

Using Deep Learning to Classify LV Scarring in Diverse Patient Populations

Ramiz S. Akhtar^a, Rohan D. Patel^a, Vignesh Valaboju^a, W. Sharon Zheng^a, Xue Feng^{a,b,1}

^a Department of Biomedical Engineering at the University of Virginia

^b Carina Medical; Charlottesville, Virginia

¹ Correspondence: Xue Feng

Email: xf4j@virginia.edu

Abstract

Artificial intelligence (AI) can be leveraged to solve modern day medical challenges such as a timely diagnostic for cardiac disease. These diagnostics are crucial for patient-specific diagnoses and treatments. Cardiologists diagnose and treat cardiovascular diseases, including hypertrophic cardiomyopathy (HCM) and acute myocardial infarction (AMI), through manual image segmentation of left ventricular (LV) scarring from cardiac magnetic resonance imaging (MRI) slices. Manual segmentation of scar is often an arduous task that is subjected to bias, error, and physician fatigue. An existing AI algorithm that automates LV scar segmentation, developed by Carina Medical, has 19% and 5% false positive (FP) and false negative (FN) rates respectively for the identification of scar. The segmentation algorithm has a scar identification accuracy of 76%. The high error rates make scar identification unreliable; thus, this paper discusses the development of a novel ensemble learning pipeline that couple segmentation and classification deep learning algorithms to help improve the robustness of scar identification in cardiac MRI. By aggregating the results of multiple deep learning algorithms, the pipeline can more confidently identify the presence of LV scar. Instances where the segmentation and classification model disagree on the presence of scar are filtered out and classified as warning cases for cardiologists to manually analyze. The coupled pipeline improved scar identification accuracy to 88.1% and reduced cardiologist workload by 68.4%. FP rates were reduced from 19% to 6.4%, while FN rates remained similar at 5% and 5.5% when comparing the segmentation model and the novel coupled segmentation and classification pipeline.

Keywords: LV Scar Classification, LV Scar Segmentation, Ensemble Learning, Myocardial Infarction, Deep Learning in Medical Imaging.

Introduction

Cardiovascular disease is the leading cause of death in the United States, accounting for 1 in every 4 deaths [1]. A strong predictor of sudden cardiac death is LV dysfunction, which can be caused by the development of myocardial scar tissue. An increased presence of myocardial scar tissue is associated with higher risk and mortality rate from cardiovascular disease. Early detection of LV scar can be treated by a ventricular reconstruction surgery or medications, but improvement of LV function with severe damage is often unlikely, and mortality remains high [2]. The paper focuses on two diseases of interest: AMI and HCM. HCM patients often have thickening of the cardiac wall paired with regions of scar due to infarction, while

AMI patients often have collagenous scar that results from the replacement of dead myocytes within the myocardium. Segmentation and quantification of the myocardial scarring in HCM patients and AMI patients is vital to guide personalized patient treatments [3].

Cardiac MRI allows clinicians to perform non-invasive and quantitative analysis of potential cardiac disease states by extracting imaging biomarkers [4]. Late gadolinium enhanced MRI (LGE-MRI) can be used to visualize myocardial infarction and identify the size and shape of LV scar [5]. LV scar is analyzed through image segmentation where the location, size, and shape of the scar is identified from the MRI scans. Current clinical image segmentation practices are insufficient because images are manually

segmented by cardiologists, often leading to human error and inter-observer biases [6]. Manual segmentation is also laborious and slow which can be detrimental during time-sensitive medical procedures [7].

Deep learning, a subset of AI, segmentation algorithms have been leveraged in several fields to automate and standardize manual image segmentation. Similar deep learning segmentation algorithms can be used on medical images to increase the accuracy and efficiency of diagnosis by obtaining quantitative and qualitative biomarkers from medical images. There are currently several deep learning algorithms in development for automated LV scar segmentation. Carina Medical previously developed an LV scar segmentation deep learning model. The model was trained on the EMIDEC dataset, which consisted of AMI patient LGE-MRI scans, and a HCM patient dataset [8]. The segmentation model used a UNet-2D architecture to segment four main regions of interest (ROIs): LV blood cavity, normal LV myocardium wall, myocardial infarction area, and no-reflow scar area [9]. Although the segmentation model can effectively segment key ROIs including infarction scar, exploratory analysis of the model shows that the model produces several FPs and FNs for scar identification, resulting in a high error rate. A FP is defined when the model segments scar when there is no scar present in the ground truth while, and a FN is defined when the model fails to segment scar when there is scar present in the ground truth. The ground truth is the segmentation map manually produced by the cardiologist. The model accurately identified the presence of scar 76% of the time; 19% of the MRI slices are FPs, and 5% of the MRI slices are FNs. A high number of FPs and FNs result in poor patient diagnoses and decreased physician and patient confidence in medical deep learning algorithms.

To improve the LV scar segmentation, several ensemble learning algorithms have been developed. Ensemble learning algorithms aggregate the results of several deep learning algorithms to improve a task's accuracy. For instance, CMPU-Net cascades two different UNet segmentation algorithms to segment LV scar [10]. Although the algorithm performs well, the model was trained on a small number of patients. For a deep learning segmentation model to be accurate and generalizable, the model must be trained on large, diverse patient populations. Generalizability is essential because cardiologists in clinical practice see patients of various age groups, races, and disease states. The CMPU-net algorithm, which couples two segmentation algorithms, is computationally expensive and subsequently difficult to transition to clinical use. Other

studies have tried to couple a segmentation model and a computationally inexpensive classification algorithm to improve segmentation tasks. For example, a YNet architecture was developed by coupling instance-level segmentation masks and instance-level probability maps from classification. Both are combined to produce a segmentation mask for a ROI [11]. YNet was tested on breast biopsy images. To our knowledge, YNet has not been tested on LV scar MRIs.

Thus, we propose a novel pipeline that uses ensemble learning by coupling a scar classification model with Carina Medical's LV scar segmentation model to lower the scar identification error rate (**Figure 1**). The pipeline uses a computationally inexpensive classification model to filter out FPs and FNs. Furthermore, both the segmentation and classification models are trained on diverse patient populations. The pipeline creates a semi-automated LV scar segmentation and classification model where only filtered FPs and FNs are manually analyzed by a cardiologist. The pipeline is confident in segmentations that are not filtered out by the classification model. In addition, the pipeline does not combine the masks and maps of the segmentation and classification models like YNet. The pipeline is a simpler aggregation of the segmentation and classification models where each deep learning model runs independently.

As shown in **Figure 1**, the classifier and the segmentation algorithms work in parallel. The classifier identifies whether scarring exists on an MRI slice while the Carina Medical LV segmentation model segments the size and location of scar. From the classifier, we do not obtain the location of the scar, size of scar, or shape of the scar. The most important outputs for our pipeline are the scar size and location from the segmentation model and the binary decision from the classifier regarding the presence of scar within a given slice. Subsequently, processing of the outputs is best visualized by the truth table shown in **Figure 1**. If the classification and segmentation models agree on the presence of scar, the pipeline is confident in the output of those slices and sends the segmentation map to the cardiologist. Specifically, the models agreeing means that the classifier identified scar and the segmentation algorithm segmented scar, or the classifier identified no scar and the segmentation algorithm segmented no scar. If the classifier and segmentation model disagree on the presence of scar, the respective MRI slice is considered a warning slice. The cardiologist must manually segment warning slices. As a result, the pipeline creates a semi-automated segmentation

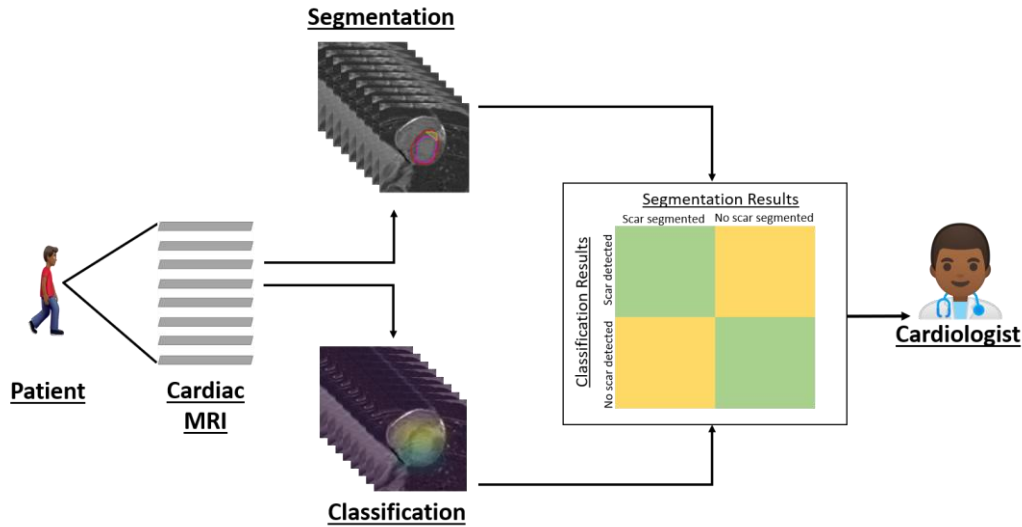


Fig. 1. Depicted is the organization of our novel coupled pipeline. The pipeline leverages a scar classifier in tandem with the Carina Medical LV scar segmentation algorithm to improve scar identification and limit the number of FPs and FNs.

algorithm where only warning slices must be manually segmented.

The project aims to couple the Carina Medical LV scar segmentation model and a scar classifier to improve the confidence of a deep learning model in identification of LV scar. The bulk of the project focuses on building and validating a deep learning classification model. Lastly, the project analyzes the effectiveness of the coupled pipeline to reduce FPs and FNs.

Materials and Methods

Dataset Description

To train the scar classifier and encourage model generalizability, two unique LGE-MRI datasets were used: the EMIDEC AMI patient dataset and a subset of the National Heart Lung and Blood Institute HCM patient dataset. Both datasets contain MRI slices that are 8 mm thick, acquired at 1.5 or 3 Tesla on MRI systems using electrocardiographic gating after the injection of gadolinium-based contrast agent. However, the acquisition protocol between the AMI and HCM datasets differ. The acquisition of EMIDEC dataset is 10 min post-contrast using the Siemens MRI system, while the HCM datasets acquire information 5-, 14-, and 29-minutes post-contrast using MRI systems from the 3 primary vendors (General Electric, Philip Medical System, and Siemens), resulting in different image contrasts and resolutions.

The AMI dataset contained 100 anonymized patients. 60 pathological patients had scarring from AMI while 40

patients were normal. The mean patient age was 61 years old. 12% of the patients were diabetic, and 57% of the patients were overweight. Race, ethnicity, and socioeconomic status are unknown [8]. The training and testing split was 60/40, where 60 patients were used for training and 40 patients were used for testing. As a result, 422 and 286 MRI slices were used for training and testing, respectively.

The subset of the HCM dataset used contained 67 anonymized patients. All patients were 65 years or younger. Data was collected from 41 hospital sites across the United States, Canada, and Europe [3]. Additional demographic information regarding the dataset is unavailable. The training and testing split was 70/30, where 46 patients were used for training and 21 were used for testing. As a result, 366 and 169 MRI slices were used for training and testing, respectively.

The training and testing MRI slices were combined across the two datasets, meaning there were a total of 106 training patients and 61 testing patients. A training and validation split of 80/20, respectively, was applied to the combined training data. As a result, 630 MRI slices were used for training, 158 MRI slices were used for validation, and 455 MRI slices were used for testing.

Classification Algorithm

The Xception transfer learning convolutional neural network (CNN) framework was used to build the classification algorithm. A CNN uses a series of convolutions to extract unique features from a given image.

Xception’s architecture is based entirely on depth wise separable convolutions. They consist of a single pointwise convolution, which down samples data from each channel within a slice, followed by a series of depth wise convolutions, which focus on aggregating data from neighboring pixels within slice [12]. By combining pointwise and depth wise convolutions in the aforementioned manner, features are extracted from every aspect of a given slice. Global average pooling, batch normalization, and dropout layers were added after the Xception network architecture. Binary SoftMax classification was performed. An Adam optimizer and sparse categorical cross entropy loss function were used.

The MRI slices were preprocessed such that each slice has a zero mean and unit variance. The slices were then center cropped to be 112×112 (width × height). Since Xception requires a 3-channel RGB input, each slice was replicated three times such that the dimension of each slice was 112×112×3 (width × height × channel). The model was trained for 45 epochs with a batch size of 64. In addition, the dataset was shuffled before and during training. Validation accuracy and loss were also monitored during training. The aforementioned parameters were found through a grid search algorithm which systematically found the best model optimizers, learning rates, and batch sizes. Due to the limited datasets, image augmentation is implemented to address different imaging modalities and populations by generating more training images. The MRI images come from a variety of imaging systems, resulting in MRI images of varying resolutions. To mitigate the impact of the aforementioned variances on the model, gaussian noise from the batchgenerators Python package was added to help the network ignore differences in resolution and other variances in the different patient populations [13]. For each slice, batchgenerators created 4 new slice images by randomly applying gaussian noise with a variance ranging from 0 to 0.5. To further increase the generalizability of the model, ImageDataGenerator API in Keras was used to apply random affine transformations, including 180-degree rotations, horizontal and vertical flips, 20% width shifts, 10% shear, and 20% zoom. Combining both gaussian noise and affine transformations, we expanded the size of the pre-existing dataset by a factor of 48. All experiments were performed using NVIDIA GPUs on the Carina Medical UNIX Server. Deep learning models were constructed using TensorFlow 2.4.

GradCAM (Interoperability Model)

Once the Xception model was trained, validated, and tested on the appropriate datasets, we used an interpretability

model to confirm the classifier’s robustness. A major question in using classification algorithms is whether the algorithm is actually looking at targeted ROI. Often the classifier could potentially be guessing whether there is scar, which is especially true for binary classification. To better understand the classifier, we explored the use of interpretability models, specifically the Gradient Class Activation Map (GradCAM) algorithm. GradCAM is an algorithm used to help quantitatively analyze where a convolutional neural network is looking. GradCAM works by finding the final layer in the network and then examining the gradient information flowing into that layer. GradCAM focuses on the final layer of the CNN to understand each neuron for a decision of interest since those final neurons are of paramount importance in determining the classification label for a given MRI slice. The output of GradCAM is a heatmap for a given class label. In our case, the only class labels are “scar present” and “no scar present.” **Figure 2** showcases an output of GradCAM. Higher ROIs, regions in yellow, are areas that the classification model considers with a higher weight in comparison to the medium ROIs in blue. GradCAM heatmaps are produced for each MRI slice passed into the classifier to visually verify the focus of the CNN on the image. In addition, GradCAM heatmaps are overlaid with the ground truth segmentation maps to quantify if the classifier is looking at the scar.

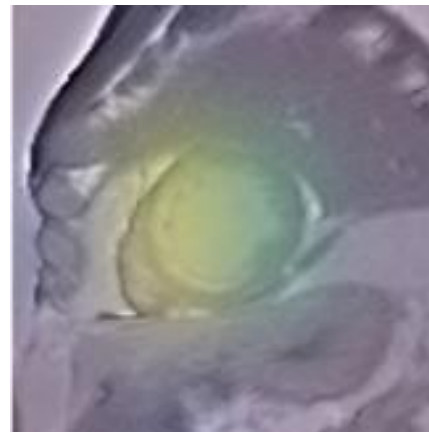


Fig. 2. *Depicted is the output of GradCAM for an AMI patient (Patient P096). Within the GradCAM output, we can see the heatmap is focused on the myocardium. In addition, the GradCAM output contains 2 primary colors: yellow and blue. The yellow showcases regions of the heat map that are higher regions of interest (ROI), while the blue showcases regions of the heat map that are medium ROI.*

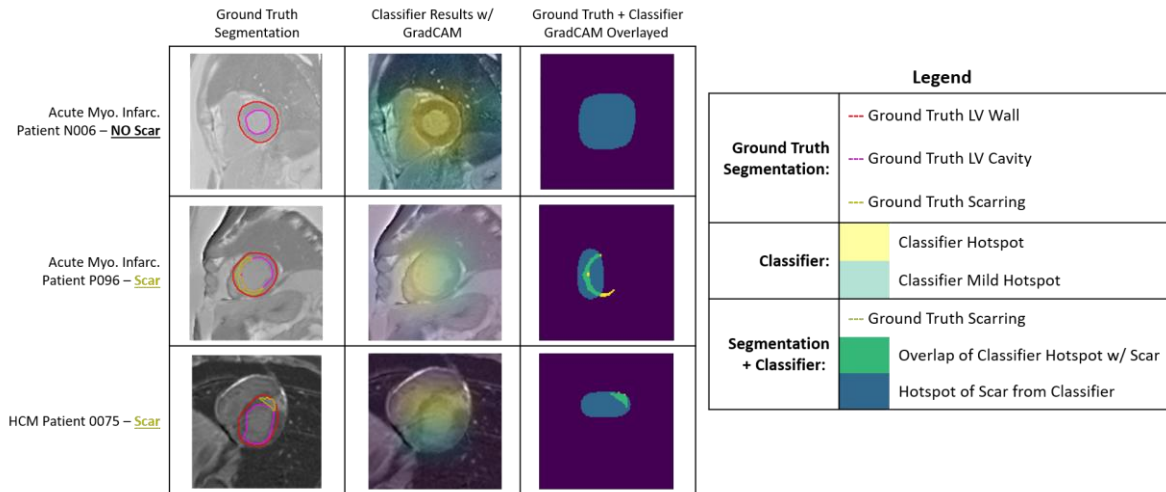


Fig. 3. Depicted is the ground truth segmentation, the classifier GradCAM heatmap results, and the overlay between the classifier's GradCAM high ROI with ground truth scar segmentation for three different patients. The first column showcases the ground truth segmentation, which is treated as the true location, size, and shape of the LV wall, LV cavity, and scar. The second column contains the GradCAM output. The third column displays the higher ROI from the GradCAM output overlaid with the ground truth scarring. Within the third column, the blue color represents the GradCAM higher ROI only, the yellow color represents the ground truth scar labels only, and the green color represents the overlap between the ground truth scar and higher ROI. The first row showcases a patient slice without scar, the second row and the third row both showcase patient slices with scar from AMI and HCM, respectively. Quantifying the overlap between the ground truth scar and GradCAM's high ROI for every slice classified to have scar, we see that $43 \pm 37\%$ of the scar will be considered.

Results

GradCAM Results

Using GradCAM we were able to visually observe the focus of the classifier. **Figure 3** suggests that the heatmap is focused on the LV myocardium, which verifies that the model is looking at scarring. 88% of the slices had the GradCAM heatmap focused on the myocardium. To test whether the classification algorithm is looking at scar, we decided to overlay the GradCAM higher ROIs with the ground truth segmentations of myocardial scarring as shown in column three of **Figure 3**. If the higher regions of interest on the heatmap encompass the ground truth values of the myocardial scarring, then it suggests that our classification algorithm is deciphering between scar or no scar. Out of all the slices, on average the GradCAM higher ROI encompasses about $43 \pm 37\%$ of the scar. This means that given for a slice, $43 \pm 37\%$ of the scar will be considered or "looked at" by the classification algorithm. When we encompass both the higher and medium regions, then on average the heatmap encompasses about $83 \pm 30\%$ of the scar. The GradCAM results suggest that our classification algorithm is looking at scar. In fact, our GradCAM results showcase a higher-than-average value for

heatmap encompassing important biomarkers. For example, in overlaying GradCAM heatmaps to find pneumothorax (a collapsed lung) within thoracic cavity images, a previous study found that the GradCAM heatmap only encompasses around 33% of pneumothorax [14].

Classifier and Segmentation Results

Once we identified that the classification model is looking at the scar, we calculated the accuracy of the classifier. In addition, we calculated the FP and FN rates of the classifier. The classifier accuracy came out to be 76%; the FP rate was 9.7% and the FN rate was 14.3%. The classification model training accuracy and loss are in **Supplementary Figure 1**. The Carina Medical LV segmentation model has a scar identification accuracy of 76%, a FP rate of 19%, and a FN rate of 5%. The results are visualized in **Figure 4**. The segmentation and classification models have similar accuracies, FP rates, and FN rates.

Coupled Pipeline Results

Figure 4 shows the increase in accuracy of the segmentation model and classifier coupled pipeline and the change in FPs and FNs. The coupled model has an accuracy of 88.1%. Slices are FPs when both the classifier and

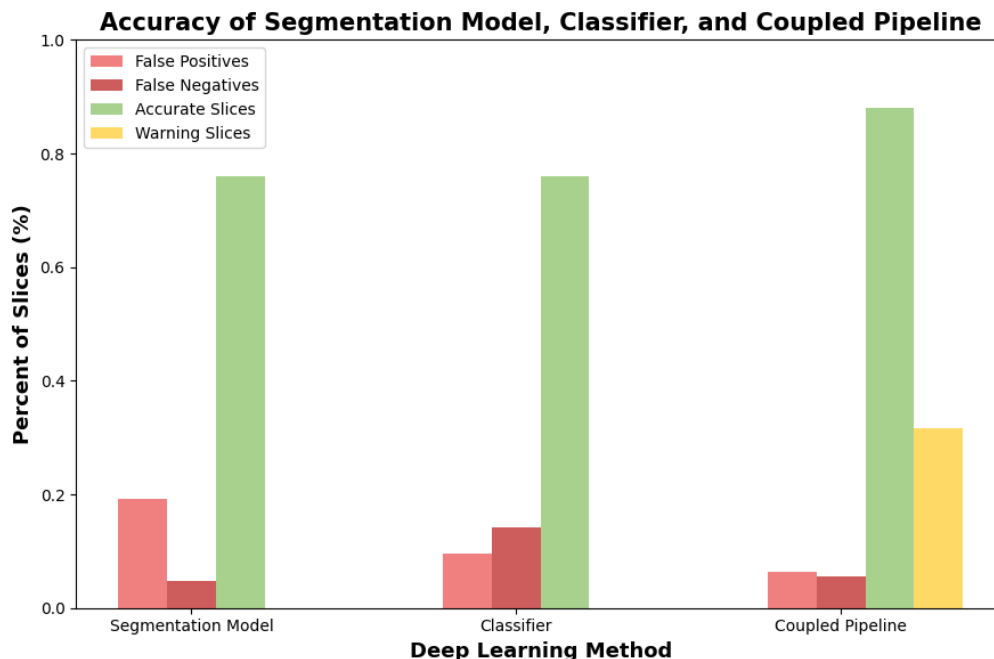


Fig. 4. The bar graph depicts the accuracies, FPs, and FNs of the testing set for the segmentation model, classifier, and coupled pipeline. Warning slices are instances where the segmentation model and classifier disagreed. The accuracy of the coupled pipeline was computed as (# of Slices Segmentation Model and Classifier agree on) / (Total MRI Slices - Warning Slices).

segmentation model individually output FPs. The same goes for FNs. For the coupled pipeline, the FP rate is 6.4% and the FN rate is 5.5%. The coupled pipeline also identifies warning slices when the segmentation model and the classifier disagree. A disagreement means that the pipeline is not confident about the presence of the scar. The pipeline suggests that warning slices be manually evaluated by the cardiologist. The cardiologist does not have to manually evaluate cases where the segmentation model and classifier agree. 31.6% of the slices are warning slices. **Table 1** shows on how many slices the classifier and segmentation model agreed upon. The warning slices are not included in the pipeline accuracy since the pipeline is not confident in the output. Example slices of FPs and FNs identified by the pipeline are depicted in **Figure 5**.

Discussion

The results shown in **Figure 4** and **Table 1** suggest that the coupled pipeline reduces FP identification of normal myocardial scar, but there is no reduction in FN identification of normal myocardial scar. There is a cost-benefit analysis to the coupled pipeline. Although the pipeline produces several warning slices, the pipeline has a higher accuracy when considering the cases that the pipeline is confident in. When only the segmentation model is used, no warning slices are produced because the segmentation

		Classifier		
		False Positive	False Negative	Accurate Slice
Total MRI Slices = 455				
Segmentation Model	False Positive	20	0	67
	False Negative	0	17	5
	Accurate Slice	24	48	274

■ Segmentation Model and Classifier agree incorrectly
■ Segmentation Model and Classifier agree correctly
■ Disagreement between Segmentation Model and Classifier

Table 1. The table shows how many MRI slices were accurately classified and inaccurately classified by the classifier, segmentation model, and coupled pipeline on the testing dataset. Green shading means that the classifier and segmentation model agree, and the models correctly classified scar. Red shading means that the classifier and segmentation model agree; however, both models incorrectly classified scar. Yellow shading indicates that the classifier and segmentation model disagree; thus, the slices are considered warning slices.

model is confident in all the segmentations; however, the model is only 76% accurate. When prioritizing patient outcomes, the pipeline poses a better error rate of 11.9% compared to the using only the segmentation model. In addition, when compared to the traditional scenario when no deep learning algorithm is used, the pipeline reduces cardiologist workload by 68.4%. The absence of a deep learning algorithm means that the cardiologist must manually segment 100% of the slices. The pipeline produces 31.6% warning slices; thus, the cardiologist only has to manually segment 31.6% of the slices. When using only the segmentation model, the cardiologist workload is decreased by 100%; however, the lower accuracy of the segmentation model leads to poorer patient outcomes.

The coupled pipeline reduces the FP rates from 19% to 6.4% when compared to the segmentation model. The pipeline has a FN rate of 5.5% while the segmentation model has a FN rate of 5%; thus, the pipeline is not able to decrease the number of FNs. In addition, the pipeline was trained on a diverse patient population. By training the algorithms on AMI and HCM patients, the pipeline is more generalizable. The experiments suggest that the coupled pipeline is an improvement over the current segmentation model in terms of overall accuracy and reduction in FPs.

Figure 5 highlights representative slices of FPs and FNs produced by the coupled pipeline. For the pathological FN slice, the segmentation model may have had trouble identifying the location of scar since the scar is relatively small compared to scar in the accurate slice. The small scar regions present in the FN slices provides a possible explanation as to why the segmentation model and classifier were not able to identify any scarring. Further fine tuning of the models and image augmentation process may help the models identify small regions of scarring. The thickening of the LV wall in the FN slice also likely made it difficult for the classifier to focus in one region. For the non-pathological FP case, the classifier had identified scarring even though there was no scar present, and the segmentation model falsely segmented a small region of scar only. For the accurate slice, the figure suggests that the classifier was looking at scar the same place the segmentation model segmented scar. The ground truth for the accurate slice suggests that the pipeline was looking at the correct place.

Limitations

A major limitation of our coupled pipeline is that the accuracy of the pipeline is limited by how well the segmentation model performs. Even if the classifier had 100% accuracy, the combined pipeline's ability to reduce

cardiologist workload is limited by the segmentation model's ability to segment scars. Therefore, improving the scar segmentation model is important because the cardiologist primarily desires scar segmentation maps. In addition, the true accuracy of the segmentation model is also limited by cardiologists' ability to create accurate ground truth segmentation maps that are used to train the segmentation model. Any biases or incorrect segmentations produced by cardiologists will propagate to the deep learning pipeline.

A key ethical limitation surrounds the question of liability in the event of the segmentation model and classifier both outputting a FP or FN result. Slices where the segmentation model and classifier agree ideally should not have to be reviewed by a cardiologist to truly reduce cardiologist workload. A FP that successfully passes through the pipeline would create alarm fatigue for the cardiologist by marking a slice with scar where no scar exists. However, a FN that successfully passes through the pipeline would result in missing a slice with scar which is potentially detrimental to patient outcomes. For our novel pipeline, while the number of FPs decreases to improve the pipeline accuracy over the segmentation model alone, the number of FNs remains approximately constant. Therefore, reducing FNs and further reducing FPs remain a central goal.

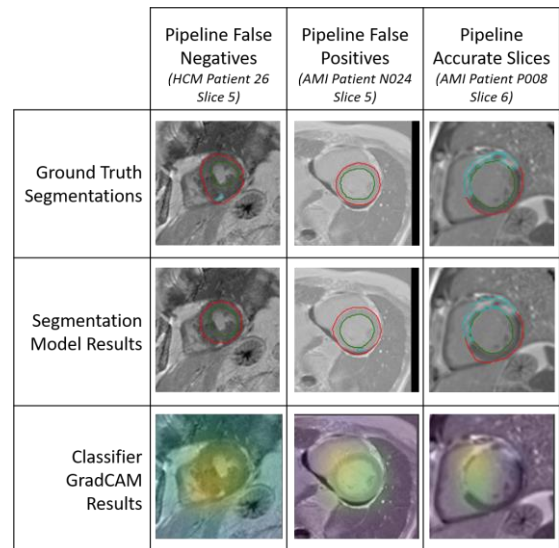


Fig. 5. Depicted are examples of FPs and FNs identified by the coupled pipeline. For each patient, ground truth segmentations, segmentation model results, and classifier GradCAM results are depicted. For the segmentations, the LV wall is shown in red, the LV blood cavity is shown in green, and normal myocardial scar is shown in cyan. The color coding of the GradCAM results matches that of Figure 2.

Future Work

Further improvements can be made to the deep learning classifier to help further reduce FNs and FPs. Fine tuning of the classification architecture and training set can be completed to amplify regions of small scarring so that fewer slices are FNs. In addition, the classification architecture can be iteratively improved in attempt to improve the accuracy of the classifier. As stated in the limitations, further improvements to the segmentation architecture need to be made to improve overall accuracy of the coupled pipeline.

To improve usability and encourage integration into a clinical workflow, our novel deep learning pipeline can be integrated into a user-friendly software package. When implementing new software in a clinical setting, cardiologists and other medical professionals must be comfortable with the newly implemented technologies to fully adopt them into their workflow [15]. A user-friendly interface is important because it discourages a negative impression of the software, ultimately increasing the software's perceived value. Additional software features can be developed to maximize the potential of the pipeline in the clinical setting. For example, MRI slices designated as containing scar by both the segmentation and classification models can be prioritized within the user interface and trigger a prominent notification to alert the cardiologist. The warning slices would be a secondary priority for the cardiologist to look at. In addition, further improvements on the interpretability model can help provide the cardiologist deeper insight into where the pipeline is looking to identify scar, ultimately increasing confidence and understanding of deep learning-based technologies for diagnosis.

Although the pipeline was trained on a diverse patient population, additional future work includes further increasing the diversity training set to remove algorithmic biases and improving model generalizability [16]. If any bias existed when patients were chosen to participate in the studies creating the MRI datasets, the deep learning pipeline's outcome could potentially reflect the bias when attempting to make a prediction from new patient data. Generalizability is critical to ensuring that the pipeline can be used on patients of all backgrounds and health states.

End Matter

Author Contributions and Notes

R.A., R.P., V.V., and W.Z. wrote the paper, designed and tested the myocardial scar classifier, and developed a

pipeline using Carina Medical LLC's existing LV scar segmentation algorithm. X.F. advised the project. The authors declare no conflict of interest.

Acknowledgments

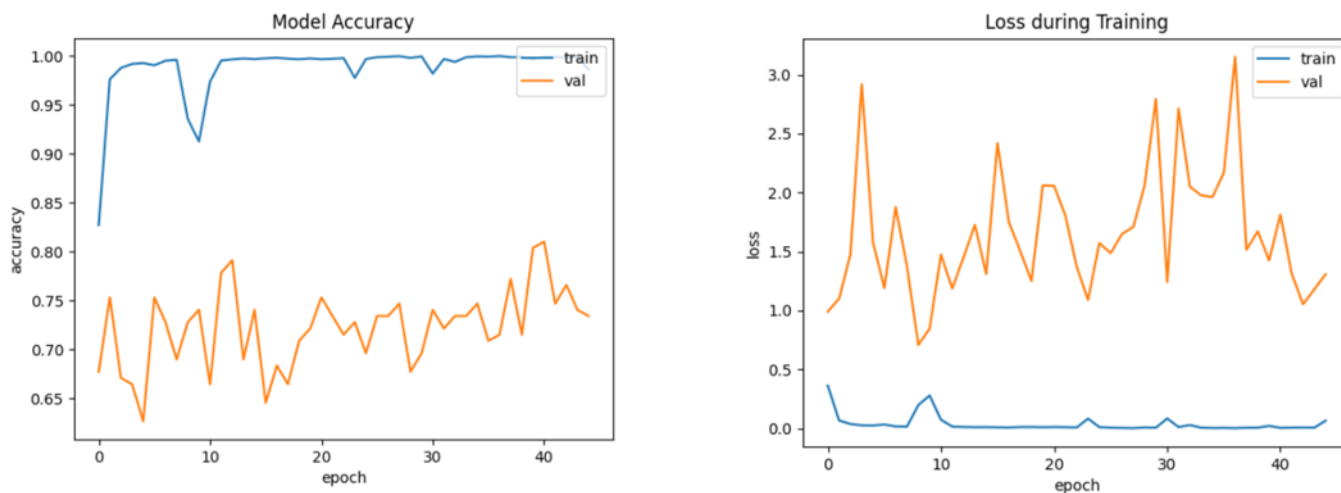
We would like to thank Dr. Xue Feng, our advisor at Carina Medical, for providing resources and advice on the project. His ceaseless and consistent support throughout the academic year made our capstone project possible. We would also like to thank Dr. Timothy E. Allen and Dr. Shannon Barker, as well as the rest of the teaching team from the University of Virginia Department of Biomedical Engineering Capstone course for advising throughout the project.

References

- [1] CDC, "Heart Disease Facts | cdc.gov," *Centers for Disease Control and Prevention*, Sep. 08, 2020. <https://www.cdc.gov/heartdisease/facts.htm> (accessed Oct. 25, 2020).
- [2] J. Bax, E. E. van der Wall, and M. Harbinson, "Radionuclide techniques for the assessment of myocardial viability and hibernation," *Heart*, vol. 90, no. Suppl 5, pp. v26–v33, Aug. 2004, doi: 10.1136/hrt.2002.007575.
- [3] C. M. Kramer *et al.*, "Hypertrophic Cardiomyopathy Registry (HCMR): The rationale and design of an international, observational study of hypertrophic cardiomyopathy," *Am. Heart J.*, vol. 170, no. 2, pp. 223–230, Aug. 2015, doi: 10.1016/j.ahj.2015.05.013.
- [4] E. P. Balogh *et al.*, *The Diagnostic Process*. National Academies Press (US), 2015.
- [5] Kim Raymond J. *et al.*, "Relationship of MRI Delayed Contrast Enhancement to Irreversible Injury, Infarct Age, and Contractile Function," *Circulation*, vol. 100, no. 19, pp. 1992–2002, Nov. 1999, doi: 10.1161/01.CIR.100.19.1992.
- [6] R. Merjulah and J. Chandra, "Chapter 10 - Classification of Myocardial Ischemia in Delayed Contrast Enhancement Using Machine Learning," in *Intelligent Data Analysis for Biomedical Applications*, D. J. Hemanth, D. Gupta, and V. Emilia Balas, Eds. Academic Press, 2019, pp. 209–235.

- [7] “Manual Segmentation Errors in Medical Imaging. Proposing a Reliable Gold Standard,” *springerprofessional.de*. <https://www.springerprofessional.de/en/manual-segmentation-errors-in-medical-imaging-proposing-a-reliab/17315296> (accessed May 03, 2021).
- [8] “Emidec - Dataset.” <http://emidec.com/dataset> (accessed Feb. 24, 2021).
- [9] X. Feng, C. M. Kramer, M. Salerno, and H. Meyer, “Automatic Scar Segmentation from DE-MRI Using 2D Dilated UNet with Rotation-based Augmentation,” p. 5.
- [10] F. Zabihollahy, M. Rajchl, J. A. White, and E. Ukwatta, “Fully automated segmentation of left ventricular scar from 3D late gadolinium enhancement magnetic resonance imaging using a cascaded multi-planar U-Net (CMPU-Net),” *Med. Phys.*, vol. 47, no. 4, pp. 1645–1655, 2020, doi: <https://doi.org/10.1002/mp.14022>.
- [11] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, “Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, vol. 11071, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 893–901.
- [12] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Jul. 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.
- [13] F. Isensee et al., batchgenerators - a python framework for data augmentation. Zenodo, 2020.
- [14] N. Arun *et al.*, “Assessing the (Un)Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging,” *Radiology and Imaging*, preprint, Jul. 2020. doi: 10.1101/2020.07.28.20163899.
- [15] M. L. Langan, A. Riera, J. C. Kurtz, P. Schaeffer, and A. G. Asnes, “Implementation of newly adopted technology in acute care settings: a qualitative analysis of clinical staff,” *J. Med. Eng. Technol.*, vol. 39, no. 1, pp. 44–53, Jan. 2015, doi: 10.3109/03091902.2014.973618.
- [16] R. Faruqi and A. Singh, “Best Practices for Addressing Risks Associated with a Lack of Diversity in Machine Learning,” p. 12.

Supplemental Figures



Supplemental Figure 1: Depicted above on the left are the training and validation accuracies of the classifier. The model was run for 45 epochs and the best model was saved at the highest validation accuracy of 81%. Depicted above on the right are the training and validation losses of the classifier. A sparse categorical cross entropy loss function was used with an Adam optimizer.