Multi-scale Feature Learning with CNN-RNN-Attention Framework for ECG-based Cancer Therapy-Related Cardiac Dysfunction Detection

Natsu Suyama[1], Akira Furui[1], Takio Kurita[1], and Kazuko Tajiri[2]

***Abstract*— Cancer therapy-related cardiac dysfunction (CTRCD) is an increasingly significant concern due to cardiac function deterioration caused by anticancer drug side effects. While echocardiography is the conventional diagnostic method for CTRCD, its accuracy heavily depends on operator expertise and the procedure is both time-consuming and costly. Electrocardiogram (ECG), being more accessible and easier to measure, presents a promising lower-cost alternative. In this paper, we propose a deep learning model for CTRCD detection from ECG signals. Our model integrates convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to capture both local and global ECG features, while incorporating an attention mechanism to comprehensively learn feature importance. To enhance model interpretability, we visualize the attention weights to identify ECG features that significantly contribute to the classification decision. Through extensive ablation studies using standard 12-lead ECG data, we demonstrate the effectiveness of our proposed architecture. This work is expected to contribute to the development of cost-effective and reliable diagnostic tools for monitoring cardiac side effects during cancer treatment.**

I. Introduction

Managing adverse effects is a critical challenge in cancer treatment. Cancer therapy-related cardiac dysfunction (CTRCD), characterized by myocardial damage and heart failure associated with cancer treatment [1], affects approximately 10% of cancer patients treated with anthracyclines [2]. Once CTRCD occurs, it can lead to heart failure and significantly deteriorate patients' quality of life. Early detection of CTRCD is crucial as it enables timely interventions for managing the side effects of anticancer drugs and reduces the overall burden of cancer treatment on patients.

The current standard for CTRCD diagnosis relies on measuring the left ventricular ejection fraction (LVEF) through echocardiography [1]. However, this approach presents several limitations in clinical practice: accuracy and reproducibility of echocardiographic assessment heavily depend on the echo-image quality, and obtaining optimal images requires specialized expertise [3]. In contrast, electrocardiogram (ECG) measurement offers a more accessible and cost-effective alternative, providing valuable information about cardiac electrical activity through waveform analysis. The potential of ECG as a screening tool for CTRCD is particularly promising for regular monitoring during cancer treatment.

Recent studies have explored various applications of deep learning to ECG data, including the detection of reduced LVEF [4], [5]. These approaches primarily employ convolutional neural networks (CNNs) for local feature extraction. While CNNs can effectively capture hierarchical features as the network deepens, they have limitations in analyzing ECG signals comprehensively. ECG analysis requires understanding both local patterns and global temporal relationships, such as PP intervals and QRS complexes, along with their long-term continuity and variability. Moreover, for practical clinical adoption, it is essential to establish the reliability of deep learning model outputs in diagnostic decisions.

To address these challenges, we propose a deep learning framework that combines CNN-based feature extraction with recurrent neural networks (RNNs) to capture both local and global temporal patterns in ECG signals. Our model incorporates an attention mechanism to identify and prioritize the most relevant ECG features for CTRCD detection. By visualizing the attention weights, our model provides insights into which regions of the ECG signals contribute most significantly to the detection process. This approach aims to contribute to the development of cost-effective ECG analysis for CTRCD detection while maintaining interpretability of the automated analysis process.

II. Proposed Method

Fig. 1 presents an overview of the proposed deep learning model for CTRCD prediction. The model architecture consists of three main components: a CNN block for feature extraction, an RNN block for temporal modeling, and an attention block for feature importance weighting. The CNN block extracts both temporal and spatial features from the ECG data in parallel, which are then processed by the RNN block to capture long-term dependencies. Then, the attention block identifies and emphasizes the most relevant temporal features. After processing through a classifier consisting of fully connected layers, the model outputs the final CTRCD prediction.

A. Notation

Let T denote the number of time frames and D represent the number of ECG electrodes. For the nth training patient data ($n = 1, 2, \ldots, N$), we define $\mathbf{X}_n \in \mathbb{R}^{D \times T}$ as their ECG data matrix. The binary label $y_n \in \{0, 1\}$ indicates the presence (1) or absence (0) of CTRCD, while $\hat{y}_n$ represents

This work was supported by the National Cancer Center Research and Development Fund (2023-A-12).

[1]Graduate School of Advanced Science and Engineering, Hiroshima University, Higashi-hiroshima, Japan. {natsusuyama, akirafurui,tkurita}@hiroshima-u.ac.jp

[2]Department of Cardiology, National Cancer Center Hospital East, Kashiwa, Japan. ktajiri@east.ncc.go.jp

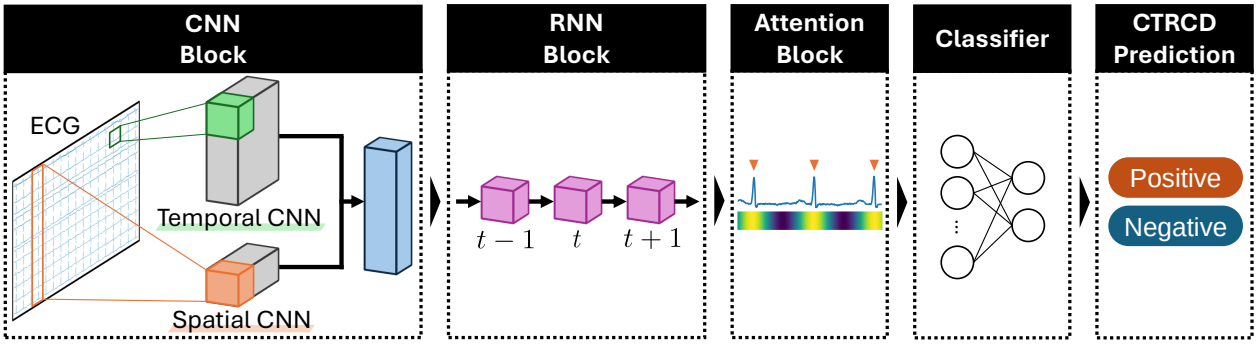

Fig. 1. Overview of the proposed deep learning framework integrating CNN, RNN, and attention mechanisms for CTRCD detection

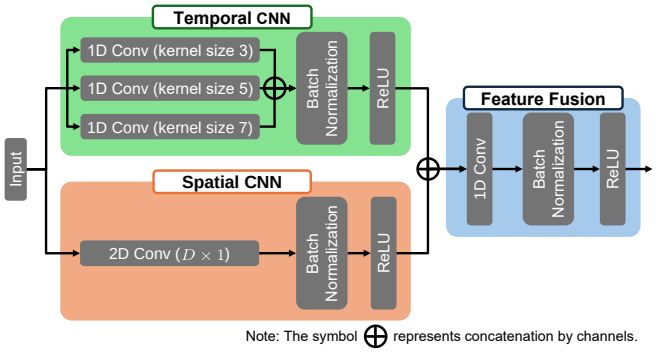

Fig. 2. Detailed architecture of the CNN block showing parallel temporal and spatial feature extraction paths followed by feature fusion

the model's predicted output. We collect these labels into vectors $\mathbf{y} = [y_1, y_2, \ldots, y_N]^\top$ and $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N]^\top$ for all N patients.

B. CNN Block

As illustrated in Fig. 2, the CNN block comprises three key components designed to capture different aspects of the ECG signals:

- **Temporal CNN:** This component processes temporal features independently for each electrode using parallel 1D convolution with varying kernel sizes (3, 5, and 7). The features extracted from different kernel sizes are concatenated along the channel dimension. This multi-scale approach enables the extraction of features at different temporal resolutions, capturing both short-term and medium-term patterns in the ECG signals.
- **Spatial CNN:** This component applies 2D convolution across all electrodes simultaneously, enabling the model to learn spatial correlations between different ECG leads. This is crucial for capturing the relationships between electrical activities recorded from electrodes at different locations on the heart.
- **Feature fusion:** The temporal and spatial features are concatenated along the channel dimension, followed by a 1D convolution to integrate the information from both branches. This fusion strategy allows the model to combine local temporal pattens with global spatial relationships.

In all convolution operations throughout the CNN block, batch normalization and ReLU activation are applied after each convolution layer.

C. RNN Block

In this block, temporal features are extracted from the features $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_t, \ldots, \tilde{\mathbf{x}}_T]$ obtained by the CNN block, where $\tilde{\mathbf{x}}_t \in \mathbb{R}^{D'}$ and D' is the feature dimension. We employ the gated recurrent unit (GRU) [6], a variant of RNN, which outputs hidden states $\mathbf{h}_t \in \mathbb{R}^H$, where H is the hidden dimension. The GRU operations are defined as follows:

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \tilde{\mathbf{x}}_t + \mathbf{U}_r \mathbf{h}_{t-1}), \tag{1}$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \tilde{\mathbf{x}}_t + \mathbf{U}_z \mathbf{h}_{t-1}), \tag{2}$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_{\tilde{h}} \tilde{\mathbf{x}}_t + \mathbf{U}_{\tilde{h}} (\mathbf{r}_t \odot \mathbf{h}_{t-1})), \tag{3}$$

$$\mathbf{h}_t = (\mathbf{1} - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t, \tag{4}$$

where σ denotes the logistic sigmoid function, and $\mathbf{W}_r, \mathbf{W}_z, \mathbf{W}_{\tilde{h}} \in \mathbb{R}^{H \times D'}$ and $\mathbf{U}_r, \mathbf{U}_z, \mathbf{U}_{\tilde{h}} \in \mathbb{R}^{H \times H}$ are learnable weight parameters.

D. Attention Block

The attention block implements temporal attention [7], enabling the model to focus on relevant parts of the input sequence. The context vector $\mathbf{v}$ is computed as a weighted sum of the features $\mathbf{h}_t$ obtained from the GRU block:

$$\mathbf{v} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t, \tag{5}$$

where the attention weight $\alpha_t \in [0, 1]$ for each time step t is calculated through a feed-forward transformation followed by softmax normalization:

$$\alpha_t = \text{Softmax}(\mathbf{m}_t) = \frac{\exp(\mathbf{w}_\alpha^\top \mathbf{m}_t)}{\sum_{i=1}^{T} \exp(\mathbf{w}_\alpha^\top \mathbf{m}_i)}, \tag{6}$$

$$\mathbf{m}_t = \tanh(\mathbf{W}_h \mathbf{h}_t). \tag{7}$$

Here, $\mathbf{W}_h \in \mathbb{R}^{H \times H}$ and $\mathbf{w}_\alpha \in \mathbb{R}^H$ are learned parameters. The attention weights α_t not only help the model focus

on the most relevant temporal features but also provide interpretability by revealing which parts of the ECG signal contribute most significantly to the model's predictions.

E. Classifier and Loss Function

The classifier consists of a two-layer multilayer perceptron (MLP) that processes the context vector from the attention block. The first layer transforms the feature representation, followed by a ReLU activation function. The second layer then maps to a single output unit with a sigmoid activation, producing the final prediction probability for CTRCD.

For training the model, we employ the cross-entropy loss function $\mathcal{L}$ as follows:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log \hat{y}_n + (1 - y_n)\log(1 - \hat{y}_n)]. \quad (8)$$

This loss function is used to calculate the difference between the correct label y_n and the output of the model $\hat{y}_n$. The model parameters are optimized by minimizing this loss function.

III. Experiments

This study was approved by the review board of the National Cancer Center (research project no. 2022-152). The requirement of informed consent was waived because this study was conducted according to a retrospective chart review protocol.

A. Dataset

The ECG data were collected from 616 patients (87 males, 529 females) treated with anthracyclines at the National Cancer Center Japan. Although multiple ECG recordings were available for each patient, we selected only one recording per patient: for CTRCD-positive cases, we selected one ECG recording per patient from the date of CTRCD detection, while for negative cases, we randomly selected one from their available longitudinal measurements. The final dataset comprised 557 patients (43 positive, 514 negative). The ECG signals were recorded at 500 Hz for 11 ± 0.2 s. To ensure uniform signal length across all samples, we truncated each recording to 10 s ($T = 5000$ samples) by removing the excess data from the end of each signal.

B. Experimental Design

Our proposed model consists of three main components: a CNN block (combining temporal and spatial CNNs), a GRU layer, and an attention mechanism. We conducted two ablation studies to validate the effectiveness of each component. The first ablation study focused on the architecture of the CNN block by replacing our proposed temporal-spatial combination with simplified variants: temporal CNN only or spatial CNN only, while retaining the GRU layer and the attention mechanism. The second ablation study examined the necessity of sequential processing components by removing either the attention mechanism alone or both the attention mechanism and GRU layer from our proposed model, while maintaining the temporal-spatial CNN block.

To interpret the model's behavior, we visualized the attention weights (α_t) as a color map, where the brightness indicates the contribution magnitude at each time point t ($1 \leq t \leq T$). By comparing these weights with the original ECG signals, we can analyze which temporal features the model emphasized in its feature extraction process.

We employed stratified 5-fold cross-validation with 10 different seeds. To address the class imbalance, we implemented oversampling during training by maintaining equal proportions of positive and negative cases within each batch, allowing repeated sampling cases.

C. Implementation Details

The CNN block consists of temporal CNNs with 20 output channels for each kernel size (3, 5, and 7) and a spatial CNN with 30 output channels. The feature fusion layer outputs 64-dimensional features, matching the hidden dimension of the subsequent GRU layer (i.e. $H = D' = 64$). In the classifier, the hidden layer consists of 64 units, with dropout ($p = 0.5$) applied before and after the hidden layer.

The model was trained for 100 epochs using Adam optimizer with a learning rate of 0.001 and a batch size of 64. All network parameters were initialized using the default PyTorch initialization scheme.

D. Evaluation Metrics

We evaluated our model using five standard performance metrics for binary classification: the area under the receiver operating characteristic curve (ROC-AUC), accuracy, precision, sensitivity, and specificity. ROC-AUC evaluates the model's discriminative ability by measuring the area under the curve of true positive rate versus false positive rate, with a range from 0 to 1. Accuracy represents the overall correct classification rate, while precision measures the reliability of positive prediction. Sensitivity and specificity quantify the model's ability to correctly identify positive and negative cases, respectively.

IV. Results and Discussion

A. Quantitative Evaluation

The ablation study results for the CNN components (Table I) reveal that temporal CNN achieves slightly better performance than spatial CNN, particularly in precision. This performance difference likely stems from the structural characteristics of ECG data, where temporal dimensions ($T = 5000$ samples) substantially exceed spatial dimensions ($D = 12$ electrodes). This finding aligns with conventional ECG diagnostic approaches, where analysis typically focuses on temporal waveform patterns at individual leads rather than cross-lead relationships at each specific time points. While temporal features demonstrated stronger individual performance, the integration of both temporal and spatial CNNs yielded more robust results across all metrics, indicating the complementary value of both feature types in ECG analysis.

Table II shows the results of the ablation study examining the contributions of GRU and attention mechanisms in the

TABLE I

RESULTS OF ABLATION STUDY ON CNN ARCHITECTURE

CNN architecture		Performance				
Temporal CNN	Spatial CNN	ROC-AUC	Accuracy	Precision	Sensitivity	Specificity
✓		0.905 ± 0.016	0.931 ± 0.008	0.596 ± 0.074	0.507 ± 0.054	0.967 ± 0.009
	✓	0.902 ± 0.017	0.925 ± 0.005	0.547 ± 0.028	0.480 ± 0.046	0.963 ± 0.005
✓	✓	**0.912 ± 0.012**	**0.933 ± 0.005**	**0.607 ± 0.029**	**0.515 ± 0.038**	**0.968 ± 0.003**

TABLE II

RESULTS OF ABLATION STUDY ON SEQUENTIAL PROCESSING COMPONENTS

Model	Performance				
	ROC-AUC	Accuracy	Precision	Sensitivity	Specificity
CNN only	0.491 ± 0.080	0.870 ± 0.048	0.494 ± 0.098	0.487 ± 0.111	0.902 ± 0.059
CNN + GRU	0.541 ± 0.100	0.878 ± 0.014	0.420 ± 0.040	**0.776 ± 0.058**	0.886 ± 0.015
CNN + GRU + Attention	**0.912 ± 0.012**	**0.933 ± 0.005**	**0.607 ± 0.029**	0.515 ± 0.038	**0.968 ± 0.003**

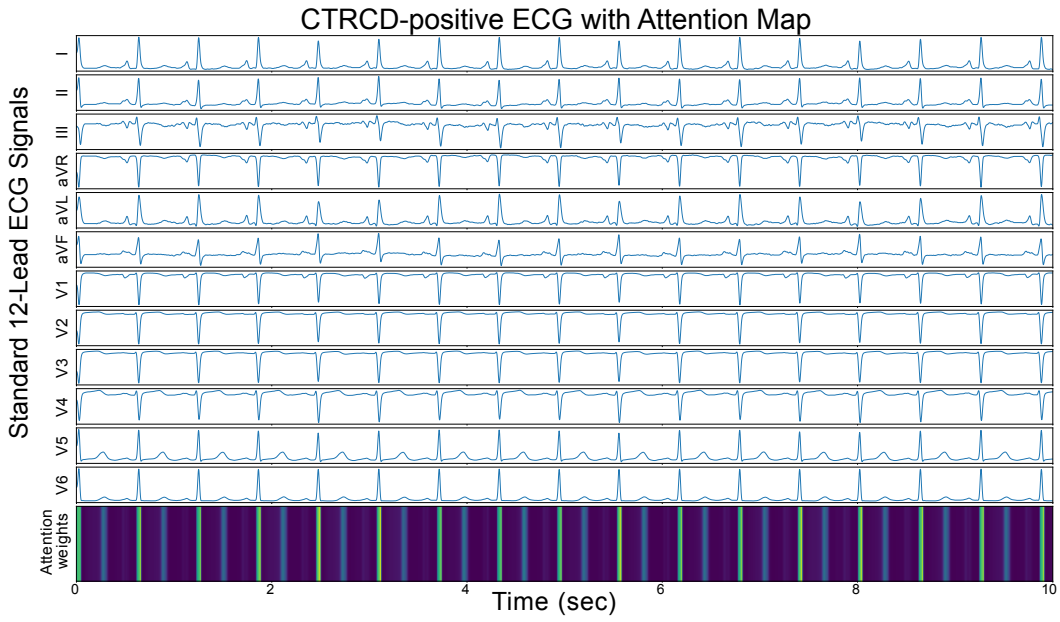

Fig. 3. Visualization of standard 12-lead ECG signals with attention weights for CTRCD-positive case. The attention map (bottom) shows the weight distribution α_t, where brighter colors indicate higher attention values.

proposed architecture. The CNN alone showed limited performance, while the addition of GRU improved sensitivity but at the cost of precision. The complete model with both GRU and attention mechanisms led to substantial improvements in ROC-AUC and accuracy compared to both the CNN only and CNN+GRU configurations. This enhancement suggests that the attention mechanism effectively captures relevant temporal patterns in the ECG signal, contributing to more accurate discrimination between positive and negative cases of CTRCD.

B. Qualitative Evaluation

The visualization of attention weights α_t for both CTRCD positive and negative cases revealed interesting patterns in how the model processes ECG signals (Fig. 3 and Fig. 4). The attention weights showed higher values in regions corresponding to QRS complexes and T waves, indicating that the model particularly focuses on these waveform components in its feature extraction process.

The identified regions of high attention weights align with established clinical knowledge of cardiac function. T waves, which reflect ventricular repolarization, as known to indicate abnormal ventricular activity in various cardiac conditions, including hypertrophy and cardiomyopathy [8], [9]. Similarly, QRS complexes represent ventricular depolarization and provide direct insight into ventricular myocyte activity [10]. The emphasis on both these components through attention weights is particularly relevant given that CTRCD is fundamentally defined by changes in LVEF.

These clinically meaningful attention patterns help explain the substantial performance improvement observed

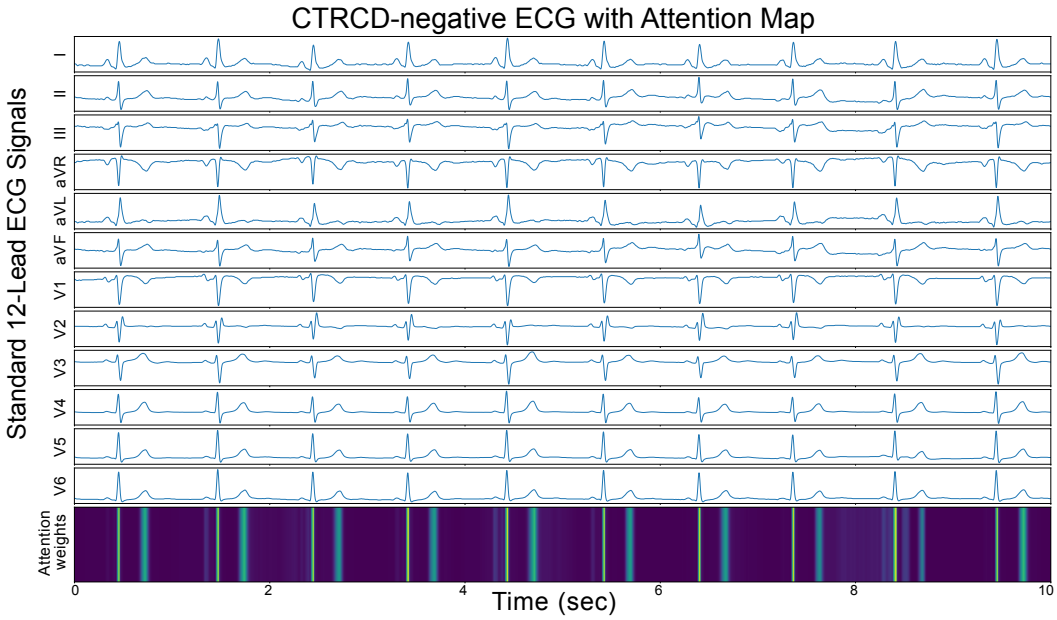

Fig. 4. Visualization of standard 12-lead ECG signals with attention weights for CTRCD-negative case. The attention map (bottom) shows the weight distribution α_t, where brighter colors indicate higher attention values.

when incorporating the attention mechanism (Table II). The automatic focus on these physiologically significant ECG components, validating our architectural design choice, resulted in notable increases in ROC-AUC (from 0.541 to 0.912) and accuracy (from 0.878 to 0.933). The highlighted regions corresponding to both T waves and QRS complexes suggest an evaluation of ventricular function during both relaxation and contraction phases, enabling a comprehensive assessment for CTRCD detection.

V. Conclusion

In this paper, we proposed a deep learning model for predicting CTRCD from ECG signals. Our method effectively combines CNN and RNN architectures to extract both local and global features from ECG data, while incorporating an attention mechanism to identify time-dependent pattens crucial for diagnosis. The model demonstrated promising performance, achieving an ROC-AUC of 0.912, an accuracy of 0.933, and a precision of 0.607. Through visualization of the attention mechanism, we identified that the model particularly focuses on QRS complexes and T waves, which aligns with clinical understanding of cardiac dysfunction detection.

Despite these encouraging results, our study has some limitations. The model exhibited an imbalance between sensitivity (0.515) and specificity (0.968), indicating room for improvement in detecting CTRCD-positive cases. Furthermore, while we implemented and visualized the attention mechanism for temporal features, our analysis did not extend to identifying the relative importance of different ECG leads.

Future work should focus on two key areas: optimizing the classification threshold to achieve a better balance between sensitivity and specificity, and extending the attention mechanism to analyze the contributions of individual ECG leads. With these refinement, our approach has the potential to enhance ECG-based CTRCD detection and contribute to more cost-effective, accessible cardiac monitoring in cancer treatment.

References

[1] A. R. Lyon *et al.*, "2022 ESC Guidelines on cardio-oncology developed in collaboration with the European Hematology Association (EHA), the European Society for Therapeutic Radiology and Oncology (ESTRO) and the International Cardio-Oncology Society (IC-OS): Developed by the task force on cardio-oncology of the European Society of Cardiology (ESC)," *European Heart Journal*, vol. 43, no. 41, pp. 4229–4361, 08 2022.

[2] D. Cardinale *et al.*, "Early detection of anthracycline cardiotoxicity and improvement with heart failure therapy," *Circulation*, vol. 131, no. 22, pp. 1981–1988, 2015.

[3] T. Onishi *et al.*, "Practical guidance for echocardiography for cancer therapeutics-related cardiac dysfunction," *Journal of Echocardiography*, vol. 19, pp. 1–20, 2021.

[4] R. Yagi *et al.*, "Artificial intelligence-enabled prediction of chemotherapy-induced cardiotoxicity from baseline electrocardiograms," *Nature Communications*, vol. 15, no. 1, p. 2536, 2024.

[5] Z. I. Attia *et al.*, "Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram," *Nature Medicine*, vol. 25, no. 1, pp. 70–74, 2019.

[6] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Oct. 2014, pp. 1724–1734.

[7] D. Bahdanau *et al.*, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014, Presented at ICLR 2015.

[8] V. M. Meijborg *et al.*, "Electrocardiographic T Wave and its Relation With Ventricular Repolarization Along Major Anatomical Axes," *Circulation: Arrhythmia and Electrophysiology*, vol. 7, no. 3, pp. 524–531, 2014.

[9] K. Yodogawa, "Approach to T wave Variability," *Japanese Journal of Electrocardiology*, vol. 42, no. 2, pp. 103–108, 2022.

[10] D. G. Strauss and D. D. Schocken, *Marriott's Practical Electrocardiography*. Wolters Kluwer Health, Inc., 2020.