Automated Training Set Generation for Aortic Valve Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

Affecting 1% of the population, bicuspid aortic valve (BAV) is the most prevalent anatomical malformation of the heart. Currently, the limited availability of labeled data hinders the development of automated detection methods. This paper presents a new method for efficiently generating training labels for the BAV classification task. We first define heuristic rules based on geometric features of phase-contrast MRI images to assign labels to the images, albeit noisily. We then define a factor graph based generative model to learn the accuracies and dependencies of the heuristics. Finally, we use our learned parameters to optimally combine the noisy labels from the heuristics into probabilistic training labels for the cardiac MRI dataset. We demonstrate how our model improves over majority vote by 0.0268 points AUC and by 18.24% accuracy.

1 Introduction

Bicuspid aortic valve (BAV) is a highly prevalent malformation of the aortic valve that occurs in 1-2% of the population, where two leaflets of the aortic valve are present instead of the normal three. BAV has a wide variety of symptoms and presentations, sometimes requiring surgery at the time of birth or going undiagnosed into middle or late adulthood [Roberts and Ko, 2005]. UK Biobank (UKBB) released a public dataset of 100,000 adult participants [Allen et al., 2014] and their associated cardiac MRI sequences [Petersen et al., 2013]. The availability of such a dataset carries the potential for conducting a variety of genetic and epidemiological studies; however, the first step is to classify each image as being normal or BAV.

Although various machine learning approaches have been used for automated image classification, these methods require large magnitudes of labeled training data to achieve state-of-the-art performance. For medical datasets, the cost of hand-labeling by certified physicians is significantly higher than that of contracted human-intelligence workers such as those available via the Amazon Mechanical Turk service. For example, we only have 112 labeled videos for our dataset, which were hand labeled by collaborating cardiologists. Moreover, only 12 videos (10.7%) of these depict BAV valves, leading to a very small subset to learn from. Therefore, there is a need for an efficient approach to labeling large magnitudes of training data that can feed the data-hungry machine learning models.

In our approach, we employ weak supervision, which relies on high level knowledge such as knowledge bases and domain expertise, to label data efficiently, albeit noisily [Mintz et al., 2009, Bunescu and Mooney, 2007, Craven et al., 1999]. For our dataset, we use Coral [Varma et al., 2017] , a weak supervision paradigm that relies on user-defined heuristic rules to imperfectly label data to address the issue of limited labels and remove the necessity for cardiologists to hand-label additional data. To develop these heuristics, we first extract geometric features of the valve by preprocessing the phase-contrast cardiac MRI images. We then develop heuristics that take these features as input and use simple if-then rules to assign labels to the MRI data. Even when developing heuristics, we only

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

rely on domain-expertise and feature value histograms and do not use any ground truth labels. We use Coral's underlying generative model to learn the accuracies and dependencies for these heuristics and assign probabilistic training labels to the data. Finally, we validate our weak supervision approach for generating training labels by evaluating our labels against the ground truth labels provided by cardiologists, obtaining an accuracy of 85.28% and AUC of 0.7376.

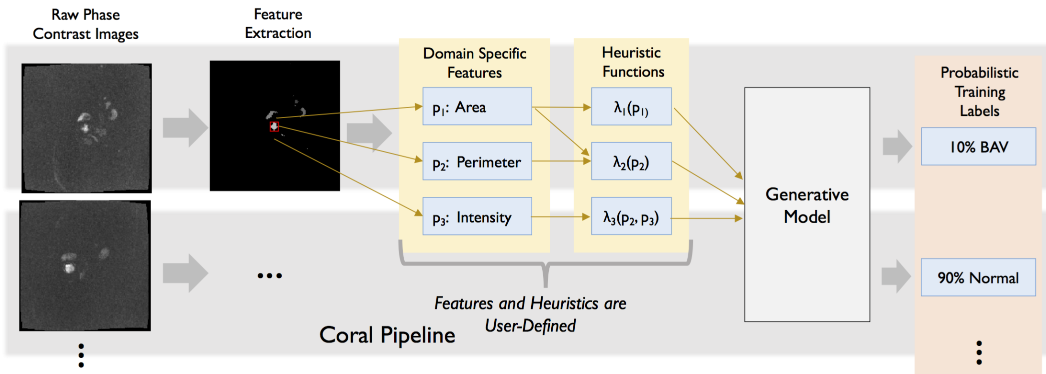


Figure 1: High-level workflow for probabilistic training label generation based on user-defined features and heuristics.

2 Methodology

We analyze phase-contrast images from the UKBB heart MRI dataset, which consists of 112 videos of the heart during the cardiac cycle. These phase-contrast videos for blood flow were captured from the short axis plane oriented to the aortic annulus. Each video consists of 30 frames that are 192×192 pixels. Since the images target the aortic valve, which is the point of concentrated blood flow during the cardiac cycle, they capture the brightest portion of the image. We can exploit this phenomenon to easily extract geometric features of the heart valve. We selected the six brightest frames from each 30-frame series for analysis, resulting in 600 frames of healthy patients and 72 frames of patients with BAV.

2.1 Preprocessing

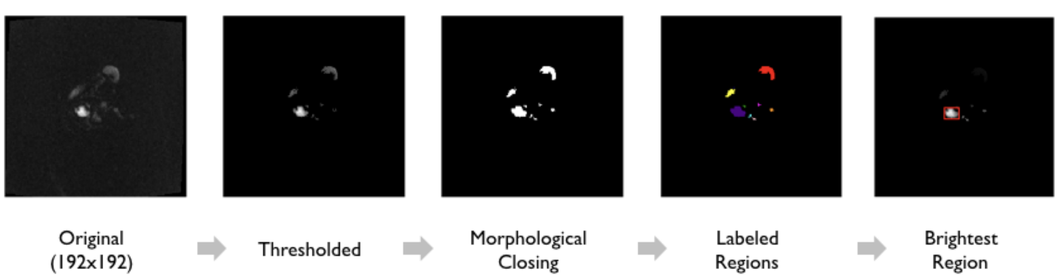


Figure 2: Pipeline from raw phase-contrast images to regions ready for feature extraction.

Since we did not have access to ground truth segmentations outlining the valve area, we experimented with several thresholding techniques to isolate regions of interest and denoise the image background. Computing the mean brightness of each image was a helpful way to approximate the general range of threshold values that could be manually set. We experimentally found an intensity threshold used to generate binary masks of each image by optimizing for thresholds that removed background noise while maintaining the integrity of the aortic valve's shape. When each of these masks were applied to the original images, the background was removed and only the primary regions of interest remained. Next, we computed the Otsu threshold to apply a morphological closing to each thresholded image, which fills holes such that each region is treated as a discrete geometric shape for feature extraction

Table 1: Feature values for target regions of patient images with ground truth labels.

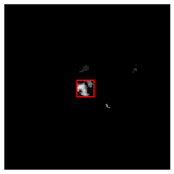
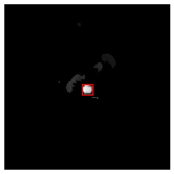
MRI	Classification	Area	Perimeter	Eccentricity	Intensity
	BAV	152	61.31	0.7861	79.76
	Normal	112	38.14	0.32	125.82

Table 2: Heuristic Function Evaluation Results

	Statistics					
Heuristic Function	Coverage	AUC	Accuracy	F1 score	Recall	Precision
`HF_Area`	0.4479	0.6325	0.8040	0.4158	0.3443	0.5250
`HF_Perimeter`	0.6235	0.6494	0.8186	0.4571	0.3478	0.6667
`HF_Eccentricity`	0.6190	0.5593	0.7380	0.2781	0.2019	0.4468
`HF_Intensity`	0.5446	0.5278	0.5574	0.2286	0.1420	0.5854

[Van der Walt et al., 2014]. Finally, we selected the region with the highest intensity as our region of interest, representing the heart valve.

2.2 Heuristic Generation

Heuristic functions (HFs) map from features to potential labels for each image in the training set. These user-defined HFs are composed of nested if-then statements that determine whether features fall above or below user-set thresholds. We collaborate with cardiologists and use histograms of feature values to develop heuristic functions without explicitly using ground truth labels.

Physiologically, we expected the area and perimeter of BAV images to be smaller than those of normal images. However, after our preprocessing steps, we noticed that it was not uncommon for the region labeling to overestimate the area of the aortic valve, as seen in 1. We suspect that this is a result of the irregularity in the shape of BAVs. Based on physiological intuition, we also expected eccentricity values to be greater for BAV, also reflected in 1. Finally, we expected intensity to be greater for the BAV images because a smaller valve typically leads to more blood flow. This is not reflected in the example provided 1, which highlights the challenge of using a single threshold across images without normalized intensity values. We provide statistics of these HFs in Table 1, evaluated on the 672 images we had access to ground truth labels for.

2.3 Weak Supervision

We use the Coral paradigm [Varma et al., 2017] to learn dependencies and accuracies of the HFs in order to generate training labels. Coral infers dependencies by performing static analysis over the source code and uses a factor graph to encode the relationships between HFs, features, and class labels. Coral uses this model to optimally combine noisy labels from the HFs and assign probabilistic labels to the data.

3 Experimental Evaluation

In order to evaluate the efficacy of our weakly-supervised approach, we compare the probabilistic training labels from our generative model to labels from majority vote, which does not take the

Table 3: Label Generation Evaluation Results

Method	Statistics					
	Coverage	AUC	Accuracy	F1 score	Recall	Precision
Majority Vote	0.9300	0.7108	0.6704	0.3125	0.4348	0.2439
Generative Model		**0.7376**	**0.8528**	**0.3947**	0.4348	**0.3614**

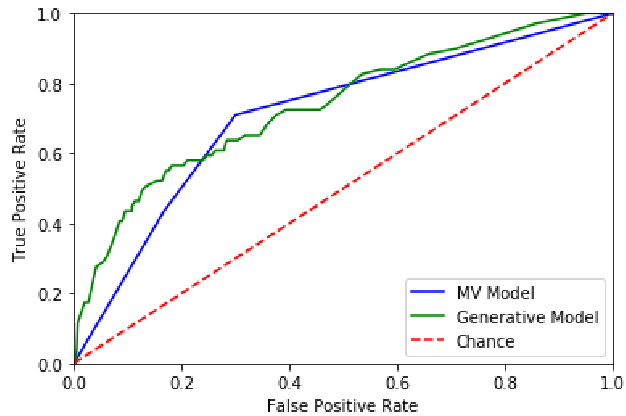


Figure 3: ROC curves for majority vote (MV) and generative models.

different accuracies and dependencies of the heuristics into account. In our evaluation, we prioritize F1 score because it captures the trade-off between precision and recall, both important metrics for our task. For the purposes of evaluating our training labels, we define a marginal threshold (`thresh`) to convert our probabilistic labels (`prob`) into true labels (y) such that $y = \mathbb{I}[\texttt{prob} \geq \texttt{thresh}]$.

The class imbalance in our data also translates into trade-offs when considering the marginal threshold for converting probabilistic labels into true labels for evaluation of the generative model.

We experimentally quantify the effectiveness of our generated training labels by considering AUC, accuracy, and F1 score and show the resulting performance in Table 3. While both methods achieve the same recall, the generative model approach outperform majority vote on every other metric, making it well-suited for generating fairly accurate training labels without requiring data with ground truth labels.

Note that as shown in Table 3, the coverage of both methods is 93%. This translates to 7% of the images not receiving a label from any heuristic function. Since we are generating training labels, not predicting final labels, the less than complete coverage is favorable since it will prevent the end model we train from learning possibly incorrect relations.

4 Conclusion and Next Steps

We propose a pipeline to preprocess phase-contrast images targeting aortic valves and generate relevant heuristic functions. In doing so, we rely on intuition about the physiological characteristics of phase-contrast images that target aortic valves to write user-defined heuristic functions. We evaluate quality of the probabilistic training labels from the generative models to labels from simple majority vote. These results lessen the semantic gap between cardiologists' diagnostic intuitions for labeling aortic valve data and a machine's ability to automate the label generation process.

We have room to improve the output of our generative model by including additional image processing steps to more accurately label regions. In addition, we can attempt additional preprocessing to normalize feature values, allowing the absolute thresholds of our heuristics to operate more effectively. Finally, we plan to use the probabilistic labels of our generative model to train deep convolutional neural networks for the BAV classification task.

References

N. E. Allen, C. Sudlow, T. Peakman, R. Collins, et al. Uk biobank data: come and get it, 2014.

R. Bunescu and R. Mooney. Learning to extract relations from the web using minimal supervision. In *ACL*, 2007.

M. Craven, J. Kumlien, et al. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, pages 77–86, 1999.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches. *Journal of Cardiovascular Magnetic Resonance*, 15(1):46, 2013.

W. C. Roberts and J. M. Ko. Frequency by decades of unicuspid, bicuspid, and tricuspid aortic valves in adults having isolated aortic valve replacement for aortic stenosis, with or without associated aortic regurgitation. *Circulation*, 111(7):920–925, 2005.

S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

P. Varma, B. He, P. Bajaj, I. Banerjee, N. Khandwala, D. L. Rubin, and C. Ré. Inferring Generative Model Structure with Static Analysis. In *Advances in Neural Information Processing Systems*, 2017.