Applied Computer Systems
ISSN 2255-8691 (online)
ISSN 2255-8683 (print)
2025, vol. 30, no. 1, pp. 12–20
https://doi.org/10.2478/acss-2025-0002
https://content.sciendo.com

Peering into the Heart: A Comprehensive Exploration of Semantic Segmentation and Explainable AI on the MnMs-2 Cardiac MRI Dataset

Mohamed Ayoob[1*], Oshan Nettasinghe[2], Vithushan Sylvester[3], Helmini Bowala[4], Hamdaan Mohideen[5]
[1–5]*Informatics Institute of Technology, Colombo, Sri Lanka*

***Abstract* – Accurate and interpretable segmentation of medical images is crucial for computer-aided diagnosis and image-guided interventions. This study explores the integration of semantic segmentation and explainable AI techniques on the MnMs-2 Cardiac MRI dataset. We propose a segmentation model that achieves competitive dice scores (nearly 90 %) and Hausdorff distance (less than 70), demonstrating its effectiveness for cardiac MRI analysis. Furthermore, we leverage Grad-CAM, and Feature Ablation, explainable AI techniques, to visualise the regions of interest guiding the model predictions for a target class. This integration enhances interpretability, allowing us to gain insights into the model decision-making process and build trust in its predictions.**

***Keywords* – Cardiac magnetic resonance imaging (CMRI) semantic segmentation, explainable AI (XAI), residual blocks, vision transformer (ViT).**

I. Introduction

A. Introduction to the Domain

According to the World Heart Report 2023, published by the World Heart Federation, over 500 million people worldwide have cardiovascular diseases, resulting in 20.5 million deaths in 2021. This is almost one-third of all deaths globally, and it is an increase from the estimated 121 million cardiovascular disease deaths that have occurred so far [1]. Cardiac MRI (CMRI) images are one of the main varieties of cardiovascular image diagnosis. It uses a magnetic field to create a set of detailed images for image analysis for assorted clinical staff. Routinely CMRI images are standard for preventive care and detection of cardiac ailments.

Historically, the right ventricle (RV) in the circulatory system has been overshadowed by the left ventricle (LV). For many years, RV dysfunction was believed to have minimal influence on cardiac output and pressures, with the LV being considered the primary driver of cardiac hemodynamics [2]. As a result, the RV received limited attention and was often referred to as the "forgotten ventricle". However, over the past few decades, our understanding has evolved, and it is now clear that RV dysfunction significantly impacts cardiac function. Extensive research has eventually revealed the crucial role of the RV in cardiac function and its importance in various high-impact diseases [3].

Ever since, RV segmentation has become a challenging task.

Artificial intelligence (AI) and machine learning (ML) have been rapidly transforming the field of medical imaging. AI/ML techniques are adept at automating tasks that are presently undertaken by radiologists, encompassing image segmentation, classification, and diagnosis [3].

By doing so, radiologists can divert their attention to more intricate cases, while simultaneously enhancing the precision and consistency of diagnoses. The progressive integration of AI/ML in the field of medical imaging is paving the way for more efficient and reliable healthcare practices.

In this paper, we will discuss the utilisation of two different AI techniques that can be used for a variety of use cases. This paper will explore how these techniques will pique the interest in utilising such techniques in a holistic level on the MnMs-2 dataset for primarily RV segmentation, and, in addition to it, LV and Myocardium (MYO) are segmented.

Section 1 of this paper introduces the two techniques and their use cases. Section 2 describes the dataset we utilised. Sections 3 and 4 are the methodology followed and the results obtained for each of the two use cases. Section 5 summarises the findings, outlines future works, and concludes the paper.

B. Semantic Segmentation

One of the most common applications of AI/ML in MRI imaging is image semantic segmentation. Image semantic segmentation is the process of dividing an image into its constituent parts, such as organs, tissues, and lesions on a per-pixel level. This can be a challenging task for radiologists, but AI/ML techniques can be used to automate the process. One of the consequential works [4] on image segmentation is using deep convolutional neural networks (CNN) to segment images from the PASCAL VOC dataset with an accuracy of 94 %.

The U-Net paper [5] introduced a new architecture for semantic segmentation that has since become one of the most popular and widely used models in the field. The U-Net model is a fully convolutional network, such that it consists of convolutional layers and pooling layers. This makes it well-

* Corresponding author's e-mail: ayoob.m@iit.ac.lk
Article received 2024-12-09; accepted 2025-01-07

suited for image segmentation tasks, as it can learn to extract features from images. Furthermore, prevalence of such layers facilitates the possibility of extracting learnt patterns and features post-hoc training for explainable AI purposes.

The U-Net architecture currently stands as the preeminent model framework for medical image segmentation. Yet, the amalgamation of recent innovations, notably the Vision Transformer (ViT) [6] and Residual Blocks [7], has markedly enhanced model efficacy [8]. Nevertheless, a scarcity of investigations on ViT applicability within the domain of cardiac magnetic resonance imaging (CMRI) exists, which necessitates further inquiry [3]. This research underscores the imperative for dedicated scrutiny of CMRI owing to its intricate biological substratum.

C. Explainable AI

Due to the nondeterministic nature of deep neural networks, Explainable AI (XAI) has gained traction as a sub-field of artificial intelligence that seeks to make AI models more interpretable and explainable to humans. This is important because AI models are often used to make decisions that have a significant impact on human lives, and it is important to be able to understand how these models work, particularly in critical sectors such as medicine, healthcare, and self-driving autonomous cars.

In the domain of biomedical image classification and segmentation, ensuring the explainability, interpretability, and transparency of AI models is of great importance to the domain experts. Due to the 'black box' nature of Deep Neural Networks (DNNs), they are inherently unable to provide human understandable explanations to physicians or radiologists [9]. Therefore, it is imperative that we employ XAI techniques to build a foundation of trust fostering an environment where these models can be utilised and made use of.

Many post-hoc techniques such as GradCAM [10], Feature Ablation [11], Layer wise Relevance Propagation (LRP) [12], Shapely Additive Explanations (SHAP) [13] have been utilised to add explainability to DNNs performing biomedical image processing related tasks.

XAI techniques like GradCAM [10] and Feature Ablation [11] are increasingly used in medical imaging for model interpretability. Studies employing these methods across brain and cardiac MRI segmentation have shown significant advancements. For example, GradCAM has facilitated interpretable visualisations in brain tumour segmentation using the BRATS dataset, while methods like D-TCAV [14] and GradXcepUNet [15] have enhanced cardiac MRI analysis and segmentation accuracy. Furthermore, Ablation-CAM's [16] integration of feature map deactivation with CAM offers detailed insights into model predictions across various medical fields, promoting responsible AI in healthcare diagnostics.

II. MnMs-2 Dataset

The MnMs-2 dataset is a collection of MRI images of the heart. The dataset was created by AI in Medicine Lab, at the University of Barcelona [3], [17]. The dataset consists of 360 sets of images where each set of images corresponds to a patient with various pathologies affecting the right ventricle and left ventricle, along with healthy subjects. The individuals underwent scanning at one of the three different clinical centres in Spain, using magnetic resonance scanners. The clinical centres comprise scanners from three different vendors, namely, Siemens, General Electric, and Philips.

Each dataset consists of magnetic resonance imaging scans captured in two distinct cardiac states: end-systolic (ES) and end-diastolic (ED). These scans include both short-axis (SA) and long-axis (LA) views. Consequently, each dataset comprises four MRI images. Accompanying these images are their respective masks, which serve to delineate specific anatomical or functional areas of interest within the cardiac structure. From the dataset, there were 1440 (360 sets × 4 images per set) annotated images gathered from the collection centres.

According to the MnMs-2 Challenge, the dataset was categorised into training, testing and validation, in the following proportions 160, 160, and 40 sets of images, respectively. These CMRI images were segmented by experienced clinicians from their respective institutions. The segmentation included contours for the LV and RV blood pools, as well as for the left ventricular MYO. The labels used were: 1 for LV, 2 for MYO, and 3 for RV in both SA and LA views, covering a wide range of challenging RV and LV pathologies. Each image consists of its corresponding mask. The SA images are three-dimensional in shape (voxels) and the LA images are two-dimensional in shape (pixels).

A. Dataset Pre-Processing

The MnMs-2 dataset, acquired from the MnMs-2 Challenge is composed of raw cardiac MRI images stored in .nii.gz format. To accommodate these images within the framework of our U-Net model architecture, we undertook a series of preparatory procedures.

Data Partitioning into Training, Testing and Validation Sets: Initially, the dataset was decompressed, and a partitioning process was employed to create distinct sets for training, testing, and validation as discussed above. This partitioning approach was guided as outlined in Table II of the publication "Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation" [3].

Data Transformation: The inherent diversity in image dimensions stemming from diverse imaging systems (scanners) and the vendors necessitated a transformation step. We structured the dataset into distinct numpy arrays, encompassing both ES and ED images within LA and SA perspectives, while adhering to the following protocols:

Long-Axis (LA) Images: Integer interpolation was undertaken when resizing the semantic masks. This was done aiming to preserve information fidelity and computational efficiency. A balance was struck between maintaining accuracy and preventing undue computational burdens. This resulted in dimensions of (200×200) for the 2D-LA images, ensuring compatibility with the training U-Net architecture while remaining divisible by two throughout the encoder and decoder stages.

Short-Axis (SA) Images: The challenge posed by SA image dimensions in height, width, and depth necessitated a slightly distinct approach. Resizing alone proved insufficient due to variations in depth (due to various scanners and vendors). To surmount this, height and width adjustments were accompanied by integer interpolation (when resizing semantic masks) and depth adjustment was accompanied by zero-padding to achieve a common depth. To this extent, the maximal depth within the dataset was identified as 28, leading us to adopt a depth of 32, which facilitates compatibility with the U-Net architecture's data dimension requirements. (exponents of 2). Consequently, the dimensions of the SA dataset were established at (240×240).

B. Persistence of Processed Data

In line with the utility of the numpy library for image manipulation, the processed LA and SA datasets were stored as separate numpy (.npy) files. This organisation permitted subsequent transformation of the datasets into tensors, subsequently facilitating their integration into data loaders designated for training, testing, and validation purposes. This separation of data pre-processing from model convergence, validation, and testing phases economises computational resources.

III. Semantic Segmentation

A. Methodology

Figure 1 depicts the methodologies taken to semantically segment the data of the SA and LA images, respectively.

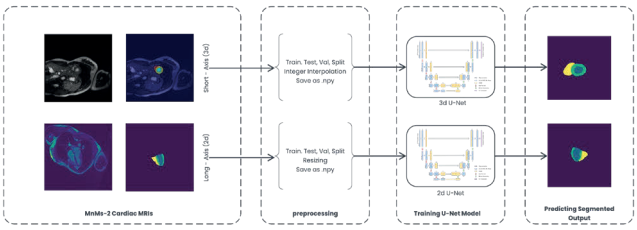


Fig. 1. Model architecture for semantically segmenting SA and LA images.

To segment CMRI images within the MnMs-2 dataset, we employed a few additional features to nnU-Net architecture. Renowned for its efficacy in medical image segmentation tasks, this modified architecture incorporates both Vision Transformer (ViT) and residual blocks. The distinctive encoder-decoder structure of the U-Net architecture is pivotal; the encoder progressively down-samples the input image to extract salient features, while the decoder subsequently up-samples these features to generate a segmented output. Additionally, the integration of residual blocks within each encoder and decoder segment serves to enhance the model performance by facilitating gradient flow, preserving spatial information, and mitigating overfitting [7]. Furthermore, we augmented the U-Net model with a multi-head self-attention mechanism and positional encoding within its intermediate layer. This augmentation enables the model to assimilate spatial information from the input data by assigning attention scores to individual embeddings. Additionally, PReLU [18] has been explored as an alternative activation function to ReLU within nnU-Net architectures, which are widely adopted [5].

Below are the architectural components corresponding to the model architecture.

Convolutional Layer: The core operation in the convolution layer is the convolution operation. The 2D and 3D convolutions have been performed for LA and SA, respectively.

Residual Blocks: Residual blocks are integrated with each block of encoder and decoder part of the model. This has been implemented in a way that input of each block relates to the output from the activation function following the convolution layers.

Vision Transformer: The multi-head self-attention is integrated with the middle layer of the U-Net where the input feature map gets converted to patch embedding and positional encodings and sent to multi-head self-attention, which extract the query, key and value to assign multi-head attention scores for each encoding.

Batch Normalization: It helps in accelerating the training of deep networks by reducing internal co-variate shifts, thereby stabilising, and speeding up the convergence of the training process.

Parametric Rectified Linear Unit: PReLU [18] helps mitigate the vanishing gradient problem and enables the network to learn from the negative input values, improving the learning capacity of the model.

The PReLU operation is defined as:

$$PReLU\left(x\right)=\max\left(0,x\right)+a.\min\left(0,x\right), \tag{1}$$

where a is a learnable parameter, and x is the input.

Max Pooling Layer: In the max pooling layer, the maximum value is selected from each region of the input tensor defined by the pooling window size.

ConvTranspose Layer (Deconvolution): This layer performs a transposed convolution operation. The operation effectively performs the gradient of a corresponding forward convolution concerning its input, which is why it is often referred to as a deconvolution.

Hyperparameters: The model was trained for 200 epochs and batch size of 2. These hyperparameters were chosen to foster a conducive learning environment for the model, ensuring a balance between computational efficiency and convergence quality.

Loss Function and Optimizers: To evaluate the quality of the segmentation, the following two measures were used.

1. Dice similarity coefficient (DSC) – degree of overlapping of the ground truth and prediction.

$$DSC\left(P,G\right)=\frac{2\left|P\cap G\right|}{\left|P\right|+\left|G\right|} \tag{2}$$

2. Hausdorff Distance (HD) – the largest distance between the ground truth and prediction.

$$HD\left(P,G\right)=\max\left\{\sup_{p\in P}d\left(p,G\right),\ sup_{g\in G}d\left(g,P\right)\right\}, \tag{3}$$

where P and G are the prediction and the ground truth, respectively.

To balance the contribution of each loss function, a parameter was introduced, computed as the ratio between the mean values of each loss function during the last epoch. This allows to be dynamic which helps evenly distribute the weight on both the loss function for each epoch [19].

$$\lambda = \frac{HD(P,G)}{DSC(P,G)} \tag{4}$$

$$\text{Loss} = \lambda \cdot DSC(P,G) + HD(P,G) \tag{5}$$

RMSprop optimizer adaptably adjusts the learning rates during training, thus, we used it for optimization. To reduce the possibility of over-fitting, a weight decay of 1×10^{-8} was applied, and the starting learning rate was set to 0.0001. Simultaneously, a CosineAnnealingLR scheduler was employed to adjust the learning rate.

B. Results

To assess the individual contributions of proposed features, we trained six U-Net models: three for Long-Axis and three for Short-Axis. Table I presents the performance scores of these models.

TABLE I
DICE SCORES AND HAUSDORFF DISTANCE OF ALL THE MODELS

No.	Image Type	BackBone	Additional Features	DSC Score	HD Score
1	Long-axis (2D)	nnU-Net	-	0.8856	15.8427
2	Long-axis (2D)	nnU-Net	PReLU activation	0.8834	16.3743
3	Long-axis (2D)	nnU-Net	ViT + ResBlocks + PReLU activation	0.8930	15.6076
4	Short-axis (3D)	nnU-Net	-	0.9111	64.6101
5	Short-axis (3D)	nnU-Net	PReLU activation	0.9126	66.1701
6	Short-axis	nnU-Net	ViT + ResBlocks + PReLU activation	0.9096	66.4372

As shown in Table I, the integration of ViT and ResBlocks with PReLU activation yielded a modest improvement in the DSC for 2D long-axis images. Specifically, our approach increased the DSC from 0.8856 to 0.8930, while also slightly reducing the HD. Although these improvements are not substantial, these results suggest that ViT-based architectures are better suited to capture global dependencies in 2D slices. The slight improvements in HD and DSC suggest a potential for increased robustness in clinical settings, especially in cases requiring precise boundary delineation. Reducing HD, even marginally, can lead to fewer segmentation errors in regions with complex structures, enhancing the model reliability.

In the 3D short-axis dataset, PReLU activations outperformed the baseline nnU-Net, yielding a modest improvement in DSC which is 0.9111 to 0.9126. This highlights PReLU's ability to adaptively model the negative activation space, which is crucial for capturing subtle variations in volumetric data. Deeper 3D networks, which handle larger volumes of data, benefit from PReLU's ability to facilitate smooth gradient flow and refined feature learning. The learnable negative slope enables PReLU to capture nuances in volumetric data that may be overlooked by ReLU. The enhancements provided by PReLU and ViT offer greater adaptability to diverse datasets, which is essential for real-world applications where MRI images often exhibit significant variations in quality, contrast, and resolution across different scanners and institutions.

However, the integration of ViT and residual blocks led to a noticeable decline in SA segmentation performance, as evidenced by a decrease in DSC and an increase in Hausdorff distance. This decline may be attributed to the increased complexity of the 3D model, which may hinder its ability to effectively capture spatial information within the dataset. Additionally, the zero-padding pre-processing technique might have negatively impacted the extraction of spatial context in volumetric data. Furthermore, the size of medical imaging datasets, including the one used in this study, is typically smaller compared to datasets used to train ViT in other domains (e.g., ImageNet). This data limitation may hinder the ViT components from learning meaningful representations, resulting in suboptimal performance.

Likewise, the marginally poorer performance of PReLU in the 2D nnU-Net setup can be attributed to the combination of limited spatial complexity, the risk of overfitting in small datasets, and the limited benefit of learnable negative slopes in the context of 2D slice-based processing. 2D typically has a shallower architecture compared to its 3D counterpart, the gradient flow requirements are less stringent. PReLU's ability to mitigate vanishing gradients is less impactful in this scenario, and the additional complexity introduced may hinder optimisation. In such scenarios, ReLU's simplicity and generalisation capabilities make it a more suitable choice. Future research could explore hybrid activation strategies or regularisation techniques to better leverage PReLU's potential in 2D setups.

Both models demonstrate good performance in each category (see Table II). Notably, the SA model exhibits a marginally superior average score of 0.91, compared to the LA model's average of 0.89 in DSC. This indicates a slightly better overall efficacy of the SA model in cardiac segmentation tasks. Notably, this is better than the 6th rank in (ES and ED are averaged) VI from the MnMs-2 [3].

TABLE II
DICE SCORES OF THE BEST PERFORMING MODEL

	RV	LV	MYO	Background	Average
SA(3D)	0.88	0.91	0.86	0.99	0.91
LA(2D)	0.84	0.91	0.81	0.99	0.89

Special emphasis must be made on the Right Ventricular (RV) segmentation. As it is deemed to be a challenging task in the MnMs-2 dataset [3], it should be noted that it is segmented

with an accuracy of nearly 88 % and 84 % in the SA and LA views, respectively.

Hausdorff distance for both the models (see Table III), is performing competitively considering the scores in the leader board of MnMs2 challenge [3], [17].

TABLE III
HAUSDORFF DISTANCE OF THE BEST PERFORMING MODEL

	RV	LV	MYO	Background	Average
SA(3D)	63.0	49.15	64.0	88.4	64.6
LA(2D)	15.42	12.4	14.1	18.7	15.6

In summary, this study has demonstrated significant results, despite avoiding data augmentation during pre-processing. While data augmentation is commonly used in benchmark models for the MnMs-2 challenge [3], it was intentionally avoided in this study due to the critical nature of medical datasets.

Figures 2–4 depict the loss curves during the training for both the SA and LA views. Since we are using a combined loss function of Dice Loss and Hausdorff Loss, all the three losses (i.e., Dice Score loss, Hausdorff Distance Loss, Combined Loss) for each of the two views of the dataset (SA and LA) are depicted.

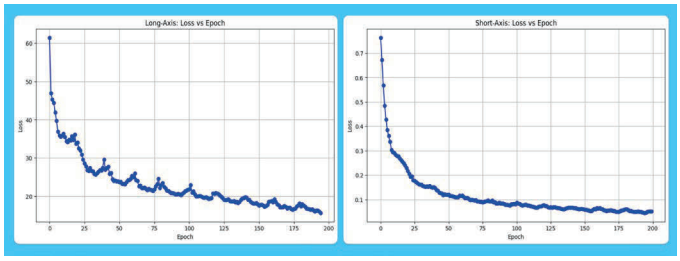


Fig. 2. Loss curve for long-axis (model 4) and short-axis (model 6).

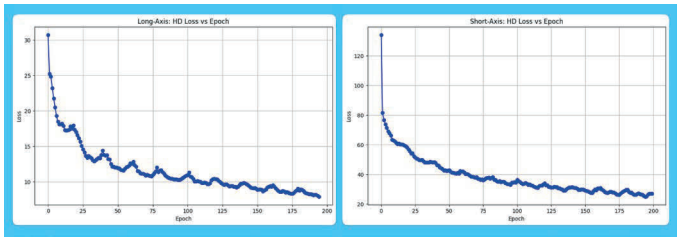


Fig. 3. HD loss curve for long-axis (model 4) and short-axis (model 6).

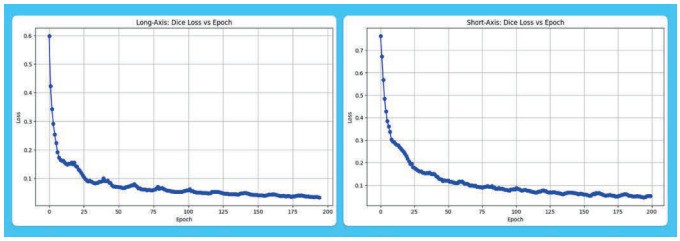


Fig. 4. DSC loss curve for long-axis (model 4) and short-axis (model 6).

Figures 5 and 6 show the few output segmentation maps from the model 2 and model 6.

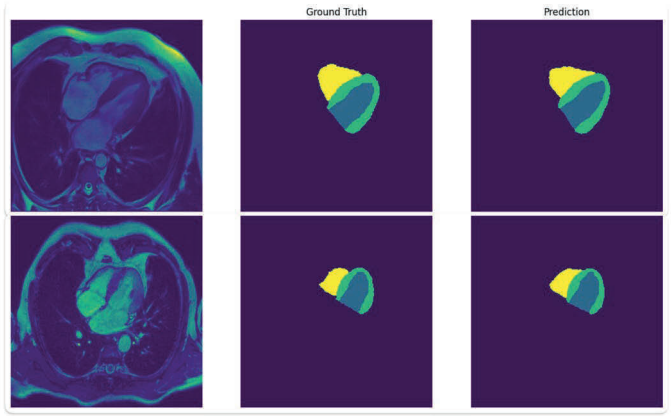


Fig. 5. Visualization of input image, ground truth and the prediction for long-axis (model 2).

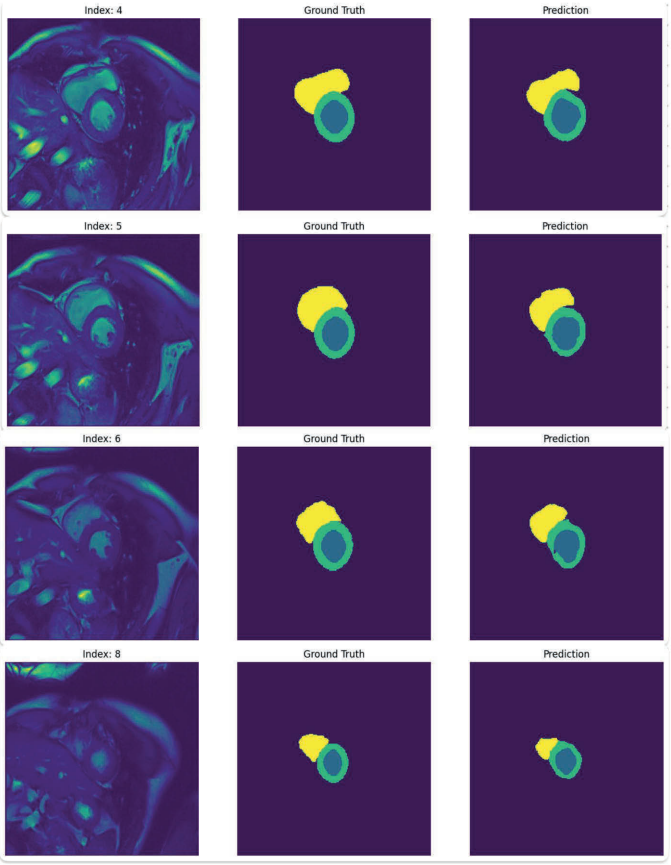


Fig. 6. Visualisation of input image, ground truth and the prediction for short axis (model 6).

IV. EXPLAINABLE AI

A. Methodology

To interpret the predictions from both 2D and 3D CMRI segmentation U-Net based models, two prominent techniques, GradCAM [10] and Feature Ablation [11] have been used.

Both methods are post-hoc XAI (i.e., both these methods are used to visualize the layers after training) that provide

qualitative analysis to some degree of quantitative analysis. GradCAM leverages the gradients flowing into the output layer and relates the significance of each feature in producing the final segmentation map of CMRI images. The other XAI technique that will be explored is feature ablation, which is a model agnostic perturbation-based approach. The method systematically evaluates the impact of each segment of the CMRI on the model's output by replacing each input feature with a reference baseline.

GradCAM offers complimentary insights to the segmentation output, particularly in localising the regions of interest in CNNs, whereas feature ablation provides a way to quantify the significance of various segments/features in CMRI segmentation. For the implementation purposes, PyTorch Captum [20] library was utilised.

B. Results of the XAI Analysis

GradCAM Heatmap Analysis: GradCAM is a gradient based approach that can be applied to many different CNN based models irrespective of the underlying architecture of the model. It produces high resolution saliency maps compared to other techniques such as occlusion [10].

Attribution scores are numerical values assigned to input classes that indicate their contribution to segmenting a target class. We are calculating the attribution scores based on the gradient flow into the last layer of the U-Net model and analysing the outcome by visualising the results with respect to each segmentation class. These visualisations are produced from the normalised attribution scores.

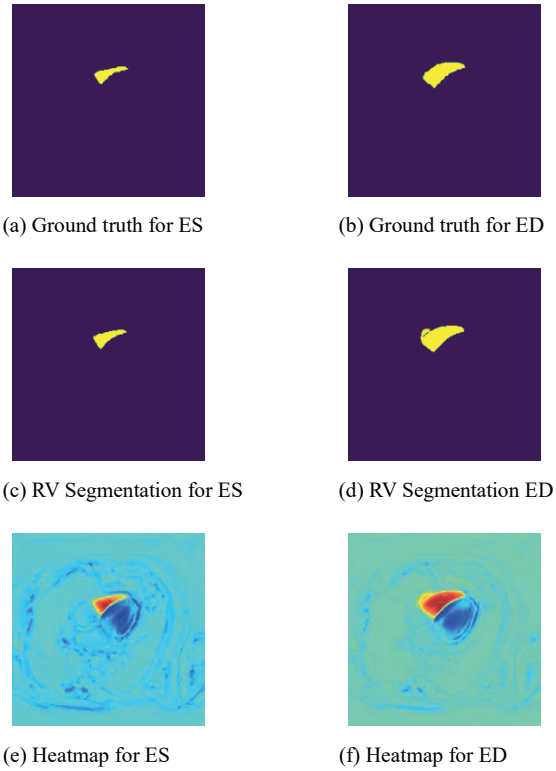
While calculating the attribution scores, we disable the ReLU operation of GradCAM attribution calculation method [20] to preserve the negative attributions which aids us in understanding the most and least impactful regions.

Fig. 7. Blue indicates the values with high negative contribution to the segmentation whereas the red regions indicate positive contributions to the segmentation. This colour bar is applicable to the figures depicting saliency maps obtained via GradCAM.

Long Axis ES and ED CMRI Saliency Maps from GradCAM for RV: Figure 8 presents Grad-CAM attribution heatmaps for two randomly selected ES and ED CMRI. The visualisations in Fig. 8 (e) and (f) highlight the salient features that contribute to RV segmentation, specifically for ES and ED phases.

Both sets of heatmaps indicate that the RV itself is the most salient class for RV segmentation, while the LV and MYO exhibit strong negative contributions. This suggests that the model's gradient flow is focused on the target class, RV in the segmentation. This increases the confidence on the model segmentation outcome.

The significant difference that can be observed in the ED heatmap compared to ES is that the ED has a larger area contributing positively. The assumption here is that the heart would be at its highest volume during ED phase and subsequently the captured anatomical regions would be visualised in a larger region. This is a proven clinical fact [21]. This provides further confidence on the model's prediction.



(a) Ground truth for ES (b) Ground truth for ED

(c) RV Segmentation for ES (d) RV Segmentation ED

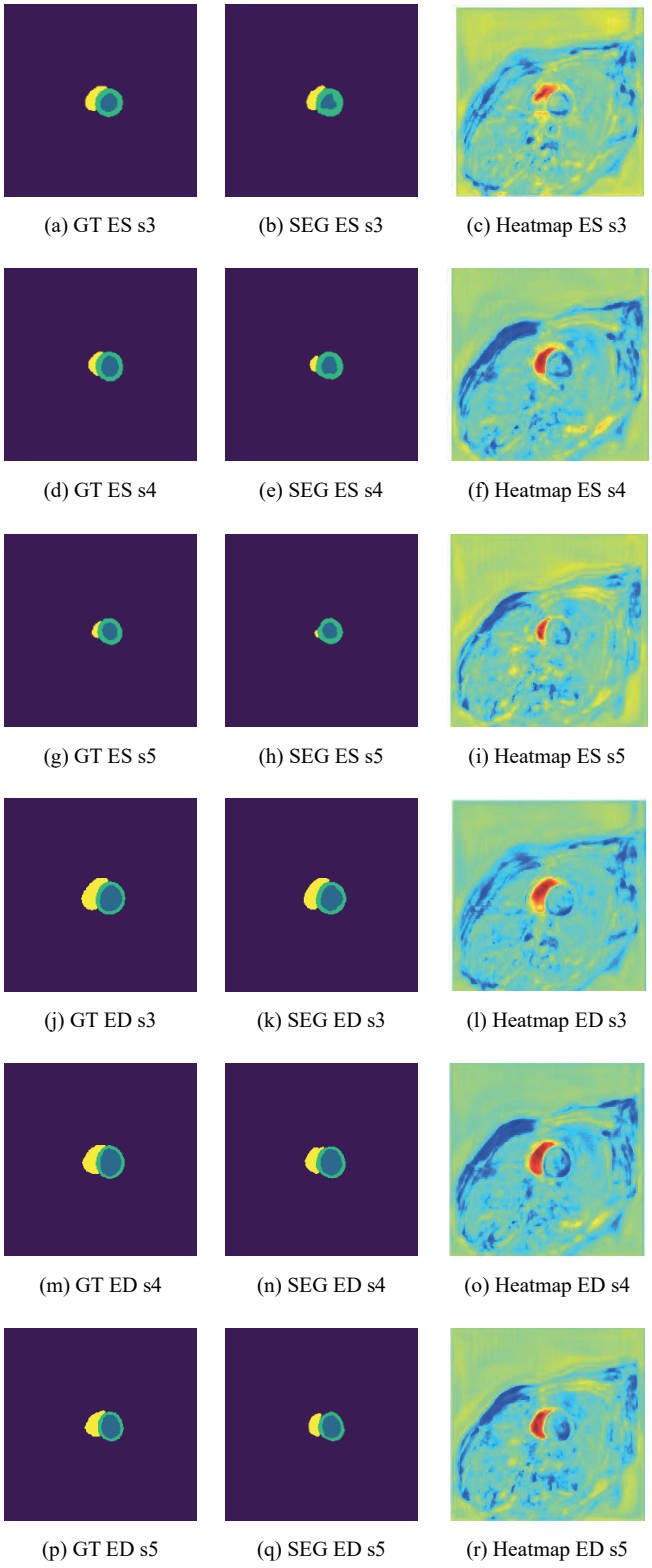
(e) Heatmap for ES (f) Heatmap for ED

Fig. 8. Composite image of GradCAM attribution heatmaps with RV as target class.

Short Axis ES and ED CMRI Saliency Maps from GradCAM for RV: Figure 9 displays a composite image of randomly selected SA, ES and ED CMRI ground truth (GT), segmentation (SEG) output, and attribution heatmaps from applying GradCAM with RV as a target class for the Slices 3-5 of the said ES and ED CMRI.

As the target class is set to RV for this GradCAM visualisation, both ES (as visualised by sub-figures in Fig. 9 (c), (f), (i)) and ED (as visualised by sub-figures in Fig. 9 (l), (o), (r)), the highest salient score appears to be RV, followed by the background. This validates the model's gradient flow in SA images as well.

A key distinction between the ES and ED attribution heatmaps here lies in the extent of the region impacting the model's predictions. In ED saliency heatmaps, the model is influenced by a larger area, causing RV to be represented with higher intensity compared to ES saliency heatmaps. This aligns with the increased cardiac volume during the ED phase [21] suggesting a fair model that distinguishes the RV from other cardiac structures.



(a) GT ES s3 (b) SEG ES s3 (c) Heatmap ES s3

(d) GT ES s4 (e) SEG ES s4 (f) Heatmap ES s4

(g) GT ES s5 (h) SEG ES s5 (i) Heatmap ES s5

(j) GT ED s3 (k) SEG ED s3 (l) Heatmap ED s3

(m) GT ED s4 (n) SEG ED s4 (o) Heatmap ED s4

(p) GT ED s5 (q) SEG ED s5 (r) Heatmap ED s5

Fig. 9. Composite image of ground truth (GT) and segmentation output (SEG) and GradCAM attribution heatmaps with RV as target class, for ES and ED Slices 3-5 (s3-s5).

Feature Ablation Heatmap Analysis: In addition to GradCAM, feature ablation has been employed from PyTorch Captum [20] to understand the decision-making process of both 2D and 3D CMRI segmentation models that were trained on the MnMs2 dataset. The approach involves the systematic ablation of each input class and observing the resulting effect on the model output. Feature ablation was used on the input images focusing on a specific target class of interest (in this case RV) on the model prediction for the said input images. The results of the ablation study, presented below, offer insights as to which classes of the CMRI were most influential in the segmentation of the target class, improving the transparency and interpretability of the behaviour of the DNNs.

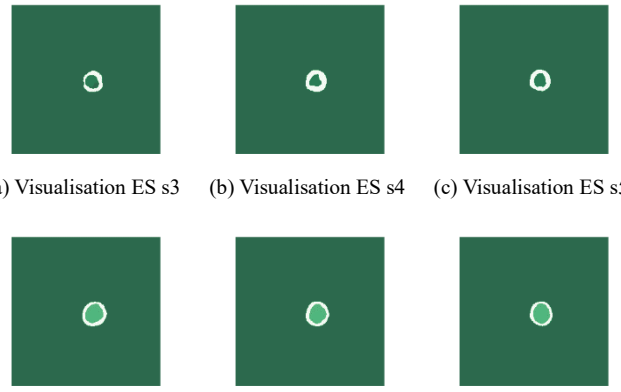
Fig. 10. Colour bar used for feature ablation maps. Dark green in the colour bar depicts the highest positive impact and white depicts the least impact.

Long Axis ES and ED CMRI Feature Ablation Visualisations: Figure 11 shows the visualisations of the attribution results gained by ablating each segment, and illustrates the impact of the segment on detecting the target class for both ES and ED images' RV. In feature ablation attribution visualisations (Fig. 11 (a) and (b)), the same classes seem to impact more in segmenting the RV. In both the images, background has the highest impact, but, in the ES image, (a) the target class itself, RV appears to have a very slightly lower contribution than the contribution of RV in the ED (b). Additionally, there is a slightly higher contribution from the LV in the ES (a) than in the (b). Overall, in both ED and ES images it appears that the same classes have a higher impact on RV segmentation.

(a) Visualisation for ES (b) Visualisation for ED

Fig. 11. Composite image of feature ablation visualisation with RV as target class.

Short Axis ES and ED CMRI Feature Ablation Visualisations: Figure 12 depicts a composite figure containing the feature ablation attribution visualisations from three slices of two randomly selected ES and ED LA CMRI for LV segmentation. This composite image only depicts three slices s (3,4,5) from each image. The attribution visualisation maps in figures (a), (b), (c), (d), (e), (f) show that regions belonging to background and RV class are most impactful. Both visualisations show that MYO is not contributing a higher impact to the final segmentation of RV; however, it can be observed that in ES images the impact from LV is higher compared to the impact from LV in ED, as depicted by the attribution maps (d), (e), (f).



(a) Visualisation ES s3 (b) Visualisation ES s4 (c) Visualisation ES s5

(d) Visualisation ED s3 (e) Visualisation ED s4 (f) Visualisation ED s5

Fig. 12. Composite image of feature ablation attribution maps with RV as target class, for ES and ED Slices 3-5 (s3-s5).

V. Conclusion

This research presented an exploration of semantic segmentation and explainable AI techniques applied to the MnMs-2 Cardiac MRI dataset. We achieved interesting results in terms of DSC and HD, demonstrating the effectiveness of our proposed approach for cardiac MRI segmentation. Furthermore, we incorporated two explainable AI techniques, Grad-CAM and feature ablation, to visualise the regions of interest guiding the segmentation model for the queried class. This enhanced interpretability allowed us to gain insights into the model's decision-making process. The authors believe that this will eventually provide more transparency and confidence of AI models in clinical segmentations tasks in real world.

Building upon these results, future endeavours will focus on expanding the reach and impact of our approach. Firstly, we aim to investigate the generalisability of our methodology by applying it to various medical imaging datasets and diverse segmentation tasks, testing its adaptability across different imaging modalities (target classes) and clinical anatomical structures while employing more advanced pre-processing techniques. Secondly, we plan to delve deeper into the model's inner workings by exploring more advanced explainable AI techniques [22]. This will provide richer insights into its reasoning, decision-making processes, and potential biases, fostering trust and understanding. Finally, we hope that our work will pique the interests of works that lead to integrating explainable AI into clinical workflows, facilitating a collaborative environment, where human expertise blends with AI capabilities. This synergy has the potential to significantly improve medical decision-making, ultimately leading to better patient outcomes.

Code and Dataset Availability

The source code for the segmentation models and XAI applications can be found in the following GitHub page [23]. The dataset can be accessed from M&Ms-2 Challenge website [17].

Contribution Statement

All authors acknowledge, that all authors contributed equally to this study.

References

[1] World Heart Federation, "World heart report 2023. Confronting the world's number one killer," 2023. [Online]. Available: https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf

[2] J. Edward, J. Banchs, H. Parker, and W. Cornwell, "Right ventricular function across the spectrum of health and disease," *Heart*, vol. 109, pp. 349–355, May 2022. https://doi.org/10.1136/heartjnl-2021-320526

[3] C. Martín-Isla *et al.*, "Deep learning segmentation of the right ventricle in cardiac MRI: The M&Ms challenge," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3302–3313, Jul. 2023. https://doi.org/10.1109/JBHI.2023.3267857

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv:1411.4038*, Nov. 2014. https://doi.org/10.48550/arXiv.1411.4038

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *arXiv:1505.04597*, May 2015. https://doi.org/10.48550/arXiv.1505.04597

[6] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, Oct. 2020. https://doi.org/10.48550/arXiv.2010.11929

[7] Y.-Z. Li *et al.*, "RSU-Net: U-net based on residual and self-attention mechanism in the segmentation of cardiac magnetic resonance images," *Computer Methods and Programs in Biomedicine*, vol. 231, Aug. 2023, Art. no. 107437. https://doi.org/10.1016/j.cmpb.2023.107437

[8] C. Fan, Q. Su, Z. Xiao, H. Su, A. Hou, and B. Luan, "ViT-FRD: A vision transformer model for cardiac MRI image segmentation based on feature recombination distillation," *IEEE Access*, vol. 11, pp. 129763–129772, Jan. 2023. https://doi.org/10.1109/access.2023.3302522

[9] K. Borys *et al.*, "Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches," *European Journal of Radiology*, vol. 162, May 2023, Art. no. 110786. https://doi.org/10.1016/j.ejrad.2023.110786

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020. https://doi.org/10.1007/s11263-019-01228-7

[11] S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass, CO, USA, Mar. 2020, pp. 972–980. https://doi.org/10.1109/WACV45572.2020.9093360

[12] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, *Lecture Notes in Computer Science*, W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and K. R. Müller, Eds., vol. 11700, Sep. 2019, pp. 193–209. https://doi.org/10.1007/978-3-030-28954-6_10

[13] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv:1705.07874*, Nov. 2017. https://doi.org/10.48550/arXiv.1705.07874

[14] A. Janik, J. Dodd, G. Ifrim, K. Sankaran, and K. M. Curran, "Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset," *Medical Imaging 2021: Image Processing*, vol. 11596, Feb. 2021. https://doi.org/10.1117/12.2582227

[15] A. Kaur, G. Dong, and A. Basu, "GradXcepUNet: Explainable AI based medical image segmentation," in *Smart Multimedia. ICSM 2022. Lecture Notes in Computer Science*, vol 13497. Springer, Cham, Jan. 2022, pp. 174–188. https://doi.org/10.1007/978-3-031-22061-6_13

[16] R. Gipiškis, D. Chiaro, D. Annunziata, and F. Piccialli, "Ablation studies in activation maps for explainable semantic segmentation in industry 4.0," in *IEEE EUROCON 2023 – 20th International Conference on Smart Technologies*, Torino, Italy, Jul. 2023, pp. 36–41. https://doi.org/10.1109/eurocon56442.2023.10199094

[17] MICCAI2021, "Multi-disease, multi-view & multi-center. Right ventricular segmentation in cardiac MRI (M&Ms-2)." [Online]. Available: https://www.ub.edu/mnms-2/ Accessed on: Feb. 27, 2024.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," *arXiv:1502.01852*, Feb. 2015. https://doi.org/10.48550/arxiv.1502.01852
[19] D. Karimi and S. E. Salcudean, "Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 39, no. 2, pp. 499–513, Feb. 2020. https://doi.org/10.1109/tmi.2019.2930068
[20] N. Kokhlikyan *et al.*, "Captum: A unified and generic model interpretability library for PyTorch," *arXiv:2009.07896*, Sep. 2020. https://doi.org/10.48550/arXiv.2009.07896
[21] UCSF, "How the heart works," UCSF Department of Surgery, 2024. [Online]. Available: https://surgery.ucsf.edu/condition/how-heart-works
[22] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *arXiv:2005.13799*, May 2020. https://doi.org/10.48550/arXiv.2005.13799
[23] Ayoob7, "Ayoob7/Peering_in_to_the_heart," GitHub, 2024. [Online]. Available: https://github.com/Ayoob7/Peering_in_to_the_heart Accessed on: Apr. 28, 2024.

Mohamed Ayoob received a B. Eng. (Hons) degree in Software Engineering from the University of Westminster, the UK, in 2020. He completed his MPhil in Computer Science at the University of Nottingham. Currently, he is a Lecturer at the Informatics Institute of Technology, Colombo, Sri Lanka. His research interests include computer vision and artificial intelligence and their associated applications in the environment and healthcare. Mr. Ayoob obtained the best final year research project award during his undergraduate studies.
E-mail: ayoob.m@iit.ac.lk
ORCID iD: https://orcid.org/0000-0003-3585-4383

Oshan Nettasinghe received a B. Eng. (Hons) degree in Software Engineering from the Informatics Institute of Technology (IIT), Colombo, Sri Lanka, affiliated with the University of Westminster, thr UK in 2023. Currently, he is pursuing an undergraduate degree in Computer Engineering at the University of Ruhuna, Faculty of Engineering, Sri Lanka. His primary field of study is computer vision, with a focus on vision transformer-based architectures for image recognition and processing. He has contributed to various projects in artificial intelligence and computer vision, exploring novel applications of machine learning algorithms.
E-mail: oshan.20191052@iit.ac.lk
ORCID iD: https://orcid.org/0009-0000-9093-1914

Vithushan Sylvester received a B. Eng. (Hons) degree in Software Engineering from the Informatics Institute of Technology (IIT), Colombo, Sri Lanka, affiliated with the University of Westminster, the UK in 2024. Currently, he is a Technical Lead at Right Data Inc. with over 5 years of experience in software development. While working with the industry he has contributed to enabling generative AI in various production systems. His research interests include computer vision, medical-imaging, generative AI and multi-agent systems. He obtained a National Ingenuity Award from SLASSCOM for best project / product – university category in 2021.
E-mail: vithushan.2019830@iit.ac.lk
ORCID iD: https://orcid.org/0009-0006-5289-8176

Helmini Bowala received a B. Eng. (Hons) degree in Software Engineering from the Informatics Institute of Technology (IIT), Colombo, Sri Lanka, affiliated with the University of Westminster, the UK in 2024. Currently, she is working as an apprentice Software Engineer at Informatics International Limited. Her main focus of study is artificial intelligence in medical imaging and explainable AI. She obtained a National Ingenuity Award from SLASSCOM for best project / product – university category in 2021.
E-mail: helmini.2019948@iit.ac.lk
ORCID iD: https://orcid.org/0009-0004-4468-902X

Hamdaan Mohideen received a B. Eng. (Hons) degree in AI & Data Science from the Informatics Institute of Technology (IIT), Colombo, Sri Lanka affiliated with Robert Gordon University, the UK in 2024. Currently, he is working as a Data Engineer at Circles, Sri Lanka. His research interests include deep learning, computer vision, graph neural network, and data science.
E-mail: hamdaan.20200977@iit.ac.lk
ORCID iD: https://orcid.org/0009-0003-3465-0976