SAF-Net: Self-Attention Fusion Network for Myocardial Infarction Detection Using Multi-View Echocardiography

Ilke Adalioglu[1], Mete Ahishali[1], Aysen Degerli[1], Serkan Kiranyaz[2], Moncef Gabbouj[1]

[1] Tampere University, Tampere, Finland
[2] Qatar University, Doha, Qatar

Abstract

Myocardial infarction (MI) is a severe case of coronary artery disease (CAD) and ultimately, its detection is substantial to prevent progressive damage to the myocardium. In this study, we propose a novel view-fusion model named self-attention fusion network (SAF-Net) to detect MI from multi-view echocardiography recordings. The proposed framework utilizes apical 2-chamber (A2C) and apical 4-chamber (A4C) view echocardiography recordings for classification. Three reference frames are extracted from each recording of both views and deployed pre-trained deep networks to extract highly representative features. The SAF-Net model utilizes a self-attention mechanism to learn dependencies in extracted feature vectors. The proposed model is computationally efficient thanks to its compact architecture having three main parts: a feature embedding to reduce dimensionality, self-attention for view-pooling, and dense layers for the classification. Experimental evaluation is performed using the HMC-QU-TAU[1] dataset which consists of 160 patients with A2C and A4C view echocardiography recordings. The proposed SAF-Net model achieves a high-performance level with 88.26% precision, 77.64% sensitivity, and 78.13% accuracy. The results demonstrate that SAF-Net model achieves the most accurate MI detection over multi-view echocardiography recordings.

1. Introduction

Early detection of Myocardial Infarction (MI), an indicator of coronary artery disease (CAD), is substantial to avoid further health complications. Echocardiography (echo) is a primary diagnostic tool used in MI investigation. Particularly, detection of the early signs of MI is often not viable using an electrocardiogram (ECG) as the disease has already progressed significantly by the time it becomes visible in ECG recordings [1].

Recently, there has been a growing interest in leveraging machine learning (ML) techniques in echo analysis to develop automated, fast, accessible, and cost-effective tools [2, 3] to assist medical doctors (MDs). Nevertheless, there is a lack of research on MI detection and publicly accessible echo datasets are limited. In fact, only the study in [3] has released the first publicly available dataset containing apical 4-chamber (A4C) view echo recordings for the detection of MI. The proposed ML-based detection framework in [3] utilizes only single-view recordings. Obviously, if the infarcted cardiac muscle segment is invisible in the chamber view, it will be missed by the detection framework. To this end, the proposed approach in [4] focuses on capturing early MI from multi-view echo using displacement features extracted by analyzing segmented endocardial boundary.

As a recent trend, the attention mechanism introduced in [5] is considerably used in various domains achieving promising outcomes in challenging computer vision tasks such as image classification, captioning, and recognition [6, 7, 8]. By operating between different parts of the input or using learned context vectors, the attention mechanism is used to make predictions from different sequences. Furthermore, the self-attention mechanism enables the model to capture long-term dependencies and contributions of the different features of different positions in a single sequence. The medical field benefits from self-attention with its capability of extracting robust and descriptive features by exploiting patterns within the data [9, 10]. The ability to extract these descriptive features improved the performance in medical classification problems such as in [11, 12]. Similarly, self-attention is utilized to detect pulmonary nodule classification in [13] using multi-view data.

In this study, we propose a self-attention fusion network (SAF-Net) for early MI detection using multi-view echo recordings. The proposed detection framework consists of a novel feature extraction technique that can capture the dependencies in the pruned feature space using a self-attention module. In summary, our contributions can be listed as follows:

- We devise a framework for capturing salient features from multi-view echo recordings that reflect the impor-

[1] The benchmark HMC-QU-TAU dataset is publicly shared at the repository https://www.kaggle.com/aysendegerli/hmcqu-dataset.

tant patterns. Our approach includes the extraction of three distinct frames in both views, complemented by the utilization of different state-of-the-art networks to extract the feature set.

- By utilizing the self-attention mechanism in the proposed SAF-Net, we improve the model's ability to learn the dependencies of extracted feature vectors, ultimately improving the detection of early MI.
- We conduct an extensive set of experimental evaluations for the comparison of the proposed approach against various classifiers. Accordingly, the proposed approach achieves the highest MI detection performance.
- We extend the multi-view echocardiography dataset, HMC-QU-TAU by incorporating data from 160 patients.

2. Proposed Method

The model performance is highly correlated to the descriptive power of the used features for the classification. In this manner, we propose the following feature extraction scheme. First, we extract the first, middle, and last frames in one-cardiac cycle from both A4C and A2C views for every recording. Let the selected reference frames be $\mathbf{X}_{\text{A4C}}$ and $\mathbf{X}_{\text{A2C}}$ for both views, where $\mathbf{X} \in \mathbb{R}^{N \times N \times 3}$ and $N = 224$. Then, for each recording, we utilize pre-trained networks initialized with their ImageNet [14] weights in order to extract representative features for each one-cardiac cycle. The selected frames are depicted in Figure 1. Accordingly, we use the following state-of-the-art deep networks for the feature extraction given as $\mathbf{X}_{\text{A4C}}$ and $\mathbf{X}_{\text{A2C}}$ inputs:

- **DenseNet-121** [15] is characterized by its fully-connected convolutional layers with skip-connections resulting in an effective information flow through the layers.
- **Inception-v3** [16] is a deep network that is designed for image classification and recognition. By utilizing parallel convolutions in modules, and keeping the dimensionality low, Inception can achieve promising performance levels with less computational complexity compared to the deeper competitors.
- **ResNet-50** [17] mainly utilizes skip-connections to build a more stable but deeper network. It aims to overcome the problem of gradient vanishing by utilizing residual connections with addition operations.

By using pre-trained ImageNet weights of the above-mentioned state-of-the-art deep networks, we aim to extract representative features to ensure state-of-the-art MI detection performance. Given a chamber view $\mathbf{X}$ for a patient, a pre-trained network provides the following mapping: $\phi : \mathbb{R}^{N \times N \times 3} \rightarrow \mathbb{R}^{d}$ where d is the feature vector dimension. The extracted features $\mathbf{f} = \phi(\mathbf{X})$ have a dimension of $d = 1024$ for DenseNet-121, whereas $d = 2048$ for Inception-v3 and ResNet-50. Then, as shown in Figure 2, we obtain the final concatenated feature vector as $\mathbf{f} = [\mathbf{f_D}, \mathbf{f_I}, \mathbf{f_R}] \in \mathbb{R}^{5120}$. The feature pair for both chamber views containing $\mathbf{f_{A2C}}$ and $\mathbf{f_{A4C}}$ forms the matrix $F^{5120 \times 2}$ that is given to the SAF-Net model as input.

The first part of the SAF-Net consists of a feature embedding step for dimensionality reduction. For obtaining latent feature space ($L^{64 \times 2}$), we use one fully-connected layer with the Rectified Linear Unit (ReLU) activation function. After the feature extraction processes, we strategically placed the self-attention mechanism in the model. By using the mapped latent feature representation as input, the self-attention mechanism utilizes the benefits of capturing dependencies between the extracted and pruned features.

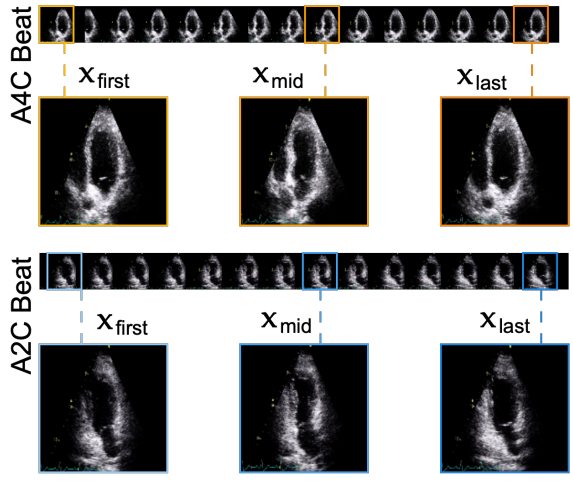

Figure 1: From one-cardiac cycle of both apical chamber views, three reference frames are extracted as the first, middle, and last.

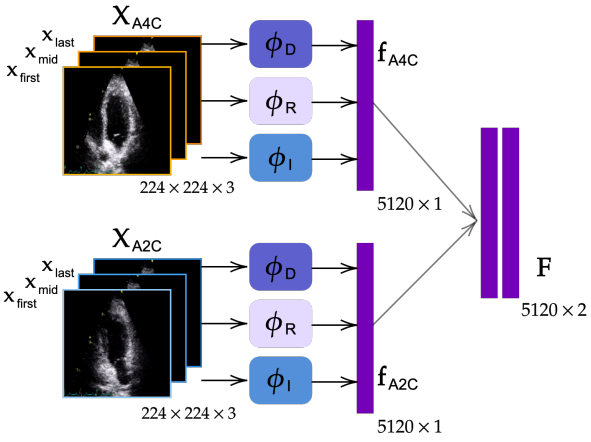

Figure 2: Features are extracted from both views frame set using pre-trained state-of-the-art networks: DenseNet-121 (ϕ_D), ResNet-50 (ϕ_R), and Inception-v3 (ϕ_I).

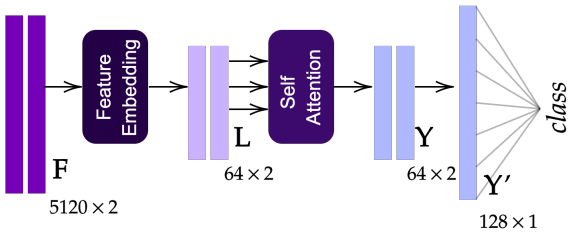

Figure 3: The proposed SAF-Net model consists of a feature embedding layer that computes latent space representation, L of the feature matrix, $\mathbf{F}$ which is followed by a self-attention layer. Lastly, a fully-connected layer is attached to perform the classification.

In the self-attention mechanism, for an input matrix $\mathbf{X} \in \mathbb{R}^{n \times d_{model}}$ where n is the latent dimension after linear embedding layer and d_{model} is the model's depth, the self-attention computation is performed using the weight matrices $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V \in \mathbb{R}^{d_{model} \times d_k}$, where d_k is the output dimension which is less than d_model. First, we calculate query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ as $\mathbf{X}\mathbf{W}_Q = \mathbf{Q}$, $\mathbf{X}\mathbf{W}_K = \mathbf{K}$, and $\mathbf{X}\mathbf{W}_V = \mathbf{V}$, where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times d_k}$. The weight matrices are learned through training to ensure a common-space representation for both views. By using the computed Q, K, and V, we calculate the attention scores as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \quad (1)$$

where d_k is used as a scaling factor. Essentially, Eq. (1) is a scaled dot product that calculates the similarity between the vectors. Hence, we try to capture the relationship and dependencies between the embedded features. The last step in the MI classification process is to employ a fully-connected layer with a sigmoid activation function, subsequent to the self-attention layer. The proposed SAF-Net model is illustrated in Figure 3.

3. Experimental Evaluation

Experimental Setup. We use the HMC-QU-TAU dataset [4] to evaluate the performance of the proposed SAF-Net. The dataset consists of 320 echocardiography recordings of A2C and A4C views from 160 individuals; **103** MI patients and **57** of non-MI subjects. Considering the positive label as MI, the performance of the proposed and competing methods are computed using metrics as follows: Sensitivity (Sen) is the ratio of correctly predicted positives to all actual positives, Specificity (Spe) is the ratio of correctly predicted negatives to all actual negatives, Precision (Pre) is the ratio of correctly predicted positives to all predicted positives, F1-Score (F1) is the harmonic mean of Sen and Pre, Accuracy (Acc) is the correctly classified samples to all samples in dataset, and Geometric Mean (GM) is the geometric mean of Sen and Pre. We perform 10-fold cross-validation and report averaged performance. To train the classification and feature embedding layer, we use the Adam optimizer [18] with a learning rate of 10^{-3} and weighted binary cross-entropy loss to overcome the challenges caused by the imbalanced dataset.

We have evaluated several methods for comparison including random forest (RF), decision tree (DT), k-nearest neighbor (KNN), support vector machine (SVM), and multi-layer perceptron (MLP) classifiers. We have performed grid-search approach for the hyper-parameter selection by performing another 3-fold CV over the training set. All classifiers are fed with the feature matrix $\mathbf{F}$ for a fair comparison with the proposed approach. The search parameters of SVM include linear, Gaussian, and polynomial kernels, where $C \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, $\gamma \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, and the degree of polynomial is set to $\{2, 3, 4\}$, respectively. The MLP classifier has the same structure as used in [19] and it is trained with Adam optimizer with a learning rate of 10^{-3} and binary cross entropy loss.

Table 1: Averaged MI detection performance results (%) obtained by the proposed SAF-Net approach and compared methods over the HMC-QU-TAU dataset.

Model	Sen	Spe	Pre	F1	Acc	GM
SAF-Net	77.64	**79.0**	**88.26**	81.57	**78.13**	**77.19**
RF	88.36	52.33	78.37	**82.11**	75.62	66.49
DT	66.18	51.33	72.52	67.28	60.63	56.52
KNN	82.54	60.33	79.21	80.49	74.37	69.69
SVM	**90.09**	39.66	75.30	80.82	72.50	43.90
MLP	76.63	64.66	81.45	77.47	72.50	67.73

Results and Discussion. Table 1 illustrates the superiority of SAF-Net over widely recognized classifiers in terms of specificity, precision, accuracy, and geometric mean for MI detection. Despite its outstanding performance in various evaluation metrics, SAF-Net is unable to achieve the highest sensitivity score. On the other hand, the compared classifiers fail to achieve any reasonable specificity level. Especially, SAF-Net excels in specificity score, which shows its capacity to minimize false positives as it can be seen in Table 2, it has achieved the highest GM value of 77.19% with a sensitivity level of 77.64%. The performance gap in GM is significant by 7.5% between the proposed approach and the closest competing method showing that the classifier has achieved the highest MI detection performance.

4. Conclusion

Early diagnosis of MI holds paramount significance to prevent further necrosis. Using a multi-view approach improves the identification accuracy, as it provides more vis-

Table 2: Cumulative confusion matrices of the proposed (a) SAF-Net, (b) RF, (c) DT, (d) KNN, (e) SVM, and (f) MLP models over test sets of each fold.

(a)

SAF-Net		Predicted	
		Non-MI	MI
Ground Truth	Non-MI	45	12
	MI	23	80

(b)

RF		Predicted	
		Non-MI	MI
Ground Truth	Non-MI	30	27
	MI	12	91

(c)

DT		Predicted	
		Non-MI	MI
Ground Truth	Non-MI	29	28
	MI	35	68

(d)

KNN		Predicted	
		Non-MI	MI
Ground Truth	Non-MI	34	23
	MI	18	85

(e)

SVM		Predicted	
		Non-MI	MI
Ground Truth	Non-MI	23	34
	MI	10	93

(f)

MLP		Predicted	
		Non-MI	MI
Ground Truth	Non-MI	37	20
	MI	24	79

ibility over myocardial segments for detecting MI. In this study, we proposed a novel feature extraction technique with frame selection for one-cardiac cycle by deploying pre-trained networks. Subsequently, the proposed SAF-Net model consisting of linear embedding, view pooling, and classification parts utilizing self-attention mechanisms exploits the dependencies in extracted features to maximize the MI detection performance. The experimental evaluations over the HMC-QU-TAU dataset consisting of multi-view echo pairs of 160 patients show that the proposed approach has achieved the highest accuracy, precision, specificity, and geometric mean with 78%, 88.26%, 79%, and 77.19%, respectively, as outperforming the competing well-known classifiers. This performance holds promise for the potential of SAF-Net in early MI detection and aiding healthcare professionals to enable their diagnostic tasks more effectively. Finally, the released HMC-QU-TAU dataset with this study lays the groundwork for future cardiology studies and our aspiration to leverage the potential of machine learning in the healthcare domain.

Acknowledgments

This work has been supported by Business Finland project AMALIA under NSF IUCRC Center for Big Learning and the Tietoevry Veturi program.

References

[1] Taban Sadeghi M, Mahmoudian B, Ghaffari S, Moharamzadeh P, Ala A, Pourafkari L, Gureishi S, Roshanravan N, Abolhasani S, Pouraghaei M. Value of early rest myocardial perfusion imaging with spect in patients with chest pain and non-diagnostic ecg in emergency department. Int J Cardiovasc Imaging 2019;35:965–971.

[2] Degerli A, Zabihi M, Kiranyaz S, Hamid T, Mazhar R, Hamila R, Gabbouj M. Early detection of myocardial infarction in low-quality echocardiography. IEEE Access 2021;9:34442–34453.

[3] Kiranyaz S, Degerli A, Hamid T, Mazhar R, Fadil Ahmed RE, Abouhasera R, Zabihi M, Malik J, Hamila R, Gabbouj M. Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. IEEE Access 2020;8:210301–210317.

[4] Degerli A, Sohrab F, Kiranyaz S, Gabbouj M. Early myocardial infarction detection with one-class classification over multi-view echocardiography. In Comput. Cardiol. (CinC), volume 498. 2022; 1–4.

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lu, Polosukhin I. Attention is all you need. In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds.), Adv. Neural Inf. Process. Syst., volume 30. 2017; .

[6] Rao A, Park J, Woo S, Lee JY, Aalami O. Studying the effects of self-attention for medical image analysis. In Proc. IEEE/CVF Int. Conf. Comp. Vis. (ICCV) Workshops. October 2021; 3416–3425.

[7] Huang L, Wang W, Chen J, Wei XY. Attention on attention for image captioning. In Proc. IEEE/CVF Int. Conf. Comp. Vis. (ICCV). October 2019; .

[8] Zhao H, Jia J, Koltun V. Exploring self-attention for image recognition. In IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). 2020; 10076–10085.

[9] Zhang Y, Liu H, Hu Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng Y, Essert C (eds.), Med. Image Comput. Comput. Assist Interv. (MICCAI). 2021; 14–24.

[10] Dong B, Wang W, Fan DP, Li J, Fu H, Shao L. Polyp-pvt: Polyp segmentation with pyramid vision transformers. CAAI Artif Intel Res 09 2023;.

[11] Wang Z, Zhang J, Zhang X, Chen P, Wang B. Transformer model for functional near-infrared spectroscopy classification. IEEE J Biomed Health Inform 2022;26(6):2559–2569.

[12] Shi L, Zhang Y, Cheng J, Lu H. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In Proc. Asian Conf. Comp. Vis. 2020; .

[13] Zhu Q, Wang Y, Chu X, Yang X, Zhong W. Multi-view coupled self-attention network for pulmonary nodules classification. In Proc. Asian Conf. Comp. Vis. 2022; 995–1009.

[14] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). 2009; 248–255.

[15] Huang G, Liu Z, Maaten LVD, Weinberger KQ. Densely connected convolutional networks. In IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). 2017; 2261–2269.

[16] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). 2016; 2818–2826.

[17] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). 2016; 770–778.

[18] Kingma D, Ba J. Adam: A method for stochastic optimization. Int Conf Learn Rep 12 2014;.

[19] Ahishali M, Degerli A, Yamac M, Kiranyaz S, Chowdhury MEH, Hameed K, Hamid T, Mazhar R, Gabbouj M. Advance warning methodologies for COVID-19 using chest x-ray images. IEEE Access 2021;9:41052–41065.

Address for correspondence:

Ilke Adalioglu

P.O. Box 553, FI-33014, Tampere Finland

ilke.adalioglu@tuni.fi