A Multi-modality Network for Cardiomyopathy Death Risk Prediction with CMR Images and Clinical Information *

Chaoyang Xia[1**], Xiaojie Li[1**], Xin Wang[2(✉)], Bin Kong[3], Yucheng Chen[4(✉)], Youbing Yin[2], Kunlin Cao[2], Qi Song[2], Siwei Lyu[5], and Xi Wu[1(✉)]

[1] College of Computer Science, Chengdu University of Information Technology, Chengdu, China
xi.wu@cuit.edu.cn
[2] CuraCloud Corporation Seattle, WA 98104, USA
xinw@curacloudcorp.com
[3] University of North Carolina at Charlotte
[4] West China Hospital, Sichuan University
chenyucheng2003@126.com
[5] SUNY Albany

Abstract. Dilated Cardiomyopathy (DCM) is one of the main worldwide causes of sudden cardiac death (SCD). Early diagnostics significantly increases the chances of correct treatment and survival. However, there are no efficient methods for mortality risk prediction from learning cardiac magnetic resonance (CMR) image and clinical data due to the poor image quality and extreme imbalanced datasets. To solve this problem, we proposed an effective multi-modality network (MMNet) for mortality risk prediction in DCM, and we firstly directly optimize the AUC to train the multimodal deep learning classifier by maximizing the WMW statistic. This can achieve significant improvements in AUC, especially under the imbalanced learning problem. MMNet consists of two branches: clinical data branch and T1 mapping CMR images branch, which allows the model to learn more comprehensive features and makes a more accurate prediction. We validated our approach on a DCM dataset, which contains 450 CMR images that only holds 34 positive samples. Experimental results show that our approach archived accuracy of 98.89%, AUC of 99.61%, sensitivity of 100% and specificity of 98.8%, demonstrating the effectiveness of the proposed method.

Keywords: Dilated cardiomyopathy · cross-modality medical data · AUC optimization.

* This study was supported by the major project of the Education Department in Sichuan (2017JQ0030 and 17ZA0063) and the National Natural Science Foundation of China (Grant No. 61602066), and in partly by the Project of Sichuan Outstanding Young Scientific and Technological Talents (19JCQN0003) and the Natural Science Foundation for Young Scientists of CUIT (J201704).

** Chaoyang Xia and Xiaojie Li have contributed equally to this work.

1 Introduction

Dilated Cardiomyopathy (DCM) is a common chronic and life-threatening cardiopathy [5]. It can lead to cardiovascular death, progressive heart failure or sudden cardiac death (SCD). Thus, immediate emergency diagnosis of DCM is critical for life saving and later recovery. For severe cardiomyopathy patients, cardiologists may consider ventricular assist devices (VAD) or heart transplants operation in the early stage of an incident. However, both are not only expensive but also can lead to serious complications, including infection, thromboembolism, and multiple organ failure. In routine clinical diagnosis, especially for early screening and postoperative assessment, visual assessment and empirical evaluation are widely used. Nevertheless, they are subject to high inter-observer variability, and the results are subjective and non-reproducible. Furthermore, utilizing multi-modality medical data (e.g, clinical text report and magnetic resonance imaging (MRI)) to assess the accurate risk of DCM patients are even more challenging due to the complex nature of medical data (e.g., height, weight, family history of cardiopathy, blood pressure, and so on).

In this regard, automatic computer-aided diagnosis systems are highly desirable. Many attempts have been made to automatically assess the mortality risk in DCM. Traditional risk assessment approaches have been mainly based on boosted ensemble algorithms (feature selection by information gain ranking) [9, 1]. However, most of these methods are based on a small subset of the clinical and imaging data. As a result, they are not able to capture sufficient information to establish intrinsic correspondences between the mortality risk and clinical and imaging data. Additionally, their performances are confined by the handcrafted descriptors. As the one of the most successful machine learning techniques today, Li et al. [7] [8] provided manifold alignment and efficient boundary point detection method, which can help in the study of the characteristics of diseases. Deep learning has been successfully applied to the recognition and prediction of prostate cancer, Alzheimers disease, and vertebrae and neural foramina stenosis. In this work, we aim to propose a deep learning based framework for DCM mortality risk assessment.

Automatic assessment the mortality risk in DCM remains a challenging problem due to two main issues. First, due to the rare occurrences of death, the imaging data as well as the discrete clinical text data are highly imbalanced. Nevertheless, most of the existing classification losses such as cross entropy are not suitable for dealing with imbalanced classes. In machine learning literautre [10], many a study has suggested that compared to simple classification losses, AUC (area under the receiver operating characteristic curve) is a robust evaluation measure for classification problem. However, it is non-differentiable and not easy to compute. Therefore, directly optimize the AUC loss to train a classifier is usually impractical. Although existing sampling, adjust class weight and data enhancement have shown great success [3, 12], the imbalanced learning problem is still challenging. Second, in routine clinical procedures, risk assessment of DCM is often carried out through evaluation of multi-modality image data (e.g., images and clinical text), in which one data modality is complementary to other

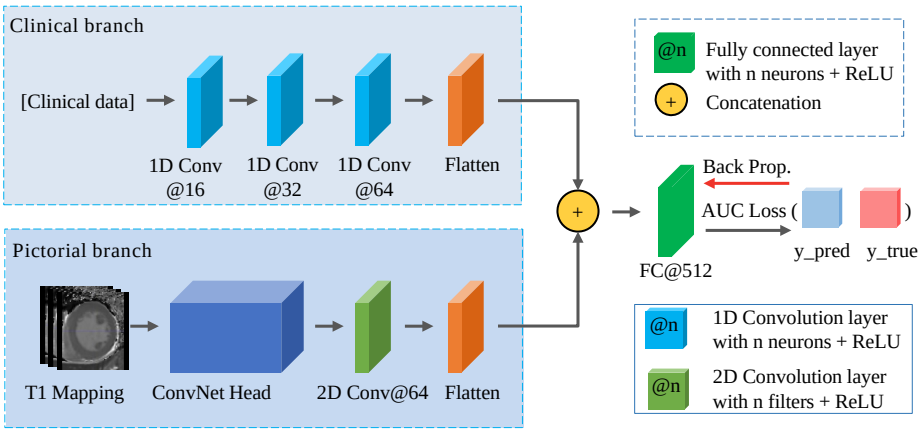

Fig. 1. The architecture of the multi-modality framework. It mainly consists of a pictorial branch and a clinical branch. The model is directly optimized by the AUC Loss.

modalities. Thus, fusing multi-modality medical data for accurate mortality risk assessment is highly desired.

In this paper, an end-to-end multimodal framework is proposed to solve the above issues. The core idea is to leverage the fused features extracted from the multi-modality data. It consists of two branches: clinical branch and image branch. The image branch is a 2D convolutional network, which extracts image features from the left ventricle T1 mapping CMR images. The clinical branch is a fully-connected network, which extracts features from clinical textual data. Afterward, these features are fused by concatenation layer for the final prediction. Furthermore, inspired by [13], we optimize the alternative Wilcoxon-Mann-Whitney (WMW) loss to directly optimize the AUC instead of other typical classification losses to address the imbalanced learning issue. Our method main contributions are as follows: 1) a multi-modal framework seamlessly fuse CMR images and clinical data features to better learn hierarchical feature representations; 2) we combine AUC optimization and multimodal framework together to train the proposed network. This approach effectively addresses the imbalanced learning issue. To the best of our knowledge, this is first work to directly use AUC to optimize complex deep neural networks.

2 Methodology

Fig. 1 shows an overview of the proposed method. Our goal is to automatically generate a prediction score for the cardiac patients. A multi-modality framework is proposed to seamlessly fuse CMR images and clinical text features, thereby generating better hierarchical feature representations for accurate DCM risk assessment. Our network is able to effectively fuse the information from

multi-modality medical data, helping it to generate more reliable predictions. We directly optimize the AUC loss to train the proposed network.

2.1 A Multi-Modality Network (MMNet)

As shown in Fig. 1, the proposed framework consists of two branches: the clinical branch and the pictorial branch. The clinical branch is used to extract 1D texture record features from clinical textual data. The first layers will learn the low level features with multiple 1D convolutional layers, which are followed by ReLU non-linearity activation layers. There are three convolutional layers (with 16, 32 and 64 filters) for feature extraction.

The pictorial branch automatically extract discriminative features from the CMR images. It consists of a ConNet Head (a convolutional network) and a convolutional layer. The ConvNet head is used to successively extract noisy-invariant high-level representations from the images. This is extremely important, especially when low-level features (e.g., color, texture) is not sufficient to represent the image due to various reasons (e.g., variation of image intensities). Note that our ConvNet Head is not limited to one kind of network. It can be VGG-16 [11], DenseNet-121, Inception-V3, and Xception, etc. To reduce the computational costs, a 2D convolution layer with 64 filters (1×1 kernels) is followed by the ConvNet head to reduce the number of feature map channels.

Finally, after flattening the feature maps generated by both branches, the high-level textual semantics are concatenated with the high-level image semantics, yielding the fused high-level representations. The fused high-level representations is fed into a fully connected (FC) layer with 512 neurons and softmax layer, generating the final predictions.

2.2 AUC Optimization

Due to the extremely low mortality of DCM, the number of observations belonging to the death class is significantly lower than those belonging to the non-fatal class. As a result, the imbalanced class problem is predominant. Notably, the rate of negative samples to positive samples is 12.23 in our dataset. In this situation, if the common loss functions such as cross entropy is employed to train the network, the prediction will be extremely biased. Although a very high accuracy can be obtained, the specificity is very low. Although AUC is non-differentiable and not easy to compute (it is usually used as a robust measure to evaluate the performance of classifiers), several works [13] had demonstrated that maximizing the alternative Wilcoxon-Mann-Whitney (WMW) loss is equivalent to directly optimizing AUC. Inspired by [13], we adopted the WMW loss to optimize our network. Formally,

$$R(x_i, y_j) = \begin{cases} (-(x_i - y_j - \gamma))^p \text{ , } x_i - y_j < \gamma \\ \\ 0 \qquad\qquad \text{, } otherwise \end{cases} \tag{1}$$

where the x_i is the predicted score for i^{th} positive sample, and y_j is the predicted generated for j^{th} negative sample. The margin γ and exponent p are two hyper-parameters. Finally, we directly optimize the AUC by minimizing the objective function U_R with $0 < \gamma \leq 1$ and $p > 1$:

$$U_R = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} R\left(x_i, y_j\right). \tag{2}$$

Note that function R is differentiable, we can use any gradient based methods to train the MMNet.

3 Experiments

Data Acquisition & Pre-processing. Our experiments are conducted on a dataset with 450 patients, which are collected from our collaborative hospital. The clinical text reports and the incidents are provided by senior cardiologists during clinical examinations and follow-up visits. As each report belongs to a cardiomyopathy patient and each patient has several CMR images, each clinical text report corresponds to several CMR images.

All CMR images are 2D short axis cine native T1 mapping MR images. The pixel spaces of those CMR images range from $1.172 \times 1.172 \times 1.0mm^3$ to $1.406 \times 1.406 \times 1.0mm^3$. The original dimension size is $256 \times 218 \times 1$ pixels. To ensure the training datasets are one and the same common space to enable improved quantitative analysis, we resample the image to a spacing of $1.0 \times 1.0 \times 1.0mm^3$ to ensure isotropy and normalize the of CMR image intensities to $[-1, 1]$. The clinical text data contains extensive patient information such as family history of cardiopathy, height, weight, blood pressure. We normalize the categorical clinical textual values to $[0, 1]$.

Unfortunately, as the mortality rate is 7.5%, our dataset is extremely imbalanced. Due to the lack of positive samples, several data augmentations are applied to them to virtually enlarge the training set. These augmentations include: random rotation from 0 to 180 degrees, random vertical or horizontal flip, and random shift along the X axis or Y axis from 0 to 2%. During the augmentation process, the corresponding text data is duplicated. Four types of criteria is used to measure the performance of classifiers: 1) accuracy; 2) sensitivity; 3) specificity; 4) AUC.

Implementation Details. We adopted 5-fold cross-validation to evaluate the performances of different methods. The final evaluation score is calculated by averaging the scores of all 5 folds.We use Adam optimizer [6] with a initial learning rate of 0.003, and leave other parameters as Keras default. Our model uses batch size of 32 training with 40 epochs.

Results and Analysis. We compared our method with the signal model methods [11, 4] that only using CMR images and traditional methods which only using clinical data to predict death risk. Table 1 shows the performance by traditional methods: Linear SVM, Decision Tree, Random Forest and other advanced CNN models: VGG-16 [11], ResNet-50 [4], DenseNet-121, Inception-V3,

Table 1. Comparison of the proposed MMNet with the other advanced CNN methods of risk assessment.

Methods	Loss	Inputs	Accuracy	Sensitivity	Specificity	AUC
Linear SVM	-	Clinical	85.29%	0.0	100%	66.21%
Decision Tree	-	Clinical	82.35%	0.0	100%	45.52%
Random Forest	-	Clinical	85.29%	0.0	100%	63.10%
VGG-16	BCE	CMR	-	-	-	-
ResNet-50	BCE	CMR	80.67%	47.06%	83.41%	73.36%
Inception-V3	BCE	CMR	89.11%	26.47%	94.23%	73.47%
Xception	BCE	CMR	86.00%	41.18%	89.66%	78.00%
MMNet	BCE	CMR+Clinical	97.04%	100%	96.81%	99.31%
VGG-16	WMW	CMR	-	-	-	-
ResNet-50	WMW	CMR	72.00%	67.65%	72.36%	76.21%
Inception-V3	WMW	CMR	90.44%	44.12%	94.23%	75.24%
Xception	WMW	CMR	84.00%	41.18%	87.50%	78.09%
MMNet	WMW	CMR+Clinical	98.89%	100%	98.80%	99.61%

and Xception [2]. For a classifier, WMW loss aims to achieve reliable improvements in the AUC measure. In order to demonstrate the effectiveness of AUC optimization on imbalanced datasets, we also compare our framework with different loss function: BCE loss and WMW loss in deep architecture (see Table 1). Table 1 illustrates that our method in general achieve better performance than all the other methods, in terms of Accuracy, Sensitivity, Specificity and AUC, and the ROC curve for each mehtod is shown in Fig. 2 (Left).

From the results of risk assessment in Table 1, we can observe three key points. First, it is difficult to train an effective classifier under imbalanced datasets. Traditional classification methods cannot identify any positive samples. Due to vanishing gradient problem, the VGG-16 cannot report valid results. Although Inception-V3 produce a high classification accuracy of 90.44%, the sensitivity actually is 44.12% and the specificity is 94.23% by contraries. This illustrates the Inception-V3 model learn too much about majority class, but it provides insufficient information about minority class. Therefore, it may predict almost every sample as majority class. Second, our proposed MMNet takes the advantage of the combined information of CMR images and clinical data, it performs much better in terms of five types of criteria than the other methods only use image information, but without clinical data. Specifically, MMNet with BCE loss achieves a classification accuracy of 97.04%, a sensitivity of 100%, a specificity of 96.81% and an AUC of 99.31%. Finally, AUC optimization is crucial for classification task when using imbalanced datasets. As shown in the rear rows in the Table 1, the MMNet with WMW loss achieve better performance than one with BCE loss. Moreover, Fig. 2 (Left) demonstrates again that our method achieves substantial improvement of the ROC curve over other advanced CNN models.

Hyper-parameter Selection. The WMW loss (Eq.(1)) has two essential hyper-parameters: the exponent p and the margin γ. Generally, we set $p = 3$

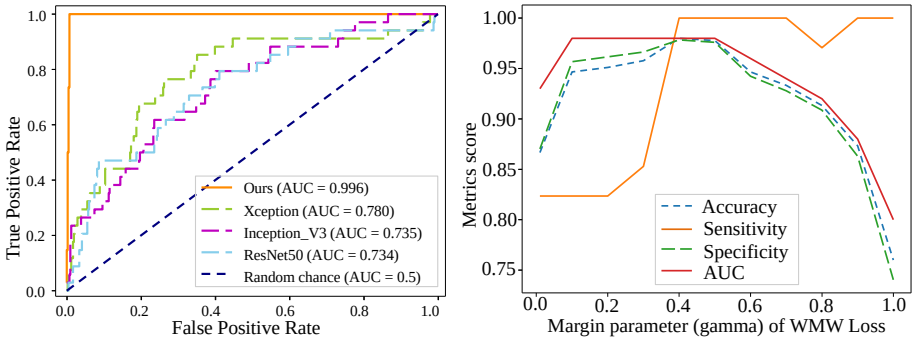

Fig. 2. Left: The ROC curves with AUC values for 5-fold cross-validation results of different methods. **Right:** The Accuracy, Sensitivity, Specificity and ROC curves of different γ for WMW Loss when $p = 3$ over 5-fold cross validation.

as suggested in [13]. Fig. 2 (Right) shows how the average performance of our model varies with the parameter γ. As we can see, when $\gamma \leq 0.4$, our model consistently increases as γ increases. When $\gamma > 0.4$, all the metric scores in the figure start to decrease. Nevertheless, the AUC is less affected when γ is near 0.4, which demonstrates that the model has high robustness when $\gamma = 0.4$. Thus, we set $p = 3$ and $\gamma = 0.4$ in our experiment for the WMW loss.

Evaluation of the ConvNet Head. To investigate the effectiveness of using different convolutional network as ConvNet Head, we evaluate several widely used ConvNets in our framework and the results are summarized in Table 2. Experimental results show that our framework is consistent robust for various ConvNet Head.

Table 2. Comparison results of the different ConvNet as image branch heads.

Head	Accuracy	Sensitivity	Specificity	AUC
ResNet-50	99.11%	100%	99.04%	99.43%
VGG-16	99.11%	100%	99.04%	99.49%
Xception	96.22%	100%	95.91%	99.53%
Inception-V3	98.89%	100%	98.80%	99.61%

Class Activation Map (CAM). To further understand how the classifier make the predictions, we visualize the class activation map (CAM) [14] of the last convolutional layer. Two group of negative and positeive examples are shown in Fig. 3. There are obvious difference on the corresponding CAMs of negative (survivors) and positive (dead) samples. We can see that the heatmap or jet color map of positive samples is more comparatively concentrated than that of negative samples.

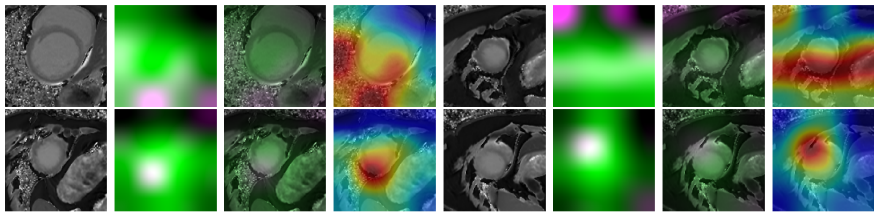

Fig. 3. Activation maps. The first row is examples of negative samples. The Second row is examples of positive samples. (a) shows the original CMR images. (b) are the activation maps, the green color indicates larger standard deviations, the pink color represents the maximum values and the white color is composed of both types. The heatmap is overlaid on the original images in (c). (d) displays heatmaps as jet color map and overlap on the original image, the red color highlights the activation region associated with the predicted class.

4 Conclusion

In this paper, we proposed an effective multi-modality framework (MMNet) for mortality risk prediction in DCM. As far as we know, we are the first to directly optimize the AUC to train the multimodal deep learning classifier by maximizing the WMW statistic. This can achieve significant improvements in AUC, especially for the imbalanced learning problems. Experiment results demonstrate the superiority of the proposed method.

References

1. Afshin, M., Ayed, I.B.T., Li, S.e.a.: Assessment of regional myocardial function via statistical features in mr images. In: MICCAI (2011)
2. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: CVPR. pp. 1251–1258 (2017)
3. Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. arXiv preprint arXiv:1705.00440 (2017)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
5. Jefferies, J.L., Towbin, J.A.: Dilated cardiomyopathy. The Lancet **375**(9716), 752 – 762 (2010)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Li, X., Lv, J.C., Zhang, Y.: Manifold alignment based on sparse local structures of more corresponding pairs. In: Twenty-Third International Joint Conference on Artificial Intelligence (2013)
8. Li, X., Lv, J., Yi, Z.: An efficient representation-based method for boundary point and outlier detection. IEEE transactions on neural networks and learning systems **29**(1), 51–62 (2018)

9. Motwani, M., Dey, D., Berman, D.S., Germano, G., Achenbach, S., et al.: Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. European Heart Journal **38**(7), 500–507 (2016)
10. Provost, F.J., Fawcett, T., Kohavi, R., et al.: The case against accuracy estimation for comparing induction algorithms. In: ICML. vol. 98, pp. 445–453 (1998)
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. pp. 1–9 (2015)
13. Yan, L., Dodier, R.H., Mozer, M., Wolniewicz, R.H.: Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In: ICML. pp. 848–855 (2003)
14. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)