Uncertainty-Based Multi-modal Learning for Myocardial Infarction Diagnosis Using Echocardiography and Electrocardiograms

Yingyu Yang, Marie Rocher, Pamela Moceri, Maxime Sermesant

Uncertainty-based Multi-modal Learning for Myocardial Infarction Diagnosis using Echocardiography and Electrocardiograms

Yingyu Yang[1], Marie Rocher[2], Pamela Moceri[2], and Maxime Sermesant[1]

[1] Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France
[2] CHU de Nice - Hôpital Pasteur, Nice, France

Abstract. Medical devices used in cardiac diagnostics typically capture only one aspect of heart function. For instance, 2D B-mode echocardiography reveals the heart's anatomy and mechanical changes, while an electrocardiogram (ECG) records the heart's electrical activity from various positions. These examinations, essential for diagnosing cardiac diseases, are usually performed sequentially rather than simultaneously, providing complementary information for the final diagnosis. Recently, the integration of multi-modal information in AI research for healthcare has gained popularity, aiming for more robust diagnostic outcomes. However, the scarcity of publicly available multi-modal data for cardiac disease diagnosis poses a significant challenge to multi-modal learning and evaluation. In this study, we propose an uncertainty-based deep learning framework that utilizes unpaired data from different modalities to improve the diagnosis of myocardial infarction (MI) using both echocardiography and ECG data. Specifically, we trained two unimodal classification models incorporating uncertainty using public single-modal datasets. We then performed multi-modal classification using uncertainty-based decision fusion on a paired dataset, without the need for transfer learning or retraining. Our experiments demonstrated that uncertainty-based multi-modal decision fusion outperforms conventional fusion strategies by 4% in accuracy and unimodal models by 7% in accuracy. This approach is both flexible and data-efficient, making uncertainty-based multi-modal fusion a sustainable and strong solution for both unpaired and paired multi-modal classification.

Keywords: Multi-modal classification · Echocardiography · Electrocardiogram.

1 Introduction

Clinicians usually combine information from different examinations and measurements to make clinical decisions. However, most current AI research for healthcare simply considers one single modality, which does not profit from the complex and heterogeneous information that one can observe from patients using different imaging modalities, sensor devices, biochemical tests, etc. Multi-modal

machine learning, which seeks to model the interactions between different modalities, brings opportunities for improving the prevention, diagnosis and therapy in AI-enabled healthcare [1–4].

One challenge in biomedical multi-modal learning is to determine how to fuse information from different medical modalities for downstream tasks. Depending on when the fusion occurs, one can distinguish: early fusion and late fusion respectively. Early fusion combines the raw modality or extracted features at the input level according to certain fusion approaches, such as concatenation, multiplicative interaction [5], polynomial fusion [6], tensor fusion [7, 8], etc. Late fusion aggregates the prediction outputs of different modalities at the decision level (e.g. using majority voting, weighted voting etc.) to generate a final decision. Early fusion usually demands paired multi-modality data to explore detailed interaction strategies, while late fusion only need single modality outputs, thus being less demanding for paired data.

In this study, we focus on detecting myocardial infarction (MI) using both echocardiography (ECHO) and eletrocardiogram (ECG) data. Researchers have explored different multi-modal approaches for MI detection, such as combining ECG with demographic features [9], using images and clinical data together [10]. Very few have investigated the combination of ECHO and ECG, while ECHO and ECG can reveal different diagnosis characteristics of MI respectively [11]. In addition, with very limited paired multi-modal data by hand, we concentrate on how to improve the late fusion strategy which can leverage on the most confident modality.

The contribution of this paper is twofold. Firstly, we have adapted a trustworthy method to fuse decisions from different modalities by considering the uncertainty of each prediction. The proposed fusion strategy is efficient and flexible, capable of fully utilising public single-modal datasets and performing test-time multi-modal fusion when paired multi-modal samples are available. Secondly, our experiments on multi-modal myocardial infarction detection using both ECHO and ECG demonstrate superior performance compared to single-modal detection or conventional fusion methods. This suggests the potential of combining ECHO and ECG for robust cardiac diagnosis.

2 Method

We first introduce how to quantify the uncertainty for unimodal classification using evidential deep learning [12]. In the second part, we present test-time multi-modal fusion strategy that takes into account the uncertainty from each modality.

2.1 Evidential Deep Learning for Unimodal Classification

Uncertainty and the Theory of Evidence Evidential deep learning (EDL) quantifies the class probabilities and overall uncertainty in a unified theoretical framework [12]. Considering a K classification problem, it introduces an idea of

evidence e_k, which represents a measure of the amount of support for k^{th} class category collected from data input. Using the evidence, the belief of possible class label assignments b_k and an overall uncertainty mass u can be obtained through

$$b_k = \frac{e_k}{S} \text{ and } u = \frac{K}{S}, S = \sum_{i=1}^{K}(e_i + 1) \tag{1}$$

The sum of the $K+1$ mass values is one, i.e. $u + \sum_{k=1}^{K} b_k = 1$. Actually, EDL associates the belief of possible class label assignments (subjective opinion) with the parameters of a Dirichlet Distribution [13], i.e. $\alpha_k = e_k + 1$, including the belief that the truth label is equally likely (i.e., *"I do not know"* for uncertainty quantification).

The Dirichlet distribution is parameterised by K parameters $\alpha = [\alpha_1, ..., \alpha_K]$. Its probability density function (pdf) is given by

$$D(\mathbf{p}|\alpha) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{K} p_i^{\alpha_i - 1} & \text{for } \mathbf{p} \in \mathcal{S}_K, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathcal{S}_K$ represents the K-dimensional unit simplex $\mathcal{S}_K = \{\mathbf{p}| \sum_{i=1}^{K} p_i = 1 \text{ and } 0 \leq p_1, ..., p_k \leq 1\}$, and $B(\alpha)$ is the K-dimensional multinomial beta function. Given an opinion, the expected probability $\hat{p}_k$ for the k_{th} class category is the mean of the corresponding distribution, $\hat{p}_k = \frac{\alpha_k}{S}$.

The above relationship reveals that the higher the evidence e_k for k^{th} class is observed, the greater the class belief b_k and the corresponding Dirichlet parameter α_k will be. Similarly, when the total evidence observed from the input data is small, i.e. $\sum e_k$ is closer to 0 and $\alpha_k, k = 1, ..K$ are closer to 1, the uncertainty of the prediction becomes higher.

Learning to form opinions Evidential deep learning replaces the last *softmax* activation in neural network classifiers with non-negative activation, such as *ReLU*. The output of this final activation layer is taken as the evidence vector. It forms class belief masses and constitutes the parameters for the estimated Dirichlet distribution (illustrated in the upper right part of Figure 1).

We assume that y_i is a one-hot vector of ground truth classification label for input data x_i. The cross-entropy loss is usually used in conventional neural network classifiers:

$$\mathcal{L}^{CE} = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij} \log(p_{ij}), \tag{3}$$

where p_{ij} is the predicted probability for sample x_i belonging to class j. Under the theory of evidence and Dirichlet distribution assumption, we can compute the Bayes risk of cross-entropy loss function as

$$\mathcal{L}_i^{UC} = \int [\sum_{j=1}^{K} -y_{ij} \log(p_{ij})] \frac{1}{B(\alpha_i)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d\mathbf{p}_i = \sum_{j=1}^{K} y_{ij}(\psi(S_i) - \psi(\alpha_{ij})), \tag{4}$$

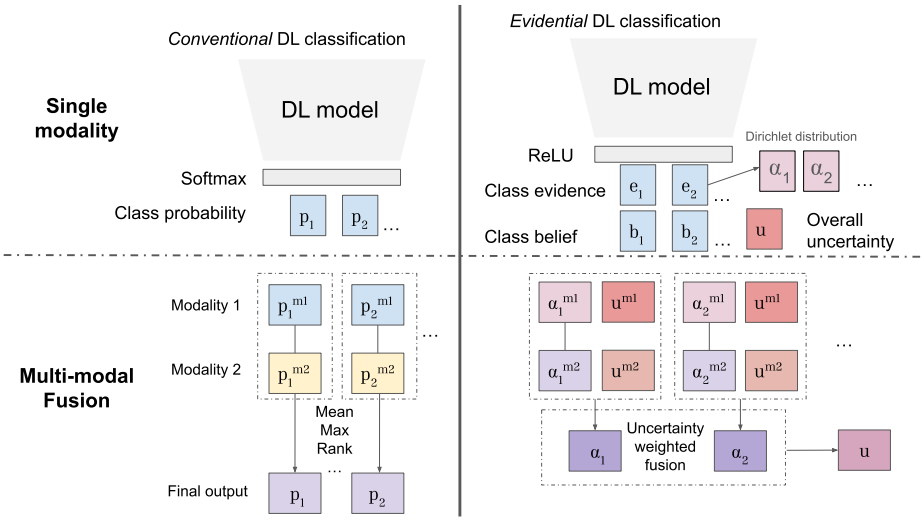

Fig. 1. Comparison of conventional fusion strategies and uncertainty based fusion.

where $\psi(\cdot)$ represents the *digamma* function.

The minimisation of the above loss does not guarantee that less evidence will be generated when the model predicts incorrect labels. To guide the network into learning zero total evidence for uncertain samples, a regularisation term is introduced. This term deploys a Kullback-Leibler divergence term to penalise the predictive Dirichlet distribution to be close to $D(\mathbf{p}|\mathbf{1})$.

$$KL[D(\mathbf{p}_i|\tilde{\alpha}_i)||D(\mathbf{p}_i|\mathbf{1})] = \log(\frac{\Gamma(\sum_{k=1}^{K}\tilde{\alpha}_{ik})}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\tilde{\alpha}_{ik})})+\sum_{k=1}^{K}(\tilde{\alpha}_{ik}-1)[\psi(\tilde{\alpha}_{ik})-\psi(\sum_{k=1}^{K}\tilde{\alpha}_{ij})], \tag{5}$$

where $\Gamma(\cdot)$ represents the *gamma* function and $\mathbf{1}$ refers to a K-dim vector of all ones. And $\tilde{\alpha}_i = y_i + (1 - y_i) \odot \alpha_i$, $\tilde{\alpha}_i$ are the parameters that has removed non-misleading evidence.

Thus, the final loss function for evidential deep learning neural networks reads:

$$\mathcal{L} = \sum_{i=1}^{N}\mathcal{L}_i^{UC} + \lambda_t \sum_{i=1}^{N} KL[D(\mathbf{p}_i|\tilde{\alpha}_i)||D(\mathbf{p}_i|\mathbf{1})], \tag{6}$$

where $\lambda_t = \min(1, t/T) \in [0, 1]$ is a balancing coefficient for regularisation and t represents the current training epoch.

2.2 Multi-modal Fusion with Uncertainty

Considering two independent sets of evidence values $\{e_k^1\}_{k=1}^K$ and $\{e_k^2\}_{k=1}^K$, the corresponding parameters of Dirichlet distribution are $\{\alpha_k^1 = e_k^1 + 1\}_{k=1}^K$ and

$\{\alpha_k^2 = e_k^2 + 1\}_{k=1}^K$. We propose to fuse opinions from all modalities through uncertainty-weighted fusion (illustrated in the lower right part of Figure 1):

$$\alpha_k = (1 - u^1)\alpha_k^1 + (1 - u^2)\alpha_k^2, u = \frac{K}{\sum \alpha_k} \tag{7}$$

3 Experiments and Results

3.1 Datasets

Two independent datasets of ECHO and ECG are involved in this study:

- HMC-QU dataset [14]: contains 130 long-axis 2-chamber view sequences (68 with MI) and 162 long-axis 4-chamber view sequences (93 with MI).
- PTB-XL dataset [15]: contains 12-lead ECG data (with 7185 samples of healthy controls and 2955 samples with 100%-certain MI).

In addition, a small number of paired ECHO and ECG data were collected retrospectively from Nice University hospital (CHU-Nice). This dataset contains data from 56 patients, with 56 paired data of ECG and 4-chamber view ECHO, along with 50 paired data of ECG and 2-chamber view ECHO. Detailed dataset information is listed in Table 1.

Table 1. Dataset statistics. *2ch: 2 chambers view, 4ch: 4 chambers view.*

Dataset	Modality	MI	non-MI	Total
HMC-QU	ECHO 2ch	68	62	130
HMC-QU	ECHO 4ch	93	69	162
PTB-XL	ECG 12-lead	2955	7185	10140
CHU-Nice	ECHO(2ch) + ECG	33	17	50
CHU-Nice	ECHO(4ch) + ECG	36	20	56

3.2 Experiments

We first extracted interpretable features from ECHO data in HMC-QU dataset and from ECG data in PTB-XL dataset (refer to Figure 2(a)). For ECHO data, we used a motion tracking model [16] to predict the temporal motion of 10 key points around the myocardium (refer to Figure 2(b)). From the temporal motion trace, we constructed a 40-dimension vector which composed of mean and standard deviation of the 10 key points along x- and y- axis. For ECG data, we followed the work [17] to decompose single-heartbeat ECG signal into 5 sub-components and used the predicted 21 parameters to constitute a 21×12-dimension vector as ECG features. We trained single modality models with 10-fold cross validation for ECG data (PTB-XL) and 5-fold cross validation for

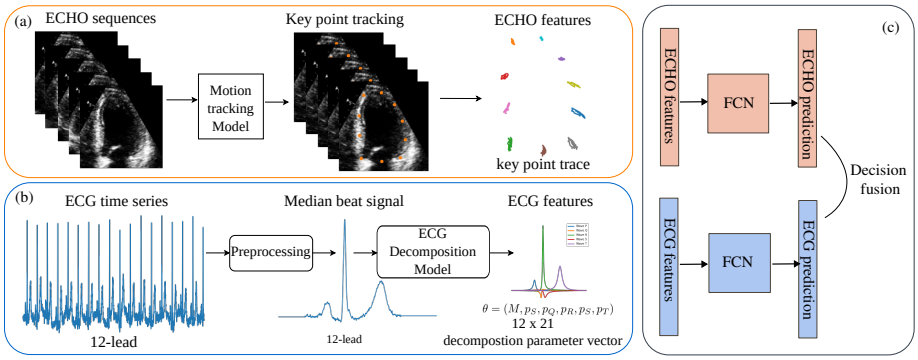

Fig. 2. ECHO and ECG feature extraction pipeline.

ECHO data (HMC-QU) using a 4-layer fully connected network (FCN) respectively. The baseline single modality model without uncertainty (w/o UC) was trained using cross-entropy loss (Equation 3) and the uncertainty model (w UC) with Equation 6.

For models without uncertainty (w/o UC), we assumed that the output of FCN after Sigmoid function was p_c^k and the prediction class was $\bar{y^k}$, where $c \in \{0, 1\}$ refers to class and $k \in \{0, 1\}$ refers to modality. The MI class was set to label 1. The following fusion strategies were included in our study:

- Max fusion: $p_c = \max\{p_c^k, k = 1, ..., K\}$, $\bar{y} = \arg\max_c p_c$;
- Mean fusion: $p_c = \text{mean}\{p_c^k, k = 1, ..., K\}$, $\bar{y} = \arg\max_c p_c$;
- Rank fusion: $\bar{y} = (\sum_k \bar{y^k}) \geq 1$;
- Multiply fusion: $p_c = \prod_k p_c^k$, $\bar{y} = \arg\max_c p_c$.

The multi-modal fusion with uncertainty was performed according to Equation 7.

3.3 Implementation

The uncertainty model for ECHO and ECG were trained using the following hyper-parameters:

- ECG: learning rate 0.01, batch size 512, total epochs 200, $T = 10$ for λ_t;
- ECHO: learning rate 0.0001, batch size 8, total epochs 200, $T = 50$ for λ_t.

We chose the model with the best validation loss during training.

3.4 Results

First, we present the cross validation results on HMC-QU and PTB-XL dataset in Table 2 and Table 3. Although evidential deep learning (EDL) model demonstrated reduced performance compared with model trained using standard cross-entropy loss, its performance was comparable when using mixed 2-chamber and

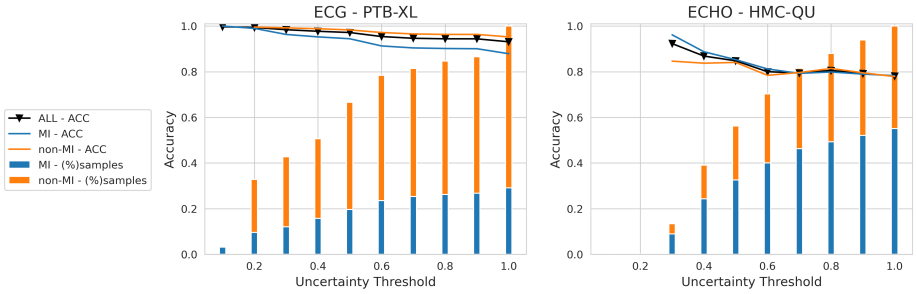

Fig. 3. The change of prediction accuracy with respect to uncertainty threshold on PTB-XL ECG dataset and HMC-QU ECHO dataset (2CH/4CH mixed). Bar plots represent the percentage of samples kept under varying uncertainty thresholds.

4-chamber (2CH/4CH mixed) views together (292 samples in total). We obtained a similar observation on ECG classification using uncertainty-based loss. Figure 3 shows how the test accuracy changes when EDL only keeps predictions under varying uncertainty thresholds. Notably, on both datasets, the accuracy increased as the uncertainty threshold decreased, which reflected the effectiveness of uncertainty quantification predicted by the model.

Table 2. ECHO classification: 5-fold CV results on HMC-QU dataset. *w/o UC: without uncertainty, w UC: with uncertainty.*

Method	View	Accuracy	Sensitivity	Specificity
KNN [14]	2CH	0.75	0.72	0.77
Ours (w/o UC)	2CH	**0.78**	**0.74**	0.82
Ours (w UC)	2CH	0.72	0.59	**0.85**
Random Forest [14]	4CH	**0.86**	**0.84**	**0.85**
Ours (w/o UC)	4CH	0.81	0.82	0.80
Ours (w UC)	4CH	0.82	0.83	0.81
Ours (w/o UC)	2CH + 4CH (mixed)	0.78	0.78	0.79
Ours (w UC)	2CH + 4CH (mixed)	0.78	0.78	0.78

Table 3. ECG classification: 10-fold CV results on PTB-XL dataset. *w/o UC: without uncertainty, w UC: with uncertainty.*

Method	Lead	Accuracy	Sensitivity	Specificity
SVM [17]	12-lead	0.96	0.93	0.96
Ours (w/o UC)	12-lead	0.95	0.89	0.97
Ours (w UC)	12-lead	0.93	0.88	0.95

Table 4. Evaluation on CHU-Nice dataset (with 2-chamber view and 4-chamber view mixed together, in total 106 paired samples). *w/o UC: without uncertainty, w UC: with uncertainty.*

Method	Modality	Accuracy	Sensitivity	Specificity
Ours (w/o UC)	ECG	0.69	0.84	0.43
Ours (w/o UC)	ECHO	**0.75**	0.75	**0.76**
Max Fusion	ECG + ECHO	0.73	0.81	0.57
Mean fusion	ECG + ECHO	**0.75**	0.86	0.54
Rank fusion	ECG + ECHO	0.73	**0.96**	0.30
Multiply fusion	ECG + ECHO	0.68	0.87	0.32
Ours (w UC)	ECG	0.72	**0.86**	0.48
Ours (w UC)	ECHO	0.71	0.68	**0.76**
Uncertain fusion	ECG + ECHO	**0.79**	0.83	0.73

We show the test-time multi-modal fusion evaluation on the CHU-Nice dataset in Table 4. The performance of conventional fusion (upper part) was limited by the best performing modality, in our case, by the ECHO modality. The mean fusion strategy outperformed the other conventional methods, with a slight improvement in sensitivity but significant reduction in specificity due to the erroneous output of the ECG prediction. In the lower part of Table 4, we observe that uncertainty-based fusion improved largely the prediction accuracy compared to single modalities with uncertainty (by 7%). In addition, this approach well combined the advantages of each modality: higher sensitivity than single ECHO output and higher specificity than single ECG output, with only a slight decrease compared with the best value of single modality outputs. As evidenced by the fusion results, uncertainty-based fusion generated multi-modal prediction according to the most trustworthy modality, therefore improving the final prediction of diagnosis.

4 Conclusion

In this study, we explored various multi-modal late fusion strategies and found that uncertainty-based fusion outperformed conventional methods, improving classification accuracy by 4%. This approach, utilizing single-modality evidential deep learning, assessed the uncertainty of each modality's prediction to prioritize the most reliable input for the final decision. Additionally, it required no sampling steps and was straightforward to implement with deep learning techniques. The test-time fusion setting maximized the use of large public single-modality datasets while preserving valuable paired multi-modal data for evaluation. Despite promising preliminary results, several limitations demand further investigation. First, we need to quantify the impact of error propagation through feature extraction on downstream classification tasks. Second, the uncertainty predicted within the evidential framework is not fully calibrated, necessitating the incorporation of uncertainty calibration for both modalities before fusion.

Acknowledgements This work has been supported by the French government through the National Research Agency (ANR) Investments in the Future with 3IA Côte d'Azur (ANR-19-P3IA-0002) and by Inria PhD funding. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

1. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. Nature Medicine 28(9), 1773–1784 (2022)
2. Soto, J.T., Weston Hughes, J., Sanchez, P.A., Perez, M., Ouyang, D., Ashley, E.A.: Multimodal deep learning enhances diagnostic precision in left ventricular hypertrophy. European Heart Journal-Digital Health 3(3), 380–389 (2022)
3. Goto, S., Solanki, D., John, J.E., Yagi, R., Homilius, M., Ichihara, G., Katsumata, Y., Gaggin, H.K., Itabashi, Y., MacRae, C.A., et al.: Multinational federated learning approach to train ecg and echocardiogram models for hypertrophic cardiomyopathy detection. Circulation 146(10), 755–769 (2022)
4. Puyol-Antón, E., Sidhu, B.S., Gould, J., Porter, B., Elliott, M.K., Mehta, V., Rinaldi, C.A., King, A.P.: A multimodal deep learning model for cardiac resynchronisation therapy response prediction. Medical Image Analysis 79, 102465 (2022)
5. Jayakumar, S.M., Czarnecki, W.M., Menick, J., Schwarz, J., Rae, J., Osindero, S., Teh, Y.W., Harley, T., Pascanu, R.: Multiplicative interactions and where to find them (2020)
6. Kefalas, T., Vougioukas, K., Panagakis, Y., Petridis, S., Kossaifi, J., Pantic, M.: Speech-driven facial animation using polynomial fusion of features. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3487–3491. IEEE (2020)
7. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)
8. Hou, M., Tang, J., Zhang, J., Kong, W., Zhao, Q.: Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
9. Xiao, R., Ding, C., Hu, X., Clifford, G.D., Wright, D.W., Shah, A.J., Al-Zaiti, S., Zègre-Hemsey, J.K.: Integrating multimodal information in machine learning for classifying acute myocardial infarction. Physiological Measurement 44(4), 044002 (2023)
10. Sharma, R., Eick, C.F., Tsekos, N.V.: Sm2n2: A stacked architecture for multimodal data and its application to myocardial infarction detection. In: Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11. pp. 342–350. Springer (2021)
11. Thygesen, K., Alpert, J.S., Jaffe, A.S., Chaitman, B.R., Bax, J.J., Morrow, D.A., White, H.D., Group, E.S.D.: Fourth universal definition of myocardial infarction (2018). European Heart Journal 40(3), 237–269 (08 2018)

12. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems 31 (2018)
13. Jsang, A.: Subjective Logic: A formalism for reasoning under uncertainty. Springer Publishing Company, Incorporated (2018)
14. Degerli, A., Kiranyaz, S., Hamid, T., Mazhar, R., Gabbouj, M.: Early myocardial infarction detection over multi-view echocardiography. Biomedical Signal Processing and Control 87, 105448 (2024)
15. Wagner, P., Strodthoff, N., Bousseljot, R.D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T.: Ptb-xl, a large publicly available electrocardiography dataset. Scientific data 7(1), 154 (2020)
16. Yang, Y., Sermesant, M.: Unsupervised polyaffine transformation learning for echocardiography motion estimation. In: International Conference on Functional Imaging and Modeling of the Heart. pp. 384–393. Springer (2023)
17. Yang, Y., Rocher, M., Moceri, P., Sermesant, M.: Explainable electrocardiogram analysis with wave decomposition: Application to myocardial infarction detection. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 221–232. Springer (2022)