Leveraging Transfer Learning with Swin Transformer to Identify Coronary Artery Disease using Cardiac MRI

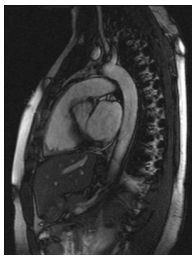
Natasha Banga
Stanford University
450 Jane Stanford Way, Stanford, CA, 94305
bnatasha@stanford.edu

Abstract

Cardiac Magnetic Resonance (CMR) is essential to diagnosis of a host of cardiovascular diseases. Coronary Artery Disease (CAD) is a common disease that can be identified using CMR technology. To aid clinical decision-making, a pretrained Swin Transformer architecture was employed in order to classify images as normal or CAD patients. CLIP Transformer was used to stratify CAD images into mild, moderate, and severe. Upon hyperparameter tuning, an AUC score of 0.98 was achieved. Results were comparable to a baseline, pretrained ResNet50 architecture. Performance was strongest on mild CAD cases and weakest on severe cases. Performance was comparable across various axes of CMR images. Future directions include increasing dataset size and modifying model architecture to evaluate performance.

1. Introduction

Cardiac Magnetic Resonance (CMR) is a critical diagnostic technique for a host of cardiovascular diseases. CMR can capture not just anatomical structure but also physiological attributes through a series of 'cine' frames. An example of a Vertical Long Axis (VLA) CMR image is below.



"Cardiac mri slice sagittal bionerd" by Bionerd is licensed under CC BY 3.0.

These images are critical to identify changes in the blood vessels, the patient's Stroke Volume (the quantity of blood pumped to the body during systole), as well as structural malformations in the heart and its valves (1). Thus, each CMR image can be considered to house many different descriptive features pertaining to the patient's cardiovascular health. For Coronary Artery Disease (CAD) which affects over 16 million Americans, CMR is key to identification of abnormalities in the heart wall and coronary vessels (2). Because CMR images can experience wide variation in quality based on the scanner used and the patient's movement while getting scanned, there is a risk for artifact that can limit interpretability (3). This along with a need for efficient analysis of results for acute clinical decision-making motivate the use of computer vision in classification of CMR images for CAD.

The current work employs transfer learning to leverage state-of-the-art Swin Vision Transformer (ViT) architectures for CMR image classification in CAD. The inputs to this model are single CMR images extracted from cine sequences and the output is a prediction of 'normal' or 'CAD.' This architecture is compared to baseline Swin ViT approaches without pretrained weights as well as ResNet50, both pretrained and non-pretrained.

2. Related Work

State-of-the-art models used in cardiac magnetic resonance (CMR) data processing include Swin Transformers. A recent paper applied VST (Video Swin Transformer) models to use understanding of specific CMR features and motion detection across the cine to classify images into categories of disease (4). They observed the model was able to outperform classification of images by cardiologists with AUC of 0.991 ± 0.0%. The paper's narrative is compelling, but the limited number of data points restricts the generalizability of their results. Other papers have applied Swin Transformers for segmentation tasks, which are often critical to effectively capture anatomical structures on CMR images (5). They were able to achieve a pixel accuracy of 93.68% and improvement of segmentation precision when compared to state-of-the-art models. Another paper took a similar approach to perform CMR image segmentation but used AnatSwin. This was clever since it integrated the Swin Transformer architecture but passed in label images as input and allowing the model to effectively train on

anatomical structures (6).

Given that CMR images often include portions of other organs, segmentation can be critical to identify relevant portions of the images. Since these papers illustrate the ability of Swin to spot differences in the minutia on images, it is promising in the realm of classification as well.

For classification tasks, an ensemble model was used integrating both a CNN and ViT framework to classify images as normal or from patients experiencing a myocardial infarction (7). They achieved an F1 score of 98.63%, which was not as successful as the other model architectures. However, it is clear from the literature that ViT models are a powerful application for this task.

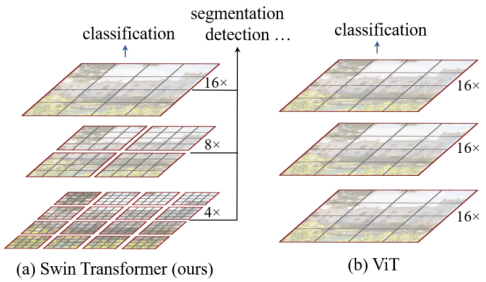
Because biomedical datasets can often be difficult to access in large quantities, transfer learning is highly relevant to this task (8). One analysis applied transfer learning to finetune a CNN-MLP framework. Though they achieved a high F1 score of 96%, the results are limited by the small size of the fine-tuning dataset—361 images.

U-Nets are another approach explored in the literature for this task, but primarily for image segmentation. One study applied Swin-Unet in order to achieve a Dice Coefficient of 91.72% on the public ACDC dataset (9). Because U-Net is specifically designed for segmentation it is not as well-suited to the current work.

While the application of ResNets to this task is currently limited, it has been explored for kidney MRI based classification of chronic kidney disease (10). While the model showed some promise, they found an accuracy of only 0.862±0.036. The current work utilizes ResNets as a baseline, comparison architecture to the ViT architecture.

3. Methods

The key architecture utilized is the Swin Transformer, specifically 'swin_tiny_patch4_window7_224.' This is a type of Vision Transformer with 12 Transformer blocks, but it is distinct from a vanilla ViT because it merges image patches through successive layers of the model. When we merge these patches, the number of channels increases. Another key attribute of the Swin Transformer is that it does not calculate global self-attention, but only does so for specific windows (unmerged patches), as depicted in red in the below figure.



Swin Transformer architecture
https://arxiv.org/abs/2103.14030v2

Each of the 12 transformer blocks include an MLP, multi-head self-attention, residual connections, and layer normalization layers.

The second algorithm deployed in this work is ResNet50. As the name suggests, this architecture consists of 50 layers that perform convolution but also include residual connections. These connections mean that the network is trained to learn F(x) + x, rather than just F(x).

An issue with comparing transformers and ResNets is the issue of the sheer greater complexity of transformer architectures. To account for this, the ResNet50 architecture was selected in order to ensure that both models have a comparable number of parameters (around 28M). An additional fully connected layer was added to the vanilla ResNet50 in order to increase the number of model parameters to be comparable to the Swin Transformer architecture.

A key difference between Swin and ResNet is the use of patches versus convolution. Swin uses patches that are processed chronologically through the self-attention transformer framework. Self-attention allows the model to compare the relevance of different patches to one another when calculating weights. Convolution, which is used in ResNets, entails simply sliding a filter of weights over all parts of the image which are then passed into the next layer. Both can be effective for image classification tasks but ViT has shown greater promise in learning finer details of CMR images.

Model inputs consisted of CMR images, x, and outputs consisted of classes 0 or 1 (normal or CAD). We utilized cross entropy loss, depicted below, where P(x) is probability of classifying the image in the correct class and Q(x) is probability of classifying the image in the class that was ultimately predicted.

$$L = -\Sigma P(x) * \log(Q(x))$$

Based on the common usage of the AdamWOptimizer in the literature, it was utilized to train both models. AdamWOptimizer applies weight decay to the weights to ensure the model is regularized, mitigating overfitting to the training set.

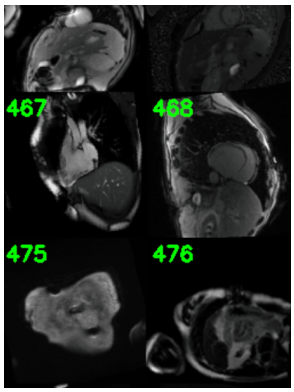
The models were trained both with pretrained ImageNet weights as well as randomly initialized weights. The models were fine-tuned using a hyperparameter search for ideal learning rates.

The current dataset did not include additional stratification of the patients beyond the 'normal' or 'sick'

labels, so additional techniques were deployed to label the images as ‘mild’, ‘moderate’, or ‘severe’ CAD. Initially, Open_AI LLM calls were attempted in order to label the images as a ‘ground truth,’ but since this was a medical image the labeling was not permitted. Similar attempts were made using the BLIP-2 ViT model, but it did not respond to the prompt well. Thus, I opted to use the CLIP transformer. I provided a set of three output labels (mild, moderate, and severe) and the algorithm (using pretrained weights) calculated cosine similarity between the labels and the images. CLIP works through contrastive learning, ensuring that cosine similarity is maximized for images that are most like their captions while it is minimized for images not like other captions. This model was deployed for classification of the validation dataset for downstream error analysis of the Swin classification architecture.

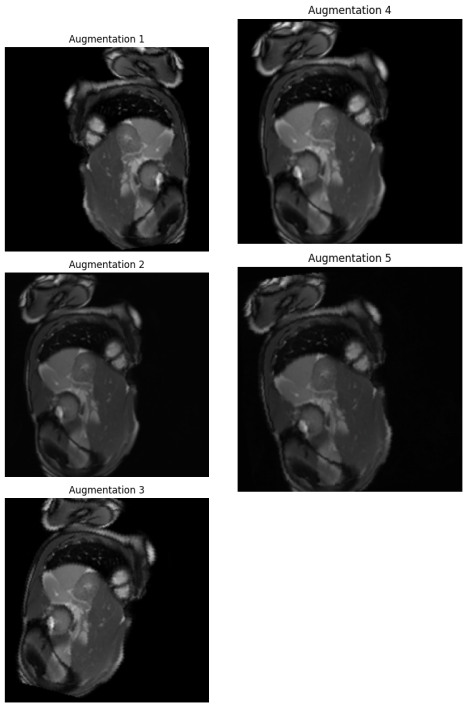
4. Dataset and Features

All CMR data comes from the CAD Dataset on Kaggle (11). The dataset consists of 51,000 images from CMR sequences. One image was extract from each CMR sequence, resulting in a total of 2875 images. Using a train/validation split of 0.8, I had 2300 images to train the model and 575 images in the validation set. Some examples of the images are below.



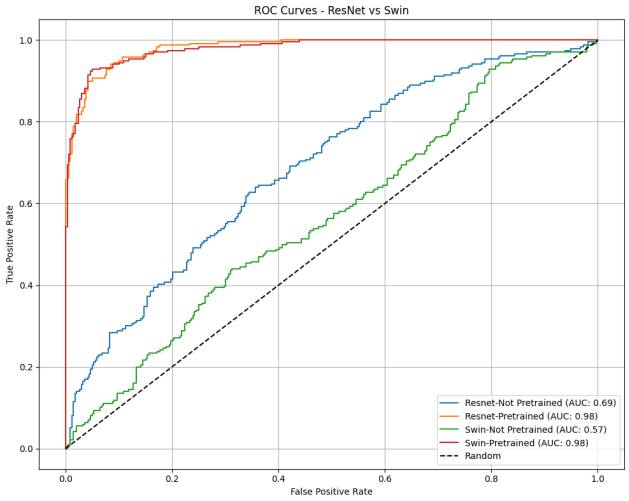
To preprocess the data, all images were resized to a resolution of 224X224, which is necessary for the Swin Transformer framework. Because CMR images are grayscale, they were converted to a three-channel tensor. All images were normalized by the mean and standard deviation of ImageNet pretrained weights.

Because of the small size of the dataset, I performed data augmentation by resizing the images, random horizontal flips, random rotation by 15 degrees, color jitter (brightness of 0.2, contrast of 0.2, saturation of 0.2, and hue of 0.05). Some examples of data augmentation are below.



5. Experiments, Results, and Discussion

Upon preprocessing, the CMR images were utilized in order to tune four models—the Swin Transformer with pretrained weights, the Swin transformer without pretrained weights, the ResNet50 with pretrained weights, and the ResNet50 without pretrained weights. Single-fold cross-validation was performed, and the mini-batch size was 32. This is because larger batch sizes caused memory issues and smaller batch sizes were less efficient. See Methods for description of the AdamWOptimizer. The ROC results after four epochs of training (when losses generally converged) are reported below.

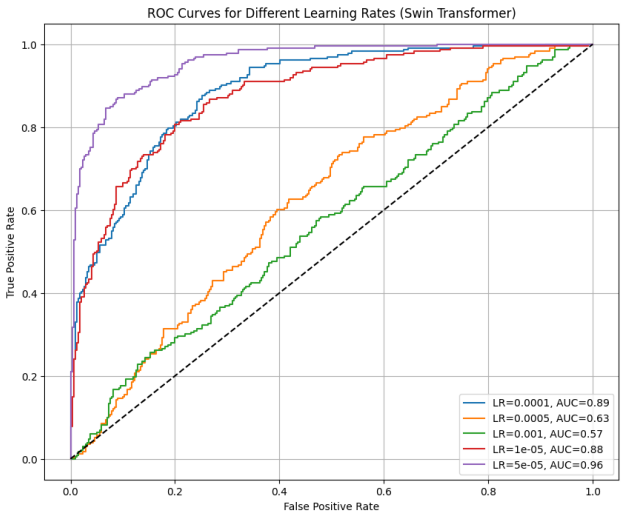


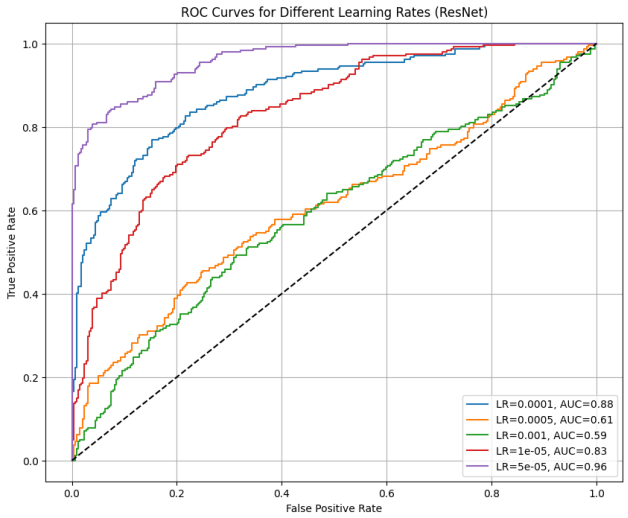
AUC scores across all four models are also summarized in this table.

Model	AUC Score after 5 Epochs
Resnet – not pretrained	0.69
Resnet – pretrained	0.98
Swin – not pretrained	0.57
Swin – pretrained	0.98

The notable result is how the pretrained weights make a substantial difference in model performance. For both the ResNet and Swin, lack of pretrained weights cause the model to suffer in performance. Additionally, it is interesting that the ResNet and Swin Transformer should reasonably similar performance. This could be due to the limited size of the dataset and lack of complexity in training examples, which prevent the transformer from outperforming the ResNet baseline.

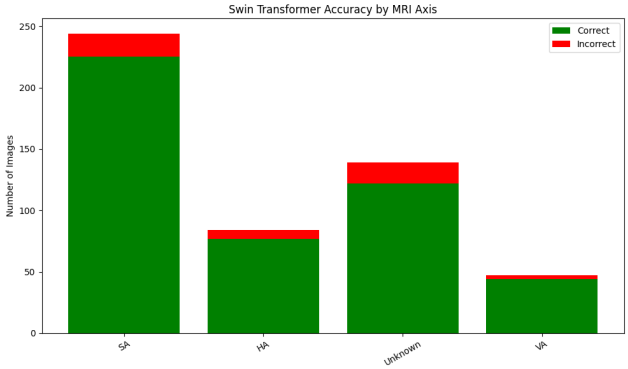
Furthermore, the pretrained ResNet and Swin were further fine-tuned using a hyperparameter search after two epochs. The results are shown below.



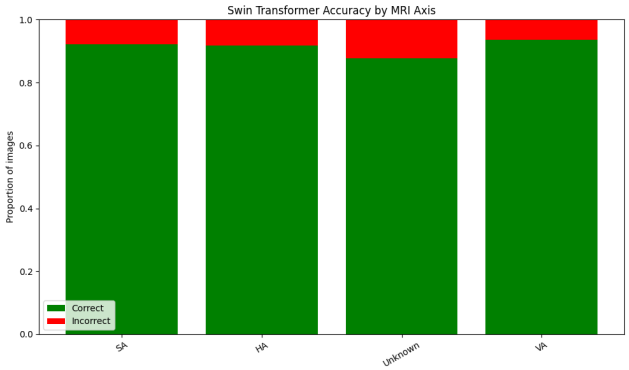


For both models, the learning rate of 5e-05 allows for the best performance, and the second-best performing learning rate is 1e-4. This is interesting because it shows that there is not a clear direct relationship between decreasing learning rate and increasing model performance. After a certain point of 1e-5, higher learning rates perform better. This makes sense because a learning rate that is too low does not allow the model to efficiently improve the weights and is too slow. An overly fast learning rate, however, may not allow the model to learn from the gradients and complexities in the dataset.

To get a sense of failure cases and highlight additional features of the images in the Swin transformer, a thorough error analysis was conducted. First, I manually labeled all 565 validation set images based on the planar axis in which the CMR image was taken. Since there was a combination of short axis, vertical long axis, and horizontal long axis images, I hypothesized that some image types may be easier or more difficult for the model to classify. After stratifying the dataset, I assessed accuracy over the various buckets, as shown below.



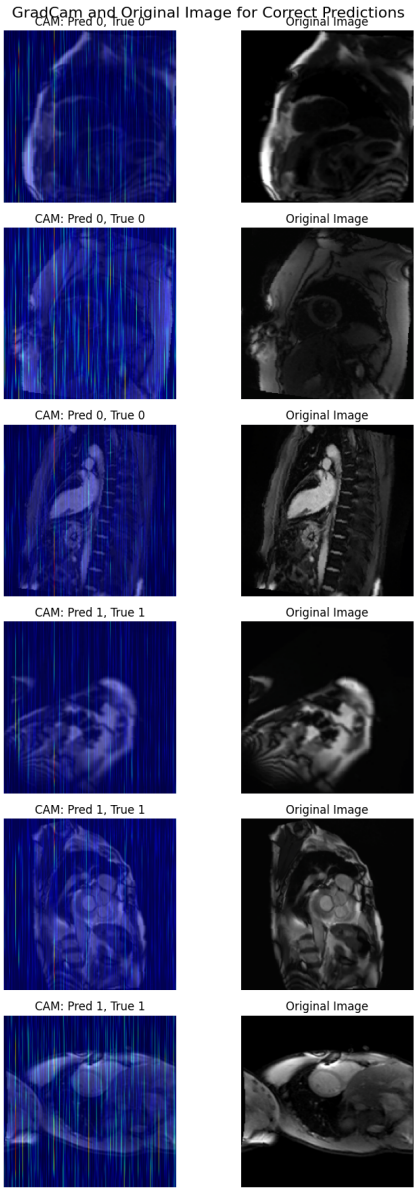
It is clear that a majority of the images are short axis images, with a large number of unknown axis images. I also plotted proportional accuracies for each bucket of image type.

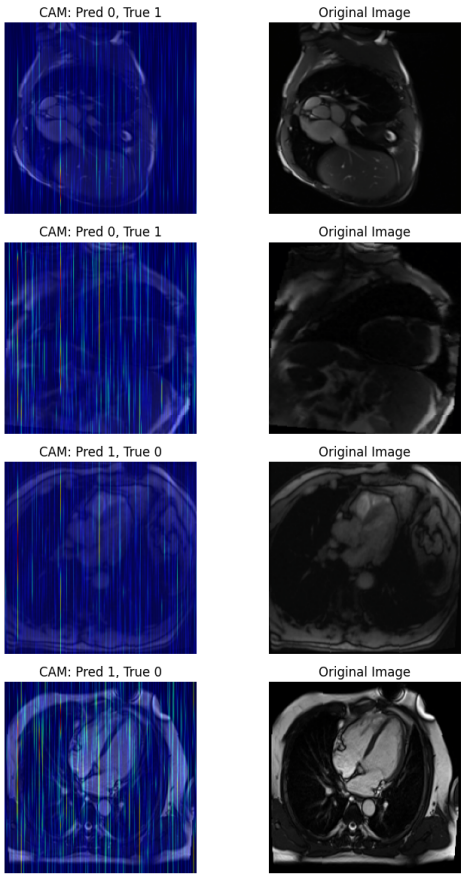


Generally, it seems that accuracy is consistent across different image types, which is certainly a key strength of this model. Despite variations in image type, the model is able to classify relatively similarly across all. From a medical relevance standpoint, it seems as if there is no particular axis view that is particularly 'salient' or more useful in CAD classification according to this analysis.

To get a sense of saliency in my images, I employed gradCam to various images in the validation set, both correctly and incorrectly classified. GradCam demonstrates salient areas in bright colors and less salient areas in blue.

Though the GradCam results should smoothly depict areas of greater aid in classification, we do not see this. Instead, the GradCam results are quite noisy. It is possible that the target layers I was analyzing had lower resolution since I had merged patches in my Swin transformer. Since GradCam is not designed for Swin Transformer, it is possible that this limitation prevented interpretability of the results. The results for incorrectly classified images are below.

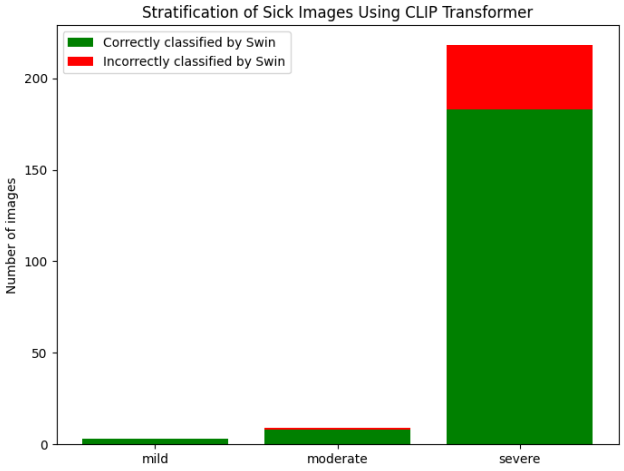




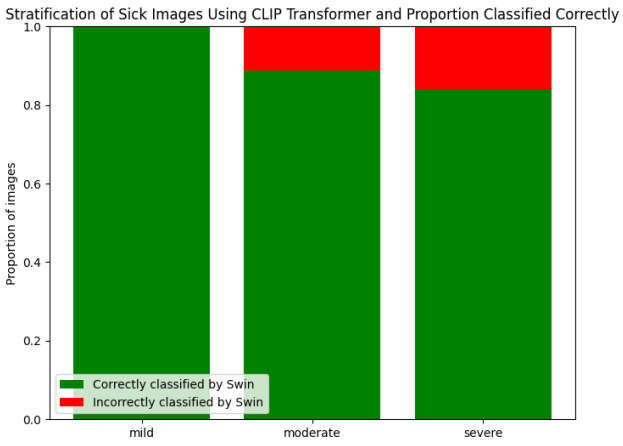
While it is difficult to say quantitatively, from a qualitative perspective, the GradCam results do appear noisier (with

many choppy colorful lines), which is reassuring since these images were unable to be correctly classified. Looking at the examples shown above, it seems that the misclassified examples exhibit less contrast in the image, which may make it harder to classify correctly.

A final assessment employed in my error analysis was the use of the CLIP transformer in order to stratify the ground truth CAD images into buckets of mild, moderate, and severe CAD. Similar to the axis analysis, I evaluated model accuracy within each of these buckets.



A vast majority of the images were classified as 'severe' by the CLIP transformer, which is the closest I could get to the ground truth since I did not have access to these bucket labels. Since CLIP is not pretrained on CMR images, this could have limited its accuracy.



It is interesting here that all mild cases were correctly classified whereas there was some error in the moderate and severe cases. One reason for this could be simply because there were many more moderate and severe cases, increasing the chance for error. Assessing this clinically, we would hypothesize that severe images are easier to classify since they show greater contrast with healthy images, but we see the opposite effect. This calls into question the classifications of the CLIP transformer, since it is not what we expect. Another explanation for this is a more complex clinical explanation—it is possible that in earlier stages of CAD, when the heart begins to fail, it exhibits signs of failure on CMR. In the moderate case, it may exhibit compensatory mechanisms such as hypertrophy (thickening of the heart walls). It is possible that this clinical mechanism is not well reflected in the CLIP model's labels, thus causing misleading results. Overall, it seems the model is most promising for mild cases.

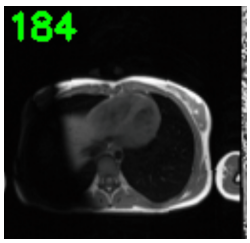
In terms of overfitting of the model, it seems unlikely because of high AUC scores on the validation set. To ensure this, future analyses could perform cross-validation on different sets of validation data. It is possible that the model is overfitting to this particular dataset; additional data from other sources is required to validate this.

6. Conclusions

The Swin Transformer and ResNet, pretrained architecture demonstrate equally highest performance (AUC=0.98) when compared to the other architectures. It makes sense that the pretrained architectures fell short because the limited size of the dataset made it difficult for the model to learn from scratch, given that there are millions of trainable parameters. Since Swin Transformers often require a larger number of data points to outperform the ResNet, we observed comparable success between the two.

Remarkably, the Swin model's classifications are consistent across varying planes in which images were taken. This is likely because the model was able to effectively identify relevant portions of the image (the ventricles and key blood vessels) respective of the background plane. Another notable finding is that the Swin model performs best on mild cases of CAD.

A key limitation in the current analysis is the quality of the dataset. I noticed some superpositions and shadows on some training images which may have made learning key parameters difficult. An example is shown below.



Future directions could include testing unique Swin architectures with varying layers and patch sizes, deploying ensemble models of combined ResNet and

Swin architectures given their roughly equivalent success, performing a classification task of different clinical features (hypertrophy, valvular dysfunction, etc.), and training on larger, diverse datasets to prevent overfitting.

7. Contributions and Acknowledgements

I'd like to thank my TA mentor, Sabri, for his guidance on the methods used in this work. I'd also like to thank Danial Sharifrazi for making his CAD dataset publicly available.

8. References/Bibliography

[1] "Cardiac Magnetic Resonance Imaging (MRI)." Www.Heart.Org, 12 Mar. 2025, www.heart.org/en/health-topics/heart-attack/diagnosing-a-heart-attack/magnetic-resonance-imaging-mri. [1]

[2] van der Wall, E. E., Vliegen, H. W., de Roos, A., & Bruschke, A. V. (1995). Magnetic resonance imaging in coronary artery disease. *Circulation*, *92*(9), 2723–2739. https://doi.org/10.1161/01.cir.92.9.2723 [2]

[3] Cau, R., Pisu, F., Suri, J. S., Mannelli, L., Scaglione, M., Masala, S., & Saba, L. (2023). Artificial Intelligence Applications in Cardiovascular Magnetic Resonance Imaging: Are We on the Path to Avoiding the Administration of Contrast Media?. *Diagnostics (Basel, Switzerland)*, *13*(12), 2061. https://doi.org/10.3390/diagnostics13122061 [3]

[4] Wang, Y. J., Yang, K., Wen, Y., Wang, P., Hu, Y., Lai, Y., Wang, Y., Zhao, K., Tang, S., Zhang, A., Zhan, H., Lu, M., Chen, X., Yang, S., Dong, Z., Wang, Y., Liu, H., Zhao, L., Huang, L., Li, Y., … Zhao, S. (2024). Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. *Nature medicine*, *30*(5), 1471–1480. https://doi.org/10.1038/s41591-024-02971-2 [4]

[5] *A Fusion-attention Swin Transformer for Cardiac MRI Image Segmentation - Yang - 2024 - IET Image Processing - Wiley Online Library*, ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12936. Accessed 5 June 2025. [5]

[6] Wang a, et al. "Anatswin: An Anatomical Structure-Aware Transformer Network for Cardiac MRI Segmentation Utilizing Label Images." *Neurocomputing*, Elsevier, 7 Feb. 2024, www.sciencedirect.com/science/article/pii/S0925231224001504. [6]

[7] Al-antari, Mugahed A., et al. "A Hybrid Segmentation and Classification CAD Framework for Automated Myocardial Infarction Prediction from MRI Images." *Nature News*, Nature Publishing Group, 23 Apr. 2025, www.nature.com/articles/s41598-025-98893-1. [7]

[8] Ben Khalifa, A., Mili, M., Maatouk, M., Ben Abdallah, A., Abdellali, M., Gaied, S., Ben Ali, A., Lahouel, Y., Bedoui, M. H., & Zrig, A. (2025). Deep Transfer Learning for Classification of Late Gadolinium Enhancement Cardiac MRI Images into Myocardial Infarction, Myocarditis, and Healthy Classes: Comparison with Subjective Visual Evaluation. *Diagnostics (Basel, Switzerland)*, *15*(2), 207. https://doi.org/10.3390/diagnostics15020207 [8]

[9] Fu, Z., Zhang, J., Luo, R., Sun, Y., Deng, D., & Xia, L. (2022). TF-Unet:An automatic cardiac MRI image segmentation method. *Mathematical biosciences and engineering:MBE*, *19*(5),5207–5222. https://doi.org/10.3934/mbe.2022244 [9]

[10] Nagawa, K., Hara, Y., Inoue, K., Yamagishi, Y., Koyama, M., Shimizu, H., Matsuura, K., Osawa, I., Inoue, T., Okada, H., Kobayashi, N., & Kozawa, E. (2024). Three-dimensional convolutional neural network-based classification of chronic kidney disease severity using kidney MRI. *Scientific reports*, *14*(1), 15775. https://doi.org/10.1038/s41598-024-66814-3 [10]

[11] Sharifrazi, Danial. "CAD Cardiac MRI Dataset." *Kaggle*, 4 Oct. 2021, www.kaggle.com/datasets/danialsharifrazi/cad-cardiac-mri-dataset. [11]

Libraries Downloaded:

[1] Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

[2] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). *PyTorch: An imperative style, high-performance deep learning library*. Advances in Neural Information Processing Systems, 32. https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[3] Wightman, R. (2020). *PyTorch Image Models*. https://github.com/huggingface/pytorch-image-models

[4] Bradski, G. (2000). *The OpenCV Library*. Dr. Dobb's Journal of Software Tools.