [64] bring in a Fully Convolutional Network model that is trained in an end-to-end, allowing it to learn features and make predictions at the pixel level. It has a cascaded approach, where a second-stage FCN is employed to focus more on boundary regions, improving the

accuracy of segmentation results. This model has shown promising results in achieving state-of-the-art segmentation performance in medical imaging applications having a Dice Score of 0,82. Zhanwei Xu et al. [65] show a model that combines Faster R-CNN and U-net Network for efficient segmentation. It uses a region proposal Network, a 3D U-net Network, and an Edge-loss head to achieve competitive segmentation performance with reduced computational cost and inference time. The model was tested in MM-WHS2017 dataset and has a Dice Score of 0,86. Mahendra Khened et al. [26] propose a model that uses multiple scales and dense residual connections for cardiac segmentation and automated diagnosis. The network uses a dual loss function that combines the advantages of cross entropy loss and Dice loss. Furthermore, an expert classifier is proposed to improve the classification accuracy of patients with specific pathologies. The model was tested in ACDC dataset and has a mean Dice Score of 0,91. Ange Lou et al. [66] introduce a model called DC-UNet which uses a dual-channel CNN block to provide more effective features with fewer parameters and replaces the skip connection between encoder and decoder with a residual module. DC-UNet has been evaluated on three datasets with tough cases and has shown a relative improvement in performance compared to the classical U-Net model. This model was tested in two datasets: ISBI-2012 Electron Microscopy dataset and the CVC-ClinicDB dataset getting an accuracy of 92.71%.

Just like CNNs, Vision Transformers models are also a type of neural network architecture. When discussing transformers, it is a widespread practice to associate them with natural language processing (NLP), rather than image-related tasks. However, transformers have already been used for image segmentation tasks, the problem or limitation is that medical image is usually different from a natural image. The most used thing about a transformer model is their attention capacity since combined with the

convolutional layers they achieve a deep analysis so, Transformers are a powerful architecture, but, since they are images, they merge with other architectures to achieve this segmentation task. The most common thing is to have a fusion between CNN and Transformers. These hybrid models are the most used because they mitigate the limitations of transformers and use their strengths for this type of computational task [67], [68]. Yutong Xie et al. [69] presented the CoTr framework which uses CNN to extract feature representations from 3D medical images and DeTrans to model long-range dependency on the extracted feature maps. The DeTrans uses a deformable self-attention mechanism to reduce computational and spatial complexities. The model was tested in the Beyond the Cranial Vault (BCV) database and achieved a Dice Score of 0,82. Yunhe Gao et al. [70] introduce UTNet, a hybrid architecture that combines Transformer's self-attention mechanism with a convolutional neural network for medical image segmentation. It captures long-range associative features and dynamically aggregates relevant features. The network is trained from scratch and evaluated on a multi-label, multi-vendor cardiac magnetic resonance imaging cohort database and achieved a 0,88. Boxiang Yun et al. [71] shows an architecture named SpecTr which uses transformers to learn contextual features across spectral bands in a U-shape architecture. It decomposes the input hyperspectral image into spectral images, applies depth-wise convolution, spectral normalization, and transformers with sparsity constraint to produce spatial-spectral contextual feature maps, which are decoded to generate the segmentation map. The model was tested un a multi-dimensional choledoch dataset and achieved a Dice Score of 0,78. Ali Hatamizadeh et al. [72] propose the UNETR model that uses transformers to capture long-range dependencies in 3D medical image segmentation. It divides the input volume into patches, projects them into an embedding space, and adds a positional embedding to preserve spatial

information. UNETR achieves state-of-the-art performance on two public datasets: the Vanderbilt Body CT (BTCV) and the Medical Segmentation Decathlon (MSD) getting a Dice score of 0,90. Chang Yao et al. [73] introduce a model called Transclaw U-Net which combines convolution and transformer operations for better medical image segmentation by using self-attention and combining encoding, upsampling and decoding parts. This model is tested in the Synapse Multi-organ Segmentation Dataset and achieved a Dice coefficient of 0,78.

Methodology

In this section, we begin by describing the CNN and ViT models selected for comparison and the hyperparameter values applied for each architecture. We then describe the dataset in which the models are tested, and evaluation metrics utilized. Finally, we present the statistical test performed to analyze the results.

Selection of the models.

After conducting research on the models that have already been applied in medical image segmentation, the most used and most mentioned in scientific literature were Convolutional Networks and Vision Transformers. The models for CNN architecture are the following: UNet++ [74], DeepLab [75], DeepLab+ [60], UNet [14], LinkNet [76]. And the models of Vision Transformers are: MissFormer [36], SegFormer [38], SwinUNet [39], ScaleFormer [37] and TransUNet [35]. These models have been selected because of their state-of-the-art results in medical image segmentation, natural images segmentation, and its code is open source.

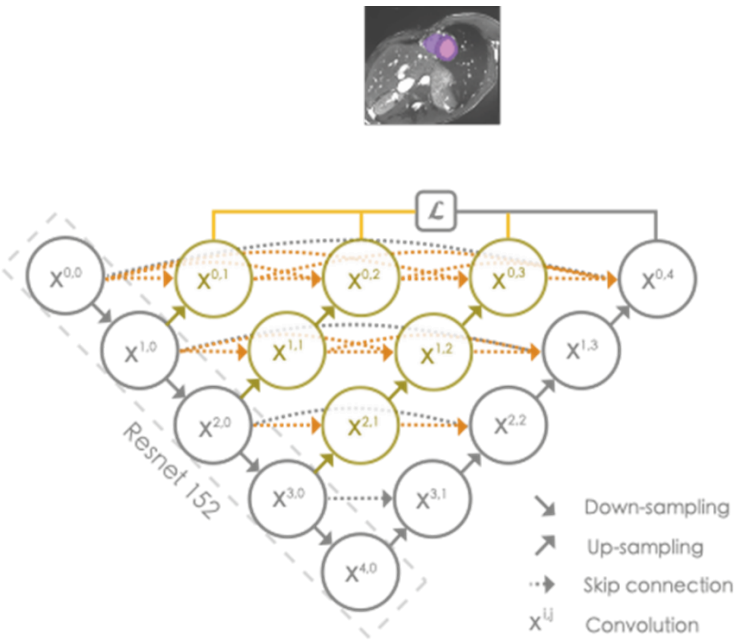
Information of each model and training details.

Convolutional Neural Networks models.

The first model is t*he U-Net++*. The base of this architecture are the nested and dense skip connections [74], which at the feature extraction stage can capture low level features such as brightness and image texture [77]. This model is recognized for its proficiency in effectively capturing intricate details within primary objects, which simplifies the learning process. UNet++ comprises two key components: an encoder and a decoder, connected through multiple interconnected blocks for processing. The primary concept behind the development of this model is to bridge the gap between the features extracted by the encoder and the prerequisites of the decoder before their fusion [74]. UNet++ operates across multiple levels of complexity, incorporating redesigned connections that link the decoder and encoder at equivalent levels of granularity [78]. By incorporating U-Nets of varying depths within its structure, all sharing a single encoder and interweaving the decoders [78], this approach eliminates the need to determine the network's depth. Consequently, it improves overall segmentation performance, accelerates prediction speed, and prevents restrictive connections that might hinder information exchange between the decoder and encoder, especially when their processing features are at the same level of detail. The model's hyperparameters were the same as in the investigation made by Zhou et al. [74]. Furthermore, a Resnet152 encoder was implemented considering the findings from Kim et al. [79]. The U-net++ model is depicted in Figure 1.

Figure 1

UNet++ architecture



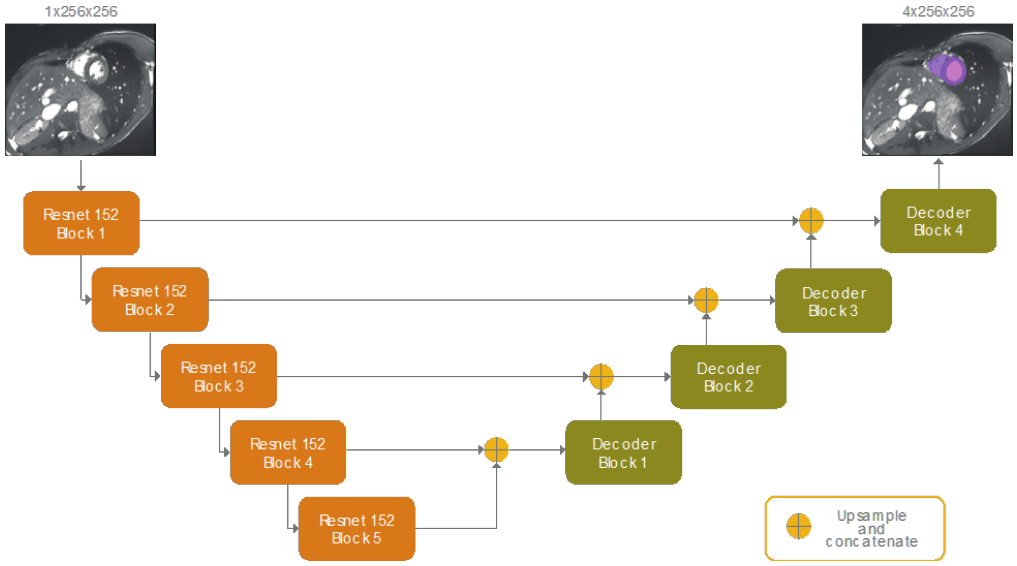
Note: Adapted UNet++ Model of Zhou et al. [57] with Resnet152 as encoder.

The second model is U-Net which is a U-shaped encoder-decoder network architecture, which consists of four encoder blocks and four decoder blocks that are connected via a bridge, it is a symmetrical architecture [80]. U-Net is a top-performing cardiovascular magnetic resonance (CMR) segmentation model [63]. U-Net which is the most important semantic segmentation framework of CNN and works well for pixel-level prediction tasks [81]. Consists of a convolutional encoder followed by a decoder composed of upward convolutions combined with skip connections [82]. The skip connections provide additional information that helps the decoder to generate better semantic features [83]. The U-Net network structure has multi-scale skip connections and a learnable up-convolution layer, which becomes a popular method for medical image segmentation. Additionally, this architecture uses an “overlap-tile strategy” to tackle large images while minimally impacting processing power [63]. For the training

details we chose the same as for UNet++ because the objective was to analyze the improvement of these models using the same training and compare the results. The U-net model is depicted in Figure 2.

Figure 2

UNet architecture



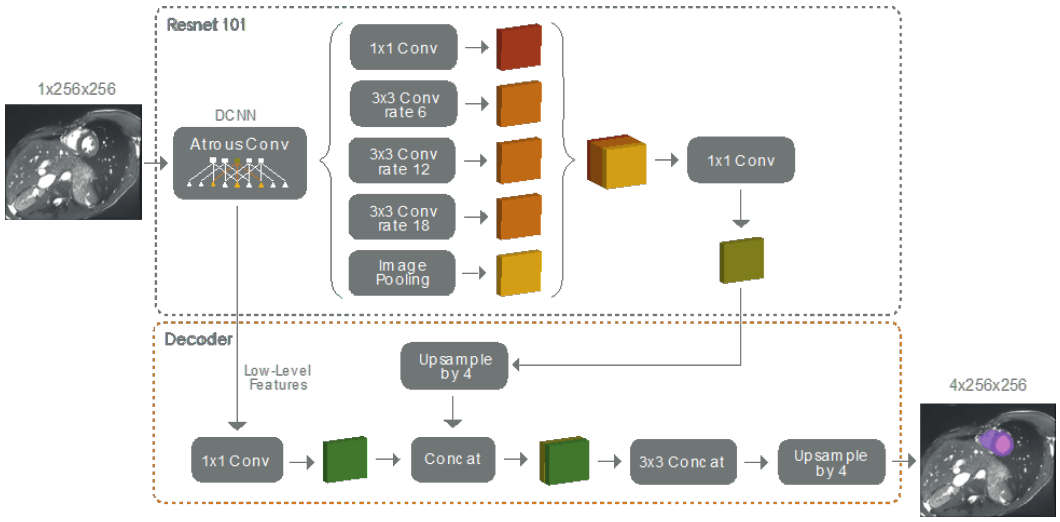
Note: Adapted UNet Model of Neven & Goedemé [84] with Resnet152 as encoder.

The third model used is DeepLabV3+ which is an enhanced network consisting of two main components: an encoder and a decoder. The encoder's role is to train the network to obtain feature maps and capture high-level semantic information [85]. While the decoder, is responsible for projecting these learned features from the encoder to the pixel space to achieve pixel segmentation. The encoder comprises a core network, often based on a traditional CNN such as a ResNet, also, employs a residual module, which enables the network to focus on learning the difference or the "residue" between the input and the desired output, simplifying the training process. Additionally, it uses the atrous convolution which is a technique that expands the convolutional feature map during convolution operations to increase the receptive field allowing each

convolutional output to encompass an impressive range of information [85]. This model significantly boosts the effectiveness of segmentation by combining with an atrous convolution [86]. We implement DeepLabV3+ with a Resnet101 encoder block and training hyperparameters as mentioned in Chen et al. [87]. The DeepLabV3+ model is depicted in Figure 3.

Figure 3

DeepLabV3+ architecture



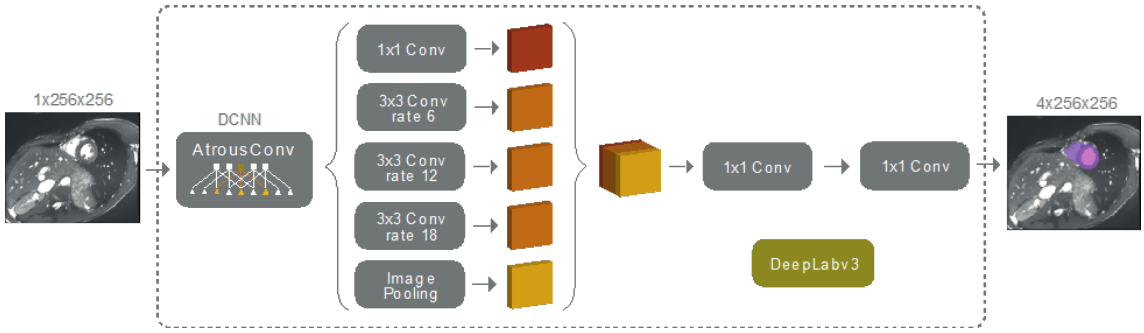
Note: Adapted DeepLabV3+ Model with Resnet101 as encoder of Chen et al. [87].

On the other hand, DeepLabV3 addresses two crucial aspects. Firstly, it deliberately sacrifices fine-grained details in features to facilitate the learning of more abstract representations. However, this capacity to overlook minor variations can be problematic in tasks that demand intricate information within specific regions. To tackle this issue, the DeepLab architecture relies on a technique called Atrous Convolution, also known as the dilated convolution, which serves as its principal component [88]. This model initiates with the training of a deep convolutional network, followed by the

transformation of all fully connected layers into convolutional layers. It enhances the feature resolution through atrous convolutional layers, which allows responses about features, for example, instead of calculating every 32 pixels, this charge will be reduced to 8 pixels, making it more efficient. Consequently, DeepLab offers several advantages: first, it exhibits enhanced speed due to the utilization of atrous convolutions; second, it delivers heightened precision by providing accurate results on challenging data; and third, it maintains simplicity by comprising two well-established modules [89]. For DeepLabV3 training details we chose the ones selected for DeplabV3+ to see if there was an improvement of the results under the same conditions. The DeepLabV3 model is depicted in Figure 4.

Figure 4

DeepLabV3 architecture



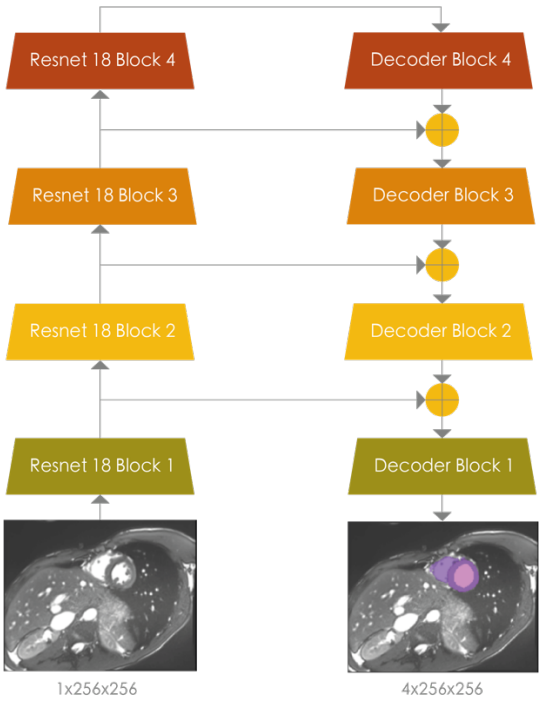
Note: Adapted DeepLabV3 Model of [90] with Resnet101 as encoder.

The last model studied is LinkNet which is a deep neural network architecture, used in problems where multi-class segmentation is required (which can be used in unmanned vehicles and some more areas) [76]. This model pushes the performance by introducing global context encoding module and geometrical layout encoding module. The objective of the LinkNet is to recover loss spatial information that can be used by the decoder and its up-sampling operations. In addition to how the decoder shares

knowledge with the encoder, this provides a network with a lower use of parameters, providing an efficient network [91]. After each data reduction phase, the feature maps of the encoder layer are combined with the feature maps of the decoder layer that have the same resolution [82]. The training details are the same as the ones mentioned by Chaurasia et al. [76]. LinkNet used RestNet18 as the encoder block, which is lighter than the others used encoders. The LinkNet model is depicted in Figure 5.

Figure 5

LinkNet architecture



Note: Adapted LinkNet Model with Resnet18 as encoder of Chaurasia et al. [76].

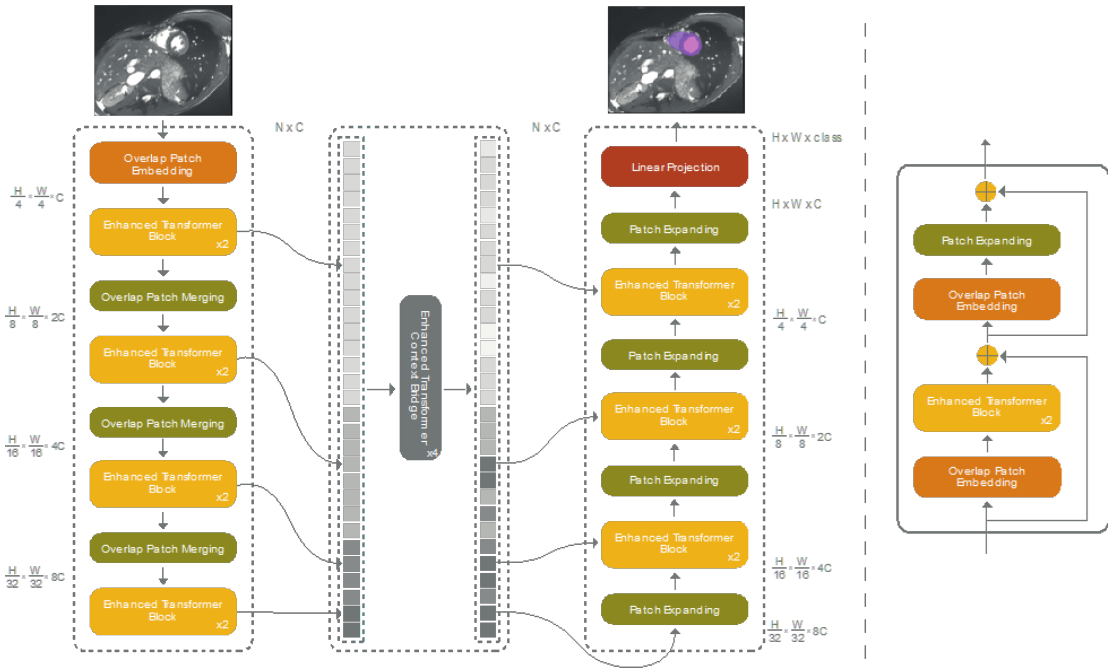
Vision Transformers models.

The first chosen ViT is MISSFormer, which was proposed by Huang et al. [36]. This model is an encoder-decoder architecture with a transformer bridge attached between them. Its main characteristic is to take the multi-scale information of the

encoder and extract both the long-range dependencies and local context; this allows the model to learn more comprehensive representations of medical image segmentation. The multi-scale features come from the overlapping patches, which capture the global and local from the input image. Then, they pass to the hierarchical enhanced transformer blocks and merging layers; this process does not need computational complexity. The merging layers oversee downsampling features by merging the overlapping patches. The resultant data goes to the bridge and then to the decoder. It upsamples and concatenates the information to recover the original image size and the segmented map. The model training details are the same as those mentioned in the investigation by Huang et al. [36]. The MissFormer model is depicted in Figure 6.

Figure 6

MISSFormer architecture

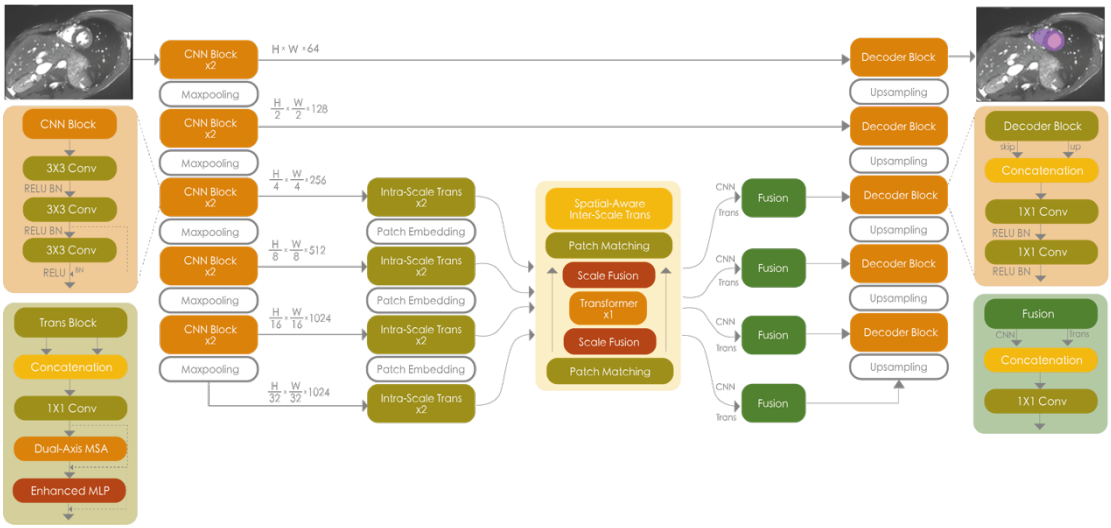


Note: Adapted MISSFormer Model of Huang et al. [36].

The second model proposed by Huang et al. [37] is ScaleFormer. It uses the ResUNet as the backbone to extract local features hierarchically. Thus, there are convolutional blocks, which makes this algorithm hybrid. The purpose of the encoder is to go deeply and find different fine-grained features in local-level details. Then, the intra-scale transformer takes global-level features from the previous information. The main characteristic of the transformer block is to compare the captured scales and model the mutual information of the objects. The output joins with the CNN's features to pass to the decoder and find the segmented map. The model training details are the same as those mentioned in the investigation [37]. The ScaleFormer model is depicted in Figure7.

Figure 7

ScaleFormer architecture



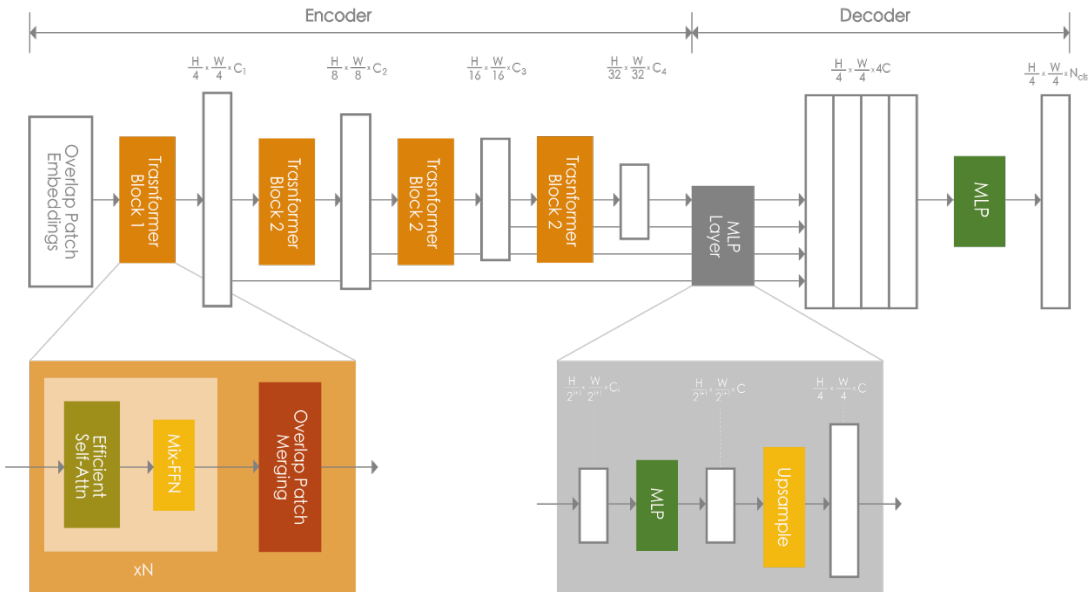
Note: Adapted ScaleFormer Model of Huang et al. [37].

On the other hand, the proposed model by Xie et al. [38] is Segformer, this model was the base for other models, such as MISSFormer [36]. The authors tried different Mix Transformer (MiT) encoders, in which the MiT-B0 is the lightweight, and MiT-B5 is the largest with better performance. In this case, the selected encoder is the first due

to its efficiency, compact architecture, and convenience for real-time application [38]. The first step of the model is to divide the input image with patches of size 4x4. This output goes to the encoder, which has hierarchical transformer blocks and generates features at multi-scale maps. These low and high-resolution features help boost the performance of the model. The resulting information goes to the All-MLP decoder, which consists of multiple layer perceptron (MLPs). This method is more lightweight than others, including the decoders from CNN. It is possible due to the skip connections that oversee combining features from different scales. This decoder aggregates the information from the encoder's layers and then produces the segmented mask. The model training details are the same as those mentioned in the investigation [38]. The SegFormer model is depicted in Figure 8.

Figure 8

SegFormer architecture

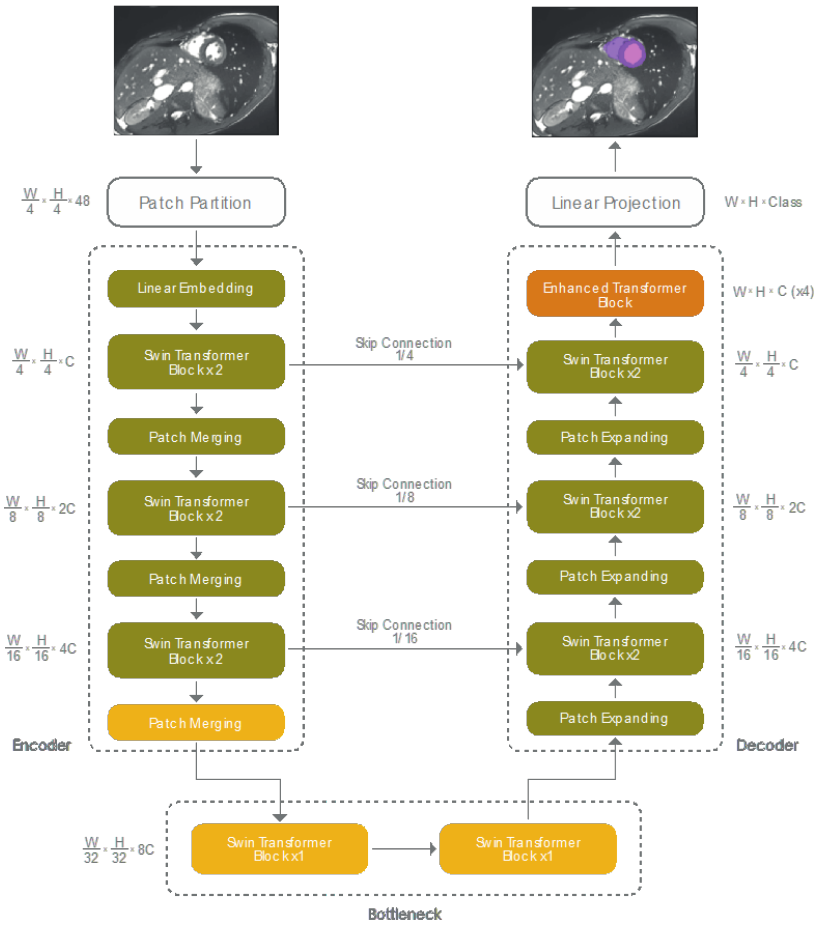


Note: Adapted SegFormer Model of Xie et al. [38].

The next model is the SwinUNet proposed by Cao et al. [39]. The architecture is an encoder, bottleneck, decoder, and skip connections. First, the encoder receives sequence embedding from splitting the input images by non-overlapping patches of size 4x4. This information passes through a linear embedding layer for projecting feature dimensions into arbitrary dimensions. Then, the encoder extracts high-level features and global and long-range semantic information. Patch merging blocks are between the Swin transformers blocks, and they oversee downsampling and increasing dimensions. The data goes to the decoder, whose u-shape is inspired by the U-Net model. This part of the model has Swin Transformers and patch expanding layers that fuse the extracted features with multi-scale features from the encoder. To do so, skip connections help to preserve spatial information. The last layer of the model returns the resolution of the input image. In this case, the output is a pixel-wise segmentation map in which each pixel is assigned to a class of the corresponding label or region. The model training details are the same as those mentioned in the investigation [39] The SwinUnet model is depicted in Figure 9.

Figure 9

SwinUnet architecture



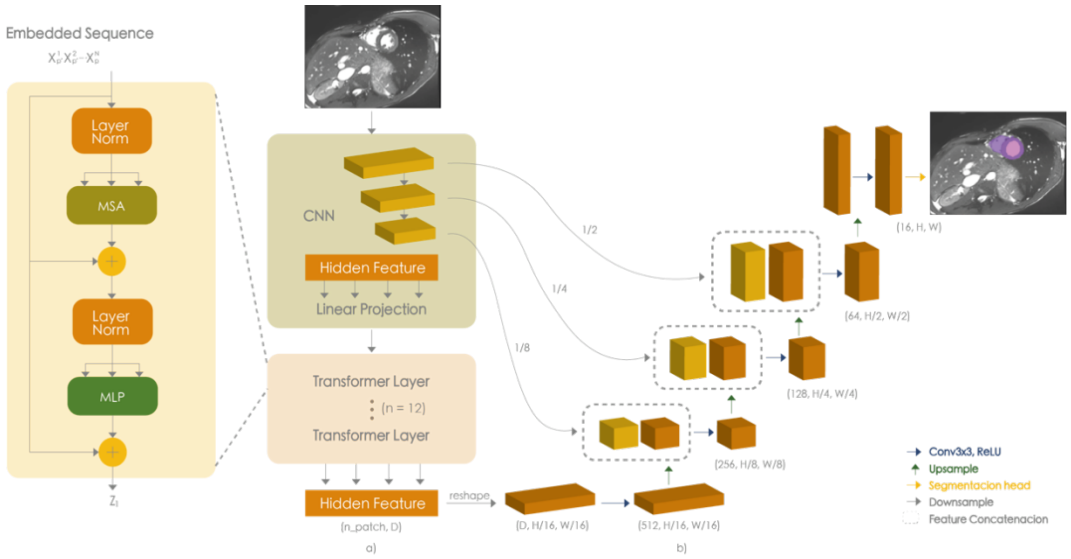
Note: Adapted SwinUNet Model of Cao et al. [39].

The last model for ViT is the TransUNet proposed by Chen et al. [35]. In this case, the model uses a self-attention mechanism in the decoder via a transformer. Due to this process in the encoder, we assigned this model as a hybrid. First, the input image passes through a CNN backbone, which extracts high-level features. The output of the first encoder does not go directly to the transformer encoder but to the patch embedding layer. On the other hand, the transformer encoder captures long-range dependencies and high-level semantic information. The features pass to the transformer decoder. The authors introduced a cascade upsampler (CUP) to output the final segmentation mask

with multiple upsampling steps. The model training details are the same as those mentioned in the investigation [35] The TransUnet model is depicted in Figure 10.

Figure 10

TransUnet model architecture



Note: Adapted TransUNet Model of Chen et al. [35].

Training Details.

For all models we used a loss function hyperparameter. It measures the difference between model's estimation from input to output, and ground truth value so its principal task is fitting the model to the given training data [92], [59]. The most used loss function for image segmentation is Dice Loss, Tversky Loss Function and Cross Entropy [93]. In fact, Dice Loss and Cross Entropy are the loss function more representative in CNN models, both use weighted loss terms to overcome class imbalance [94]. For the CNN models, the selected loss function was Dice Loss. On the other hand, for ViT models, the loss function depends on how the authors trained their models (e.g., Huang et al. trained Miss Former with weights of 0.6 and 0.4 for Dice Loss and Cross Entropy Loss, respectively. [36])

To train any model, it is also necessary to use a non-linearity or activation function. For this study, the selected one is SoftMax, which is used for the task of multiclass classification. It modifies a vector of real numbers into a probability distribution, each one of them associated to a class. This function is typically used in the final layer of the neural network [95], [96].

Besides, we used two optimizers and applied one of them depending on the architecture of each model. The two optimizers are Adaptive Moment Estimation (Adam) and Stochastic Gradient Descent (SGD). The first one computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. Also, includes bias correction and momentum term [97]. The second one (SGD) is a variant of gradient descent that uses a random subset of the training data to estimate the gradient of the loss function. Instead of computing the gradient of the entire dataset, SGD computes the gradient of a randomly selected mini batch of the dataset. This makes it computationally efficient and allows it to scale to large datasets [98].

To train the models, 200 epochs were established as the standard due to the usage of non-pretrained models [36] [37]. Like the hyperparameters mentioned above, the learning rate depended on how each author trained the model. For the CNNs the chosen optimizer was Adam, and for ViT, it depended on the methods used by the authors, in general the chosen one was SGD. It is important to mention that data augmentation was not performed to avoid randomness. Table 1 presents the hyperparameters and number of parameters for CNNs and Table 2 for ViT. It is worth mentioning that all the models are implemented with Python 3.10 and run in the Google Collab environment using an NVIDIA V100 GPU.

Table 1

Hyperparameters for CNNs

Model	Optimizer	Learning Rate	Epochs	Batch Size			Activation Function	Loss Function	Parameters
				Train	Validation	Test			
UNet	Adam	3e-4	200	16	16	1	SoftMax	Dice Loss	83'615,684
Unet++	Adam	3e-4	200	16	16	1	SoftMax	Dice Loss	67'151,044
DeepLabV3	Adam	7e-3	200	6	6	1	SoftMax	Dice Loss	58'620,356
DeepLabV3+	Adam	7e-3	200	6	6	1	SoftMax	Dice Loss	45'664,212
LinkNet	Adam	5e-4	200	10	10	1	SoftMax	Dice Loss	11'657,604

Table 2

Hyperparameters for ViT

Model	Optimizer	Learning Rate Momentum Weight Decay	Epochs	Batch Size			Activation Function	Loss Function	Parameters
				Train	Validation	Test			
SwinUNet	SGD	1e-3 0.9 1e-4	200	24	24	1	SoftMax	Dice Loss (0.6) + Cross Entropy (0.4)	2’7168,420
TransUNet	SGD	1e-2 0.9 1e-4	200	6	6	1	SoftMax	Dice Loss (0.5) + Cross Entropy (0.5)	105'912,260
SegFormer	AdamW	6e-5 - -	200	16	16	1	SoftMax	Dice Loss (0.6) + Cross Entropy (0.4)	7'718,244
ScaleFormer	SGD	3e-3 0.9 1e-4	200	8	8	1	SoftMax	Dice Loss (0.6) + Cross Entropy (0.4)	113'814,164
MISSFormer	SGD	1e-3 0.9 1e-4	200	24	24	1	SoftMax	Dice Loss (0.6) + Cross Entropy (0.4)	42'462,212

Database & Preprocessing.

The Automated Cardiac Diagnosis Challenge (ACDC) comprises of 4D cardiac cine-MR images from 150 patients with the corresponding ground truth segmentation of the left ventricle cavity (LVC), left ventricle myocardium (LVM), and right ventricle cavity (RVC) on end diastolic and end systolic phases. The images were collected over a 6-year period from the University Hospital of Dijon using two scanners with different intensities, which resulted in a variation in image resolution [99]. The patients are evenly divided into 5 medical groups, which are dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), myocardial infarction with altered left ventricular ejection fraction (MINF), abnormal right ventricle (ARV), and patients without cardiac disease (NOR) [16]. The 150 scans are divided into 90 images for training, 10 for validation, and 50 images for testing. The images for the validation set were selected randomly for each case (i.e., one for each patient type). The ACDC database was selected because it is one of the largest data sets available for CMR evaluation and has been widely used for computational analysis of medical images.

Image preprocessing was essential for training and validating the models. The first step was to understand the metadata: the mode for the image size was (216x256x10, and the spacing mode was (1.56x1.56x10). It was necessary to standardize the image as a square; thus, the final image size was set to 256x256x10. The size of the images ranged from 154 to 256 in length, 154 to 512 in width, and 6 to 18 in slice number. This data was practical for rescaling the images without losing information and calculating the scaling factor. The method applied for input images was Linear, while for ground truth was KNearestNeightbor. Each ground truth image has a number assigned to a segmented label; 0 corresponds to the resume, 1 for the right ventricle, 2 for the myocardium, and 3 for the left ventricle. The min-max normalization

task and elimination of outliers were applied to the raw images at pixel level; the pixels higher than the mean plus three times the standard deviation, were eliminated. For train CNNs, we decided to use the mode of all the spatial axes; however, for each transformer, we decided to use the image sizes the authors recommended in their investigations. For example, Chen et al. trained MissFormer using an initial image size of 512x512, which led them to have better results corresponding to the metrics they used to evaluate the model [35].

Evaluation Metrics.

Two widely used evaluation metrics in image segmentation have been chosen, which correspond to the Dice Coefficient and ASSD [100].

Dice Coefficient measures the pixel-to-pixel similarity between the segmentation mask predicted by the model and the ground truth segmentation [63]. The score ranges between 0 and 1, where 1 means a perfect similarity and 0 no similarity [80]. The Dice Coefficient is presented in Eq. 1, where true positive (TP) represent the pixel that is classified correctly from the predicted mask to a class of the ground truth, false positive (FP) is the pixel that is incorrectly assigned to a label; and false negative (FN) means that the pixel was incorrectly assigned to a different class or the background.

$$Dice\ coefficient = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{1}$$

The Average Surface Distance (ASSD) measures the average distance between the surface pixel S(A) of the gold standard A or the ground truth and the corresponding

surface pixel S(B) of the segmentation result B or the predicted mask [74]. ASSD is presented in Eq 2.

$$ASSD(A,B) = \frac{1}{S(A)+S(B)}\left(\sum_{S_A \in S(A)} s\left(S_A, S(B)\right) + \sum_{S_B \in S(B)} d(S_B, S(A))\right) \quad (2)$$

Statistical Test

We perform a statistical test between the models considering the Dice Coefficients and ASSD evaluation metrics per label and for means in general. Ending with 8 groups of data (Label1 – Dice, Label2 – Dice, Label3 – Dice, Label1 – ASSD, Label2 – ASSD, Label3 – ASSD, Resume - Dice, Resume - ASSD) and each group had a sample size of 50. The statistical test applied is the One-Way ANOVA, followed by the Tukey test, which are parametric tests that compare the mean values of each metric. To accomplish this, we use the statistical software Minitab and followed the next steps:

1. *The first step is to define hypothesis for the One-Way ANOVA test:*

Null Hypothesis (H_0): There is no significant difference among the group means.

Alternative Hypothesis (H_1): At least one group mean is different from the others.

2. *The second step is to choose a statistical test*:

For comparing means, first we must know the distribution of our data to either choose a parametric test or a non-parametric test. To do so, we do a test to check the normality and a test to check equal variances. The Anderson-Darling test was used and, with a p-value > 0,005 in all our groups, we can say that our data are not normally distributed. Later, Levene's test was applied and with a p-value > 0,05 in all the groups, we say the rest of our has not equal variances.

3. *The third step is to perform the test and analyze the results with the significance level (α) of 0.05.*

We applied a One way-ANOVA for each of our 8 groups and analyze if the p-value is < 0,05 so we can establish either to accept the Null hypothesis or reject it.

4. *The fourth step is to perform the Tukey test.*

We applied a Tukey test for each of our 8 groups. The Tukey test is not a hypothesis test so there is not a p-value. Instead, aims to determine which pairs of means exhibit significant differences from one another. Minitab put in descending order the groups analyzed according to their means and then group them giving them a letter. Means that do not share a letter are significantly different.

Results and discussion

The quantitate evaluation on the results are presented in Table 3. From these results, it is observed that the leading models are CNNs. LinkNet Dice and ASSD perform and SegFormer computational capacity is significantly better that the others 9 models. LinkNet achieved a mean Dice score of 0,90 and a mean ASSD of 0,30.

As established in section 3.5, Dice Coefficient measures the overlap between the predicted and ground truth segmentations, providing a measure of segmentation accuracy while the ASSD measures the distance between the predicted and ground truth surfaces, providing insight into the spatial accuracy of the segmentation. So, in this study we prefer having a good Dice score than having a good ASSD because we must analyze the whole anatomy structure and not only the border of this structure. Given the above, it does not matter if SwinUnet achieved the best ASSD since it also has the worst Dice.

The discussion of the current research starts with papers such as the literature review from Maurício et al. [101], which show an equilibrated analysis of ViT and CNN for image segmentation. They mentioned that visual transformers perform better due to their self-attention mechanism. Despite the slight difference in the results of some investigations, ViT's are more robust than convolutional neural networks. The datasets of the cited research ranged from 168 patches to 2.9 million images. One investigation concludes that ViT could learn patterns in small datasets. However, Deininger et al. [102] shows that to surpass CNN's performance, ViT might need more challenging tasks to benefit its characteristics. In the current investigation, we've trained the models with 1800 images for the training set and 200 for the validation set, which is not massive data compared to others. The models were not pre-trained, supporting the conclusion mentioned by Wu et al. [103], which is that ViT fails to generalize when training with fewer images. Also, this investigation did not use data augmentation. Coccomini et al. [104] show that with data augmentation, CNNs could generalize better, but ViT reduces the bias in identifying anomalies when one or more techniques are used. The reason stands out due to significant changes in the attention mechanism of ViT models. Another study says that CNNs were more robust than ViT in patch-based attacks with data perturbation as input [105]. The authors replaced the ReLU activation function with the transformed-based architecture activation function GELU. They concluded that the self-attention mechanism of a ViT is the key to its robustness.

Table 3

Comparative results in terms of metrics and computational capacity. Best results are in bold.

Method	Right Ventricle		Myocardium		Left Ventricle		Resume		Computational Cap
	Dice Coeff	ASSD	Dice Coeff	ASSD	Dice Coeff	ASSD	Dice Coeff	ASSD	
LinkNet	**0,8909**	0,3249	0,8694	0,3043	**0,9293**	**0,2616**	**0,8965**	**0,2960**	11'657,604
Unet	0,8852	0,3162	**0,8712**	**0,3033**	0,9243	0,2860	0,8936	0,3018	67'151,044
DeepLab V3+	0,8791	0,3333	0,8605	0,3066	0,9255	0,2696	0,8883	0,3031	45'664,212
Unet++	0,8730	0,3562	0,8627	0,3107	0,9214	0,2856	0,8857	0,3175	83'615,684
ScaleFormer	0,8585	0,3679	0,8469	0,3381	0,9134	0,2906	0,8729	0,3322	113'814,164
DeepLab V3	0,8608	0,3730	0,8376	0,3781	0,9056	0,3044	0,8680	0,3518	58'620,356
TransUnet	0,8600	0,3808	0,8117	0,4246	0,8883	0,3515	0,8533	0,3856	105'912,260
SegFormer	0,6638	0,3309	0,7283	0,4445	0,8189	0,4111	0,7370	0,3955	**7'718,244**
MissFormer	0,5874	0,3453	0,5876	0,4594	0,7293	0,4534	0,6347	0,4193	42'462,212
SwinUnet	0,5676	**0,2983**	0,5679	0,3889	0,6911	0,4336	0,6088	0,3736	271'684,20

On the other hand, we have the statistical analysis for the models which are shown in Tables 4, 5 and 6. The p-value of each Anderson-Darling and Levene´s tests were < 0,05 so we know that our data does not have a normal distribution and that the variances are not equal. Due to this, a non-parametric test is needed. But since the Minitab Support established that if the data contains 2 to 9 groups and the sample size for each group is at least 15 is recommended to use ANOVA. This is suggested because it will work fine with symmetric and non-normal distributions and, will have more power [101].

So, using a confidence level (α) of 0.05 in the One-Way ANOVA test we can conclude that apart from Label1-ASSD (Right Ventricle), in all data groups at least one mean is different. This can be seen in Table 4, where all p-values are < 0,05 rejecting the null hypothesis that was raised previously. For the right ventricle with the ASSD metric, the p-value >0.05 therefore the null hypothesis is accepted, and it is determined that no mean is different so, no model achieves a statistically different ASSD.

Table 4

Summary of results One way-ANOVA. Statistically significant results are in bold.

Groups	p-value
Label 1-Dice	0,000
Label 2-Dice	0,000
Label3-Dice	0,000
Label1-ASSD	**0,203**
Label2-ASSD	0,000
Label3-ASSD	0,000
Mean-Dice	0,000
Mean-ASSD	0,000

After performing the Tukey test the results were summarized in Table 5 and 6. As we explained, Minitab classified the means and gave each group a letter to establish if it is different from one another. For the Dice analysis we show in Table5 the groups that achieved the higher and equal results. Table 5 shows us Linknet, UNet, DeepLab V3+, Unet++, DeepLab V3, TransUnet, ScaleFormer achieve the best Dice mean for image segmentation of Right Ventricle. On the other hand, UNet, Linknet, Unet++, DeepLab V3+ achieve the best Dice mean for image segmentation of Myocardium. Conversely, Linknet, DeepLab V3+, UNet, Unet++, ScaleFormer, DeepLab V3, TransUnet achieve the best Dice mean for image segmentation of Left Ventricle. Finally, Table 5 presents Linknet as the model with the best Dice mean in the segmented resume.

Table 5

Summary of results Tukey test Dice means.

Dice Coefficient			
Right Ventricle	Myocardium	Left Ventricle	General
Linknet	UNet	Linknet	Linknet
UNet	Linknet	DeepLab V3+	
DeepLab V3+	Unet++	UNet	
Unet++	DeepLab V3+	Unet++	
DeepLab V3		ScaleFormer	
TransUnet		DeepLab V3	
ScaleFormer		TransUnet	

For the ASSD analysis we show in Table 6 the groups that achieved the best and equal results and demonstrate all the models achieve the same ASSD mean for the image segmentation of Right Ventricle. Conversely, UNet, Linknet, DeepLab V3+, Unet++ achieve the best ASSD mean for image segmentation of the Myocardium. On the other hand, Linknet, DeepLab V3+, Unet++, UNet, ScaleFormer and DeepLab V3 achieve the best ASSD mean for image segmentation of Left Ventricle. Finally, Linknet, UNet and DeepLab V3+ are the models with the best ASSD mean in the segmented resume. These could be because LinkNet had Nuclei segmentation; by incorporating nuclei segmentation into an architecture, the model can better identify and outline individual cell nuclei, which is important in various medical applications where precise delineation of nuclei is essential such as medical imaging modalities like microscopy or radiology. If we apply this quality to the orders methods could achieve better results.

Table 6

Summary of results Tukey test ASSD means.

ASSD			
Right Ventricle	Myocardium	Left Ventricle	General
No difference	UNet	Linknet	Linknet
	Linknet	DeepLab V3+	UNet
	DeepLab V3+	Unet++	DeepLab V3+
	Unet++	UNet	
		ScaleFormer	
		DeepLab V3	

A qualitative examination of the different method's performance is undertaken. The best and worst segmentation and edge prediction results are shown in Figures 11 – 12. As illustrated in Figure 11, LinkNet segment and predict almost perfectly by comparing to the other models that do not predict with precision; some can capture the desire areas, but pixels are not the same as the ground truth. This figure shows the best predictions for dice coefficient. In Figure 12, LinkNet also demonstrates good segmentation and prediction compared to the other models; even though, the image shows the worst predictions per model. In this case, TransUNet shows a good performance, but it only predicted two out of three desired areas.

Figure 11

All tested models' examples in order of dice coefficient results (best predictions).

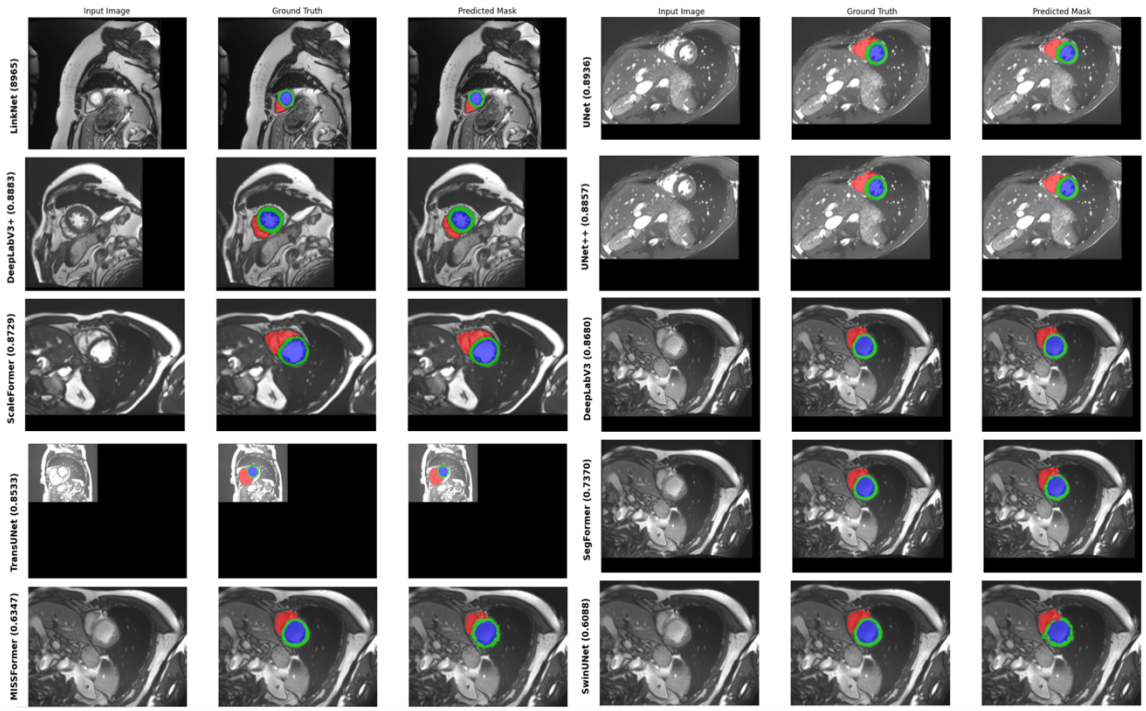
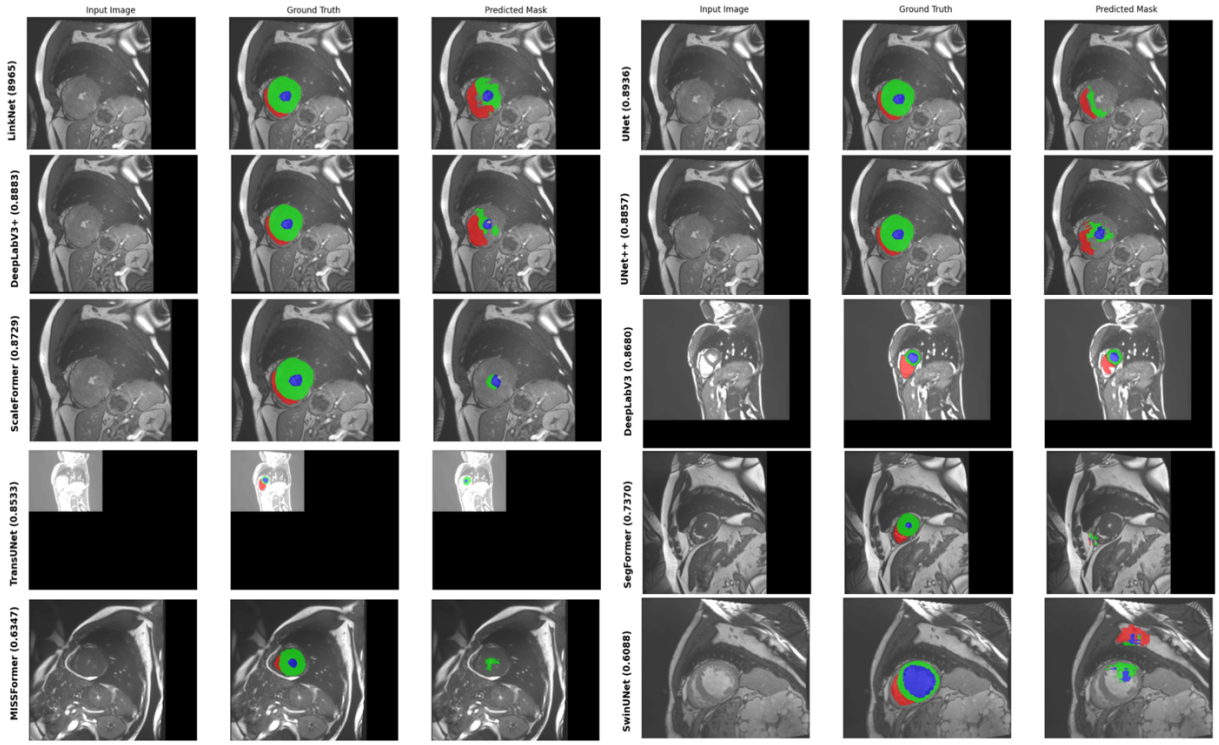


Figure 12

All tested models' examples in order of dice coefficient results (worst predictions).



In conclusion, based on the information gathered from this investigation, the quantitative, qualitative, and statistically results, we attempt to present the next inferences: From Table 3, CNNs architectures have the best Dice and ASSD means and also require less computing capacity on average. From Table 5 and Table 6, CNNs architectures achieved the best Dice and ASSD predicting the Right Ventricle, Myocardium and Left Ventricle given that they had higher scores without significance difference. Also, CNNs achieve better Dice and ASSD in the segmented resume. Finally, in the Figures 15 and 16 we can check what was previously established.

CONCLUSIONS

Cardiovascular diseases are considered one of the deadliest diseases. It is vital to diagnose CVDs in time to control them and in some cases save the patient's life. In this paper, we compared the performance of five CNN networks and five Transformer networks on the ACDC dataset. Results have evidenced that the Convolutional Neural Network architectures outperforms the other 10 models both qualitatively and quantitatively in terms of Dice Coefficient, ASSD and computational capacity. Moreover, it has been presented that LinkNet is an efficient and effective medical image segmentation method. As a future avenue, the models of cardiac image segmentation can be combined with hyperparameters optimization like Gradient Boosting, Random Forest or they also can be combined with data augmentation to improve the dataset and have better results in both architectures given that some architectures need more data to achieve better train. Doing this study demonstrates that comparisons could lead to detect the characteristics that made a model be an accurate tool for cardiac image segmentation and then apply these features in other models and achieves better scores.

REFERENCES

[1] *World Heart Report 2023: Confronting the World's Number One Killer. Geneva, Switzerland. World Heart Federation. 2023.*

[2] *Arregui, R. (2022). Las enfermedades cardiovasculares son la primera causa de muerte en el mundo — Dr. Roberto Arregui - cardiólogo, Quito. Dr. Roberto Arregui - Cardiólogo, Quito..*

[3] *Maron, B. J., & Maron, M. S. (2013). Hypertrophic cardiomyopathy. The Lancet, 381(9862), 242-255..*

[4] *Reed, G. W., Rossi, J. E., & Cannon, C. P. (2017). Acute myocardial infarction. The Lancet, 389(10065), 197-210..*

[5] *Schultheiss, H. P., Fairweather, D., Caforio, A. L., Escher, F., Hershberger, R. E., Lipshultz, S. E., ... & Priori, S. G. (2019). Dilated cardiomyopathy. Nature reviews Disease primers, 5(1), 32..*

[6] *Bartelds, B., Borgdorff, M. A., Smit-van Oosten, A., Takens, J., Boersma, B., Nederhoff, M. G., ... & Berger, R. M. (2011). Differential responses of the right ventricle to abnormal loading conditions in mice: pressure vs. volume load. European journal of.*

[7] *Wigle, E. D. (2001). The diagnosis of hypertrophic cardiomyopathy. Heart, 86(6), 709-714..*

[8] *Professional, C. C. M. (s. f.). Right-Sided heart failure. Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/21494-right-sided-heart-failure#diagnosis-and-tests.*

[9] *Japp, A. G., Gulati, A., Cook, S. A., Cowie, M. R., & Prasad, S. K. (2016). The diagnosis and evaluation of dilated cardiomyopathy. Journal of the American college of cardiology, 67(25), 2996-3010..*

[10] *Reddy, K., Khaliq, A., & Henning, R. J. (2015). Recent advances in the diagnosis and treatment of acute myocardial infarction. World journal of cardiology, 7(5), 243..*

[11] *Liu, L., Wolterink, J. M., Brune, C., & Veldhuis, R. N. (2021). Anatomy-aided deep learning for medical image segmentation: a review. Physics in Medicine & Biology, 66(11), 11TR01..*

[12] *Cai, W., Chen, S., & Zhang, D. (2007). Fast and robust fuzzy C-Means clustering algorithms incorporating local information for image segmentation. Pattern Recognition, 40(3), 825-838. https://doi.org/10.1016/j.patcog.2006.07.011.*

[13] *Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., & Rueckert, D. (2020). Deep learning for cardiac image segmentation: a review. Frontiers in Cardiovascular Medicine, 7, 25..*

[14] *Zotti, C., Luo, Z., Lalande, A., & Jodoin, P. M. (2018). Convolutional neural network with shape prior applied to cardiac MRI segmentation. IEEE journal of biomedical and health informatics, 23(3), 1119-1128..*

[15] *Chen, Y., Lu, X., & Xie, Q. (2023b). Collaborative networks of transformers and convolutional neural networks are powerful and versatile learners for accurate 3D medical image segmentation. Computers in Biology and Medicine, 164, 107228. https://doi.org/1.*

[16] *Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P. A., ... & Jodoin, P. M. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. IEEE transactions on medical i.*

[17] *Dar, A. S., & Padha, D. (2019). Medical image segmentation: A review of recent techniques, advancements, and a comprehensive comparison. Int J Comput Sci Eng, 7(7), 114-124..*

[18] *Preim, B., & Botha, C. P. (2013). Visual computing for medicine: theory, algorithms, and applications. Newnes..*

[19] *Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25..*

[20] *Al-Hammuri, K., Gebali, F., Kanan, A., & Chelvan, I. T. (2023). Vision transformer architecture and applications in digital health: a tutorial and survey. Visual Computing for Industry, Biomedicine, and Art, 6(1), 1-28..*

[21] *Yao, W., Bai, J., Liao, W., Chen, Y., Liu, M., & Xie, Y. (2023). From CNN to Transformer: A Review of Medical Image Segmentation Models. arXiv preprint arXiv:2308.05305..*

[22] *El-Taraboulsi, J., Cabrera, C. P., Roney, C., & Aung, N. (2023). Deep neural network architectures for cardiac image segmentation. Artificial Intelligence in the Life Sciences, 4, 100083..*

[23] *Liu, Z., He, X., & Lu, Y. (2022). Combining UNET 3+ and Transformer for left ventricle segmentation via signed distance and focal loss. Applied sciences, 12(18), 9208. https://doi.org/10.3390/app12189208.*

[24] *Zotti, C., Luo, Z., Humbert, O., Lalande, A., & Jodoin, P. M. (2018). GridNet with automatic shape prior registration for automatic MRI cardiac segmentation. In Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th Inte.*

[25] *Fan T, Wang G, Li Y, et al. Ma-net: a multi-scale attention network for liver and tumor segmentation. IEEE Access. 2020;8:179656-65.*

[26] *Khened, M., Kollerathu, V. A., & Krishnamurthi, G. (2019). Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Medical image analysis, 51, 21-45..*

[27] *Baccouch, W., Oueslati, S., Solaiman, B., & Labidi, S. (2023). A comparative study of CNN and U-Net performance for automatic segmentation of medical images: application to cardiac MRI. Procedia Computer Science, 219, 1089-1096..*

[28] *Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587..*

[29] *Li, K., Zhang, L., Li, B., Li, S., & Ma, J. (2022). Attention-optimized DeepLab V3+ for automatic estimation of cucumber disease severity. Plant Methods, 18(1), 1-16..*

[30] *Ruba, M. T., Tamilselvi, R., Beham, M. P., & Gayathri, M. (2023). Segmentation of a Brain Tumour using Modified LinkNet Architecture from MRI Images. Journal of Innovative Image Processing, 5(2), 161-180..*

[31] *Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., ... & Hong, Q. (2023). Lvit: language meets vision transformer in medical image segmentation. IEEE transactions on medical imaging..*

[32] *Zhang, Y., Liu, H., & Hu, Q. (2021). Transfuse: Fusing transformers and cnns for medical image segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October.*

[33] *Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., & Li, J. (2021). Transbts: Multimodal brain tumor segmentation using transformer. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France,.*

[34] *Deng, K., Meng, Y., Gao, D., Bridge, J., Shen, Y., Lip, G., ... & Zheng, Y. (2021). TransBridge: A lightweight transformer for left ventricle segmentation in echocardiography. In Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, H.*

[35] *Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306..*

[36] *Huang, X., Deng, Z., Li, D., & Yuan, X. (2021). Missformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162..*

[37] *Huang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X., Chen, Y. W., & Tong, R. (2022). ScaleFormer: revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation. arXiv preprint arXiv:2207.14552..*

[38] *Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems, 34, 12077-12090..*

[39] *Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022, October). Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision (pp. 205-218). Cham: Springer Nature Switzerland..*

[40] "Mauricio, J., Domingues, I., & Bernardino, J. (2023). Comparing vision Transformers and convolutional neural Networks for image classification: A literature review. Applied sciences, 13(9), 5521. https://doi.org/10.3390/app13095521," *applied sciences,* vol. 13, no. 5521, pp. 1-17, 2023.

[41] *Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929..*

[42] *Rustamy, F., PhD. (2023, 23 agosto). Vision transformers vs. convolutional neural networks. Medium. https://medium.com/@faheemrustamy/vision-transformers-vs-convolutional-neural-networks-5fe8f9e18efc#:~:text=The%20Vision%20Transformer%20(ViT)%20outperform.*

[43] *Deininger, L., Stimpel, B., Yuce, A., Abbasi-Sureshjani, S., Schönenberger, S., Ocampo, P., ... & Gaire, F. (2022). A comparative study between vision transformers and CNNs in digital pathology. arXiv preprint arXiv:2206.00389..*

[44] *Springenberg, M., Frommholz, A., Wenzel, M., Weicken, E., Ma, J., & Strodthoff, N. (2023). From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. Medical Ima.*

[45] *Matsoukas, C., Haslum, J. F., Sorkhei, M., Soderberg, M., & Smith, K. (2021). Can transformers replace cnns for medical image classification?..*

[46] *Elizondo, D. C., Amador, K. A., Ureña, F. S., Robledo, A., & de Microbiología, D. (2020). Factores de riesgo cardiovascular. CIENCIA Y SALUD..*

[47] *Sultana, F., Sufian, A., & Dutta, P. (2020). Evolution of image segmentation using deep convolutional neural network: A survey. Knowledge-Based Systems, 201, 106062..*

[48] *Martin-Isla, C., Campello, V. M., Izquierdo, C., Raisi-Estabragh, Z., Baeßler, B., Petersen, S. E., & Lekadir, K. (2020). Image-based cardiac diagnosis with machine learning: a review. Frontiers in cardiovascular medicine, 7, 1..*

[49] *Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC medical imaging, 15(1), 1-28..*

[50] *Yeghiazaryan, V., & Voiculescu, I. (2018). Family of boundary overlap metrics for the evaluation of medical image segmentation. Journal of Medical Imaging, 5(1), 015006-015006..*

[51] *Bandyopadhyay, H. (2023b, abril 24). An Introduction to Image Segmentation: Deep Learning vs. Traditional [+Examples]. V7. https://www.v7labs.com/blog/image-segmentation-guide#:~:text=Image%20segmentation%20originally%20started%20from,idea%20of%20the%20se.*

[52] *Merzougui, M., & El Allaoui, A. (2019). Region growing segmentation optimized by evolutionary approach and Maximum Entropy. Procedia Computer Science, 151, 1046-1051..*

[53] *Inik, Ö., & Ülker, E. (2022). Optimization of deep learning based segmentation method. Soft Computing, 26(7), 3329-3344.*

[54] *Offical, E. F. (2018, October 30). History of Image Segmentation - EAI Fund Offical - Medium. Medium. https://medium.com/@eaifundoffical/history-of-image-segmentation-655eb793559a.*

[55] *Sanchez-Ortiz, Gerardo I., and Alison Noble. "Fuzzy clustering driven anisotropic diffusion: enhancement and segmentation of cardiac MR images." Nuclear Science Symposium, 1998. Conference Record. 1998 IEEE. Vol. 3. IEEE, 1998.*

[56] *Santiago, Carlos, Jacinto C. Nascimento, and Jorge S. Marques. "Segmentation of the left ventricle in cardiac MRI using a probabilistic data association active shape model." Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual Internationa.*

[57] *Liu, T., Tian, Y., Zhao, S., Huang, X., & Wang, Q. (2019). Automatic whole heart segmentation using a two-stage u-net framework and an adaptive threshold window. IEEE Access, 7, 83628-83636..*

[58] *Galea, R. R., Diosan, L., Andreica, A., Popa, L., Manole, S., & Balint, Z. (2021). Region-of-interest-based cardiac image segmentation with deep learning. Applied Sciences, 11(4), 1965..*

[59] *Kumar, A. (2023, 26 marzo). CNN Basic Architecture for Classification & Segmentation - Analytics Yogi. Analytics Yogi. https://vitalflux.com/cnn-basic-architecture-for-classification-segmentation/.*

[60] *Zhang, Y., Lu, W., Ou, W., Zhang, G., Zhang, X., Cheng, J., & Zhang, W. (2020). Chinese medical question answer selection via hybrid models based on CNN and GRU. Multimedia tools and applications, 79, 14751-14776..*

[61] *Feng, N., Geng, X., & Qin, L. (2020). Study on MRI medical image segmentation technology based on CNN-CRF model. IEEE Access, 8, 60505-60514..*

[62] *An, F. P., & Liu, Z. W. (2019). Medical image segmentation algorithm based on feedback mechanism CNN. Contrast media & molecular imaging, 2019..*

[63] *Jasmine El-Taraboulsi, Claudia P. Cabrera, Caroline Roney, Nay Aung. (2023). Deep neural network architectures for cardiac image segmentation.Artificial Intelligence in the Life Sciences. Volume 4, 100083, ISSN 2667-3185. https://doi.org/10.1016/j.ailsci..*

[64] *Roth, H. R., Oda, H., Zhou, X., Shimizu, N., Yang, Y., Hayashi, Y., ... & Mori, K. (2018). An application of cascaded 3D fully convolutional networks for medical image segmentation. Computerized Medical Imaging and Graphics, 66, 90-99..*

[65] *Xu, Z., Wu, Z., & Feng, J. (2018). CFUN: Combining faster R-CNN and U-net network for efficient whole heart segmentation. arXiv preprint arXiv:1812.04914..*

[66] *Lou, A., Guan, S., & Loew, M. (2021, February). DC-UNet: rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation. In Medical Imaging 2021: Image Processing (Vol. 11596, pp. 758-768). SPIE..*

[67] *Yuan, F., Zhang, Z., & Fang, Z. (2023). An effective CNN and Transformer complementary network for medical image segmentation. Pattern Recognition, 136, 109228..*

[68] *Li, S., Wu, C., & Xiong, N. (2022). Hybrid architecture based on CNN and transformer for strip steel surface defect classification. Electronics, 11(8), 1200..*

[69] *Xie, Y., Zhang, J., Shen, C., & Xia, Y. (2021). Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France,.*

[70] *Gao, Y., Zhou, M., & Metaxas, D. N. (2021). UTNet: a hybrid transformer architecture for medical image segmentation. In Medical Image Computing and*

Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27.

[71] *Yun, B., Wang, Y., Chen, J., Wang, H., Shen, W., & Li, Q. (2021). Spectr: Spectral transformer for hyperspectral pathology image segmentation. arXiv preprint arXiv:2103.03604..*

[72] *Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 574-584)..*

[73] *Chang, Y., Menghan, H., Guangtao, Z., & Xiao-Ping, Z. (2021). Transclaw u-net: Claw u-net with transformers for medical image segmentation. arXiv preprint arXiv:2107.05188..*

[74] *Zhou, Z., Siddiquee, M. R., Tajbakhsh, N., & Liang, J. (2018). UNET++: a nested U-Net architecture for medical image segmentation. En Lecture Notes in Computer Science (pp. 3-11). https://doi.org/10.1007/978-3-030-00889-5_1.*

[75] *Yang, T. J., Collins, M. D., Zhu, Y., Hwang, J. J., Liu, T., Zhang, X., ... & Chen, L. C. (2019). Deeperlab: Single-shot image parser. arXiv preprint arXiv:1902.05093..*

[76] *Chaurasia, A., & Culurciello, E. (2017). LinkNet: Exploiting encoder representations for efficient semantic segmentation. Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/vcip.2017.8305148.*

[77] *Tao, R., Zhang, Y., Wang, L., Cai, P., & Tan, H. (2020). Detection of precipitation cloud over the Tibet based on the improved U-Net. Computers, materials & continua, 65(3), 2455-2474. https://doi.org/10.32604/cmc.2020.011526.*

[78] *Zhou, Z., Siddiquee, M. R., Tajbakhsh, N., & Liang, J. (2020). UNET++: Redesigning Skip Connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging, 39(6), 1856-1867. https://doi.org/10.1109/tmi.2019.2959609.*

[79] *Kim T., Oh K., Kim J., Lee Y. & Choi J.(2023) Development of ResNet152 UNet++-Based Segmentation Algorithm for the Tympanic Membrane and Affected Areas. IEEE Access, 11, 56225-56234. doi: 10.1109/ACCESS.2023.3281693..*

[80] *Baskaran, L., Al'Aref, S. J., Maliakal, G., Lee, B. C., Xu, Z., Choi, J. W., Lee, S. E., Sung, J. M., Lin, F. Y., Dunham, S., Mosadegh, B., Kim, Y. J., Gottlieb, I., Lee, B. K., Chun, E. J., Cademartiri, F., Maffei, E., Marques, H., Shin, S., . . . Shaw,.*

[81] *Wafa Baccouch, Sameh Oueslati, Basel Solaiman, Salam Labidi. (2023). A comparative study of CNN and U-Net performance for automatic segmentation of medical images: application to cardiac MRI. Procedia Computer Science,Volume 219,Pages 1089-1096,ISSN 1877-.*

[82] *Bonechi, S., Andreini, P., Mecocci, A., Giannelli, N., Scarselli, F., Neri, E., Bianchini, M., & Dimitri, G. M. (2021). Segmentation of aorta 3D CT images based on 2D convolutional neural networks. Electronics, 10(20), 2559. https://doi.org/10.3390/electr.*

[83] *Tomar, N. (2021, 27 diciembre). What is UNET? - Analytics Vidhya - Medium. Medium. https://medium.com/analytics-vidhya/what-is-unet-157314c87634.*

[84] *Neven, R. & Goedemé, T. (2021). A Multi-Branch U-Net for Steel Surface Defect Type and Severity Segmentation. Metals. 11(870). 10.3390/met11060870..*

[85] *Xu, W., Shi, J., Lin, Y., Liu, C., Xie, W., Liu, H., Huang, S., Zhu, D., Su, L., Huang, Y., Ye, Y., & Huang, J. (2023). Deep learning-based image segmentation model using an MRI-based convolutional neural network for physiological evaluation of the heart..*

[86] *Ren, Y., Yu, L., Tian, S., Cheng, J., Guo, Z., & Zhang, Y. (2021). Serial attention network for skin lesion segmentation. Journal of Ambient Intelligence and Humanized Computing, 13(2), 799-810. https://doi.org/10.1007/s12652-021-02933-3.*

[87] *Chen, L., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1802.02611v3.*

[88] *Yuan, H., Zhu, J., Wang, Q., Cheng, M., & Cai, Z. (2022). An improved DeepLab V3+ deep learning network applied to the segmentation of grape leaf black rot spots. Frontiers in Plant Science, 13. https://doi.org/10.3389/fpls.2022.795410.*

[89] *Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. (2018b). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4.*

[90] *Singh, V., & Singh, V. (2023, November 6). DeepLabv3 & DeepLabv3+ The Ultimate PyTorch Guide. LearnOpenCV – Learn OpenCV, PyTorch, Keras, Tensorflow With Examples and Tutorials. https://learnopencv.com/deeplabv3-ultimate-guide/.*

[91] *P. Sameer, V. A. R. Kaushik, R. Indrakanti and C. M. K. S, "Brain Tumor Segmentation using LinkNet," 2022 Second International Conference on Next*

Generation Intelligent Systems (ICNGIS), Kottayam, India, 2022, pp. 1-5, doi: 10.1109/ICNGIS54955.2022.100797.

[92] *Yathish, V. (2022). Loss functions and their use in neural networks - towards data science. Medium. https://towardsdatascience.com/loss-functions-and-their-use-in-neural-networks-a470e703f1e9.*

[93] *Darth Espressius. (2022). 3 Common loss functions for image segmentation. DEV Community. https://dev.to/_aadidev/3-common-loss-functions-for-image-segmentation-545o.*

[94] *Zhai, X., Qiao, F., Ma, Y., & Lu, H. H. (2022). A novel fault diagnosis method under dynamic working conditions based on CNN with an adaptive learning rate. IEEE Transactions on Instrumentation and Measurement, 71, 1-12. https://doi.org/10.1109/tim.2022..*

[95] *Baheti, P. (2023, 24 abril). Activation Functions in Neural Networks [12 Types & Use Cases]. V7. https://www.v7labs.com/blog/neural-networks-activation-functions#:~:text=Here's%20why%20sigmoid%2Flogistic%20activation,choice%20because%20of%20its%20range..*

[96] *Wang, M., Lu, S., Zhu, D., Lin, J., & Wang, Z. (2018, October). A high-speed and low-complexity architecture for softmax function in deep learning. In 2018 IEEE asia pacific conference on circuits and systems (APCCAS) (pp. 223-226). IEEE..*

[97] *Zhang, Z. (2018, June). Improved adam optimizer for deep neural networks. In 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS) (pp. 1-2). Ieee..*

[98] *Duda, J. (2019). SGD momentum optimizer with step estimation by online parabola model. arXiv preprint arXiv:1907.07063..*

[99] *Bernard. (s. f.). ACDC Challenge. https://www.creatis.insa-lyon.fr/Challenge/acdc/.*

[100] *Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., ... & Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. Academic radiology, 11(2), 178-189.*

[101] J. Maurício, I. Domingues and J. Bernardino, Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review, applied sciences, 2023.

[102] L. Deininger, B. Stimpel, A. Yuce, S. Abbasi-Sureshjani, S. Schönenberger, P. Ocampo, K. Korski and F. Gaire, "A comparative study between vision transformers and CNNs in digital pathology," 2022.

[103] Y. Wu, S. Qi, Y. Sun, S. Xia, Y. Yao and W. Qian, "A vision transformer for emphysema classification using CT images," *Institute of Physics and Engineering in Medicine,* vol. 66, no. 24, 2021.

[104] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro and G. Amato, "Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection," *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation,* pp. 1-7, 2022.

[105] Y. Bai, J. Mei, A. Yuille and C. Xie, "Are Transformers More Robust Than CNNs?," *35th Conference on Neural Information Processing Systems,* 2021.

[106] *Gonzales, J. (2018). Prueba de Kruskall Wallis..*