

# Automated Cardiovascular Pathology Assessment using Semantic Segmentation and Ensemble Learning

Tony Lindsey · Jin-Ju Lee

Received: date / Accepted: date

**Abstract** Cardiac magnetic resonance imaging provides high spatial resolution, enabling improved extraction of important functional and morphological features for cardiovascular disease staging. Segmentation of ventricular cavities and myocardium in cardiac cine sequencing provides a basis to quantify cardiac measures such as ejection fraction. A method is presented that curtails the expense and observer bias of manual cardiac evaluation by combining semantic segmentation and disease classification into a fully automatic processing pipeline. The initial processing element consists of a robust dilated convolutional neural network architecture for voxel-wise segmentation of the myocardium and ventricular cavities. The resulting comprehensive volumetric feature matrix captures diagnostic clinical procedure data and is utilized by the final processing element to model a cardiac pathology classifier. Our approach evaluated anonymized cardiac images from a training data set of 100 patients (4 pathology groups, 1 healthy group, 20 patients per group) examined at the University Hospital of Dijon. The top average Dice index scores achieved were 0.940, 0.886, 0.849 for structure segmentation of the left ventricle (LV), myocardium and right ventricle (RV) respectively. A 5-ary pathology classification accuracy of 90% was recorded on an independent test set using the trained model. Performance results demonstrate potential for advanced machine learning methods to deliver accurate, efficient and reproducible cardiac pathological assessment.

---

Tony Lindsey

Intelligent Systems, Room 250, NASA Ames Research Center, M/S N269-2, Mountain View, 94035, CA, Tel.: +1 (510) 750-0923

Biomedical Informatics, x-215, Stanford School of Medicine, 1265 Welch Road, MC:5479, Stanford, 94035, CA, Tel.: +1 (650) 723-1398

E-mail: antonia.e.lindsey@nasa.gov, nasatony@stanford.edu

Jin-Ju Lee

NASA Ames Research Center, M/S N269-2, Mountain View, 94035, CA, 94035.

Tel.: +1 (650) 604-4517

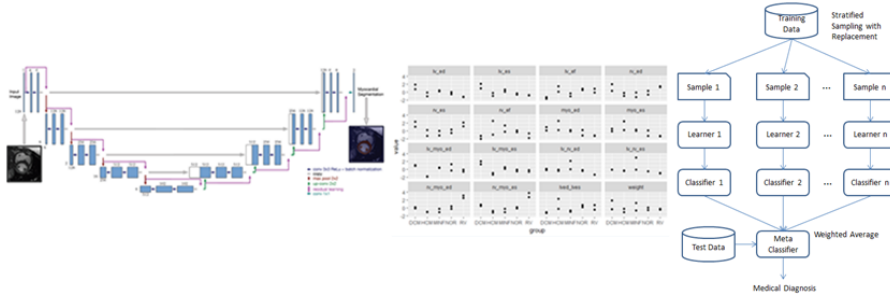
E-mail: jinjulee119@gmail.com

**Keywords** 2D U-Net · Cardiac Cine-MRI · Semantic Segmentation · Feature Selection · Classification

## 1 Introduction

A common cause for heart failure is an inability to control etiologies such as hypertension or coronary artery disease. Cardiac genome expression resulting in molecular, cellular and interstitial changes may also be a determinant in the progressive course of this pathology [1]. Clinical manifestations of ventricular remodeling include changes in heart size, mass, geometry, function and regional wall motion. Magnetic resonance imaging (MRI) is used to detect or monitor cardiac disease as well as evaluate heart anatomy and function in patients with both congenital and idiopathic cardiovascular pathologies. This modality may provide the best images for certain conditions such as cardiac amyloidosis, hemochromatosis and pericarditis.

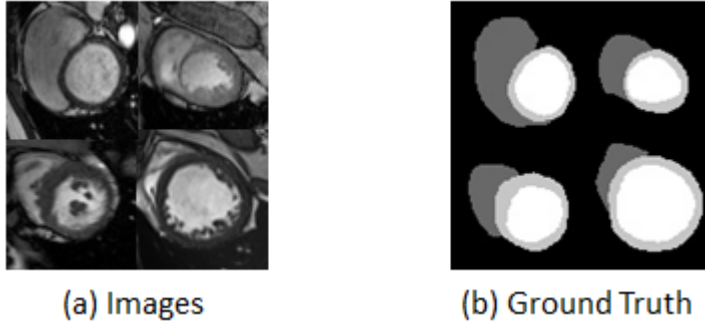
Cardiovascular MRI employs non-ionizing tomography for clinical assessment of infarcted segment expansion, late wall thinning of infarcted regions, LV volumes, distortion of LV shape and compensatory hypertrophy of non-necrotic myocardium tissue [2]. In today’s modern clinical setting the tremendous benefits of comprehensive quantitative measurements remain largely untapped due to assessment costs, interpretation variability, and lack of reproducible medical scenarios. Therefore, accurate automatic approaches that differentiate multiple coronary event features with computer-aided diagnosis are desirable assets for a large spectrum of cardiovascular diseases [3]. Computers have become progressively capable of enhancing medical imaging, morphological tissue characterization, and molecular classification with diagnostic, prognostic, and theragnostic predictions. The evidence suggests ample opportunity exists for computer-aided interpretation and multi-modality integration to provide new insights into myocardial disease mechanism, severity and prognosis [4].



**Fig. 1** Voxel-wise segmentation quantifies volumetric cardiac features used to assess pathology by an ensemble meta-classifier.

In this paper an approach for automatic classification of cardiac pathologies associated with ventricular remodeling is presented. A deep neural network al-

gorithm employed semantic segmentation and mapped each image voxel to a class label thus enabling cardiac volume dynamics quantification. The volumetric features were used to train and cross validate a classical machine learner for prediction of various cardiac diseases. More specifically, a U-Net convolutional neural network architecture [5] performed multi-structure semantic segmentation and was applied to each short-axis slice increment of cardiac MRI. The extracted domain-specific features, which were motivated by a cardiologist’s workflow, trained a weighted average ensemble classifier for disease prediction see Figure 1. Classification of five categories (normal case, heart failure with infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle) was performed at each cine-MRI time step see Figure 2. The cardiac health interpretation methods were evaluated as part of an international automated cardiac diagnosis challenge [6].



**Fig. 2** Representative end diastolic cardiac images with masks for various pathologies, from top left to bottom right: abnormal right ventricle, myocardial infarction, hypertrophic cardiomyopathy, dilated cardiomyopathy.

## 2 Methods

A modified 2D U-Net was implemented such that both transpose convolution feature maps and class labels were equal in number. The Dice similarity coefficient (DSC) was optimized to produce stable gradients by calculating a soft Dice loss on the Softmax output:

$$Dice_c = \frac{\sum_i^N R_c(i) A_c(i)}{\sum_i^N R_c(i) + \sum_i^N A_c(i)}$$

where  $R_c$  and  $A_c$  are the binary reference image and probability map respectively for class  $c$  with number of voxels  $N$  [7]. The DSC was computed for seven classes: background pixel resolution, diastolic and systolic phases for myocardium mass, left & right ventricle volumes. The volumes were averaged to ensure joint optimization for all classes.

Data augmentation with affine transforms during each training epoch was performed to improve network localization and mitigate overfitting. The Adam optimizer with L2-regularization, and Xavier initialization were used for model training [8] [9]. The network was optimized with default learning rate of 0.01 and image size 16 mini-batches containing 16 images during each iteration. Pairs of short axis cine-MR cardiac cycle training images were processed for 10,200 iterations over 30 hours on a Tesla K80 GPU to generate a model. Cardiac features for each patient were then diagnosed using an ensemble system for cardiac disease classification.

## 2.1 Preprocessing

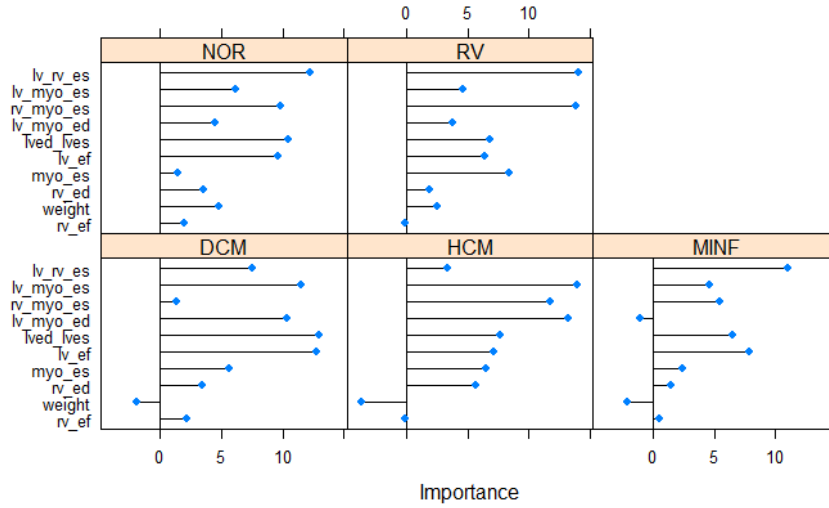
Images were normalized by subtracting the minimum pixel intensity from each channel and dividing by the mean pixel intensity to represent pixels in the range 0 to 1. A filtering method located and removed defective images with noninterpretable pixel color distributions from our data set. Resolution recording variability was mitigated by resampling images and segmentations to a common in-plane resolution of  $1.37 \times 1.37 \text{ mm}^2/\text{pixel}$ . Each resampled image was placed centrally with constant size, zero padded where necessary and intensity-normalized to zero mean. The Pearson correlation coefficient measure was applied to our volumetric feature matrix constructed from voxel segmentation class mappings. Any pair-wise correlation exceeding a conservative cutoff threshold of 0.9 was removed from the data set. Finally, the cardiac feature data was z-score normalized by determining the probability of a score occurring within our normal distribution with respect to the mean.

## 2.2 Data Augmentation

The number of images were augmented in real-time to improve network localization capability and reduce overfitting. During each epoch, a random augmentation of images that preserve collinearity and distance ratios was performed. Random padding with zeros, zoom, rolling and rotation was also implemented. The training data set was further augmented by adding new contrast limited adaptive histogram equalization filtered images randomly selected from each pathology category. There wasn't a noticeable improvement in segmentation performance using these synthetic techniques. One possible explanation is that significantly larger amounts of augmentation are required for improvement. However, it should be noted that the gap between training error and test error percentages steadily decreased as the amount of augmentation was increased, thus indicating a reduction in overfitting.

### 2.3 Feature Selection

The semantic segmentation pixel to label mapping generated a volumetric feature matrix containing 20 records for each of 5



**Fig. 3** Feature importance ranked by Gini impurity decreases. Variables lv\_rv\_es, lv\_myo\_es, rv\_myo\_es, lv\_myo\_ed, lved\_lves, lv\_ef selected.

classes and 16 quantitative covariates including weight, height, ventricular volumes, calculated ejection fractions, myocardium mass and permuted ratios for both end diastolic and systolic time points of the cardiac cycle. Feature selection determined a subset of maximally discriminatory predictors for 5-ary classifier model construction. Random forests ranked feature importance based on amount of mean Gini impurity decrease. Breiman’s 1-SE rule was utilized to select six features for building the parsimonious pathology classification model see Figure 3.

### 3 Results

A deep learning GPU training system (DIGITS) with pretrained convolutional neural networks for image classification enabled rapid prototype training, real-time performance monitoring and visualizations. The data set contained 934 parasternal short axis cine-MRI image slices partitioned into 50% training, 25% validation and 25% test subsets. The GoogLeNet architecture using the stochastic gradient descent (SGD) optimizer and transfer learning, which retains initial model weights and extracts image features via a final network layer, was chosen as our baseline classifier see Figure 4.

| Rapid Prototyping Results - Transfer Learning |         |               |                   |                      |                 |                    |
|-----------------------------------------------|---------|---------------|-------------------|----------------------|-----------------|--------------------|
| 5-ary Classifier                              | Solver  | Learning Rate | Policy            | Validation Accuracy% | Test Set 95% CI | Median Specificity |
| AlexNet                                       | RMSProp | 1e-4          | Exponential Decay | 98.96                | (0.319, 0.461)  | 84.71              |
| AlexNet                                       | SGD     | 1e-3          | (17,33) Step Down | 98.44                | (0.384, 0.529)  | 84.88              |
| AlexNet                                       | Adam    | 1e-4          | (25,43) Step Down | 98.96                | (0.329, 0.472)  | 85.63              |
| AlexNet                                       | AdaGrad | 1e-4          | (17,33) Step Down | 92.71                | (0.271, 0.408)  | 82.08              |
| AlexNet                                       | NAG     | 1e-3          | Exponential Decay | 90.10                | (0.334, 0.477)  | 83.43              |
| GoogLeNet                                     | Adam    | 1e-4          | (33,66) Step Down | 94.27                | (0.379, 0.524)  | 86.06              |
| GoogLeNet                                     | SGD     | 1e-3          | (33,66) Step Down | 95.31                | (0.502, 0.646)  | 91.67              |
| GoogLeNet                                     | AdaGrad | 1e-3          | (33,66) Step Down | 92.71                | (0.310, 0.451)  | 83.14              |
| GoogLeNet                                     | NAG     | 1e-3          | (33,66) Step Down | 91.67                | (0.379, 0.524)  | 85.63              |
| GoogLeNet                                     | RMSProp | 1e-3          | (50,85) Step Down | 95.31                | (0.415, 0.560)  | 86.42              |

**Fig. 4** Hyperparameter optimization of the ACDC data set using pretrained models with transfer learning.

We targeted comparison of this result with a pipeline consisting of semantic segmentation followed by classical machine learning of volumetric features. The initial processing model was trained on a mini-batch of images and corresponding ground truth masks with the Softmax classifier at the top. During training, masks were downsampled to match the output network size. Bilinear upsampling was applied during inference to acquire equivalent output and input dimension sizes. The final segmentation mask was computed as argmax of the classification network output tensor. A U-Net convolutional neural network was used trained to predict a 3D voxel-wise label map  $T_v \in \{0, 1, 2, 3\}$  [10]. The labels were tallied to quantify background tissues, left ventricle endocardium, epicardium and right endocardium volumes respectively.

Clinical evaluation resulting from a 2D U-Net architecture cost function optimization measured the mean symmetric surface distance (ASSD) and Dice coefficient (DSC). Performance was assessed using cross entropy (CE), weighted CE and Softmax Dice cost values averaged over both cardiac phases shown in Figure 5. Optimizing the Dice loss function directly produced the best results, followed by weighted and unweighted CE respectively.

| Cost Function | DSC (LV)             | ASSD (LV)            | DSC (RV)             | ASSD (RV)            | DSC (MYO)            | ASSD (MYO)           |
|---------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| CE            | 0.905 (0.069)        | 1.171 (1.295)        | 0.842 (0.111)        | 2.388 (5.361)        | 0.843 (0.068)        | 0.958 (0.722)        |
| Weighted CE   | 0.919 (0.060)        | 0.906 (1.035)        | 0.844 (0.088)        | <b>1.761 (1.200)</b> | 0.855 (0.064)        | 0.833 (0.546)        |
| Soft Dice     | <b>0.940 (0.042)</b> | <b>0.622 (0.655)</b> | <b>0.849 (0.098)</b> | 1.903 (1.490)        | <b>0.886 (0.036)</b> | <b>0.625 (0.366)</b> |

**Fig. 5** Segmentation accuracy optimization of a 2D U-Net architecture using three cost functions.

The Softmax Dice was utilized as our architecture loss function based on superior performance. Clinical measures of our 2D U-Net architecture evaluated over both cardiac phases and with a third surface metric, Hausdorff distance (HD) are shown in Figure 6.

Ventricular ejection fraction and all volumetric ratio permutations, for both end diastolic and systolic time points, were computed from these basic quantities and averaged over all mini-batches. The resulting feature matrix contained

| <b>Metric</b> | <b>LV (ED)</b>       | <b>LV (ES)</b>       | <b>RV (ED)</b>       | <b>RV (ES)</b>       | <b>MYO (ED)</b>      | <b>MYO (ES)</b>      |
|---------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| ASSD          | 0.359 (0.414)        | 0.885 (0.896)        | 0.987 (1.324)        | 2.816 (1.647)        | 0.536 (0.281)        | 0.713 (0.450)        |
| DSC           | <b>0.962 (0.021)</b> | <b>0.917 (0.063)</b> | <b>0.914 (0.075)</b> | <b>0.783 (0.120)</b> | <b>0.881 (0.037)</b> | <b>0.891 (0.034)</b> |
| HD            | 6.072 (3.529)        | 9.703 (5.144)        | 14.90 (9.158)        | 21.38 (9.868)        | 11.05 (7.991)        | 13.32 (9.019)        |

**Fig. 6** Architecture evaluation measures representing mean (std) value accuracy obtained for specific structures and cardiac phase.

13 cardiac volume features including permutation ratios, and 2 ejection fractions plus weight for each of 50 patients. A parsimonious model was built and trained using ensemble stacking of several learning algorithms (SVM, Random Forests, PCA and Bagging). The average performance measures across all classes include 90% accuracy, 90% sensitivity and 97.5% specificity for the held-out test set data.

### 3.1 Error Analysis

We performed error analysis during each of the major processing elements within our fully automated pipeline. The semantic segmentation element was analyzed by tuning weight decay, batch size, learning rate and momentum strength. The difference in Dice coefficient similarity error was computed over separate runs for each hyperparameter. Learning rate and the optimizer type, Adam outperformed SGD with momentum, had the largest impact in reducing error according to the analysis. The choice of selected features substantially affected misclassification error during the second processing pipeline element. Following data preprocessing, the difference in classification accuracy varied by as much as 10% depending on the cardiac quantitative features selected.

## 4 Discussion

Although the relatively small sample size is adequate to train a semantic segmentation network, it is improbable that 20 cases per disease category would be sufficient to train a study-level classification algorithm that reliably distinguishes all 5 categories. Hence, a clear limitation of our study is data sparsity and the pilot nature of this paper should be emphasized. Several approaches will be explored to develop an enhanced medical analytics pipeline including Bayesian optimization for determining the best deep neural network hyperparameter configuration that minimizes model validation loss. In addition, several image processing techniques for contrast enhancement will be evaluated with the intent of improving precision and medical imaging feature discrimination.

The evaluated data set containing a relatively sparse amount of patient images is a characteristic that typically negatively affects machine learning classifier accuracy and generalizability. Consequently, an assessment of gener-

ative adversarial networks as a data augmentation mechanism for supplying high quality synthetic images will be investigated.

Finally, a novel technique that attempts to improve interpretability of automatic computer-aided diagnosis and simultaneously increase multi-class model accuracy of medical image interpretation will continue development. The approach (Reductio ad Modicum) decomposes label classification and builds simpler  $n$ -ary models with ( $n \leq 3$ ) for each subclass. The models are executed in series to produce composite output that is more interpretable.

## 5 Conclusion

In this paper a fully automatic processing pipeline for pathology classification on cardiac cine-MRI was presented. The system achieved higher accuracy on 5-ary classification compared to a baseline robust pretrained architectures with transfer learning. The image pipeline includes semantic segmentation with a 2D U-Net convolutional neural network, volumetric feature extraction and stacked ensemble model classification. We trained the neural network, tuned hyperparameters and computed held-out test set partition performance metrics for a 5-ary classifier. Rigorous error analysis revealed that learning rate, optimizer type and feature selection were the greatest contributors to overall improved pipeline processing element performance. Data augmentation, weight decay and momentum strength variance failed to influence semantic segmentation performance by an appreciable amount according to our experiments. Dice index scores themselves may be potentially confounding, hence conduction of a clinical reader study or utilization of a more objective algorithm performance measure should be thoroughly investigated.

## References

1. Cohn J, Ferrari R, Sharpe N. Cardiac remodeling concepts and clinical implications: A consensus paper from an international forum on cardiac remodeling. *JACC*. 2000;35:569–582.
2. Saeed M, Van T, Krug R, Hetts S, Wilson M. Cardiac MR Imaging: Current Status and Future Direction. *Cardiovascular Diagnosis Therapy*. 2015;p. 290–310.
3. Doi K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *CMIG*. 2008;31(4):198–211.
4. Ritter S, Margulies K. Emerging Tools for Computer-Aided Diagnosis and Prognostication. *Journal of Clinical Trials*. 2014;4(2):1–120.
5. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *Proc of MICCAI*. 2015;p. 234–241.
6. Dublois J. Automated cardiac diagnosis challenge; 2017. Available from: <https://www.creatis.insalyon.fr/Challenge/acdc>.

7. Sorensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Kongelige Danske Videnskabernes Selskab. 1948;5(4):1–34.
8. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. In: 3rd International Conference for Learning Representations. vol. arXiv:1412.6980; 2015. .
9. Glorot X, Bengio Y. Understanding the difficulty of training deep feed-forward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics; 2010. p. 249–256.
10. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: arXiv:1505.04597; 2015. .

### Author biography

**Tony Lindsey** Chief Data Scientist for the Autonomous Medical Operations project at NASA Ames Research Center.

**Jin-Ju Lee** NASA Summer Student Intern.