university of groningen / faculty of science and engineering

Evaluating the quality of semantic segmentation in multi-source MRI images of the heart

Bachelor's Project Thesis

Gonçalo Hora de Carvalho, s3450295, g.hora.de.carvalho@student.rug.nl,
Supervisors: Prof. Dr. L.R.B. Schomaker Artificial Intelligence, University of Groningen
& Ming Wai Yeung PhD Student, University Medical Center Groningen
& Jan Walter Benjamins PhD Student, University Medical Center Groningen

Abstract: Recent research has shown advances in the analysis of cardiovascular MRI images using deep learning. However, two problems are apparent: How to measure the quality of the result of semantic segmentations and how to expose dependencies on the actual MRI apparatus used in obtaining the image data sets. The proposed method is based on traditional evaluations at the pixel level. Admittedly, it would be convenient to judge incoming samples on their familiarity in relation to the training data. This would allow for filtering out inadequate samples. In order to solve this conveniently, it is proposed to compare incoming samples to prototypical centroid vectors in an embedding (sub space), by using dimensionality reduction. MRI images used for this experiment are fed through a fully connected network model trained on short-axis MRI's of left ventricles. The machine learning model was tested using two different data sets collected from two different MRI devices, one generating the UK Biobank data and another, UMCG's data. The raw MRI's and the resulting segmentations are used for investigating the problem of finding a reliable comparison method for judging whether an input sample meets the expectations that are represented by the statistics of the training data. To achieve this, a dimensionally reduced representation of the data is calculated with which centroids can be computed for classes. Both are then used as dimensionally reduced representations of the data and averaged to represent the centroid of their embedding. An optimal measurement is discovered among three standard distance calculations (SAD, SSD and mean correlation), that is, SAD. This was the best measurement of similarity in raw MRIs (non-segmented) as well as serving as a predictor of segmentation quality, as verified by the Dice metric.

1 Introduction

Explainable AI has been receiving increasing attention in the last 10 years. This branch of AI tries to tackle problems that exist in the interface between users and AI systems. The aim revolves around describing model behaviour or decision making, understanding when a system is mostly right or very wrong in its output, how to correct this error and how trustworthy a system is [2]. This is especially relevant in the field of Medical Sciences, where the margin of error can be the threshold between life and death. This makes it an extremely sensitive area to deploying AI systems in practice or amidst existing workflows. But also, one that would prospectively benefit many times from doing so due to data dependence and abundance, technological involvement and many opportunities to automate and improve existing medical techniques.

Explainability often means a trade-off between prediction accuracy and model interpretability. Many technologies have been developed in the last decade that auxiliate model performance and deployment. For example, in deep learning a technique called deep explanation has been developed that maps an NLP model (RNN) capable of generating written explanations of models like CNNs and their decisions thus teaching models how to learn semantic associations in data features [11]. The two models output a classification together

with the associated image descriptions of discriminating features that led to the output label [9]. In model interpretation, stochastic And-Or-Graphs (AOG) [26] has been developed with the same aim as the previous research. This technique creates a five-layer AOG that maps semantic meanings between learned patterns [22]. Another technique called Probabilistic Program Induction attempts to describe parameters involved in character generation (e.g. in written symbols) by using a generative model that recognises these symbols by developing an explanation of how a new character may be created through a sequence of probable strokes that is deduced from the training cases [12]. Finally, there are model induction techniques. These refer to techniques that attempt to infer explainable models from other more complex models that are viewed as black boxes [10] (uninterpretable models). One such technique is LIME [19] (Local Interpretable Model-agnostic Explanations). This algorithm explains classifier predictions using locally approximate interpretable models that are then selected by a method that chooses a set of locally faithful representative instances of explanations of the whole model. In this project an understandable method is provided to judge the familiarity of an input sample in comparison to the input data before a prediction or classification is made regarding said data.

To illustrate an example of an interpretability problem tackled by this work, take a 250 by 250 image. Initially, it will be represented by 62500 features in a CNN. These are reduced through convolutions and max-pooling, for example, into a smaller feature representation. In the case of W. Bai, et al.'s model [5] used in this work, 16 convolutional layers downsample the feature map every two or three convolutions by a factor of 2. This means that the model will have reduced the image to approximately 245 features at the bottleneck layer. At this point, the model reached the minimal feature representation of the data that it can use to reliably classify an input image. These are 245 reasons why the image is classified with a particular label as the output of the model, all possibly contributing differently to the classification. If a visualisation technique is taken to try and make sense of this feature space, 245 dimensions would be seen - an explanation space that is made too complex to interpret by the 242 extra dimensions added on top of the graspable Cartesian plane.

Statistics can be used to reduce facts about complex data composed by inordinate amounts of singular data points to readily understood numbers (e.g. as a distribution's mean, the t-value of a t-test or the p-value used in hypothesis testing). There is then a need for such summarising values or concepts that work as sense-making mechanisms that allow for peering into the machine's mind, figuratively speaking. Especially in a field where most explanatory techniques involve model training and deployment which can easily lead to a recursion problem (producing a model that aims at explaining another model that is supposed to explain the production or main model).

This is one of the central problems regarding machine learning model interpretability - they are regarded as black boxes [10]. In such models, independently of knowledge concerning the mathematical principles involved in the system or model, there is no explicit knowledge representation that is readily understood by humans. Nor is there knowledge regarding underlying reasoning, the causal chains that lead to the output or their explanations that do not involve model making [2]. There is then a need to work towards making AI systems transparent, explainable and error [25]. Moreover, in the case of Medicine as a domain of application, this is especially relevant, where data complexity and heterogeneity is rampant. Medicine can then be seen as one of the best thresholds for AI explainability because the field will not accept anything below a gold standard. This implies that models have been extensively tested, error margins must be known and as low as possible. It has to be possible for professionals involved to understand the reasoning behind the model's probabilistic output.

In this work, a training-independent technique will be deployed to further substantiate the quality of model output, its generalisability, and interpretability. This is done using a MRI (magnetic resonance imaging) pipeline consisting of image processing and a CNN model [5] trained on the UK Biobank data set [7]. The technique involves simple measurements that could elucidate how familiar the model is with a new image by measuring the difference between the image and the training data set used (the feature knowledge of the model). This is done so as to enable some interpretation of the quality of the models' output as well as measur-

ing its generalisability before trying a large batch of new images from a novel data set. The resulting technique is supposed to deploy simple mathematics, no model production nor training. If the measurement - the distance between images - is too large, then the model probably doesn't generalise to the novel data set. The reason for this will not be self-evident and can occur for many reasons, including differences in the image generating apparatus.

A good dimensional space is required to enable feature extraction. PCA is the method of choice for many when it comes to dimensionality reduction of data (including image data). This is the benchmark used in this work. To check if using the bottleneck layers of a trained model results in a better representation than PCA when it comes to dimensionality reduction steps. These two methods will be deployed to obtain a feature vector of an input cardiac magnetic resonance image. Then, the similarity measurements will be used on both to compare them to a prototype vector that represents the data used for training. This can be obtained by averaging the image vectors used in the training data set, element-wise. This is the same as calculating the centroid of the data. The vector that generates the results that are closer to the ground truth (the centroid of the training data set) is the best according to this quantitative measure as long as it is in line with the resulting Dice metric. Raw images not included in the UK Biobank data set [8] should have a larger distance from the prototype and vice-versa: images from the UK Biobank data set (UKBB) should show a shorter distance from the prototype. While the segmentation data set that yielded the best Dice metric should be the closest to the segmentations of the training data set centroid.

Two research questions are then presented, along with two hypothesis.

Research Question 1: Does Bai et al.'s model generalise (achieve gold-standard 94% classification accuracy on the Dice metric) to a UMCG data set from the UKBB?

Hypothesis 1.1: Bai et al.'s model will not generalise to UMCG data.

And to test the distance measurement as a preventive check of model generalisability to new data sets, a second research question is asked. **Research Question 2**: Does a training independent technique that produces a similarity measurement of novel input to average training data point have a better accuracy (as measured by predicting the segmentation quality) when deploying a feature vector read from the bottleneck of the trained model than when using PCA for image feature reduction as measured by benchmark distance computations?

Hypothesis 2.1: Deploying a feature vector read from the bottleneck of the trained model is more accurate than using PCA for image feature reduction as measured by benchmark distances (the sum of absolute differences (SAD), the sum of squared differences (SSD) or the correlation coefficient).

2 Methods

The data sets of cardiac MRI used in the following experiments were fed to the model as loose sequences. In appendix G, two organised sequences of images used for training can be seen. In the case of the aforementioned data sets, the relevant images follow a sequence loosely and were picked to represent the various stages of the cycle instead of representing end diastolic relaxed periods or end systolic contraction periods specifically.

To test these hypothesis the Euclidean distance will be used on standardised, equal shaped image vectors after a normalisation algorithm is used (Appendix H). This is done so as to compare performance between using PCA-generated vector images and a feature vector directly read from the trained model's bottleneck layer when it comes to segmentation results as predicted by the Dice metric. The standard images of the left ventricle will also be analysed from this perspective so as to understand the distance of input images to the images used to train the model.

The *Euclidean* distance is the norm of the element wise difference between two arrays. A contiguous flattened array is calculated for each image on which the Euclidean distance will be determined:

$$d_{Euclid}(u, v) = \sqrt{\sum_i (u_i - v_i)^2}, \qquad (2.1)$$

where the distance d is the distance from some point $u(x_1, y_1)$ to some point $v(x_2, y_2)$, or between two feature vectors.

The aforementioned ML pipeline [5] was deployed in processing and analysing cardiovascular magnetic resonance images from the UK Biobank [7]. The model for LV, short-axis developed in the pipeline is reported to have gold-standard classification accuracy in the UK Biobank data set [5]. Regardless of this claim, a new sample of the data set was fed to the model and the resulting classification analysed (table 3.4). The sample consisted of 13 images and was evaluated as one sequence.

For evaluating the accuracy of the model segmentation on UMCG data, the same quantitative measurements deployed by W. Bai, et al. [5] will be used, namely, the Dice metric. This calculation outputs a value between 0 and 1 and is defined as:

$$Dice(u, v) = \frac{2|u \bigcap v|}{|u| + |v|} \tag{2.2}$$

It is used to calculate the overlap between an automated segmentation u and its respective manual segmentation v. The higher the output value the closer the contours, thus, the better the result.

Three more measurements are used and compared so as to obtain a good predictive metric in the context of the present work, namely that of comparing images and segmentations. In digital image processing, the sum of absolute differences (SAD) is a measure of the similarity between pixels. It is calculated by taking the absolute difference between each pixel in the original image and the corresponding pixel in the image being used for comparison.

$$d_{Abs}(u, v) = \sum_{i=0}^{n} \mid (u_i - v_i) \mid \tag{2.3}$$

Where u_i is the i^{th} item in an image vector and v_i is the i^{th} item in the other image vector.

The sum of squared differences (SSD) measures similarity based on pixel by pixel intensity differences between the two images.

$$d_{SSD} = \sum_{i=0}^{n} (u_i - \bar{u})^2 \tag{2.4}$$

Where u_i is the i^{th} item in the vector, $\bar{u}$ is the mean of all the pixels in the array and $(u_i - \bar{u})^2$ is the deviation of each pixel from the mean.

The correlation coefficient is a statistical measure of the strength of the relationship between the relative comparison of two images. The values range between -1.0 and 1.0, with a meaningful but weak relationship thresholded at 0.3 or 30%.

$$r(u, v) = \frac{Cov(u, v)}{\sigma_u \sigma_v} \tag{2.5}$$

Where $Cov(u, v)$ is the covariance of the variables u and v. And σ_u is the standard deviation of u and σ_v is the standard deviation of v.

3 Results

The Dice metric is used to compare automatic and expert drawn segmentations of the left ventricle. As such, it will be used for the same purpose as in Bai et al., to check automatic segmentation quality by comparing segmentation correlation.

Table 3.1: UKBB Dice Metric comparison (MDM - "Mean Dice Metric") calculated from 13 Dice distances.

	New UKBB data LV, SA (short-axis)	Bai et al. score LV, SA
MDM	0.132	0.94

Table 3.4 shows a very poor Mean Dice Metric result in overlap for the novel UKBB (UK Biobank) segmentations. This result might be induced by different intensity scales between the novel UKBB data set and the training data set. Indeed, it can be verified in Figure 3.9, where the overall image is much brighter than the UMCG image in Figure 3.8.

To check to what extent the model is applicable to UMCG data, a small data set consisting of 13 UMCG left ventricle short-axis images was also fed to the model. Below (table 3.2) is the Mean Dice Metric results.

Table 3.2: UMCG Dice Metric comparision (MDM - "Mean Dice Metric") calculated from 13 Dice distances.

	New UMCG data LV, SA	Bai et al. score LV, SA
MDM	0.574	0.94

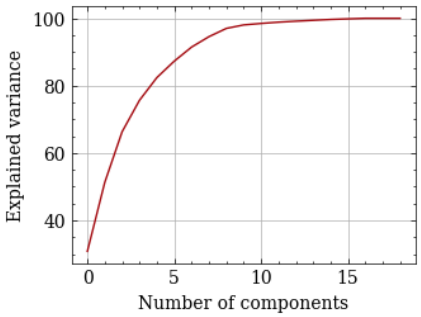
The results in table 3.2 show a much better dice metric than for the novel UKBB sample data set,

albeit it being lower than the 0.94 gold standard reported by Bai et al on the UKBB training set. It is seen then that even though hypothesis 1.1 is verified - that the model does not have a gold standard classification on UMCG data - the latter still shows a significantly higher value than the novel UKBB data set. This unexpected result will be addressed further below. If the measurements show that the used UKBB data set is indeed further from the average embedding representative of the training data of the model than the UMCG data set, hypothesis 2.1 will also hold as long as these are better predictors (smaller values or distances) than the PCA equivalents. This is the case if the UKBB novel data set shows a distance from the UKBB training data set greater than the UMCG data set and the PCA equivalents do not. Otherwise the results are inconclusive.

In appendix F the standard deviations of each vector can be seen as a preliminary comparison between PCA and the averaged vectors. These are quite similar between PCA and averaged vectors both for real images and the segmentations of said images. Overall low standard deviations can be observed, independently of the data set; this means centred distributions around the mean for the image arrays.

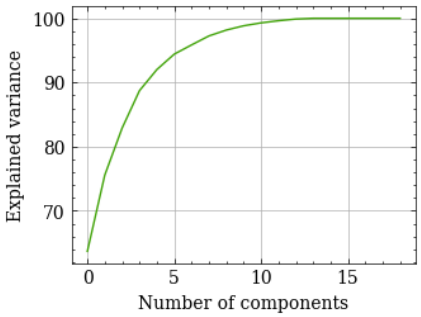
Plotting the PCA explained variance (check appendix E) reveals how many components can be discarded. This aids comparison with CNN bottleneck layer since the latter is the minimal representation possible while still allowing for reconstruction on the output. These fitted PCAs are then used as described in the methods section, to generate standardised vectors of the reduced data sets of images and segmentations. These images are used to compare distances by calculating average embeddings per data set (segmentation and normal images UKBB and UMCG).

Figure 3.1: Visualising the explained variance of a PCA fitted on 19 UMCG images



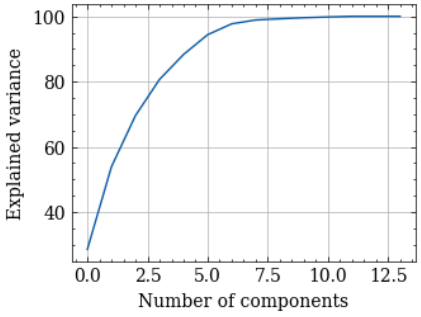
Using 12 components for PCA fitted on UMCG images.

Figure 3.2: Visualising the explained variance of a PCA fitted on 14 UMCG segmentations.



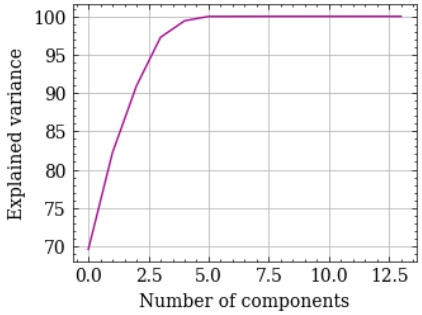
Using 12 components for fitted UMCG PCA on segmentations.

Figure 3.3: Visualising the explained variance of a PCA fitted on 14 novel UKBB images.



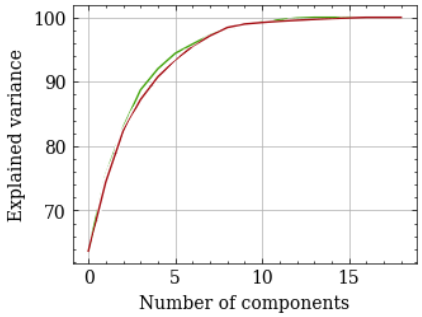
Using 8 components for fitted PCA on UKBB images.

Figure 3.4: Visualising the explained variance of a PCA fitted on 14 novel UKBB segmentations.



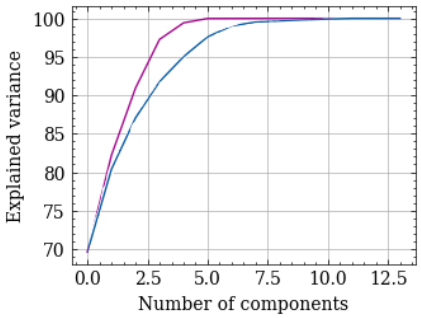
As can be seen in Figure 3.1, 12 components is all it takes to rebuild the images in UMCG PCA non-segmented data and so 12 components are used to draw image vectors and pack these in a data set. In UMCG segmented data (Figure 3.2), 12 components seem to be sufficient for image reconstruction as well. For UKBB non-segmented data (Figure 3.3) 8 components are used and 5 are deployed for the segmented data (Figure 3.4). Further comparison can be drawn between segmentations and MRI for the fitted PCAs.

Figure 3.5: Visualising and comparing the explained variance of the UMCG fitted PCA MRI and segmentations (segmentations in green, MRI in red).



The overlap is clear in the case of the UMCG data (Figure 3.5). These results serve as an indication for a good similarity between the segmentations and the MRIs, the two centroid classes used to build image array data.

Figure 3.6: Visualising and comparing the explained variance of the UKBB fitted PCA MRI and segmentations (segmentations in purple, MRI in blue).



In the case of the UKBB novel data (Figure 3.6), it can be noted that segmentations and MRI deviate before three components are available. This can be explained by the small number of images in both data sets. Notwithstanding this fact, the fleeting curves indicate dissimilarity for which further evidence will be collected.

The following tables show the distances between segmentation and image fitted PCA average embeddings and the average embedding that represents the UKBB data set used to train the model, produced by averaging 550 UKBB images and their respective segmentations separately.

Table 3.3: Distance measurements between PCA vectors and averaged vectors including segmentations from bottleneck layer (UMCG SEG) to averaged training vector from UMCG data set

	PCA seg	PCA img	UMCG seg	UMCG img
SAD	49.361	168.350	46.816	170.458
SSD	1.841	0.780	1.855	0.787
mean corr	0.071	0.323	0.064	0.324

Table 3.4: Distance measurements between PCA vectors and averaged vectors including segmentations from bottleneck layer (UKBB SEG) to averaged training vector from UKBB novel data set

	PCA seg	PCA img	UKBB seg	UKBB img
SAD	43.441	216.694	219.497	169.339
SSD	1.996	1.927	1.844	0.772
mean corr	-0.003	0.007	-0.021	0.105

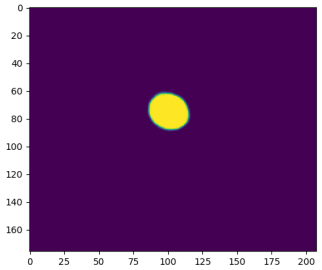
Table 3.5: Distance measurements between UMCG and UKBB vectors to averaged training vector.

	UMCG seg	UKBB seg	UMCG img	UKBB img
SAD	46.816	219.497	170.458	169.339
SSD	1.855	1.844	1.927	0.772
mean corr	0.064	-0.021	0.324	0.105

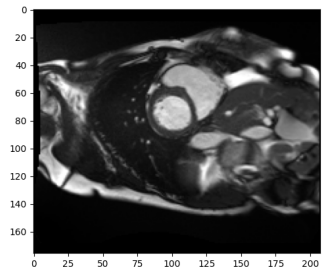
In order for hypothesis 2.1 to hold, that is, that deploying a feature vector read from the bottleneck of the trained model is better than using PCA for image feature reduction as measured by benchmark distances (the sum of absolute differences (SAD), the sum of squared differences (SSD) or the correlation coefficient), UMCG distances would have to be smaller than novel UKBB and PCA. This is because UMCG had the best Dice metric (best segmentation agreement between expertly drawn segmentations and model output) and thus these should be the most similar to the training set. The results in table 3.3 read as follow: regarding the real segmentation, the UMCG segmentation or bottleneck output is closer to average training set than the PCA segmentation, as measured by the sum of absolute differences. The SSD or sum of squares is higher for the UMCG data indicating that these vectors would fit UKBB's training set slightly worse than the PCA. A lower mean correlation indicates less correlation with the UKBB training data averaged vector than the PCA. It should be noted that these differences are very small when compared to the SSD.

Figure 3.7: Visualising a left ventricle segmentation used in the training UKBB data set.

(a) Segmentation



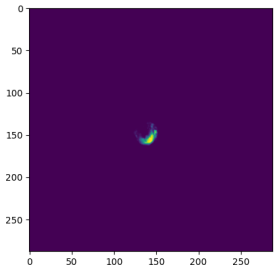
(b) LV SAX MRI (Left Ventricle Short-Axis Magnetic Resonance Image)



Taking the real images, under *UMCG img* and *PCA img*, a slight inversion of results occurs, favouring PCA averaged vectors over UMCG averaged vectors in all measures. These are smaller than for the difference across the segmentation data.

Figure 3.8: Visualising a left ventricle segmentation from the data used in the UMCG data set.

(a) Segmentation



(b) LV SAX MRI

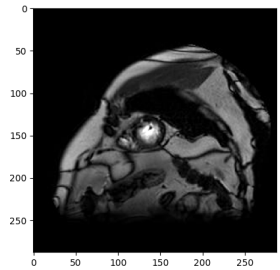
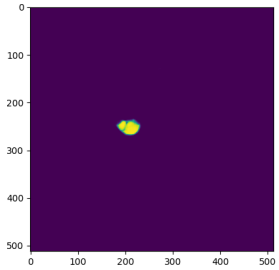
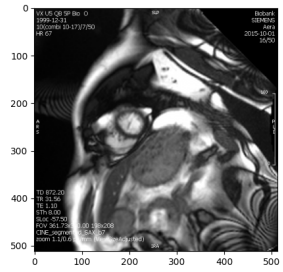


Figure 3.9: Visualising the data used in the UKBB novel data set.

(a) Segmentation



(b) LV SAX MRI



In comparing novel UKBB data and the averaged bottleneck segmentations and image vectors to PCA, a higher SAD for UKBB segmentation indicates a much worse similarity to the average training set than the PCA. The difference in fit as measured by SSD is almost neglectable in comparison, but the PCA is still worst off. Regarding the mean correlation, the PCA has less negative correlation than the novel UKBB segmentation average. In terms of raw images and not segmentations, the UKBB novel data set averaged vectors do much better in approximating the training UKBB data set than the PCA.

The UMCG data set is more similar to the average training set than the PCA in segmentation data, while the PCA fits the data slightly better in the case of raw images. The UKBB novel data set is more distant from the average training set when it comes to the segmentations than the PCA as measured by SAD.

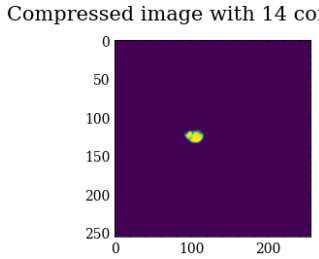
Comparing UMCG and novel UKBB in 3.5 and following from the Dice metric results, SAD (Sum of Absolute Distances) is the best measurement. The other measurements are unreliable as can be seen in Figure 3.5: SSD shows close scores for distances from UKBB training segmented set and the segmentations of UMCG and UKBB novel data sets. Since the Dice metric's base score (the ground truth) indicates that UMCG sample data is more similar to the UKBB training set than the novel UKBB sample data, this cannot be the case. That is, UMCG segmentations are much more similar than the novel UKBB data set and this is reflected in SAD but not in SSD. Mean correlation was inconclusive for all comparisons since none was above

the weak correlation threshold of 0.3. SAD yielded correct predictions for both UMCG segmentations and novel UKBB sample dataset raw images. Since novel UKBB images are still from the UKBB distribution, these raw images are expected to be more similar to the 550 UKBB data set than the UMCG ones and this is shown to be the case. The SAD score for novel UKBB images is smaller than for the UMCG data set, thus the former data set is more similar than the latter. The resulting SAD score for segmentations then predicts the Dice score as well as the objective expected difference between raw images.

The following compiled PCA images can be contrasted with the previous data set examples.

Figure 3.10: Visualising the data used in the UKBB PCA novel data set.

(a) Segmentation



(b) LV SAX MRI

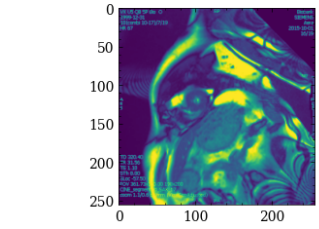
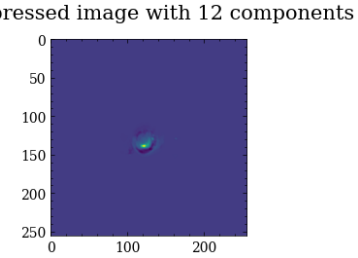
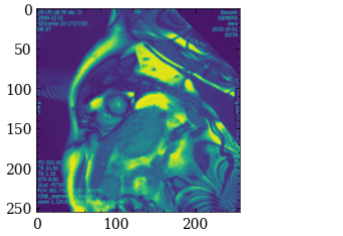


Figure 3.11: Visualising the data used in the UMCG PCA data set.

(a) Segmentation



(b) LV SAX MRI



4 Discussion

Deep learning (DL) is the most widely used Machine Learning (ML) approach, seeing plenty of applications across virtually every field, most prominently in visual object recognition and natural language processing (NLP). Models learn useful representations and features automatically, directly from data, bypassing an otherwise manual and conceptually difficult step of feature design and extraction. These automatic techniques have advantages particularly for high-dimensional data such as statistical images used in medical research. MRI is one of the fastest-growing research fields where ML meets a better-than-human performance in a clinical application, more specifically in the analysis of the output of a complex imaging technique [23]. CNNs, or convolutional neural networks, are the preferred algorithm in MRI analysis [14]. In a nutshell, automated MRI analysis using a fully convolutional network (FCN) architecture predicts a pixel-wise image segmentation (semantic segmentation) by applying a number of convolutional filters onto an input image. The convolutional layer maps local connectivity allowing the network to learn the spatial

local correlation of the input. Where for each pixel there is a higher correlation to its neighbouring pixels than to distant pixels. The model then learns image features from fine to coarse levels using these convolutions and combines multi scale features for predicting the label class at each pixel.

One problem that assails the multidisciplinary field of ML and more specifically the use of CNNs, in general, is the problem of model interpretability. Prospective applications of AI systems and efforts by the ML research community that could improve and automate diagnosis, for example, see these efforts cut short. Medical governing bodies and representatives do so on the basis of model interpretability and lack of generalisability. These experts claim that such a system is a 'black-box', where it is ultimately impossible, not to know, but to completely understand causal links and processes that generate the output so as to justify it.

SVMs (Support Vector Machines) and other automatic techniques often serve preliminary data analysis before feeding data to the main classifier in order to aid model interpretation [5]. By comparing new input to the training set used to fit the model in question, these methods usually give a dichotomous answer (yes or no) to the question "is this data similar enough to the training set to rely on model output?". This is so that researchers know when further learning is required. Two problems are relevant in this context: these methods need to be retrained on new data every time they are used - just like the classification model they serve. This is a problem because data is scarce, expensive and time-consuming to handle, as is the training involved in such preliminary models. The second problem arises from the fact that this binary check, like the served classifier, does not give reasons for its result, thus making it what some in the field refer to as a 'black-box'. Again, this is analogous to not knowing why the result is what it is, remaining unexplained or too complex to make sense out of.

SAD and other statistical summaries try to tackle the first problem by answering the same question without needing training time - they are training independent. And they tackle the second problem through their simplicity - no longer does a model need to be explained or understood since the model is removed and in its place simpler well known mathematical computations are used instead (e.g. Euclidean distance).

The goal of these experiments was to work towards AI explainability while reinforcing the use of techniques that are more general and can be applied across data sets regardless of what ML model is being deployed when checking the quality of the output of another model. If such a method is discovered, an increase in generalisability will be reached since the method is training-independent, thus reducing the resources that would have been necessary for creating a preliminary model that is not reusable (e.g. time, data collection, money and computational resources).

And perhaps more importantly, model-expert disagreement, where the model classifies a certain input as x but a human expert disagrees, taking the input to be a y instead, can be tackled by techniques that help elucidate, simplify and show how certain the model is of its conclusion. This can solve the disagreement between the human and the machine and further aid the pursuit of embedding AI in different industry wide workflows without fear for what goes on inside the machine.

5 Conclusions

The output of the bottleneck layer of a fully connected layered model was used to generate vectors representative of segmented sections of the heart, more specifically the left ventricle of a short-axis view. Using these, differences to another vector were measured. The latter vector was obtained from a sample data set consisting of data used to train the model on segmented MRI left ventricle short axis images of the heart. These efforts culminated in a data-independent technique that produces a similarity measurement in line with the expected model's output as measured by the Dice metric.

The distances measured on these averaged vectors are shown to be comparable to the state of the art PCA reduced vectors. These are comparable since the latter does not involve model training. In regards to the distance measurements, these are useful in comparing two vectors of the same domain in terms of their relative difference.

There are readily available computational methods for calculating this difference, including ML algorithms. But deploying a new model generates a black box problem and this is not acceptable in

medical science. This problem can potentially be avoided by deploying a PCA or a PCA equivalent dimensionality reduction method such as the bottleneck layer output of a model, together with said distance measurement, as long as this yields a reliable similarity measurement that predicts model outcome. This was the theoretical experiment proposed in the present work.

The Euclidean distance computes the distance of two real-valued vectors. It is the square root of the sum of squared differences across two vectors, element-wise. This method can be summarised in readily understood metrics which are widely used in statistics without the complex modelling layers that are rampant in ML algorithms. As such, the distance calculation strategy presented in this work gathered data that supports hypothesis 2.1, which has not been falsified. That is, that using a standard difference measurement - a distance measure between vectors based on simple arithmetic computations - can reliably predict the resulting Dice metric or segmentation quality of a model without the use of an extra model. More specifically, SAD or sum of absolute distances can be deployed for this purpose.

The averaged vectors from the UMCG data set have yielded the most similar measurements to the ground truth or vector averaged from the 550 MRIs UKBB data set used in training the model. This is evidenced by the Dice metric. Since the Dice metric was higher for the UMCG data set than for the novel UKBB sampled data set, a good distance measurement should result in a higher similarity (shorter distance) between UMCG and the training data set segmentations. This reasoning applies to raw images in the same way: the fact that novel UKBB raw MRI images belong to the same distribution as the training set used as ground truth means that novel UKBB raw MRI images should show a shorter distance than UMCG raw MRI images to said ground truth (under this measurement).

The results from the SSD metric do not predict the expected segmentation quality for either images, nor does mean correlation which yielded less than significant results (¡ 0.3) in both data set comparisons. The raw MRI novel UKBB data set, which was objectively closer in similarity, features and shape to the 550 image sample training set used to calculate the ground-truth or average embedding, had a shorter SAD result in raw image comparison. And UMCG segmented images had a shorter SAD as well.

In conclusion, SAD both shows that the novel UKBB data set is closer to the ground truth in terms of real image differences (the non-segmented MRIs) as well as predicting that UMCG segmentations are more similar to ground truth than the UKBB novel data set's as evidenced by the Dice Metric, which was higher for UMCG data than for the novel UKBB data set segmentations.

Regarding the poor SSD and mean correlation results, these could be caused by the differences in image production techniques, like image standardisation algorithms used in the making of the data sets as well as during the production phase (e.g. different MRI machines or underlying preprocessing algorithms). Future work should be done in rectifying that dissimilarity did not increase due to these intermediate processes. Furthermore, establishing a meaningful threshold for the SAD metric per data set is a process that still requires further analysis and development. This could be done by calculating average differences using images belonging to the same distribution as described in the present paper. Finally, upon establishing such a threshold, an image can be vectorised or reduced after which SAD should be calculated. Thresholding can then be deployed to substantiate model output and interpret it further.

References

[1] Michel Jose Anzanello and Flavio Sanson Fogliatto. "Learning curve models and applications: Literature review and research directions". English. In: *International Journal of Industrial Ergonomics* 41.5 (2011), pp. 573–583. DOI: 10.1016/j.ergon.2011.05.001.

[2] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI". In: *CoRR* abs/1910.10045 (2019). arXiv: 1910.10045. URL: http://arxiv.org/abs/1910.10045.

[3] W. Bai et al. "Recurrent Neural Networks for Aortic Image Sequence Segmentation with Sparse Annotations". In: *Medical Im-*

age Computing and Computer Assisted Intervention ? MICCAI 2018. Vol. 11073. Lecture Notes in Computer Science. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-00937-3_67. Cham: Springer, Sept. 2018, pp. 586–594. DOI: 10.1007/978-3-030-00937-3_67. URL: https://openaccess.city.ac.uk/id/eprint/23149/.

[4] Wenjia Bai et al. "A population-based phenome-wide association study of cardiac and aortic structure and function". In: *Nature Medicine* 26.10 (Oct. 2020), pp. 1654–1662. ISSN: 1546-170X. DOI: 10.1038/s41591-020-1009-y. URL: https://doi.org/10.1038/s41591-020-1009-y.

[5] Wenjia Bai et al. "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks". In: *Journal of Cardiovascular Magnetic Resonance* 20.1 (Sept. 2018), p. 65. ISSN: 1532-429X. DOI: 10.1186/s12968-018-0471-x. URL: https://doi.org/10.1186/s12968-018-0471-x.

[6] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.12.012. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[7] C. Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562 (2018), pp. 203–209.

[8] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (Oct. 2018), pp. 203–209. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0579-z. URL: https://doi.org/10.1038/s41586-018-0579-z.

[9] Lisa Anne Hendricks et al. "Generating Visual Explanations". In: *CoRR* abs/1603.08507 (2016). arXiv: 1603.08507. URL: http://arxiv.org/abs/1603.08507.

[10] Andreas Holzinger et al. "What do we need to build explainable AI systems for the medical domain?" In: *CoRR* abs/1712.09923 (2017). arXiv: 1712.09923. URL: http://arxiv.org/abs/1712.09923.

[11] Rakesh Kumar et al. "Aerial video surveillance and exploitation". In: *Proceedings of the IEEE* 89 (Nov. 2001), pp. 1518–1539. DOI: 10.1109/5.959344.

[12] B. Lake, R. Salakhutdinov, and J. Tenenbaum. "Human-level concept learning through probabilistic program induction". In: *Science* 350 (2015), pp. 1332–1338.

[13] Hao Li et al. *Pruning Filters for Efficient ConvNets*. 2017. arXiv: 1608.08710 [cs.CV].

[14] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (Dec. 2017), pp. 60–88. ISSN: 1361-8415. DOI: 10.1016/j.media.2017.07.005. URL: http://dx.doi.org/10.1016/j.media.2017.07.005.

[15] Adnan Mustafa. "A Complete Probabilistic Model for the Quick Detection of Dissimilar Binary Images by Random Intensity Mapping". In: *WSEAS Transactions on Signal Processing* 13 (Oct. 2017), pp. 208–214.

[16] Adnan Mustafa. "A Modified Hamming Distance Measure for Quick Detection of Dissimilar Binary Images". In: Jan. 2015. DOI: 10.1109/ICCVIA.2015.7351897.

[17] Steffen E. Petersen et al. "Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort". In: *Journal of Cardiovascular Magnetic Resonance* 19.1 (Feb. 2017), p. 18. ISSN: 1532-429X. DOI: 10.1186/s12968-017-0327-9. URL: https://doi.org/10.1186/s12968-017-0327-9.

[18] Steffen E. Petersen et al. "UK Biobank's cardiovascular magnetic resonance protocol". In: *Journal of Cardiovascular Magnetic Resonance* 18.1 (Feb. 2016), p. 8. ISSN: 1532-429X. DOI: 10.1186/s12968-016-0227-4. URL: https://doi.org/10.1186/s12968-016-0227-4.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[21] H. Shin et al. "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning". In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1285–1298. DOI: 10.1109/TMI.2016.2528162.

[22] Zhangzhang Si and Song-Chun Zhu. "Learning AND-OR Templates for Object Recognition and Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2013), pp. 2189–2205.

[23] K. R. Siegersma et al. "Artificial intelligence in cardiovascular imaging: state of the art and implications for the imaging cardiologist". In: *Netherlands Heart Journal* 27.9 (Sept. 2019), pp. 403–413. ISSN: 1876-6250. DOI: 10.1007/s12471-019-01311-1. URL: https://doi.org/10.1007/s12471-019-01311-1.

[24] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[25] IT GOVERNANCE PRIVACY TEAM. *EU General Data Protection Regulation (GDPR): An Implementation and Compliance Guide - Second edition*. 2nd ed. IT Governance Publishing, 2017. ISBN: 9781849289450. URL: http://www.jstor.org/stable/j.ctt1trkk7x.

[26] Tianfu Wu, Yang Lu, and Song-Chun Zhu. "Online Object Tracking, Learning and Parsing with And-Or Graphs". In: *CoRR* abs/1509.08067 (2015). arXiv: 1509.08067. URL: http://arxiv.org/abs/1509.08067.

A Classification performance measurements in Bai et al.'s pipeline

A.0.1 Dice Metric

The DICE metric requires the **MIRTK** tool to be installed. Installing CMAKE (a compiler tool) to get MIRTK to run and solve dependency problems might be required. The relevant code is in **common/cardiac_utils.py** (lines 939:961). The result is saved as a pandas dataframe. This function utilizes **common/image_utils.py** to calculate the actual metric. Two functions exist, one that solves the dice metric using the Tensorflow library and another that uses the Numpy library instead (namely, **tf_categorical_dice()** and **np_categorical_dice()**).

These three metrics were implemented again so as to process images outside of the pipeline. These methods can be seen bellow.

Listing 1: Python updated segmentation quality measurements

```
def compute_dice_coefficient(mask_gt, mask_pred):
  volume_sum = mask_gt.sum() + mask_pred.sum()
  if volume_sum == 0:
    return np.NaN
  volume_intersect = (mask_gt & mask_pred).sum()
  return 2*volume_intersect / volume_sum
```

B UK Biobank's cardiovascular magnetic resonance protocol[18]

B.1 The protocol

The reliability of most image processing networks, indeed, even their generalisabilty is dependent on details regarding the data used during training, testing and verification. Furthermore, when the origin of these data are advanced medical imaging machines, protocol differences across hospitals (e.g. at production level regarding image cleaning work flows) and manufacturers (e.g. regarding production level work flows) may change the resulting data processing strategies and the model output. Regarding the CMRs used in the UK Biobank dataset, 20-min CMRs were taken by a 1.5 Tesla scanner (fabricated by MAGNETOM Aera, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany). This was used together with the Cardiac Dot Engine (Siemens Healthcare, Erlangen, Germany) for quality control and consistency of image acquisition throughout the study.

On top of the machine's standard cardiac package, a Shortened Modified Look-Locker Inversion recovery technique (ShMOLLI, WIP780B) was implemented to allow native (non-contrast) myocardial T1 mapping.

In regards to UK Biobank's CMR acquisitions, these include piloting and sagittal, transverse and coronal partial coverage of the chest. For cardiac function, horizontal long axis, vertical long axis, left ventricular outflow tract, and cines, both sagittal and coronal, were obtained. The relevant stack for the work at hand would be the short axis (SA) stack of balanced steady state free precession (bSSFP) cines, that cover the left ventricle (LV) and right ventricle (RV).

Immediately before and after this bSSFP acquisition of the aorta, brachial blood pressure readings are being obtained using a manual sphygmomanometer used for calibrating peripheral waveforms and immediately afterwards a brachial pressure wave trace is digitally computed by the Vicorder (Skidmore Medical, Bristol, UK) with the cuff statically inflated to 70 mmHg using a volume displacement technique. The Vicorder software calculates values for central blood pressure by ap-

plying a brachial-to-aortic transfer function. Aortic distensibility represents the relative change in area of the aorta per unit pressure, measurement which is taken into account by Bai and used to check segmentation validity.

B.2 Access

Researchers are able to access the DICOM CMR image files through the UK Biobank portal by authenticated request. The automated inline ventricular function option is enabled on the scanner providing automatic assessment of LV contours and volumes. There is also a multitude of bio-statistical data available regarding subjects.

B.3 Funding and manual contours

Manual analysis to create a CMR reference standard for the UK Biobank imaging resource in 5000 CMR scans was funded by the British Heart Foundation (BHF) project grant (PG/14/89/31194, PI Petersen until 2018. A UK Biobank CMR Image Analysis Consortium exists and has been working towards standardization and automating CMR image analysis.

C NIfTI images

Medical images have 4 key constituents: Pixel Depth, Photometric Interpretation, Metadata, and Pixel data. These constituents are responsible for the size and resolution of the image. The medical image expected by Bai et al.'s pipeline is of NIfTI type. This is not a popular type in machine learning nor in cardiology. NIfTI stands for Neuroimaging Informatics Technology Initiative. A major feature is that the format contains two affine coordinate definitions. In Euclidean geometry, an affine transformation is a geometric transformation that preserves lines and parallelism. These relate each voxel index (i,j,k) to a spatial location (x,y,z) - a single sample, or data point representing a single piece of data, such as opacity, or multiple pieces of data, such as a color in addition to opacity. The main difference between the more common ML type for image manipulation - DICOM - and NIfTI is that the raw image data in NIfTI is saved as a 3d image, whereas in DICOM an image is saved in 2d slices. This makes NIFTI more preferable for some machine learning applications over DICOM, because it is modelled as a 3d image by definition. Handling a single NIFTI file instead of several hundreds of DICOM is easier for MRI purposes or modeling the corresponding 3d body parts. NIfTI stores 2 files per 3d image as opposed to dozens in the more popular DICOM. Its use is aimed at establishing a technical solution to the problem of multiple data formats used in fMRI (functional magnetic resonance imaging) research.

D Regarding Bai et al.'s model

The pipeline and associated scripts can be found in the following Github repository: `https://github.com/BlueVelvetSackOfGoldPotatoes/medicalMRIcnn`. The network is adapted from the VGG-16 network[24] and it consists of a number of convolutional layers for extracting image features. Each convolution uses a 3 * 3 kernel and it is followed by batch normalisation and ReLU. After every two or three convolutions the feature map is down sampled by a factor of 2 so as to learn features at a more global scale. Feature maps learnt at different scales are up sampled to the original resolution using transposed convolutions and the multi-scale feature maps are then concatenated. Finally, three convolutional layers of kernel size 1 * 1, followed by a softmax function, are used to predict a probabilistic label map. The segmentation is determined at each pixel by the label class with highest softmax probability. The mean cross entropy between the probabilistic label map and the manually annotated label map is used as the loss function. Excluding the transposed convolutional layers, this network has in total 16 convolutional layers. An image representing the network architecture can be found in figure G.2. This architecture is similar to the U-Net[20]. The main difference is that U-Net performs up sampling step by step. It iteratively up samples the feature map at each scale by a factor of 2 and concatenates with the feature map at the next scale. In contrast to this, the proposed network may be simpler on the up sampling path. It up samples the feature map from each scale to the finest resolution in one go and then concatenates all of them. In sum:

- Each convolution uses a 3 * 3 kernel and it is followed by batch normalisation and ReLU;
- After every two or three convolutions, the feature map is down sampled by a factor of 2 or 3;
- Feature maps learnt at different scales are up sampled to the original resolution using transposed convolutions and the multi-scale feature maps are then concatenated;
- Three convolutional layers of kernel size 1 * 1, followed by a softmax function, are used to predict a probabilistic label map;
- This network has in total 16 convolutional layers.

D.1 Pre-processing

DICOM images as well as manual annotations were converted into NIfTI format. For short-axis images, 4875 subjects (93500 annotated image slices) were available - randomly split into three sets of 3975/300/600 for training/validation/test. 3,975 subjects for training the neural network, 300 validation subjects for tuning model parameters, and finally 600 test subjects for evaluating performance.

E Regarding PCA Analysis

The goal in principal component analysis or PCA is to extract the important information from the data and to express this information as a set of summary indices called principal components. These can be taken to be the lines that best fit the data out of all the points that make up the image. In appendix G, figure G.1 there is a small visualisation analysis that was done using Jupyter notebooks. The less components used the less information is available, thus yielding a worst reconstruction of the original image.

It can be seen in appendix G, figure G.1, that with just 50 components (half of the fitted PCA) a fully recognisable image can still be reconstructed.

F Std Tables

	STD DEV
Real Imgs	0.248
Seg Imgs	0.112

Table F.1: STD DEV of avg vector representative of dataset used to train the model constructed from 550 Short Axis images and their segmentations

	Novel UKBB	PCA
Real Imgs	0.236228	0.236229
Seg Imgs	0.1077	0.1081

Table F.2: STD DEV of avg vector and PCA representative of novel ukbb dataset used to test the model constructed from 13 Short Axis images and their segmentations

	UMCG	PCA
Real Imgs	0.2822	0.2847
Seg Imgs	0.11117	0.11115

Table F.3: STD DEV of avg vector and PCA representative of UMCG dataset used to test the model constructed from 18 Short Axis images and their segmentations

	Brain Tumor MRI	Flowers	Hand Digits
Avg Vector	0.3292	0.220	0.226
PCA	0.3291	0.239	0.223

Table F.4: STD DEV of avg vector and PCA representative of three other small datasets that serve as more evidence for the quality of the distance metric

G Images

List of Figures

3.1 Visualising the explained variance of a PCA fitted on 19 UMCG images . 5
3.2 Visualising the explained variance of a PCA fitted on 14 UMCG segmentations. 5
3.3 Visualising the explained variance of a PCA fitted on 14 novel UKBB images. 5
3.4 Visualising the explained variance of a PCA fitted on 14 novel UKBB segmentations. 6
3.5 Visualising and comparing the explained variance of the UMCG fitted PCA MRI and segmentations (segmentations in green, MRI in red). . . 6
3.6 Visualising and comparing the explained variance of the UKBB fitted PCA MRI and segmentations (segmentations in purple, MRI in blue). 6
3.7 Visualising a left ventricle segmentation used in the training UKBB data set. 7
3.8 Visualising a left ventricle segmentation from the data used in the UMCG data set. 8
3.9 Visualising the data used in the UKBB novel data set. 8
3.10 Visualising the data used in the UKBB PCA novel data set. 9
3.11 Visualising the data used in the UMCG PCA data set. 9
G.1 Visualising the effects of diminishing components in image fitted PCA. . . 20
G.2 Visualising the general model deployed by Bai et al. [5] 20
G.3 Visualising the UKBB LV and RV. . 21
G.4 Visualising the UKBB sequence of end diastolic relaxation period where the chamber fills itself. 21
G.5 Visualising the UKBB sequence of end systolic contraction where the chamber empties itself. 22

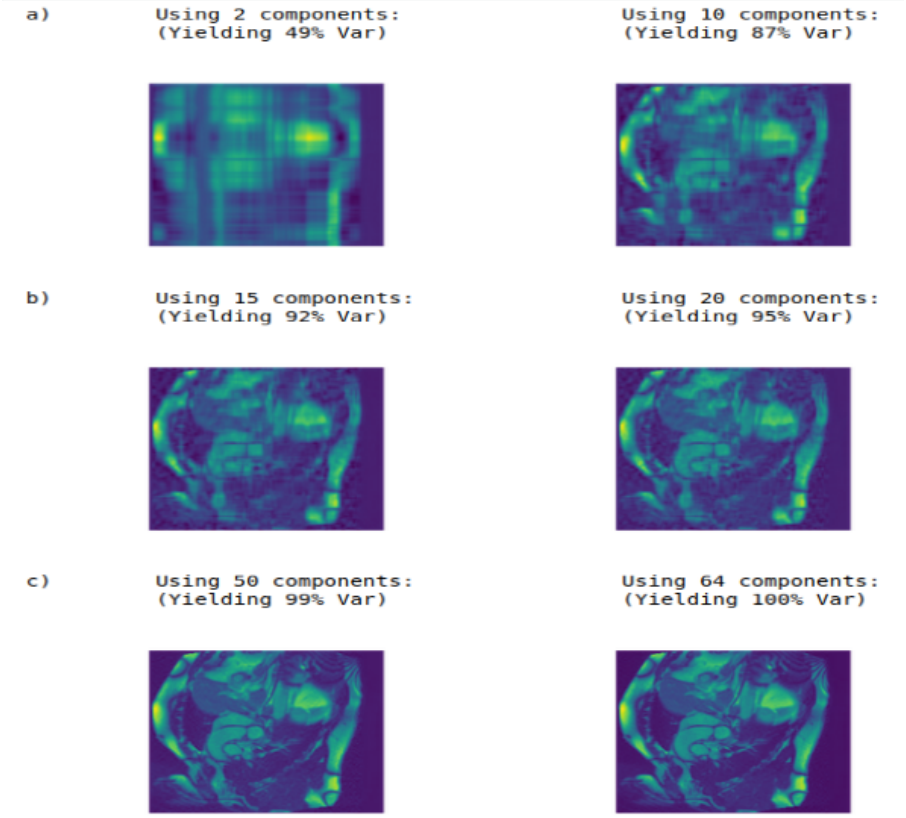


Figure G.1: Visualising the effects of diminishing components in image fitted PCA.

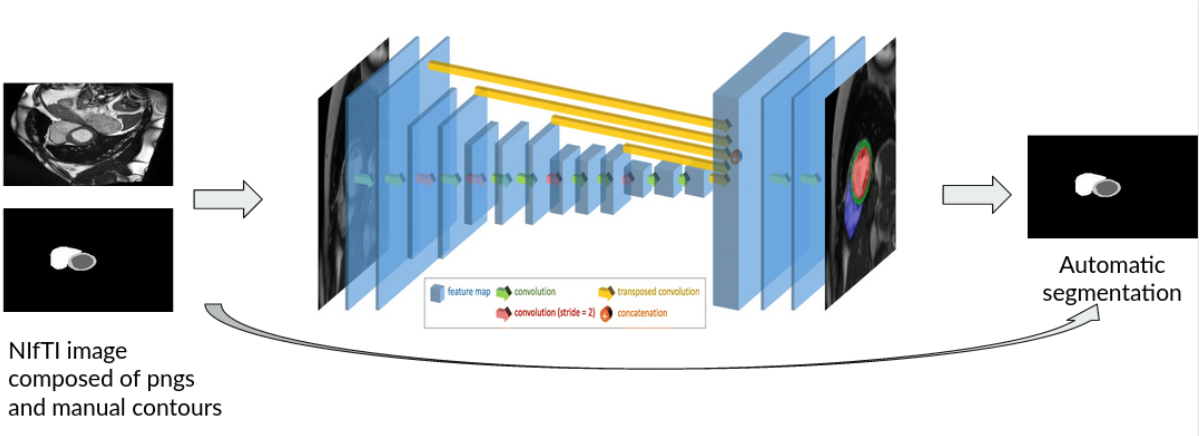


Figure G.2: Visualising the general model deployed by Bai et al. [5]

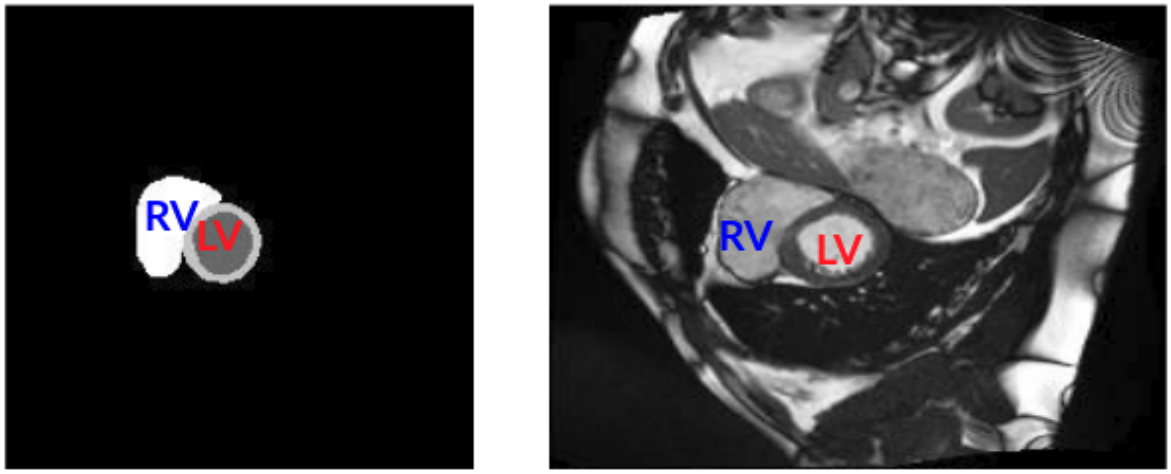


Figure G.3: Visualising the UKBB LV and RV.

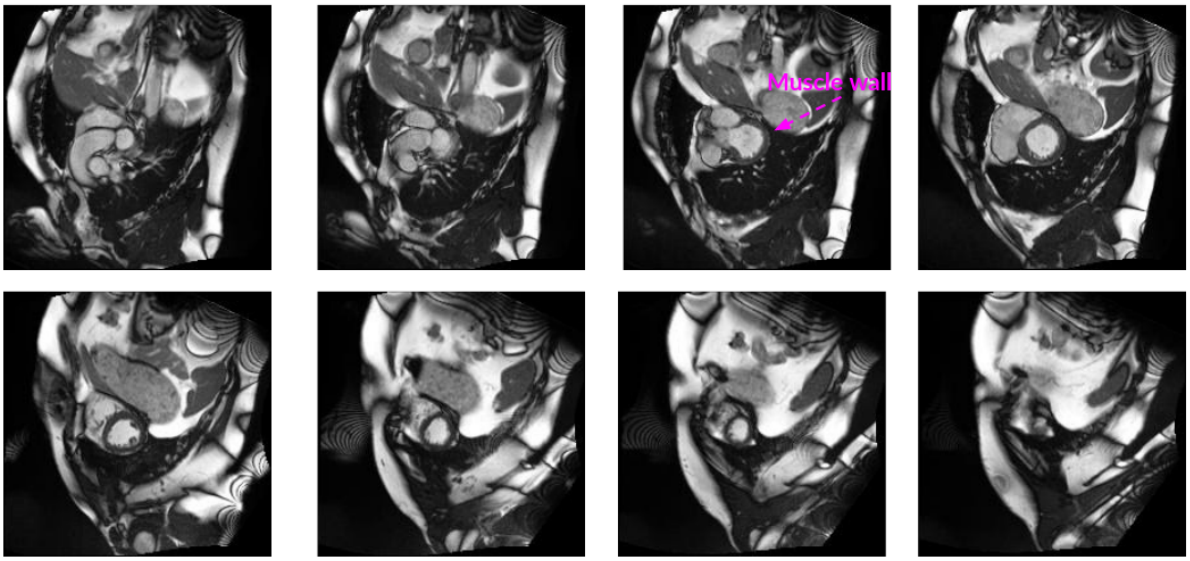


Figure G.4: Visualising the UKBB sequence of end diastolic relaxation period where the chamber fills itself.

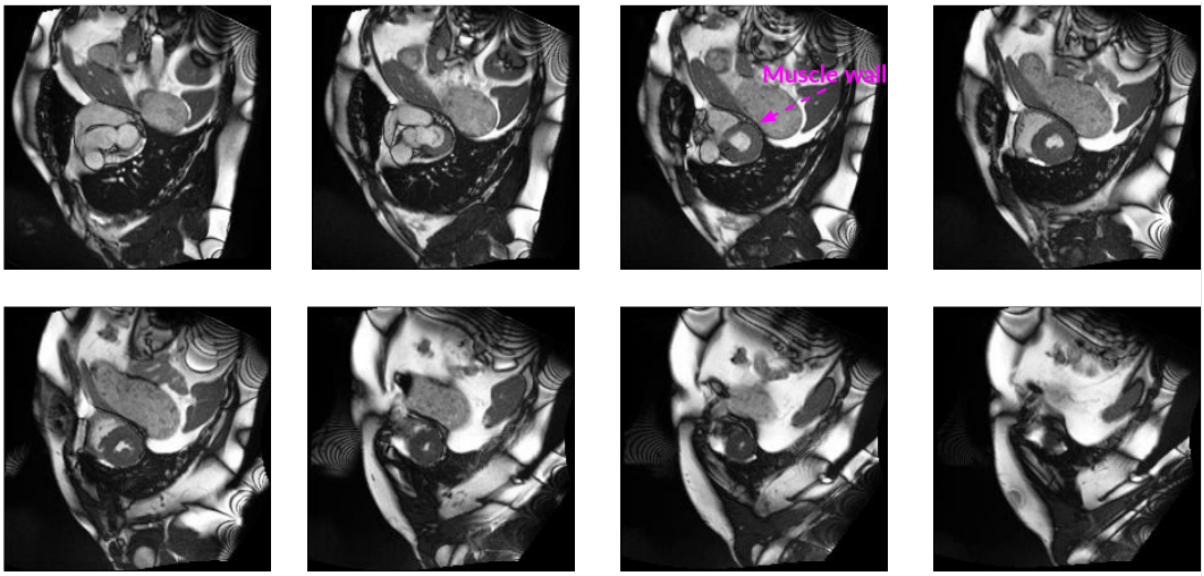


Figure G.5: Visualising the UKBB sequence of end systolic contraction where the chamber empties itself.

H Algorithmic transformations of images

Listing 2: Python standard image loading and resizing

```
'''
INPUT: img - path to image.
OUPUT: final_img - the resized, vectorized and
    grey scaled image.
'''
def standardize(img, resize=True):
    img_grey = cv2.imread(img, cv2.
        IMREAD_GRAYSCALE)
    if resize:
      resized = cv2.resize(img_grey, (256, 256))
    thresh = 1
    img_binary = cv2.threshold(resized, thresh,
        255, cv2.THRESH_BINARY)[1]

    final_img = center_image(img_binary)

    return final_img
```

Listing 3: Python standardisation of image vector before adding image to vectorized dataset

```
'''
INPUT: img - a vectorized image, vectorized
    using cv2.
OUPUT: pixels - the resulting standardized
    vector
'''
def standardize(img):
    pixels = img.flatten()
    pixels = pixels / np.linalg.norm(pixels, ord
        =1)
    pixels = normalize(pixels[:, np.newaxis], axis
        =0).ravel()
    return pixels
```