Unsupervised shape and motion analysis of 3822 cardiac 4D MRIs of UK Biobank

Qiao Zheng, Hervé Delingette, Kenneth Fung, Steffen Petersen, Nicholas Ayache

HAL Id: hal-02043380

https://hal.inria.fr/hal-02043380

Preprint submitted on 21 Feb 2019

Unsupervised shape and motion analysis of 3822 cardiac 4D MRIs of UK Biobank

Qiao Zheng[a,*], Hervé Delingette[a], Kenneth Fung[b,c], Steffen E. Petersen[b,c], Nicholas Ayache[a]

[a]*Université Côte d'Azur, Inria, 2004 Route des Lucioles, 06902 Sophia Antipolis, France*
[b]*William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, London EC1M 6BQ, UK*
[c]*Barts Heart Centre, St Bartholomews Hospital, Barts Health NHS Trust, London EC1A 7BE, UK*

Abstract

We perform unsupervised analysis of image-derived shape and motion features extracted from 3822 cardiac 4D MRIs of the UK Biobank. First, with a feature extraction method previously published based on deep learning models, we extract from each case 9 feature values characterizing both the cardiac shape and motion. Second, a feature selection is performed to remove highly correlated feature pairs. Third, clustering is carried out using a Gaussian mixture model on the selected features. After analysis, we identify two small clusters which probably correspond to two pathological categories. Further confirmation using a trained classification model and dimensionality reduction tools is carried out to support this discovery. Moreover, we examine the differences between the other large clusters and compare our measures with the ground-truth.

Keywords: Cluster analysis, Feature extraction, Cine MRI, UK Biobank, Cardiac pathology, Cardiac motion, Gaussian mixture model

1. Introduction

In recent years, more and more data are made accessible for research in medical image analysis. For instance, the UK Biobank study of Petersen et al. (2017) has released a dataset containing the cardiac cine MRI images of thousands of volunteers, from which various key cardiovascular functional indexes can be extracted for analysis (Attar et al. (2019)). The Alzheimer's Diseases Neuroimaging Initiative (ADNI, Toga and Crawford (2015)) has accumulated brain scan images of about two thousand participants. The abundant data available in the community is certainly a highly valuable resource (Rueckert et al. (2016), Suinesiaputra et al. (2016)). Researchers are hence less constrained by the scarcity of data which has been a prevailing challenge for a long time. Further re-

search is necessary (Zhang and Metaxas (2016), Barillot et al. (2016)) on new topics associated with big data. For example, one major challenge is how to make good use of unlabeled data (de Bruijne (2016), Weese and Lorenz (2016)). In fact, while there are more and more labeled data available, an important part of medical images are still unlabeled. This is understandable as it is in general expensive and tedious to diagnose and label cases by human experts. Methods that can extract useful information from unlabeled data are hence interesting and might potentially save a lot of time and effort.

Many research projects have been developed to perform pathology-related analysis using features extracted from medical images. Many of these works focus on brain scan images. For example, in Parisot et al. (2018), feature vectors extracted from brain images are used for the prediction of autism spectrum disorder and Alzheimer's disease. An anatomical landmark based deep feature representation for MRI is proposed in Liu et al. (2018) for

---

[*]Corresponding author
Email address: `qiao.zheng@inria.fr` (Qiao Zheng)

diagnosis of brain disease. Some other studies are based on digital histopathological images. For instance, Madabhushi and Lee (2017) discusses the predictive modeling of digital histopathological images from a detection, segmentation, feature extraction, and tissue classification perspective. Komura and Ishikawa (2018) reviews the machine learning methods for histopathological image analysis. But there are less pathology-related and feature-based researches on cardiac images than on brain scan images and digital histopathological images. And currently, this research (Zheng et al. (2018a), Khened et al. (2018), Khened et al. (2017), Isensee et al. (2017), Wolterink et al. (2017), Cetin et al. (2017)) is mostly about pathology classification in the dataset of Automatic Cardiac Diagnosis Challenge (ACDC) of MICCAI 2017, which contains 100 cases with labels. The work of Attar et al. (2019) is one of the very first projects to propose a fully automatic, high throughput image parsing workflow for the analysis of cardiac MRI in UK Biobank with systematic tests of the performance. As an extension of the previous works and a challenge to ourselves, we wish to conduct unsupervised analysis on large unlabeled cardiac image datasets.

Clustering, an unsupervised machine learning technique that groups similar entities together, might be suitable for analyzing large unlabeled datasets. Up to now, clustering has been widely used on image segmentation in medical image analysis. For example, the authors of Kinani et al. (2017) develop a tool based on clustering to outline brain lesion contours. Unsupervised segmentation of 3D lung CT images is proposed in Moriya et al. (2018) based on clustering and deep representation learning. Some studies show that clustering is also a powerful tool for classification. For instance, a clustering method is applied to classify the analyzed brain images into healthy and multiple sclerosis disease in Moldovanu et al. (2015). The authors of Kawadiwale and Rane (2014) introduce various clustering techniques to classify brain MR images into normal and malformed. While most of the application of clustering in the domain is on brain images, we aim to extend its application to cardiac images.

In this paper, we perform a cluster analysis of a group of features extracted from the cardiac MR images of the UK Biobank dataset. Our main contributions are threefold:

• We conduct a cardiac-pathology-related analysis on a large unlabeled dataset.

• As a novel application of a classic method in medical image analysis, clustering is used in our analysis to group cases without supervision.

• Among the resulting clusters, two can indeed be identified as leaning toward pathological categories.

2. Data

2.1. UK Biobank

The proposed method was applied to the very large UK Biobank cardiac MRI dataset, see Petersen et al. (2016)[1]. It comprises short-axis cine MRI of about five thousand participants from the general population. More details of the magnetic resonance protocol are available in Petersen et al. (2016). Each time series consists of 3D volumes with slice thickness of 8mm for short-axis images and 6mm for long-axis images. The in-plane resolution is 1.8mm × 1.8mm. Volumes at end-diastole (ED) and end-systole (ES) and ejection fraction for left ventricle cavity (LVC) were derived from InlineVF analysis algorithm (Jolly et al. (2013), Lu et al. (2010)) performed by UK Biobank (Field 22421-22422). Those values are considered in this paper as ground-truth (or reference) values. To be consistent with our previous research such as Zheng et al. (2018b) and Zheng et al. (2018a), we exclude roughly one thousand cases that are provided with incomplete or unconvincing ground-truth. The remaining 3822 cases are then used for cluster analysis. For part of these cases, the measures of LVC volumes at ED and ES and LVC ejection fraction are provided as ground-truth by UK Biobank.

As pointed out on the website of UK Biobank[2] and in Fry et al. (2017), while UK Biobank participants are not representative of the general population with evidence of a 'healthy volunteer' selection bias (and hence cannot be used to provide representative disease prevalence and incidence rates), valid assessment of exposure-disease relationships are nonetheless widely generalizable and does not require participants to be representative of the population at large.

---

[1] Application Number 2964.
[2] https://www.ukbiobank.ac.uk/scientists-3/

2.2. ACDC

In the experiment part, we will show the correspondence between some resulting clusters and the definition of some pathology categories defined in the ACDC challenge. Furthermore, a classification model trained on ACDC by Zheng et al. (2018a) will be applied on UK Biobank for comparison with the clustering method proposed in this paper. The ACDC challenge dataset consists of 100 cases, which are divided into 5 pathological groups of equal size according to their pathology on either the left ventricle (LV) or the right ventricle (RV):

- dilated cardiomyopathy (DCM): left ventricle cavity (LVC) volume at ED larger than 100 mL/m^2 and LVC ejection fraction lower than 40%
- hypertrophic cardiomyopathy (HCM): left ventricle (LV) cardiac mass higher than 110 g/m^2, several myocardial segments with a thickness higher than 15 mm at ED and a normal ejection fraction
- myocardial infarction (MINF): LVC ejection fraction lower than 40% and several myocardial segments with abnormal contraction
- RV abnormality (RVA): right ventricle cavity (RVC) volume higher than 110 mL/m^2 or RVC ejection fraction lower than 40%
- normal subjects (NOR)

3. Methods

There are mainly three steps in the proposed method: feature extraction, feature selection and cluster analysis.

3.1. Feature Extraction

The feature extraction method used in this paper is the same as the one proposed in our previous work published by Zheng et al. (2018a). We briefly describe its principal steps again below.

The first part of the feature extraction method generates 7 shape-related features. Segmentation with spatial propagation has been proven to be consistent and robust (Zheng et al. (2018b), Zheng et al. (2018c)). With the cardiac segmentation method proposed in Zheng et al. (2018b), the cardiac images are segmented such that we obtain the masks of LVC, left ventricle myocardium (LVM) and RVC on both ED and ES frames. Then the

Table 1: The 9 features generated by our feature extraction method. Among them 8 are selected for cluster analysis.

Feature	Notion	Selected
RVC volume at ED	$V_{RVC,ED}$	yes
LVC volume at ES	$V_{LVC,ES}$	yes
RVC ejection fraction	EF_{RVC}	yes
LVC ejection fraction	EF_{LVC}	no
Ratio between RVC and LV volumes at ED	$R_{RVCLV,ED}$	yes
Ratio between LVM and LVC volumes at ED	$R_{LVMLVC,ED}$	yes
Maximal LVM thickness in all the slices at ED	$MT_{LVM,ED}$	yes
Radius motion disparity	RMD	yes
Thickness motion disparity	TMD	yes

volumes of LVC, LVM and RVC at both ED and ES can be computed directly, as can the thickness of LVM. Finally, 7 shape-related features are generated (the first 7 terms in Table 1).

The second part of the method extracts 2 motion-characteristic features. Using a neural network which outputs apparent flow maps given image pairs, we get a series of apparent flow maps characterizing the in-plane motion for each MRI slice of each case. Combined with the LVM segmentation mask obtained as described above, the motion of each myocardium pixel is hence available. Eventually, 2 features are computed to present the disparity of the radial myocardial motion and the myocardial thickening respectively (the last 2 rows in Table 1).

In total, from the images of each case, 9 features characterizing the shape and the motion of the heart are extracted.

3.2. Feature Selection

As shown in Zheng et al. (2018a), these extracted features can be used for cardiac pathology classification in the ACDC dataset with performances comparable to the state-of-the-art. However, these features are not necessarily independent. Some might be redundant if there are highly correlated feature pairs. In cluster analysis, if too many variables are used simultaneously, the redundant

ones serve only to create noise that harms the clustering. So it is helpful to select a sub-group of features by removing highly correlated feature pairs.

For each pair among the 9 extracted features, we compute the Pearson correlation coefficient and the maximal information coefficient (MIC) (Reshed et al. (2011)). The former measures the linear correlation between two features, while the latter measures the mutual information between features. If there is any highly correlated pair according to these measures (i.e. Pearson correlation coefficient of absolute value above 0.8, or MIC above 0.5), we will exclude one feature in this pair. The remaining features are then considered as selected.

3.3. Cluster Analysis

We perform a model selection of Gaussian mixture model using the Bayesian information criterion (BIC). Then the selected Gaussian mixture model is applied to cluster the 8 selected features.

3.3.1. Gaussian Mixture Model Selection

A Gaussian mixture model (Reynolds (2009)) is a probabilistic model which assumes that the data points are generated from a mixture of a certain number of Gaussian distributions with unknown parameters. An expectation-maximization algorithm is used to iteratively estimate its parameters from data. Then the fitted model can assign to each sample the Gaussian component it most likely belongs to.

We use the Gaussian mixture model as implemented in scikit-learn (Pedregosa et al. (2011)). It has two major parameters, the type of covariance matrix and the number of components, upon which a selection is necessary. For this purpose, we calculate the Bayesian information criterion (BIC, Wit et al. (2012)) for Gaussian mixture models with different types of covariance matrix and numbers of components. In theory, BIC recovers the true number of components approximately. We fit the Gaussian mixture models with the following types of covariance matrix:
- 'tied': all components share the same covariance matrix;
- 'diag': each component has its own diagonal covariance matrix;
- 'full': each component has its own covariance matrix.

The number of components is also varied. By looking for models with the smallest BIC scores, we wish to select the most simple model that can fit the data thereby identifying the most suitable type of covariance matrix and a range of reasonable numbers of components.

The number of components will finally be determined by examining the sizes of resulting clusters of the Gaussian mixture models. More details will be provided in the Experiments and Results section.

3.3.2. Analysis of the Resulting Clusters

The clusters generated by the selected model will be examined. In particular, we verify if the cases in any of the clusters correspond to a pathological category according to the definitions of pathologies given by the ACDC challenge.

4. Experiments and Results

4.1. Feature Extraction

With the feature extraction method introduced in the Methods section, for each of the 3822 UK Biobank cases, 9 feature values are extracted.

4.2. Feature Selection

We calculate the Pearson correlation coefficient and MIC for each pair of features among the 9 extracted features. In Figure 1, the plot of Pearson correlation coefficient versus MIC, it is clear that the absolute value of the Pearson correlation coefficient and MIC are positively correlated. There is only one point on the upper left corner of the plot representing a highly correlated pair. It corresponds to $V_{LVC,ES}$ and EF_{LVC}, which are of Pearson correlation coefficient -0.80 and MIC 0.51. The strong negative correlation between these two features is reasonable, since by definition $EF_{LVC} = 1 - V_{LVC,ES}/V_{LVC,ED}$, in which $V_{LVC,ED}$ is the LVC volume at ED. Therefore, $V_{LVC,ES}$ and EF_{LVC} appear to be redundant. Hence we exclude EF_{LVC} and select the remaining 8 features for cluster analysis (Table 1).

4.3. Cluster Analysis

4.3.1. Gaussian Mixture Model Selection

The BIC scores of the Gaussian mixture models with various types of covariance matrix and numbers of components are plotted in Figure 2. It is clear that the 'full'
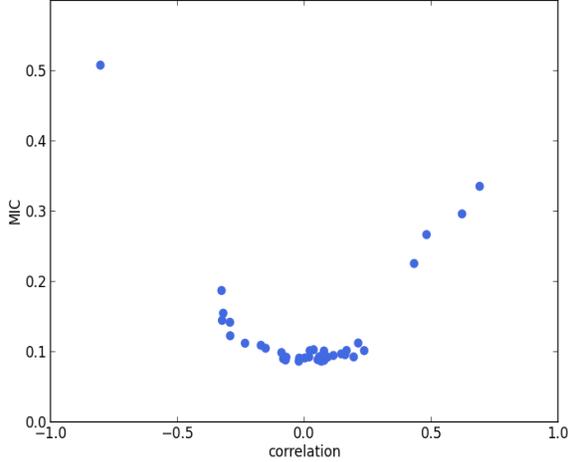
Figure 1: Pearson correlation coefficient versus MIC. Each point corresponds to a pair of features. The point in the upper-left corner corresponds to $V_{LVC,ES}$ and EF_{LVC}. The strong negative correlation between these two features is reasonable, since by definition $EF_{LVC} = 1 - V_{LVC,ES}/V_{LVC,ED}$, in which $V_{LVC,ED}$ is the LVC volume at ED.
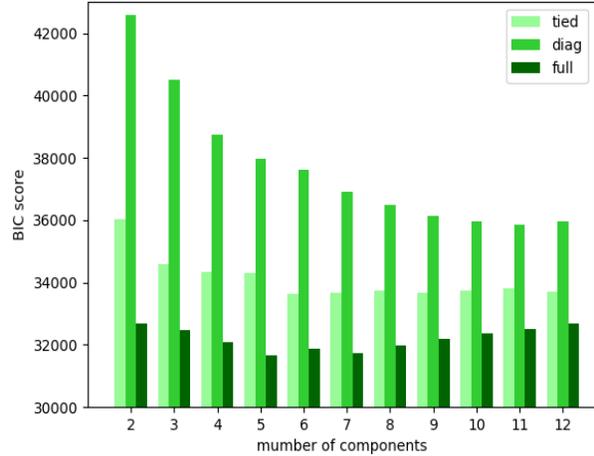

Figure 2: BIC scores of Gaussian mixture models with various types of covariance matrix and numbers of components

covariance matrix type is the best among the three. The 'full' covariance matrix type is hence selected.

And in terms of the number of components, the Gaussian mixture models with the 'full' covariance matrix type of 3 to 10 components have the smallest BIC scores. Among them, we find that:
• The models of 3 to 6 components only generate large clusters, each of which contains at least about one hundred cases.
• The models of 7 and 8 components bring about only one small cluster (less than a dozen cases).
• The models of 9 and 10 components give rise to two small clusters (less than a dozen cases).

According to the statistics[3] provided by the British Heart Foundation , about 7 million people in the UK are living with cardiovascular diseases, which is about 10.6% of the total population. More specifically, if we look at the most common cardiovascular disease categories, the per-

centages of UK population living with myocardial infarction, atrial fibrillation and heart failure are about 1.5%, 2.0% and 1.4%, respectively. This means that most of the cases in the general population do not have a cardiac pathology. Taking the 'healthy volunteer' selection bias of UK Biobank mentioned in Section 2.1 into account, the cases of cardiovascular diseases are hence probably exceedingly rare in UK Biobank. Thus, if there is any cluster that is related to a specific pathological category in an interpretable manner, its size should be small, say, no more than 76 (2% of the 3822 UK Biobank cases).

So we can now suggest that a component number of 9 or 10 is probably most suitable. We choose the model of 9 components for further analysis. But we would like to point out that the two resulting small clusters of the models of 9 and 10 components are very similar in terms of size and cases. So the results and the conclusions shown below will be roughly the same if we use the model of 10 components.

To summarize, the Gaussian mixture model with the 'full' covariance matrix type and 9 components is selected.

---

[3]https://www.bhf.org.uk/what-we-do/our-research/heart-statistics (accessed January 20, 2019)
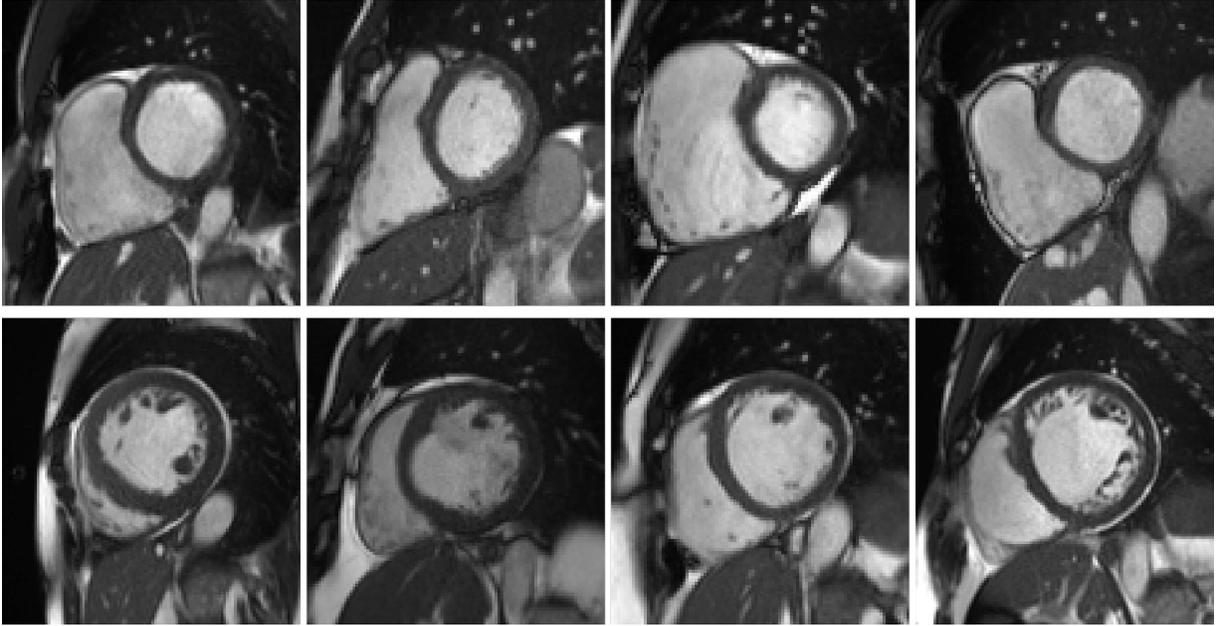
Figure 3: Examples of the cases in clusters #5 and #8. First row: example cases in cluster #5, of which the RVs appear to be exceptionally large. Second row: cases in cluster #8, of which the LVs seem to be dilated.

4.3.2. Analysis of the Resulting Clusters

Among the 9 resulting clusters (termed cluster #1 to #9) of the selected model, two are of small sizes (clusters #5 and #8). We find that they actually correspond to two pathological categories according to the definition given by the ACDC challenge (RVA and DCM respectively).

Cluster #5 has 11 cases (examples are given in Figure 3). As listed in Table 2, these cases have exceptionally large right ventricles, which are above 130 mL/m^2. In the ACDC challenge, the RVA cases are described as of RVC volumes higher than 110 mL/m^2 or RVC ejection fraction lower than 40%. Hence according to the definition of ACDC, cluster #5 is a group of cases belonging to RVA.

Cluster #8 has 4 cases (examples are given in Figure 3). As shown in Table 3, these cases have large LVC volumes at ED (above 130 mL/m^2) and low LVC ejection fractions (below 30%). In the ACDC challenge, DCM cases are those with LVC volumes larger than 100 mL/m^2 and LVC ejection fraction lower than 40%. So cluster #8 is a group of DCM cases according to ACDC. In addition, we find that the ground-truth measures of LVC volume at ED and LVC ejection fraction are available for all 4 cases in UK

Table 2: RVC volumes and ejection fraction at ED of the cases of cluster #5 based on our feature extraction method.

ID	RVC volume at ED (mL/m^2)	RVC ejection fraction
2512949	133.13	63.61%
2628396	175.77	43.91%
3423847	140.50	65.24%
3713328	169.65	71.59%
3874816	183.96	56.22%
4366978	134.68	52.53%
4681487	139.82	54.39%
4710306	144.86	29.69%
5101726	145.93	43.82%
5319688	151.30	51.93%
5561149	180.48	41.88%

Table 3: LVC volumes at ED and ejection fraction of the cases of cluster #8 based on our feature extraction method (the 2nd and 3rd columns). The same measures provided by the UK Biobank dataset are also shown (the 4th and 5th columns). The two sets of measures are quite close to each other.

ID	LVC volume at ED (mL/m^2)	LVC ejection fraction	Ground-truth LVC volume at ED (mL/m^2)	Ground-truth LVC ejection fraction
2432774	189.28	19.74%	208.24	20%
3378112	213.28	18.75%	213.03	15%
4879002	133.09	27.03%	144.59	29%
5618713	192.87	26.74%	192.43	27%

Biobank (last two columns in Table 3). It is straightforward to see in Table 3 that the measures generated by our feature extraction method are quite close to the ground-truth.

For the other 7 clusters, which are of much larger sizes (above 70), we do not identify any clear correspondence between them and the pathological categories defined in the ACDC challenge.

4.4. Further Analysis for Confirmation

To further confirm the discovered correspondence between the two small clusters and the two pathological categories, as well as to verify whether the large clusters represent normal cases, in addition to manual verification of the segmentation masks and apparent flow maps to ensure the exactness of the features, we also conduct the following analysis.

4.4.1. Interpretation of the Results of an ACDC Classification Model

We apply a pathology classification model (Zheng et al. (2018a)) trained using the ACDC dataset on the cases of clusters #5 and #8.

Seven of the eleven cases of cluster #5 are predicted to be RVA, which is as expected. However, the other 4 cases (2512949, 3423847, 4681487 and 5319688) are predicted to be NOR (i.e. normal). We suggest that this is partially due to the difference in the distributions of RVC ejection fraction. In ACDC, a great majority of the RVA cases are of RVC ejection fraction well below 50%. So the trained model has learned to rely on this feature to determine RVA cases. Yet in UK Biobank, some RVA cases, including the 4 listed above, are of RVC ejection fraction above 50%. They are not as severe cases as in ACDC.

All four cases of cluster #8 are predicted to be DCM by the classification model, which supports the correspondence between cluster #8 and DCM. In addition, by manually checking the motion, we can confirm areas of hypokinesia and akinesia for these cases but also dyskinesia for one case (3378112). For case ID 2432774, we also observe discoordinate movement of the LV myocardium suggestive of bundle branch block, which is a type of electrical conduction disease commonly associated with structural heart disease and heart failure. These observations suggest that these cases might also have some relation to MINF. In fact, as pointed out in the ACDC challenge, the increase of LVC volume can be a consequence of the adaptation of LV due to MINF (also called cardiac remodeling).

4.4.2. Reduced Dimensionality Visualization Using Principal Component Analysis

To better visualize the two isolated clusters (#5 and #8), we perform a principal component analysis to reduce the dimensionality of the 3822 vectors of size 8 (8 selected features of 3822 cases) of UK Biobank to 2. Furthermore, the centers of the 9 clusters are also projected to the sample space of the 2 principal components. As can be seen in Figure 4, the points corresponding to the cases of clusters #5 and #8, as well as the centers of the two clusters, are indeed located far away from most of the other points. This supports the suggestion that the cases in clusters #5 and #8, which are pathological, are quite different from most of the cases in the general population.

4.4.3. Visualization using t-SNE

Similarly, another tool to visualize high-dimensional data called t-SNE (t-distributed stochastic neighbor em-
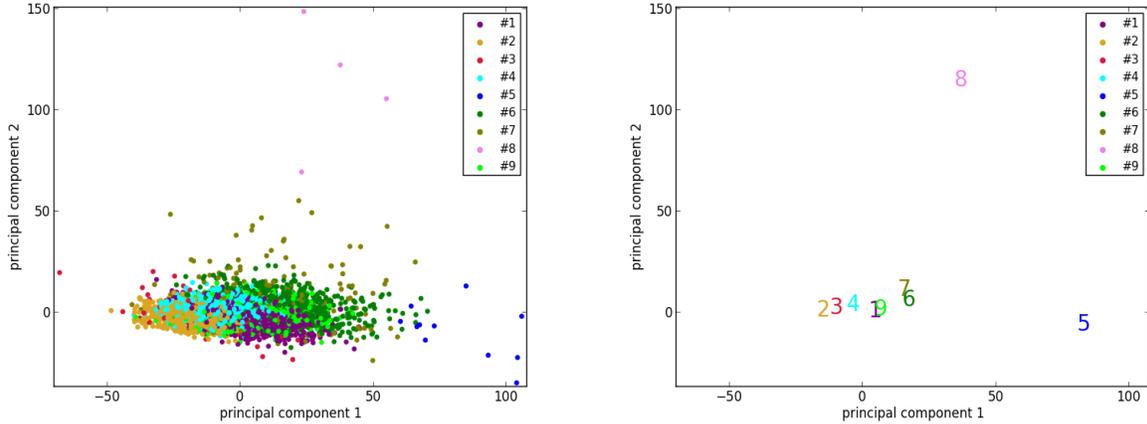
Figure 4: The results of dimensionality reduction by principal component analysis. (Left) The data points of the 3822 UK Biobank cases projected to the space of the 2 principal components. Each data point is colored according to its cluster. (Right) Projection of the centers (marked by the corresponding indexes and colors) of the 9 clusters to the same space.

bedding, van der Maaten and Hinton (2008)) is applied. Its main advantage is the ability to preserve local structure. So roughly speaking, points which are close to one another in the high-dimensional space will still be close to one another after the dimensionality reduction. t-SNE is applied to the set of the 3822 vectors of the UK Biobank cases, as well as to the set of 3831 vectors which consists of the 3822 UK Biobank cases and the 9 cluster centers. Before applying t-SNE, a normalization is performed for each feature of the original data. The purpose is to make sure that each feature is on the same scale and hence has the same importance in t-SNE. As shown in Figure 5, the points of the cases and the centers of clusters #5 and #8 are at the edge of the ensemble of points in the embedding space. This phenomenon is again consistent with the suggestion that clusters #5 and #8 correspond to pathological cases which are rather different from the other cases in the general population.

4.4.4. Examination of the Two Largest Clusters

As pointed out previously, while the pathological categories of clusters #5 and #8 are identifiable, we do not see how the other seven large clusters correspond to any cardiac pathology. In particular, the largest clusters which are of several hundreds or even more cases probably represent groups of normal cases. To verify this, we further examine the two largest clusters (#1 and #4, 889 and 1075 cases, respectively).

We plot the histograms of their ventricle volumes and ejection fractions, as well as their maximal myocardial thicknesses (Figure 6). The distributions of #1 and #4 look pretty similar in terms of LVC volume and LVC ejection fraction. But they are different on RVC volume, RVC ejection fraction and maximal myocardial thickness. On average, the cases of #4 have larger RVCs with higher ejection fractions. And their myocardiums also tend to be thicker than that of the cases of #1. Furthermore, we perform the unpaired unequal variance t-test to prove that the corresponding means of the distributions of #1 and #4 are different. Under the null hypotheses that the corresponding distributions have the same mean, the p-values for LVC volume, LVC ejection fraction, RVC volume, RVC ejection fraction and maximal myocardial thickness are all much below 0.05 (lower than 10^{-7}), which are small enough to reject the null hypotheses. This means that clusters #1 and #4 actually exhibit significant different values of the 5 features (LVC volume at ED, LVC ejection fraction, RVC volume at ED, RVC ejection fraction and maximal myocardial thickness).

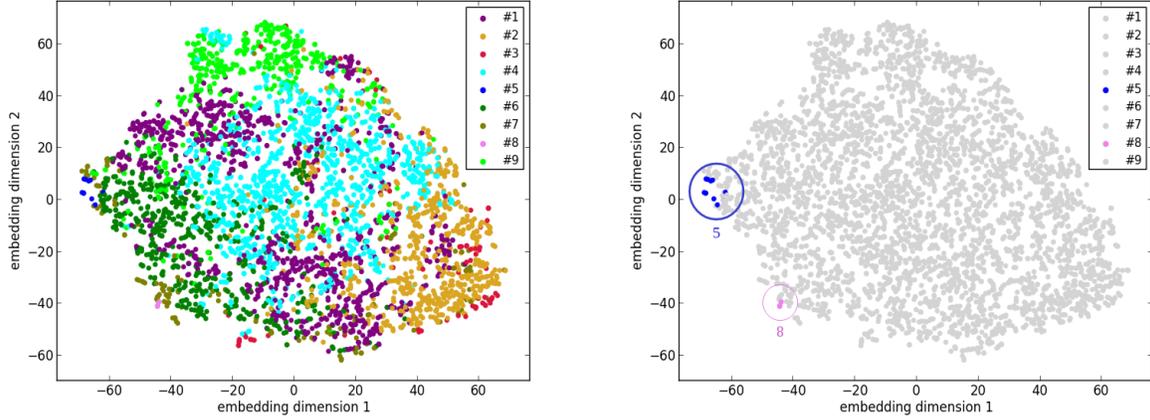For both clusters, at least a great majority of the cases

Figure 5: The results of dimensionality reduction by t-SNE. (Left) The data points of the 3822 UK Biobank cases in the space of the 2 embedding dimensions after t-SNE. Each data point is colored according to its cluster. (Right) A plot similar to the left one with only differences on coloring. Only the points of clusters #5 and #8 are highlighted with colors and circles.

satisfy:

• LVC volumes at ED less than 100 mL/m^2

• LVC ejection fraction above 40%

• RVC volumes at ED less than 110 mL/m^2

• RVC ejection fraction above 40%

• Maximal myocardial thickness less than 15 mm

Hence according to the definitions in ACDC, these two clusters do not correspond to any of the 4 pathological categories (DCM, HCM, MINF, RVA).

4.4.5. Examination of the Seven Large Clusters

To further understand the seven large clusters, we first systematically perform the unpaired unequal variance t-test. For each pair of clusters in the seven large clusters, and for each of the 8 extracted features, under the null hypothesis that the distributions of the feature has the same mean for both clusters, the p-value is computed. In this way 21×8=168 p-values are obtained. In total, 149 p-values among them are below 0.05, which are small enough to reject the corresponding null hypotheses. This confirms that the clusters have different distributions on the features. Nineteen p-values among them are above 0.05, which signify a kind of similarity between pairs of clusters (Table 4). Similarly, we perform the unpaired two-sided Mann-Whitney rank tests, under the null hy-

potheses that the corresponding distributions of the features are the same for both clusters. And we find again that a great majority (147) of the p-values are below 0.05 such that the corresponding null hypotheses can be rejected.

4.4.6. Measures by the Automatic Pipeline versus the Ground-Truth

As mentioned previously, for part of the UK Biobank cases, the ground-truth measures given by the InlineVF analysis algorithm of LVC volumes at ED and ES and LVC ejection fraction are available. In particular, among the 3822 cases used in this paper, we have access to all of the three ground-truth measures for 3212 cases. The comparison between the means and standard deviations of the measures generated by the automatic pipeline used in this paper and the ground-truth measures are shown in Table 5. It is clear that the ground-truth measures of the volumes are higher and of larger standard deviations than those estimated by the automatic pipeline.

To better understand the cause of these differences, we plot the points of the measures in Figure 7. We can see that the ground-truth values contain some obvious outliers, which are often of values well above the realistic range of LVC volumes. This explains the
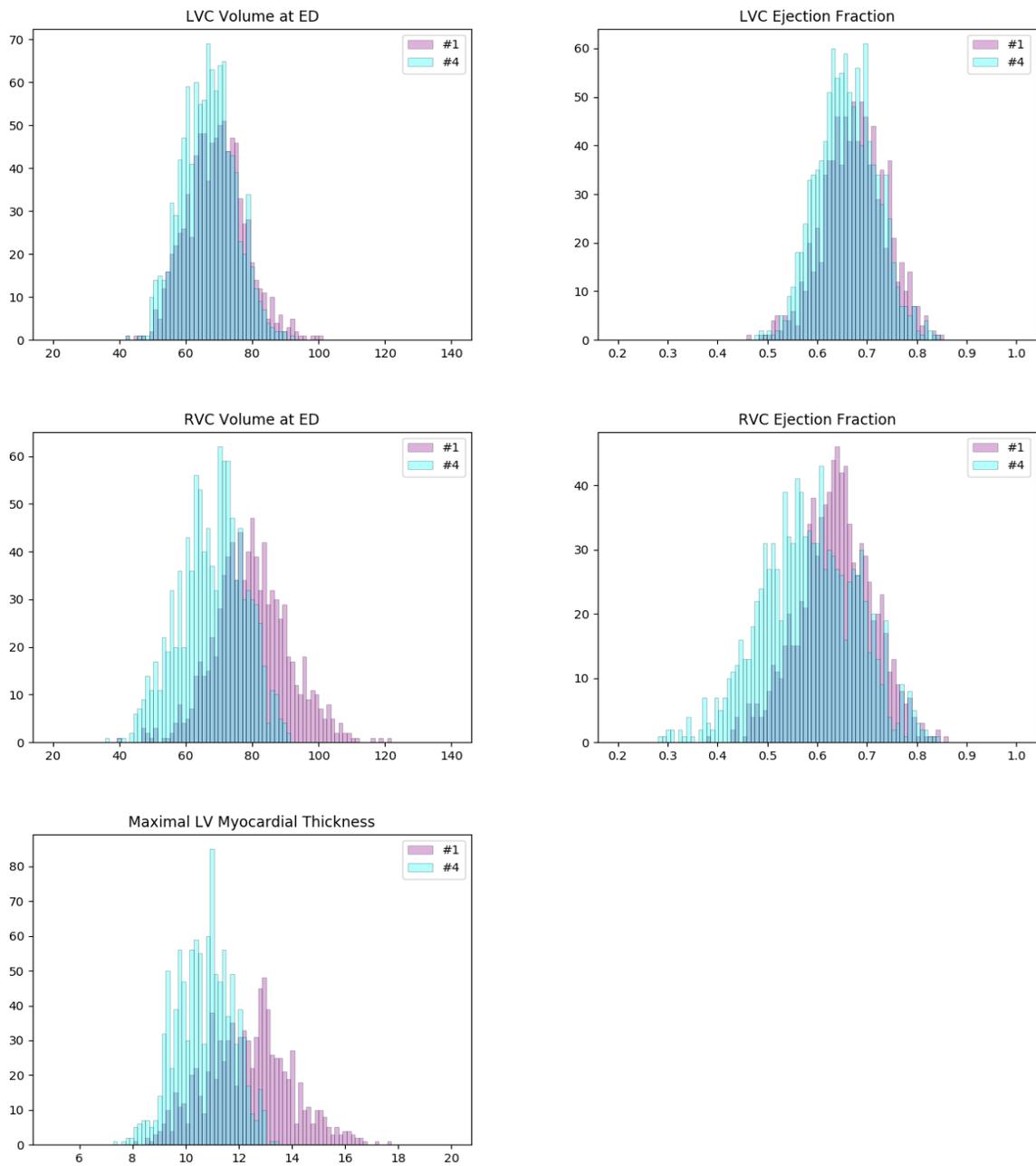
Figure 6: Histograms of some important measures of the cases in clusters #1 (pink) and #4 (cyan). The colors of the columns are set to be partially transparent such that their overlaps appear to be of color dark blue. The distributions of #1 and #4 are pretty similar in terms of LVC volume and LVC ejection fraction (1st row). But they are different on RVC volume, RVC ejection fraction and maximal myocardial thickness (2nd and 3rd rows). On average, the cases of #1 have larger RVCs with higher ejection fractions. And their myocardiums also tend to be thicker than that of the cases of #4. For both clusters, the measures are well in normal ranges according to the definitions given by ACDC.

Table 4: The large p-values of the unpaired unequal variance t-tests for the 21 pairs of clusters in the seven large clusters, and for the 8 extracted features, under the null hypothesis that the distributions of the feature has the same mean for both clusters. For most of the cluster pairs and features, the p-values are below 0.05

cluster pair	p-values above 0.05 (and the corresponding features)
(#1, #4)	0.07 ($V_{LVC,ES}$)
(#1, #6)	0.56 (RMD), 0.05 (TMD)
(#1, #9)	0.55 ($V_{RVC,ED}$), 0.76 ($R_{RVCLV,ED}$)
(#2, #3)	0.17 ($R_{RVCLV,ED}$), 0.80 ($R_{LVMLVC,ED}$)
(#2, #4)	0.31 (TMD)
(#2, #7)	0.85 ($R_{RVCLV,ED}$), 0.76 (RMD)
(#3, #4)	0.29 ($R_{RVCLV,ED}$)
(#3, #6)	0.12 (EF_{RVC})
(#3, #7)	0.07 (EF_{RVC}), 0.28 ($R_{RVCLV,ED}$), 0.61 ($MT_{LVM,ED}$), 0.25 (TMD)
(#4, #6)	0.70 ($R_{LVMLVC,ED}$), 0.14 (TMD)
(#6, #7)	0.27 (EF_{RVC})

Table 5: The means and standard deviations of the measures (in mL/m^2) by the automatic pipeline versus the ground-truth.

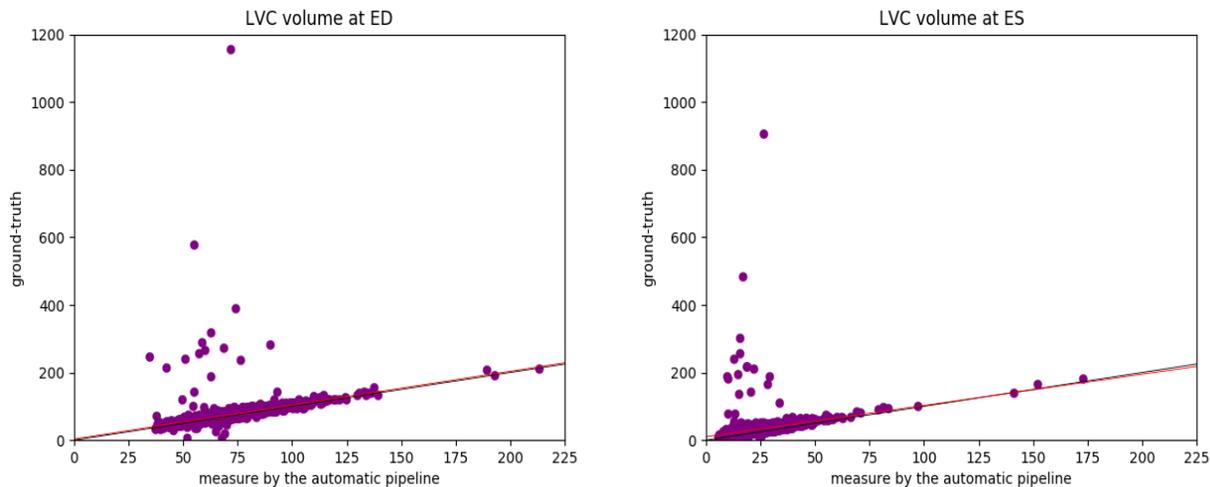|	Automatic pipeline	Ground-truth
LVC volume at ED (mL/m^2)	70.56 (13.91)	75.48 (28.62)
LVC volume at ES (mL/m^2)	24.06 (9.02)	33.87 (22.82)
LVC ejection fraction	66.41% (7.33%)	56.04% (6.53%)

Figure 7: The plots of the measures (in mL/m^2) generated by the automatic pipeline against the ground-truth for the LVC volume at ED (left) and at ES (right). We can see that the ground-truth values contain some obvious outliers, which are often of values well above the realistic range of LVC volumes. This explains the fact that the ground-truth volumes have higher means and larger standard deviations than those estimated by the automatic pipeline. The lines corresponding to the robust linear regression models (red) and the lines corresponding to *ground-truth=automatic-pipeline* (black) are also plotted. The red line and the black line almost overlap with each other.

fact that the ground-truth volumes have higher means and larger standard deviations than those estimated by the automatic pipeline. Moreover, proportionally, the mean of the ground-truth values of LVC volume at ED is 7.0% (= 75.48/70.56 - 1) above that of the estimates by the automatic pipeline, while for LVC volume at ES the ground-truth is on average 40.8% (= 33.87/24.06 -1) higher than the values obtained via the automatic pipeline. This also explains why the ground-truth of LVC ejection fraction is on average lower than that given by the automatic pipeline. The models obtained by the robust linear regression using Huber's criterion for LVC volume at ED and ES are *ground-truth=1.002×automatic-pipeline+3.373* and *ground-truth=0.923×automatic-pipeline+10.303*, respectively. The lines corresponding to the robust linear regression models (red) and the lines corresponding to *ground-truth=automatic-pipeline* (black) are plotted in Figure 7. On both graphs in Figure 7, the red line and the black line almost overlap with each other. This means that our regression lines are near the lines of identity, which signifies a similarity between the measures by our method and

those based on the InlineVF algorithm. By comparing the regression lines and identity lines in Fig. 4 of Suinesiaputra et al. (2018), we can also conclude a similarity between the measures derived from manual segmentation and those based on the InlineVF algorithm. Hence our method actually generates measures which are close to both manual and InlineVF values.

We believe that the differences between the measures by the automatic pipeline used in this paper and the ground-truth are partially due to the lack of quality control on the ground-truth. In fact, as pointed out in Suinesiaputra et al. (2018), the ground-truth is generated by the InlineVF algorithm, which may fail and hence make unreliable predictions on some cases. Without quality control, these failures causes the outliers in Figure 7.

5. Conclusion and Discussion

In this paper, we proposed a method of unsupervised cluster analysis on a large unlabeled dataset (UK Biobank) of the general population to identify pathological cases based on shape-related and motion-characteristic features extracted from cardiac cine MRI

images. As far as we know, this is a topic that has rarely been studied before. In our cluster analysis, a Gaussian mixture model is applied to cluster similar cases together without supervision. As a result, among the generated clusters, we identify two that probably correspond to two cardiac pathological categories. This idea is further supported by the observations on the results of a trained classification model and of the dimensionality reduction tools including principal component analysis and t-SNE.

As more and more large and unlabeled datasets are available in the community, researchers will be able to extract interesting information by data mining. Identification of cardiac pathology is just one among other topics such as the analysis of motion patterns, the relationship between motion and shape features, etc. In the future, more research may be carried out by including more data and different types of data (Kohli et al. (2017)), using more features, targeting other abnormalities or phenotype properties, etc. Various unsupervised learning methods (Raza and Singh (2018)) other than a Gaussian mixture model can also be applied.

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

References

Attar, R., Pereanez, M., Gooya, A., Alba, X., Zhang, L., Piechnik, S., Neubauer, S., Petersen, S., Frangi, A., 2019. High throughput computation of reference ranges of biventricular cardiac function on the UK Biobank population cohort. arXiv preprint arXiv:1901.03326 .

Barillot, C., Edan, G., Commowick, O., 2016. Imaging biomarkers in multiple sclerosis: from image analysis to population imaging. Medical Image Analysis 33, 134–139.

de Bruijne, M., 2016. Machine learning approaches in medical image analysis: from detection to diagnosis. Medical Image Analysis 33, 94–97.

Cetin, I., Sanroma, G., Petersen, S., Napel, S., Camara, O., Ballester, M., Lekadir, K., 2017. A radiomics approach to computer-aided diagnosis with cardiac cine-MRI, in: Proc. Statistical Atlases and Computational Models of the Heart (STACOM), ACDC challenge, MICCAI'17 Workshop.

Fry, A., Littlejohns, T., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., Allen, N., 2017. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. Am J Epidemiol. 186(9), 1026–1034.

Isensee, F., Jaeger, P., Full, P., Wolf, I., Engelhardt, S., Maier-Hein, K., 2017. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features, in: Proc. Statistical Atlases and Computational Models of the Heart (STACOM), ACDC challenge, MICCAI'17 Workshop.

Jolly, M., Guetter, C., Lu, X., Xue, H., Guehring, J., 2013. Automatic segmentation of the myocardium in cine MR images using deformable registration. Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges , 98–108.

Kawadiwale, R., Rane, M., 2014. Clustering techniques for brain tumor detection. Proc. of Int. Conf. on Recent Trends in Information , 299–305.

Khened, M., Alex, V., Krishnamurthi, G., 2017. Densely connected fully convolutional network for short-axis cardiac cine MR image segmentation and heart diagnosis using random forest, in: Proc. Statistical Atlases and Computational Models of the Heart (STACOM), ACDC challenge, MICCAI'17 Workshop.

Khened, M., Alex, V., Krishnamurthi, G., 2018. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. arXiv preprint arXiv:1801.05173 .

Kinani, J., Silva, A., Funes, F., Vargas, D., Diaz, E., Arellano, A., 2017. Medical imaging lesion detection based on unified gravitational fuzzy clustering. Journal of Healthcare Engineering 2017.

Kohli, M., Summers, R., Geis, J., 2017. Medical image data and datasets in the era of machine learning-whitepaper from the 2016 C-MIMI meeting dataset session. Journal of Digital Imaging 30, 392–399.

Komura, D., Ishikawa, S., 2018. Machine learning methods for histopathological image analysis. Computational and Structural Biotechnology 16, 34–42.

Liu, M., Zhang, J., Nie, D., Yap, P., Shen, D., 2018. Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. IEEE Journal of Biomedical and Health Informatics 22, 1476–1485.

Lu, X., Georgescu, B., Jolly, M., Guehring, J., Young, A., Cowan, B., Littmann, A., Comaniciu, D., 2010. Cardiac anchoring in MRI through context modeling. Med Image Comput Comput Assist Interv 13(1), 383–390.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. J. Mach. Learn. Research 9, 2579–2605.

Madabhushi, A., Lee, G., 2017. Image analysis and machine learning in digital pathology: Challenges and opportunities. Med Image Anal 33, 170–175.

Moldovanu, S., Obreja, C., Moraru, L., 2015. Threshold selection for classification of MR brain images by clustering method. AIP Conference Proceedings 2015 1694.

Moriya, T., Roth, H., Nakamura, S., Oda, H., Nagara, K., Oda, M., Mori, K., 2018. Unsupervised segmentation of 3D medical images based on clustering and deep representation learning. Proceedings of the SPIE 10578.

Parisot, S., Ktena, S., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., Rueckert, D., 2018. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. Medical Image Analysis 48, 117–130.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Petersen, S., Aung, N., Sanghvi, M., Zemrak, F., Fung, K., Paiva, J., Francis, J., Khanji, M., Lukaschuk, E., Lee, A., Carapella, V., Kim, Y., Leeson, P., Piechnik, S., Neubauer, S., 2017. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. J Cardiovasc Magn Reson 19(1), 18+.

Petersen, S., Matthews, P., Francis, J., Robson, M., Zemrak, F., Boubertakh, R., Young, A., Hudson, S., Weale, P., Garratt, S., Collins, R., Piechnik, S., Neubauer, S., 2016. UK Biobank's cardiovascular magnetic resonance protocol. J Cardiovasc Magn Reson 18:8, 8+.

Raza, K., Singh, N., 2018. A tour of unsupervised deep learning for medical image analysis. arXiv preprint arXiv:1812.07715 .

Reshed, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M., Sabeti, P., 2011. Detecting novel associations in large data sets. Science 334, 1518–1524.

Reynolds, D., 2009. Gaussian mixture models. Encyclopedia of Biometrics , 659–663.

Rueckert, D., Glocker, B., Kainz, B., 2016. Learning clinically useful information from images: past, present and future. Medical Image Analysis 33, 13–18.

Suinesiaputra, A., McCulloch, A., Nash, M., Pontre, B., Young, A., 2016. Cardiac image modelling: breadth and depth in heart disease. Medical Image Analysis 33, 38–43.

Suinesiaputra, A., Sanghvi, M., Aung, N., Paiva, J., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A., Carapella, V., Kim, Y., Francis, J., Piechnik, S., Neubauer, S., Greiser, A., Jolly, M., Hayes, C., Young, A., Petersen, S., 2018. Fully-automated left ventricular mass and volume MRI analysis in the UK Biobank population cohort: evaluation of initial results. Int J Cardiovasc Imaging 34(2), 281–291.

Toga, A., Crawford, K., 2015. The Alzheimer's disease neuroimaging initiative informatics core: a decade in review. Alzheimers Dement. 11, 832–839.

Weese, J., Lorenz, C., 2016. Four challenges in medical image analysis from an industrial perspective. Medical Image Analysis 33, 44–49.

Wit, E., van den Heuvel, E., Romeijn, J., 2012. 'All models are wrong ...': an introduction to model uncertainty. Statistica Neerlandica 66, 217–236.

Wolterink, J., Leiner, T., Viergever, M., Isgum, I., 2017. Automatic segmentation and disease classification using cardiac cine MR images, in: Proc. Statistical Atlases and Computational Models of the Heart (STACOM), ACDC challenge, MICCAI'17 Workshop.

Zhang, S., Metaxas, D., 2016. Large-scale medical image analytics: Recent methodologies, applications and future directions. Medical Image Analysis 33, 98–101.

Zheng, Q., Delingette, H., Ayache, N., 2018a. Explainable cardiac pathology classification on cine mri with motion characterization by semi-supervised learning of apparent flow. arXiv preprint arXiv:1811.03433 .

Zheng, Q., Delingette, H., Duchateau, N., Ayache, N., 2018b. 3D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. IEEE Trans Med Imaging 37(9), 2137–2148.

Zheng, Q., Delingette, H., Duchateau, N., Ayache, N., 2018c. 3D consistent biventricular myocardial segmentation using deep learning for mesh generation. arXiv preprint arXiv:1803.11080 .