Classifying the orientation of the heart in a series of MRI scans

Samuel R D'Souza[1], James P Howard[1]

[1]National Heart and Lung Institute, Imperial College London, Hammersmith Hospital, London, UK

This research report is submitted in partial fulfilment of the requirements for the degree of BSc in Medical Biosciences.

Abstract

Cardiac magnetic resonance imaging (CMR) is considered to be the current state of the art technique for cardiac tissue characterisation and chamber quantification. Automated diagnosis of CMR images or video data is challenging and relies upon the correct identification of the 'view' (imaging plane and orientation). This paper defines the development of a machine learning algorithm that accurately classifies the 'view' of a CMR image into one of seven classes. To this effect, we compare the accuracy of multiple state-of-the-art image classification neural network architectures to evaluate which one best fits this task. The EfficientNet-B5 architecture is the highest performing architecture, with an overall accuracy of 98.4%. We assessed the model's categorical accuracy to account for the dataset's uneven distribution between categories, and then validated the model's performance against that of an expert cardiologist. The model was shown to have learned, by use of saliency mapping, the same key features in each view that a cardiologist does, thus indicating that it both semantically and functionally mimics a clinician. Building an ensemble model combining different neural network architectures further boosted performance. Our solution has approached human expert-level performance and may form the basis of future AI solutions which rely on view recognition. Future work potentially involves training neural networks to diagnose various cardiac conditions from labelled CMR data. Once these networks are integrated into a diagnostic procedure, this software will facilitate cardiology patient's diagnosis, and enable hospitals to provide care to more people.

Key Words: Cardiovascular magnetic resonance imaging, deep learning, neural networks

Abbreviations: CMR - cardiovascular magnetic resonance imaging; CH2 - 2 chamber view; CH3 - 3 chamber view; CH4 - 4 chamber view; RV - right ventricle view; AOV - aortic valve view; SAX - short-axis view; LVOT - left ventricular outflow view

Introduction

Cardiac magnetic resonance imaging is increasingly being employed in hospitals to evaluate cardiovascular health. It can assess biventricular function, volumes, and mass. The strong tissue characterisation capabilities of CMRs mean that oedema, fibrosis, and myocardial tissue perfusion are visible. This level of detail allows doctors to identify conditions such as myocardial ischemia or infarction[1].

The two most common alternatives to CMR for cardiovascular evaluation are echograms and computerised tomographies. CMRs allow for better tissue characterisation than echograms[2]. Compared to computerised tomographies of the heart, CMRs can capture cardiac motion with video data and don't expose subjects to harmful radiation[3]. CMRs image a 2D plane within the heart. Depending on the angle and depth of this plane, the image will visualise different cardiac volumes and valves. Each of these different angles is labelled as a 'view'. At some depths, these views appear to look the same given they show the same structures at slightly different angles. Various views contain different diagnostic information, depending on the structures they visualise. For example, when assessing for regional wall motion abnormalities, which could indicate a previous myocardial infarction, a cardiologist examines the 2-chamber, 3-chamber, 4-chamber and short-axis views as they display the left ventricle[4].

A new approach to streamline cardiac diagnostics procedures involves using neural networks. A neural network will assess CMR data in seconds. This offers multiple advantages over manually reviewing data in terms of speed and efficiency. Normally, once the correct view is identified, clinicians must make time-consuming measurements of the image to diagnose the subject. As an example of this, the process for identifying hypertrophic cardiomyopathy from CMRs involves calculating the septal to lateral wall thickness ratio as well as the left ventricular wall thickness to make a judgement[5]. Implementing neural networks in hospitals expedites diagnosis, allowing cardiologists to examine more patients daily. These networks may even discover novel features humans have failed to quantify and allow doctors to more accurately diagnose diseases.

A key component of automating CMR-based diagnosis is recognising the 'view' presented in the data. Without this information, neural networks are inefficient and error-prone[6]. It is better to segment the diagnostic pipeline into discrete tasks, each with bespoke neural networks, to reduce the complexity of the challenge as well as making it technically feasible. This paper assesses a machine learning classifier that recognises the 'view' of a static CMR image. The end goal is for this classifier to be the first step in a cardiac diagnostics pipeline consisting of multiple neural networks. This 'ViewDetector' will be built using cutting-edge neural network architectures known to perform well on image classification tasks. The hypothesis of this paper is as follows:

A machine learning classifier should be able to accurately recognise the orientation of the heart in MRI images, to facilitate automated CMR analysis with AI.

Once the best neural network architecture is decided, further analysis will be done to investigate its decision-making process. This will involve quantifying the model's accuracy for each 'view', reviewing the model's saliency maps with a trained cardiologist, and, finally, testing the model and a cardiologist on the same dataset to validate the model. It is vital this component of the pipeline performs accurately so that it can be endorsed for use in medical diagnostic procedures.

Materials and Methods

Data

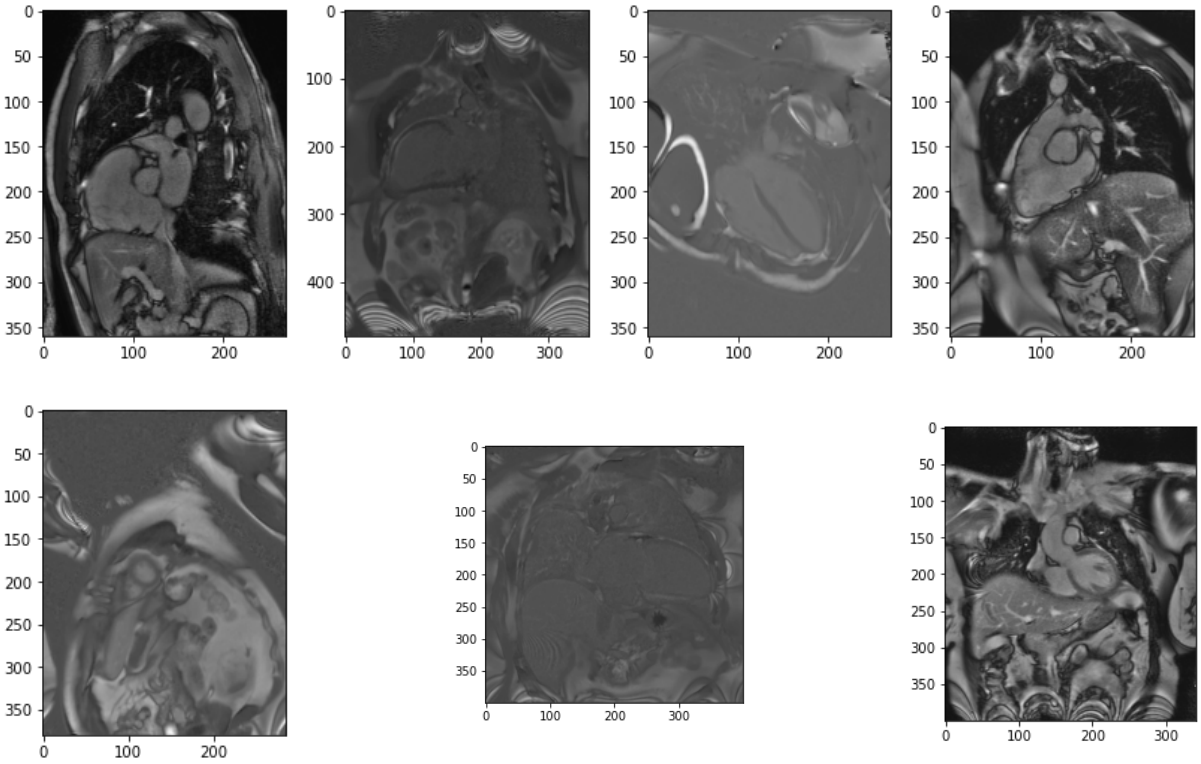

Figure 1: Sampled images of various views from the dataset. From top to bottom, right to left, they are labelled as AOV, CH2, CH4, RV, SAX, CH3, and LVOT.

There are 7 labelled orientations of the heart that have been identified by cardiologists at Hammersmith Hospital. These are the classes used to train the machine learning model. They are the 2 chamber view (CH2), 3 chamber view (CH3), 4 chamber view (CH4), right ventricle view (RV), aortic valve view (AOV), short-axis view (SAX), and the left ventricular outflow view (LVOT).

Methods

The CMR data was converted to multi-dimensional arrays termed 'Tensors' using the PyTorch machine learning framework, which allows rapid training of neural networks using Graphical Processing Units (GPUs). Transformations are used on the images to augment the available data. The transformations used on the training subset include: random resize cropping, shift scale rotation, horizontal flipping, normalisation, and vertical flipping. The test subset was only normalised. The transformations' implementation was handled by the Albumentations library[7]. These transformations supplement the dataset, effectively providing more training examples for the algorithm to learn from. They, additionally, reduce the likelihood of the model 'overfitting'. Overfitting is when the model memorises patterns instead of learning them. Ground truth annotations for the views of each image were derived from the name of the sequence entered at scan time by the clinician.

There are several competitive architectures for image classification[8]. This paper will assess the performance of these state of the art architectures rather than develop custom neural networks for the task.

Figure 2 displays a comparison of several architectures' performance on the ImageNet classification challenge. There is a clear progression in accuracy for image classification models over the past decade, from AlexNet and VGG[9] to ResNet[10] and, most recently, EfficientNet[11]. Each model will be trained and then evaluated thrice, to get an accurate understanding of its performance. The 'best run' from each model will be used as it represents the network's potential.

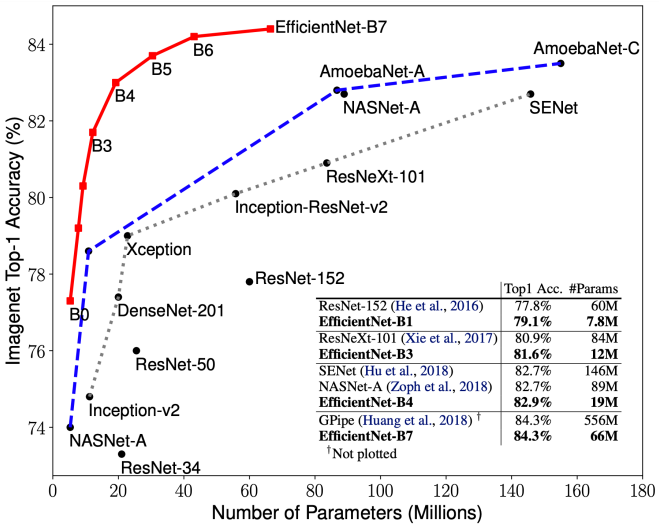

Figure 2: A comparison of neural network architectures. It compared their showing their complexity (# of parameters) against their accuracy on the ImageNet classification challenge (8).

The techniques and algorithms used to train these models include: an AdamW optimiser with a learning rate (LR) and weight decay of 5.0e-4, a cross-entropy loss function, a 60 epoch limit, the one cycle learning rate scheduler (max LR of 1.0e-3, and final LR of 1.0e-10)[12], and the aforementioned Albumentation transforms.

Some elements, such as activation functions and dropout layers, will be fixed with preloaded neural network architectures. Following the selection of the best network, the remaining parameters, such as batch size, loss function, transforms, and learning rate can be optimised to improve the network's accuracy[13].

Once the best architecture is patent, its function will be analysed using more complex methods. The model's accuracy within each category needs to be verified. The overall and class-wise accuracies will be compared to that of a cardiologist's to assess whether the model is a viable substitute. We report neural network performance in terms of both a raw accuracy statistic and Cohen's Kappa. Cohen's Kappa is a measure of agreement that adjusts for imbalanced classes[14].

Saliency maps are used to understand how the network uses features of the image to make a decision. Saliency maps go through each pixel and, using differential calculus, derive the effect of changing the pixel's value on the network's prediction, thus quantifying each pixel's importance. This data can be transformed into a heat-map which is then superimposed onto the original image. The overlay highlights regions important to the neural network's decision matrix.

The performance of an expert cardiologist and the model was assessed on a specially curated dataset. This was done using 350 randomly sampled images evenly distributed across the seven 'views'.

A model ensemble will be trained using 4-fold cross-validation. This involves splitting up the training data into quarters and training four of the same architecture on different test-train folds. Alternatively, one could train four different architectures on the same data. Different networks have different strengths and may learn slightly different features. Averaging the results of these models has been shown to improve accuracy. The ensemble's accuracy will be assessed on the validation dataset so it can be statistically compared to the single model and human expert.

Ethical Considerations

The classifier must be accurate enough that there is an equal or less-than-human chance of misclassification. This ensures that the diagnostic pipeline does not adversely affect the quality of care provided by hospitals. Additionally, keeping the MRI views dataset anonymised is important for ensuring patient privacy. The subjects gave their consent for this data to be stored in a database and used for medical research. Beyond these considerations, there is limited scope for ethical malfeasance.

Results

Dataset

In development, the model was only evaluated on a portion of the available data. The dataset this classifier was trained on consisted of a train and test subset. The dataset was randomly split 10% into testing and 90% into training. There were 2812 labelled train images: 348 LVOT views, 497 SAX views, 367 RV views, 404 CH4 views, 435 CH3 views, 402 CH2 views, and 359 AOV views. There were 249 test images: 49 LVOT views, 1 SAX view, 39 RV views, 52 CH4 views, 21 CH3 views, 42 CH2 views, and 42 AOV views.

Performance of different architectures

The best performing network on the dataset was EfficientNet-B5, with an accuracy of 98.8%. Figure 3 shows the best test accuracy of each architecture at epoch 60. Generally, as the complexity of the neural network increased, it was able to more accurately predict CMR views. ResNet50 was able to perform similarly to EfficientNet-B3, as they have a similar number of trainable parameters. Following this, more complex EfficientNets better classified the data up until B5. EfficientNet-B5, even with a reduced batch size of 32, boasted 98.795% accuracy. The Cohen's Kappa value for EfficientNet-B5 was 98.54%. EfficientNet B6 and B7, despite their increased complexity, do not achieve higher accuracy than EfficientNet-B5 by epoch 60. This suggests that the machine-learning models have reached a "glass ceiling" in terms of their ability to learn from this task or dataset. The risk is that more complex networks

Architecture	Accuracy (%)	Cohen's Kappa
ResNet50	96.39%	0.9562
ResNet101	94.779%	0.9369
EfficientNet-B3	96.79%	0.961
EfficientNet-B4	97.59%	0.9708
EfficientNet-B5	98.8%	0.9854
EfficientNet-B6	98.39%	0.9805
EfficientNet-B7	98.8%	0.9854

Figure 3: A table showing each architecture's accuracy and Kappa Cohen value on the test subset of the data. These values are extracted from the model at the 60th epoch, beyond which the model's stopped improving. If Kappa Cohen's is significantly divergent from the accuracy, the model is biased towards a single class.

tend towards overfitting the training data with this relatively straightforward classification task[15]. To further assess EfficientNet-B5's function, its categorical accuracy for each 'view' was investigated.

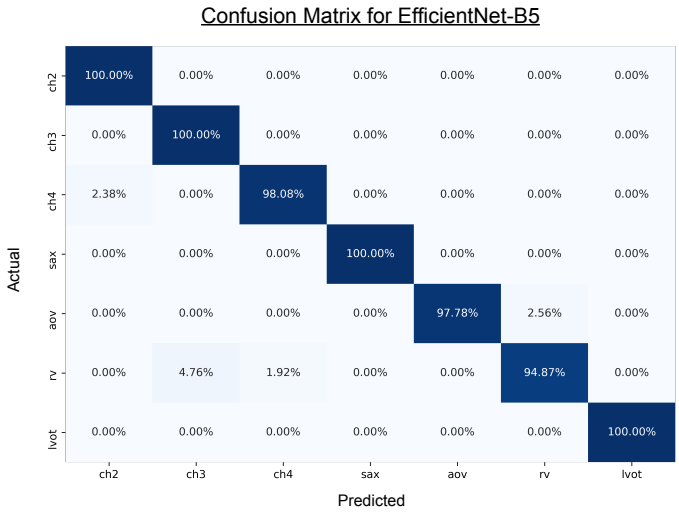

Figure 4: A confusion matrix which visualises the categorical accuracy of the network. Each value represents the normalised percentage (2 s.f.) of the testing dataset that was classified by EfficientNet-B5 following 60 epochs.

Class-wise accuracy for EfficientNet-B5

There was minimal confusion between classes, as evidenced by the confusion matrix in Figure 4. The RV class was most challenging for the model to accurately predict. CH4 and CH3 were also confused by the model. In the less complex networks, RV was frequently misclassified as CH2, CH4, and AOV. As architectural complexity increased, the error rate greatly diminished. RV inaccuracies stabilised across EfficientNets B5, B6 and B7 at 5.128%.

To compare the analysis done by the B5 model and that of a human expert, saliency maps were developed to investigate which features the model 'learned'.

Saliency maps demonstrate that the model uses similar logic to that of cardiologists

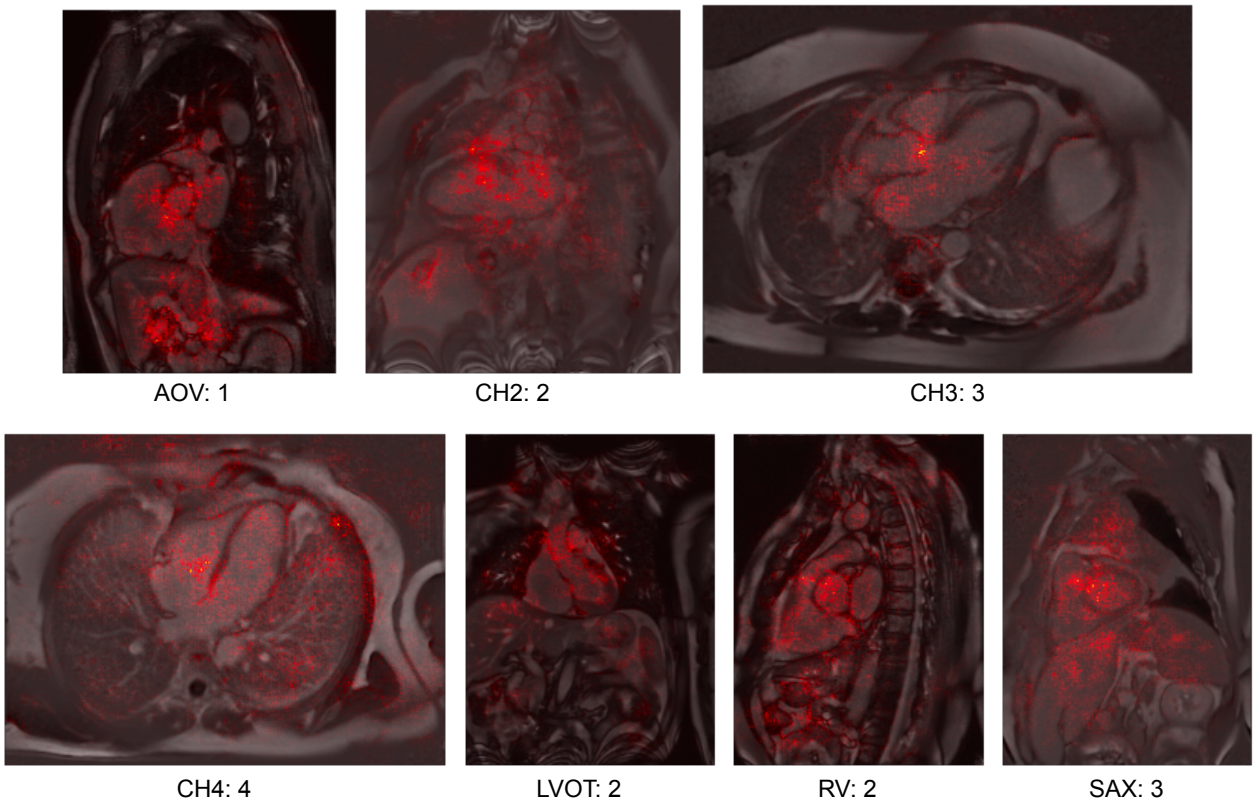

Figure 5: Saliency maps of the CMR views. The black and white image is the CMR scan. Overlaid onto of this image is a saliency map. The red dots indicate which pixels had a significant image on the final class prediction made by the model. The number after the view label indicates the position of the image in the Kaggle notebook.

Several saliency maps of each view were reviewed by a cardiologist. This allowed for comparison between the features a cardiologist is taught to recognise and those the model learned from the training dataset. This verified that the algorithm utilised the correct structures to predict views. These maps are laid out in Figure 5.

In the aortic valve view (AOV), the aortic valve itself appears particularly salient regarding the neural network's final decision. The recognition of the aortic valve, and its orientation, are the characteristic features used by cardiologists and radiographers in orientating this view during scanning. In the two-chamber view (CH2), the model focuses on the left ventricle, mitral valve, and left atrium; all are structures indicative of the view. The three-chamber view (CH3) is a more challenging case, as it is similar to a sagittal LVOT view. Nevertheless, the model recognises the side-on “sack-like” shape of the aortic valve as unique to the CH3 view. The activation on the walls of the valve indicates the model is successfully distinguishing this view from the similar LVOT view which also shows the aortic valve.

In the four-chamber (CH4) view, the model is highly active around the tricuspid valve and inter-atrial and inter-ventricular septum which are specific to this view, therefore indicating the model is semantically sound. LVOT views look at the heart from an end-on perspective, slicing through the aortic valve. The model correctly focuses on the aorta and somewhat uses the left ventricle as predictive structures. Though, in this view, the model’s semantics are less perspicuous. The right ventricle (RV) view is the most troublesome view for cardiologists to identify due to its similarity to other views at different depths. In the example, the model works surprisingly well. It focuses on the aortic valve, but also the right ventricular outflow tract into the pulmonary artery, which is specific to RV views. The short-axis (SAX) view is, again, a more challenging image to deconstruct. The model appears to look at the left ventricle (the ring-like shape), and the interventricular septum which would identify SAX views. This view is infrequently used, and therefore less common in the dataset so it follows that the model’s activation in this saliency map is less specific.

In summary, throughout each of these classes, the model’s activation is localised to structures that indicative of the view, though the concentration of this activation varies. These findings may indicate the algorithm has learned to look for the same structures and indicators as a trained cardiologist would, and so is functioning in the “correct” way.

Human expert accuracy compared to EfficientNet-B5

	CH2	CH3	CH4	AOV	RV	LVOT	SAX	Total	P-Value
Human Expert:	96%	94%	96%	86%	96%	100%	96%	94.86%	N/A
B5 Model:	90%	98%	96%	100%	94%	86%	98%	94.29%	0.735
B5 Model Ensemble:	84%	96%	94%	100%	96%	88%	98%	93.71%	0.556
DiffArch Model Ensemble:	88%	96%	96%	100%	98%	86%	100%	94.86%	0.307

Figure 6: The table shows the accuracy of a human expert’s label prediction compared to that of the EfficientNet-B5 model’s and the model ensembles’ predictions. The dataset consisted of 350 randomly selected samples from the entire dataset (50 from each view). The DiffArch ensemble consisted of two EfficientNet-B5s, an EfficientNet-B6, and a ResNet101. The P-value was derived using a McNemar’s test (Appendix III and V).

The accuracy of EfficientNet was 94.3%. In comparison, the accuracy of the expert was 94.9%. Differences between AI and human accuracy were judged using McNemar's test, with a p=0.05 value used as a threshold for statistical significance. There was no significant difference between the 2 classes (p=0.735).

Model ensemble improves performance

Finally, model ensembles were developed to see if performance could be further improved. The ensembles were validated across the n=350 balanced dataset and compared to a single EfficientNet-B5 as well as a human expert. As seen in figure 6, both of the model ensembles' performance is not statistically different from that of a human expert. The EfficientNet-B5 ensemble performed worse on this dataset. The DiffArch Model Ensemble, however, achieved almost the exact accuracy as that of a human expert improving beyond the single B5 model.

Discussion

This study supplies an accurate machine learning algorithm for classifying the view of the heart in a CMR scan. Advanced neural network architectures for image classification were assessed on a subset of a labelled CMR dataset. The best performing architecture, EfficientNet-B5, stopped improving at an accuracy of 98.8%.

The ImageNet classification challenge has, for the past decade, been the benchmark for neural networks in image classification[16]. This is a highly developed field with a wealth of different high-performance networks. EfficientNet's are at the top of the rankings. They utilise sophisticated techniques in a scalable and condensed manner. It utilises 1x1 convolutions, dropout, dense connections, inverted residual blocks, RMSProp, squeeze-and-excitation blocks, as well as Swish activation functions. ResNet and other competitors use a lesser portion of these techniques in a less efficient and scalable manner[11 17]. It was important to validate that assumption, however, as complexity does not necessarily translate into performance. This occurred in our use case, where ResNet101 performed worse than ResNet50. EfficientNet-B5, B6, and B7 performed well, but as complexity increased the network's performance did not scale. Techniques, such as model ensembles, can be used to increase performance by a small degree, though it is likely that the networks have "learned" as much as is possible from the dataset. Therefore there is nothing significant they can improve on. The remaining inaccuracies are likely due to poor image quality or views that are indistinguishable from each other. The 5% residual error may be genuine, or indeed it may indicate errors in the ground truth labels. This warrants further investigation to elucidate whether humans and the network are getting the same cases wrong or different ones.

The network's categorical accuracy was examined using a confusion matrix and an accuracy statistic calculated using Cohen's Kappa coefficient. This was to ensure accuracy is not being biased by a single class having a larger sample size[18]. There was minimal diversion between test accuracy, 98.795%, and Cohen's Kappa, 98.54%. This means that there is no significant bias in the algorithm. Accuracy was respectable across all of the classes of view, as indicated in the confusion matrix.

Both of these assessments of accuracy and categorical accuracy indicated that the model was performing strongly, by generic machine-learning metrics. The algorithm also needed to be assessed, however, in the context of its clinical application.

To do this, the model's performance was compared to that of a practising cardiologist. This demonstrated that the algorithm exhibited a similar level of competency to that of a standard clinician.

Machine learning is often referred to as a "black box" because it is challenging to derive what features the hidden layers of a neural network analyse. Saliency mapping is a tool that has been developed

to elucidate the decision-making mechanisms neural networks use within the field of image processing[17]. In this paper, it is used to validate that the algorithm is using relevant cardiac structures to identify the view. This indicates that the model has generalised relevant features from each category in the dataset, and, in doing so, semantically mimics a trained cardiologist. It also potentially points out secondary features that help indicate view.

In conclusion, this study proves that machine learning classifiers can accurately identify the 'view' of CMR images. Furthermore, the classifier can do this task with the same level of efficiency as a trained cardiologist.

Limitations

The neural networks were trained and evaluated using Kaggle, a free online data science platform. The freely available NVidia K80 GPU from Kaggle limited either the size of the neural network architecture or the batch size that could be used. From EfficientNet-B5 and higher, the number of parameters in the model meant that batch size had to be halved from 64 to 32 to train the model, otherwise, a memory error was thrown. For EfficientNet-B7, with 66 million trainable parameters, the batch size was reduced to 8 in order to run. If batch size falls too low, it impedes the training of the model. This is because there is not enough data for batch normalisation to normalise the sample in relation to the whole dataset[15]. Therefore no networks with more parameters than EfficientNet-B7 were assessed.

There is a risk, that with more computational power and data the algorithm will 'overfit' the data. The algorithm, with too many trainable parameters and too much data, may end up "memorising" the training data rather than "learning" key features. There are methods for minimising this risk using "regularisation" including dropout and weight decay, both of which are used in these networks.

There are some experimental limitations as well. The "true" labels from the dataset are not all the ground truth. This is due to some images being ambiguous combined with human error. Several "correctly labelled" images were discerned to be incorrect in the dataset. This potentially limits the performance of the neural network, confusing the features it is attempting to learn. However, it could also mean that network accuracy is higher than reported if the network correctly identifies mislabelled images in the "test" subset.

There are rarer CMR views that are not represented in this dataset, limiting the algorithm's ability to recognise views. Furthermore, people with highly abnormal cardiac anatomies, e.g. congenital heart diseases, may cause the system to fail as it has not learnt to generalise to this subcategory of people.

Future Work

Various methods could be tested to improve the accuracy of this model. These include using "model ensembles", which were evaluated as part of this paper. Multiple networks are developed and their probability distributions summated to get the most accurate result. As we saw in this paper, using an ensemble consisting of different neural networks yielded slightly improved performance. Alternatively, large batch training and transfer learning could be used to increase performance. However, given the model is already equalling the human benchmark, and that this is a relatively simple classification challenge, it is likely the algorithm has already reached the upper bounds of its potential. Further optimisation will likely have an exponentially decreasing effect on the model's accuracy.

The end goal of this project is to develop a pipeline that can diagnose cardiac diseases or overall cardiac health. As the first step to any diagnostics procedure, the 'view' needs to be recognised. Cardiologists do this intuitively, however neural networks require a discrete labelling process. Now that this

labelling process is developed, future cardiologists and machine-learning experts can produce neural networks that evaluate CMR images for various diseases or features.

This CMR diagnostics pipeline would provide benefits to hospitals across the United Kingdom (UK). It would speed up referral times for patients, allow cardiologists to treat more people daily, and reduce the chance of subjects with critical conditions being backlogged. A centralised API from the NHS diagnosing CMRs would improve cardiac health throughout the UK as well as becoming a potential source of data. This dataset could be leveraged in several ways. It could be fed back into machine-learning algorithms, improving automated diagnostics. By logging both patient's health and their CMR scans, doctors could use unsupervised learning techniques to develop new models that diagnose conditions using features difficult for humans to identify. These networks could identify new salient features for cardiac disease, feed this data back to cardiologists, therefore improving doctors' ability to diagnose.

Acknowledgement

We thank Dr James Howard for his input as an expert cardiologist.

We also thank Hammersmith Hospital for providing the CMR Views database.

References

(1) Lurz P, Luecke C, Eitel I, Föhrenbach F, Frank C, Grothoff M, et al. Comprehensive Cardiac Magnetic Resonance Imaging in Patients With Suspected Myocarditis. *Journal of the American College of Cardiology.* 2016; 67 (15): 1800-1811. Available from: doi: 10.1016/j.jacc.2016.02.013 Available from: https://search.datacite.org/works/10.1016/j.jacc.2016.02.013 .

(2) Gardner BI, Bingham SE, Allen MR, Blatter DD, Anderson JL. Cardiac magnetic resonance versus transthoracic echocardiography for the assessment of cardiac volumes and regional function after myocardial infarction: an intrasubject comparison using simultaneous intrasubject recordings. *Cardiovascular ultrasound.* 2009; 7 (1): 38. Available from: doi: 10.1186/1476-7120-7-38 Available from: https://search.datacite.org/works/10.1186/1476-7120-7-38 .

(3) Dweck MR, Williams MC, Moss AJ, Newby DE, Fayad ZA. Computed Tomography and Cardiac Magnetic Resonance in Ischemic Heart Disease: State-of-the-Art Review. *Journal of the American College of Cardiology.* 2016; 68 (20): 2201-2216. Available from: doi: 10.1016/j.jacc.2016.08.047 Available from: https://search.proquest.com/docview/1846419793 .

(4) Hendel RC, Patel MR, Kramer CM, Poon M, Carr JC, Gerstad NA, et al. ACCF/ACR/SCCT/SCMR/ASNC/NASCI/SCAI/SIR 2006 Appropriateness Criteria for Cardiac Computed Tomography and Cardiac Magnetic Resonance Imaging**Developed in accordance with the principles and methodology outlined by ACCF: Patel MR, Spertus JA, Brindis RG, Hendel RC, Douglas PS, Peterson ED, Wolk MJ, Allen JM, Raskin IE. ACCF proposed method for evaluating the appropriateness of cardiovascular imaging. J Am Coll Cardiol 2005;46:1606–13. *Journal of the American College of Cardiology.* 2006; 48 (7): 1475-1497. Available from: doi: 10.1016/j.jacc.2006.07.003 Available from: https://search.datacite.org/works/10.1016/j.jacc.2006.07.003 .

(5) Noureldin RA, Liu S, Nacif MS, Judge DP, Halushka MK, Abraham TP, et al. The diagnosis of hypertrophic cardiomyopathy by cardiovascular magnetic resonance. *Journal of cardiovascular magnetic resonance.* 2012; 14 (1): 17. Available from: doi: 10.1186/1532-429x-14-17 Available from: https://search.datacite.org/works/10.1186/1532-429x-14-17 .

(6) Howard JP, Tan J, Shun-Shin MJ, Mahdi D, Nowbar AN, Arnold AD, et al. Improving ultrasound video classification: an evaluation of novel deep learning methods in echocardiography. *Journal of Medical*

Artificial Intelligence. 2020; 3 4. Available from: doi: 10.21037/jmai.2019.10.03 Available from: https://search.proquest.com/docview/2384831024 .

(7) Albumentation.ai. *Albumentation.* Available from: http://www.albumentation.ai.

(8) Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik.* 2019; 29 (2): 102-127. Available from: doi: https://doi.org/10.1016/j.zemedi.2018.11.002 Available from: http://www.sciencedirect.com/science/article/pii/S0939388918301181 .

(9) Simonyan K, Zisserman A. *Very Deep Convolutional Networks for Large-Scale Image Recognition.* 2014. Available from: https://explore.openaire.eu/search/publication?articleId=od_____18::817f05f7aff35beeac308419b9b028c3 .

(10) He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition.* 2015. Available from: https://explore.openaire.eu/search/publication?articleId=od_____18::e7235b2295e7fd00c3555a8bfeb2c6b0 .

(11) Tan M, Le QV. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.* 2019. Available from: https://explore.openaire.eu/search/publication?articleId=od_____18::b2017de2c9fdf0efd4efcfd5f70ee6e4 .

(12) Smith LN. *A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay.* 2018. Available from: https://explore.openaire.eu/search/publication?articleId=od_____18::c5686481eb8483daaf80b5e4d4f648c9 .

(13) Liashchynskyi P, Liashchynskyi P. *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS.* 2019. Available from: https://explore.openaire.eu/search/publication?articleId=od_____18::4f07db30b0b438314f787049502ba53a .

(14) scikit-learn. *Cohen_Kappa_Score.* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html: .

(15) Wu Y, He K. *Group Normalization.* Lecture Notes in Computer Science Cham: Springer International Publishing; 2018. Available from: https://search.datacite.org/works/10.1007/978-3-030-01261-8_1.

(16) Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. *International journal of computer vision.* 2015; 115 (3): 211-252. Available from: doi: 10.1007/s11263-015-0816-y Available from: https://search.datacite.org/works/10.1007/s11263-015-0816-y .

(17) He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition.* 2015. Available from: https://explore.openaire.eu/search/publication?articleId=od_____18::e7235b2295e7fd00c3555a8bfeb2c6b0 .

(18) Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks.* 2008; 21 (2-3): 427-436. Available from: doi: 10.1016/j.neunet.2007.12.031 Available from: https://search.datacite.org/works/10.1016/j.neunet.2007.12.031 .

Appendix

I. Codebase and Data Analysis: https://www.kaggle.com/samueldsouza/viewdetector
II. Neural Network Metrics Logging: https://wandb.ai/samsrd/viewdetector/reports/ViewDetector-Final-Report--VmlldzozNTUxODA?accessToken=tpxjt50ed8ial3ty8503y81j1e52xms4fm6bovrv3c4sq51kbgdoertyg5rja7sf

III. Statistical Comparison to Human Expert: https://www.kaggle.com/samueldsouza/viewdetector-mcnemar-s-comparison
IV. Model Ensemble Training: https://www.kaggle.com/samueldsouza/viewdetector-ensemble-training
V. Model Ensemble Testing: https://www.kaggle.com/samueldsouza/viewdetector-ensemble-testing