ECG-Based Detection of Acute Myocardial Infarction using a Wrist-Worn Device

Karolina Jančiulevičiūtė, Daivaras Sokas, Justinas Bacevičius, Leif Sörnmo, *Life Fellow*, *IEEE*, and Andrius Petrėnas

***Abstract*— Background: A wrist-worn wearable device for acquiring limb and chest ECG leads (wECG) may constitute a promising approach to detection of acute myocardial infarction (AMI). However, it remains to be demonstrated whether the information conveyed by the wECG is sufficient for AMI detection.**
Objective: To explore explainable machine learning models for detecting AMI using the wECG.
Methods: Two types of machine learning models are explored: a convolutional neural network (CNN) using the raw ECG as input and a gradient-boosting decision tree (GBDT) using clinically informative features. 123 participants were included, divided into patients with AMI, patients with other cardiovascular diseases, and healthy individuals. A wrist-worn device equipped with three biopotential electrodes was used to acquire two ECG leads with a single touch: limb lead I and another lead involving a specific body site, i.e., either the V3 or V5 electrode positions, or the abdomen.
Results: The best performance on the test dataset is obtained using models that incorporate all four leads. The CNN model performs slightly better than the GBDT model, with a sensitivity of 0.77 and specificity of 0.75 compared to 0.77 and 0.72, respectively. When distinguishing between AMI and healthy participants, the specificity increases to 0.94 for the CNN model and 0.90 for the GBDT model. Feature importance analysis shows that the GBDT model primarily relies on the J point, while the CNN model primarily relies on the QRS complex.
Conclusions: wECG-based AMI detection shows considerable promise in out-of-hospital settings. However, caution is needed as CNN explanations rarely agree with the ECG intervals typically analyzed in clinical practice.

***Index Terms*— Wearable device, interpretability, explainability, decision tree, convolutional neural network, ST elevation.**

I. Introduction

Effective health management and timely intervention in individuals under 75 years of age could prevent two out of three deaths [1], with acute myocardial infarction (AMI) as the leading cause [2]. Considering that the risk of mortality within one year increases by 8% for every half-hour delay [3], timely treatment is crucial. Therefore, developing widely accessible, easy-to-use technology for AMI detection in out-of-hospital settings is essential when early symptoms like chest pain appear [4], [5]. With the growing adoption of telemedicine, smart device-based AMI detection offers a promising solution [6]. However, no such technology is currently available commercially.

When diagnosing AMI, a single-lead ECG, commonly available in smart devices, is insufficient because infarct-related changes may appear in different leads depending on infarct location. Therefore, sequential acquisition of a multi-lead ECG has been considered using portable smartphone accessories [7] and smartwatches with integrated electrodes [8], [9]. The multi-lead ECG is obtained by acquiring sequential, single-lead ECGs at different electrode sites, a procedure which is not only time-consuming but prone to diagnostic errors [10], [11]. An alternative approach to increasing the number of leads is to use a device with three electrodes, enabling the simultaneous acquisition of lead I and an additional lead involving a specific touch site, such as a precordial [12].

Machine learning models have been explored for detecting myocardial infarction, often without accounting for whether the infarction is acute or past although differently manifested in the ECG. Some models rely solely on clinically meaningful features like ST segment deviation [13], [14], while others incorporate additional features which have not been established in clinical practice [15]. Another increasingly more common approach is to bypass feature engineering and use the raw ECG, employing, e.g., convolutional neural networks (CNNs) [16]–[25], autoencoder deep learning [26], and residual neural networks [27]. Since most models were developed using the 12-lead ECG, their performance on few-lead ECGs obtained from wrist-worn devices remains unclear.

An important limitation of CNN models is their lack of explainability, hindering their adoption in clinical settings. Despite ongoing efforts to enhance explainability of such models [28], [29], no major breakthrough has been achieved thus far. Numerous studies indicate that machine learning models that bypass feature engineering, particularly those trained on small datasets, may explore peculiarities in the ECG, such as artifacts or device-specific properties, rather than clinically relevant features, see, e.g., [30], [31]. For instance, a dense CNN employing gradient-weighted class activation mapping (Grad-CAM) was used to detect myocardial

Karolina Jančiulevičiūtė, Daivaras Sokas, and Andrius Petrėnas are with the Biomedical Engineering Institute, Kaunas University of Technology, Kaunas, Lithuania (e-mail: karolina.janciuleviciute@ktu.lt, daivaras.sokas@ktu.lt, andrius.petrenas@ktu.lt).

Justinas Bacevičius is with the Institute of Clinical Medicine, Faculty of Medicine, Vilnius University, and Center of Cardiology and Angiology at Vilnius University Hospital Santaros Klinikos, Vilnius, Lithuania (e-mail: justinas.bacevicius@santa.lt).

Leif Sörnmo is with the Department of Biomedical Engineering, Lund University, Lund, Sweden (e-mail: leif.sornmo@bme.lth.se).

infarction [31]. However, the activation maps were not always clinically informative, as the highlighted ECG intervals lacked relevance to infarction. Similarly, the local interpretable model-agnostic explanations method identified the QRS complex as the most influential interval in a random forest model [30], thus contradicting the well-established understanding that the ST segment and T wave are the most relevant.

For the first time, the present study explores AMI detection using a wrist-worn device capable of acquiring a two-lead ECG (hereafter referred to as"wECG") with a single touch. Another novelty is that detection performance is evaluated on a tailored dataset comprising patients with AMI and those with other cardiovascular diseases (CVDs) that cause infarction-like ECG deviations or pose challenges in ECG interpretation. Due to the lack of databases containing the necessary wECG leads, an unconventional approach is adopted in this study in which the wECGs are converted from a publicly available 12-lead ECG dataset and used for training and validating machine learning models. This approach has the potential to facilitate the creation of training databases for non-standard ECGs, resembling those acquired from chest touch sites using commercial smartwatches [9]. Moreover, the study examines whether a wrist-worn device can provide valuable diagnostic information when the wECG is acquired by the patient without technician assistance. To compare their explanatory capabilities, two types of machine learning models are examined: one that relies on clinically relevant features and another that bypasses feature engineering.

The paper is organized as follows. Section 2 describes the study population and the training dataset derived from a public 12-lead ECG database to match the morphology of wECGs. Section 3 details the architecture of the explainable machine learning models used for AMI detection and Sec. 4 covers hyperparameter selection. Section 5 presents the performance results, followed by a discussion in Sec. 6.

II. Materials

A. Test dataset

A total of 123 participants were enrolled in the present study to create a test dataset. Patients were recruited from the inpatient wards of the Cardiology Department at Vilnius University Hospital Santara Clinics, Lithuania. Participants provided signed, written informed consent, adhering to the ethical principles in the Declaration of Helsinki. Approval of the study was obtained from the regional bioethics committee under the reference number 158200-18/7-1052-557.

The participants were aged 18 or older, without an implanted cardiac device or cognitive/linguistic deficits. They were classified into three groups: (1) The AMI group consisting of patients diagnosed with ST-elevation (STEMI) or non-ST-elevation myocardial infarction (NSTEMI), with the wECG recorded within 24 hours of percutaneous coronary intervention; (2) The other CVD group consisting of patients with heart conditions causing infarction-like ECG deviations or challenging interpretations, e.g., acute pericarditis, left ventricular hypertrophy, and bundle branch blocks; (3) The healthy group consisting of participants with no history of heart disease. The AMI and other CVDs groups were well-matched in terms of age, height, weight, and BMI ($p > 0.05$); however, participants in the healthy group were significantly younger, weighed less, and had a lower BMI compared to the AMI and other CVDs groups.

The test dataset was acquired using a wrist-worn wearable device (Biomedical Engineering Institute, Kaunas University of Technology), equipped with three biopotential electrodes: one at its base, another at its upper end, and a third on the strap. Using this configuration, two wECG leads were acquired with a single touch. One lead corresponds to limb lead I as one electrode is touched by the right index finger, while another contacts the left arm near the wrist (LA) where the wrist-worn device is attached. The other lead is nonstandard and obtained between the LA electrode and the strap electrode, which contacts a specific body part. The wECG was acquired at a sampling rate of 500 Hz.

To assess the utility of different touch sites, the dataset was collected by touching specific body sites under the guidance of a technician. The three selected sites correspond to the V3 and V5 electrode sites and the abdomen (A), positioned 2 cm to the left of the umbilicus. As a result, leads between the LA electrode and specific sites were established, labeled as V3-LA, V5-LA, and A-LA. The placement of A-LA was chosen based on the perceived ease for patients to accurately touch the abdomen compared to the V3 and V5 sites. After acquiring lead A-LA, the participants acquired by themselves the same lead without technician assistance, then denoted A-LA^w. In total, four pairs of leads were obtained for each participant. During wECG acquisition, participants were positioned in a supine posture with their upper bodies slightly elevated at an angle not exceeding 30 degrees. Each recording lasted approximately one minute, with at least a 1-min interval in-between successive recordings.

Some participants, particularly elderly patients, had problems with maintaining constant pressure on the electrode for the required duration, leading to fewer ECGs of acceptable quality. The participant demographics and the test dataset composition are presented in Table I.

The wECG dataset used in this study is publicly available through the open-access portal Zenodo [32].

B. Training and validation datasets

To train the proposed models, a subset of recordings from the Physikalisch-Technische Bundesanstalt PTB-XL dataset (version 1.0.3) was used [33]. This subset includes 238 12-lead ECGs, each of 10-s duration and sampled at 500 Hz, from patients with AMI (100 women, 138 men, age: 68.5 $\pm$ 12.2). Only recordings labeled "stadium I", "stadium I-II", and "stadium II", corresponding to acute and subacute stages of infarction, were used. Among these, 215 recordings (90%) had a confidence level of 100%, indicating that the diagnosis reported in clinical records is certain, while only 20 recordings had a confidence level of 50% or lower, indicating a probable diagnosis. The subset also includes 685 controls (392 women, 293 men, age: 52.3 $\pm$ 16.6) labeled as normal in the PTB-XL dataset. The ECGs were randomly divided into training and validation sets with an 80:20 ratio.

TABLE I: Participant demographics and test dataset composition.

	AMI	Other CVD	Healthy
Men	36	15	29
Women	16	5	22
Age, yrs	62.3 ± 12.8	58.7 ± 15.8 ($p < 0.01$)	27.0 ± 7.8
Height, m	1.74 ± 0.10	1.73 ± 0.16 ($p = 0.51$)	1.76 ± 0.11
Weight, kg	87.2 ± 15.9	87.3 ± 19.8 ($p < 0.01$)	72.6 ± 12.9
BMI, kg/m^2	28.6 ± 3.9	29.0 ± 5.6 ($p < 0.01$)	23.4 ± 3.0
AMI type			
NSTEMI	11		
STEMI	41		
AMI location			
Anteral	27		
Lateral	16		
Inferior	14		
Other CVDs			
Right bundle branch block	7	8	
Left bundle branch block	3	6	
Atrial fibrillation		2	
Myocarditis		2	
Pericarditis		2	
Conduction disorder		3	
Left ventricular hypertrophy		3	
Prior infarction	6	4	
Aortic stenosis		2	
Test dataset composition			
I & V3-LA	42	17	45
I & V5-LA	42	17	45
I & A-LA	42	17	45
I & A-LA^w	50	18	50

Note: In some STEMI patients, infarction affected multiple heart regions, resulting in a larger number of locations. p-values among the three groups are based on the Kruskal–Wallis test.

Given that the wECG lead between LA and a specific site is nonstandard, ECGs from the PTB-XL dataset were converted to match the wECG. The nonstandard leads of the wECG are obtained as follows:

$$\text{V3-LA} \approx \frac{\text{III}_\text{s} - \text{I}_\text{s}}{3} + \text{V3}_\text{s}, \tag{1}$$

$$\text{V5-LA} \approx \frac{\text{III}_\text{s} - \text{I}_\text{s}}{3} + \text{V5}_\text{s}, \tag{2}$$

where leads I_s, III_s, V3_s, and V5_s correspond to leads I, III, V3, and V5, respectively, of the 12-lead ECG in the PTB-XL dataset.

Lead A-LA resembles lead III because the electrode's contact site on the abdomen is farther from the heart and its voltage potential is closer to that of the left leg electrode in the 12-lead ECG system. Therefore, no conversion was applied. The resemblance between the acquired wECG and the converted wECG, obtained from a simultaneously acquired 12-lead ECG, is illustrated in Fig. 1.

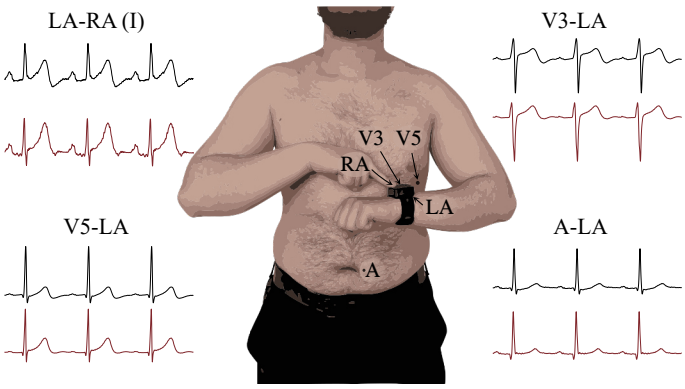

Fig. 1: Example of wECGs acquired using a wrist-worn device (black) and converted from a simultaneously acquired 12-lead ECG (red). To facilitate comparison, a wECG of higher than usual quality is used.

III. Methods

A. Preprocessing

The main aim of preprocessing is to address the fact that the signal quality of wECGs is often poor. This is done by reducing the influence of baseline wander and high-frequency noise using zero-phase bandpass filtering (cutoff frequencies at 0.5 and 100 Hz) [34], and excluding poor-quality segments. Signal quality was determined in each segment using the *bsqi* index [35], defined as the fraction of beats detected by one QRS detector matching those detected by another detector [36], together with an acceptance threshold of 0.65 [34].

To further suppress noise, an average beat is computed; for details, see [34]. The following two-step procedure is used to ensure that only representative beats are included for averaging: 1. Beats whose energy exceeds the 90th percentile of the energy distribution of all beats are excluded as such beats are most likely influenced by substantial finger movement artifacts. 2. Each beat that remains after the first step is correlated to the average of the other remaining beats to further exclude distorted beats, including premature ventricular beats. Provided the correlation coefficient exceeds 0.5, the beat is included for averaging. If less than 10 beats are accepted, the entire wECG is excluded from further analysis. Given that fiducial points are more accurately identified in lead I, beat averaging was performed using the synchronously acquired pairs, with lead I serving as reference. Beats from the wECG were extracted by splitting at the midpoint of the TP interval. To ensure uniform length, each beat was padded at the start and end to reach a fixed length of 500 samples, with 33% of the samples preceding the R peak and 67% following it.

In each lead, wave onset and end are determined using a validated wavelet-based delineator [37]. To improve the accuracy of wave delineation, two synchronously acquired wECG leads are used in pairs: I and V3-LA, I and V5-LA, and I and A-LA.

B. Detection using a convolutional neural network

Deep learning models extract features from the raw ECG, thus eliminating the need for expert-crafted features. This capability makes deep learning well-suited for tasks such as detecting AMI-related changes in the preprocessed wECG.

1) Detector structure: Four different models are explored, where each model assumes a certain lead set as input. Each

lead of the preprocessed wECG is condensed into an average beat so that the model input consists of a lead set of average beats. The first model, denoted D_0, assumes a four-lead set obtained by sequentially touching specific body sites:

$$\mathcal{D}_0 = \{\text{I, V3-LA, V5-LA, A-LA}\}, \tag{3}$$

where lead I is selected from the pair that includes A-LA.

The other three models, denoted D_1, D_2, and D_3, assumes two-lead sets obtained from a single touch:

$$\mathcal{D}_1 = \{\text{I, V3-LA}\}, \tag{4}$$
$$\mathcal{D}_2 = \{\text{I, V5-LA}\}, \tag{5}$$
$$\mathcal{D}_3 = \{\text{I, A-LA}\}. \tag{6}$$

Each model employs a 1D CNN that uses a 500×2 matrix as input to D_1, D_2, and D_3, and a 500×4 matrix to D_0, corresponding to the number of samples in the average beat and the number of wECG leads, respectively. The CNN architecture consists of three convolutional blocks, each comprising a 1D convolutional layer, batch normalization, dropout, and max pooling [38]. Each convolutional layer uses the rectified linear unit (ReLU) activation function to introduce non-linearity. The output from the final convolutional block is flattened and passed through a fully connected layer with ReLU activation function, followed by a dropout layer for regularization. A second fully connected layer produces the output for each class, i.e., AMI and non-AMI, with a softmax activation function applied to generate probability distributions over the output.

2) Explanation: Grad-CAM is used to explain CNN decisions [39]. The Grad-CAM highlights intervals of the wECG that contribute the most to the detection by analyzing the outputs of the convolutional filters in the final convolutional layer, see Fig. 2. Each filter produces a vector of values, known as a feature map, which forms an activation map after applying the activation function. The activation map is then weighted by the gradients.

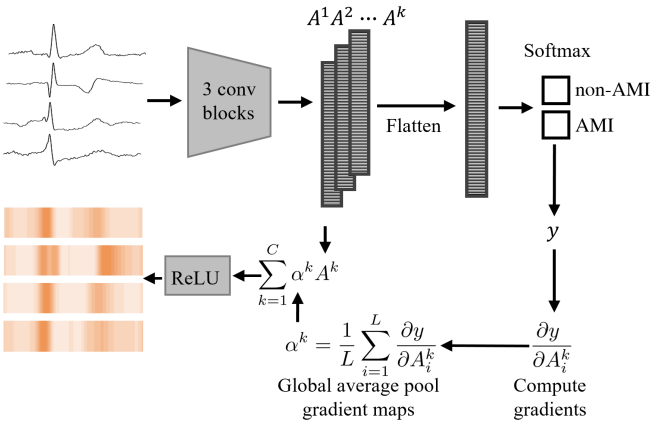

Fig. 2: Gradient-weighted class activation mapping (Grad-CAM) for determining the intervals that contribute the most to the decision.

The importance weights of the k:th activation map in the output convolutional layer for the AMI class, denoted α^k, are computed by applying global average pooling to each activation A^k and taking the gradient of the output y:

$$\alpha^k = \frac{1}{L}\sum_{i=1}^{L} \frac{\partial y}{\partial A_i^k}, \tag{7}$$

where i index the element of the activation map and L is the total number elements in the activation map.

The localization map, which highlights the importance of intervals in the wECG, is then computed as a weighted combination of the activation maps, followed by the ReLU function to preserve only non-negative contributions:

$$\text{Grad-CAM} = \text{ReLU}\left(\sum_{k=1}^{C} \alpha^k A^k\right), \tag{8}$$

where C refers to the total number of activation maps in the final convolutional layer.

Since the localization map contains fewer samples in time than the input, it is upsampled to have the same length as the wECG.

To determine the interval contributing the most to the decision, each average beat is divided into intervals representing the P wave, the QRS complex, the ST segment, the T wave, and the TP interval. The interval with the largest average value in the localization map is chosen as the one contributing the most to the decision.

C. Detection using gradient-boosted decision trees

For expert-crafted features, the use of gradient-boosted decision trees (GBDTs) is known to be efficient, especially when the training dataset is small. Therefore, this technique is considered in the following.

1) Detector structure: The GDBT technique combines decision trees to improve the detection performance by learning from previous mistakes. Each tree recursively splits the dataset based on the feature values, reducing impurity in a subdivided dataset as measured by the reduction in entropy between AMI and non-AMI after each split. The process continues until a stopping criterion is fulfilled, either the maximum number of decision splits or the minimum number of leaf node observations.

Gradient boosting builds a sequence of decision trees, where each tree is trained to correct the misclassifications of the previous tree by fitting the negative gradient of the logarithmic loss function. Thanks to gradient boosting, each tree can focus on previously misclassified wECGs so that the classification error is reduced iteratively.

The GBDT models corresponding to the CNN models are denoted $\mathcal{G}_0$, $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$, respectively.

2) Feature selection and extraction: The ECG waves are characterized by their respective amplitudes, measured relative to the baseline in the PR segment. Features relevant to AMI, i.e., T wave amplitude, the amplitudes at the J point and at the point 80 ms after the J point to characterize the ST segment, and Q wave amplitude [40], are extracted from the average beat of each lead. In total, eight features are used as inputs to $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$, while sixteen features are used to $\mathcal{G}_0$.

3) Explanation: To assess the importance of different features, the Shapley additive explanations (SHAP) method is employed [41]. With this method, the average additive contribution of each feature to the model's output can be quantified. For a given feature j, its additive contribution ϕ_j^m at the m:th iteration is computed by

$$\phi_j^m = f(l_{+j}^m) - f(l_{-j}^m), \quad m = 1, \ldots, M, \tag{9}$$

where $f(\cdot)$ is the GBDT, l_{+j}^m is the feature vector with feature j included, and l_{-j}^m is the feature vector with feature j excluded, and M is the total number of possible feature combinations. The final SHAP value of feature j, denoted ϕ_j, is then obtained by averaging the contributions across all M iterations:

$$\phi_j = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m. \tag{10}$$

The magnitude of ϕ_j reflects the importance of feature j: a positive ϕ_j suggests that the presence of feature j has an increasing effect on the detection, while a negative ϕ_j indicates a decreasing effect. To accelerate the computations, the interventional Tree SHAP algorithm is used instead of computing ϕ_j for every value of M [42].

To determine the feature contributing the most to the decision, the average SHAP value of a particular feature across all leads is computed. The feature associated with the largest value is considered the one contributing the most.

D. Performance evaluation

The area under the receiver operating characteristic (ROC), denoted A, is used to assess how different hyperparameters influence the performance of the GBDT and CNN models when using the validation dataset. The following performance measures are used: sensitivity (Se), defined as the number of correctly detected infarcts divided by the total number of infarcts, and specificity (Sp), defined as the number of correctly detected non-infarcts divided by the total number of non-infarcts.

Agreement between the detection performance on wECGs acquired with and without technician assistance is quantified using Fleiss' kappa, κ.

The models were initialized 100 times, and the average results are reported.

IV. Hyperparameter selection

A. CNN hyperparameter selection

Each convolutional layer employs eight filters with a kernel size of 1 × 15 with a stride of 1, followed by a 1 × 2 average-pooling layer with a stride of 2. The fully connected layer contains 128 neurons using the ReLU activation function and two output neurons a softmax activation function. To prevent overfitting, dropout layers with a probability of 0.1 are applied after each layer. Batch normalization is performed on the outputs of the convolutional layers. The detector is trained using the Adam optimizer [43], with a learning rate of 0.001.

Using the validation dataset, Fig. 3(a) shows A as a function of batch size for the four models. For models $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$, performance no longer improves beyond a batch size of 32, which is therefore used in the following.

Figure 3(b) presents the ROC for each model. The detection thresholds, determined by the point closest to the upper left corner of the ROC curve, are as follows: 0.08 for $\mathcal{D}_0$, 0.14 for $\mathcal{D}_1$, 0.18 for $\mathcal{D}_2$, and 0.11 for $\mathcal{D}_3$. The models $\mathcal{D}_1$, $\mathcal{D}_0$, and $\mathcal{D}_2$ exhibit similar performance for the chosen batch size, resulting in A = 0.98, Se = 0.95, Sp = 0.97, A = 0.98, Se = 0.95, Sp = 0.96, and A = 0.98, Se = 0.93, Sp = 0.95, respectively. The model $\mathcal{D}_3$ performs slightly worse, with A = 0.96, Se = 0.92, and Sp = 0.95.

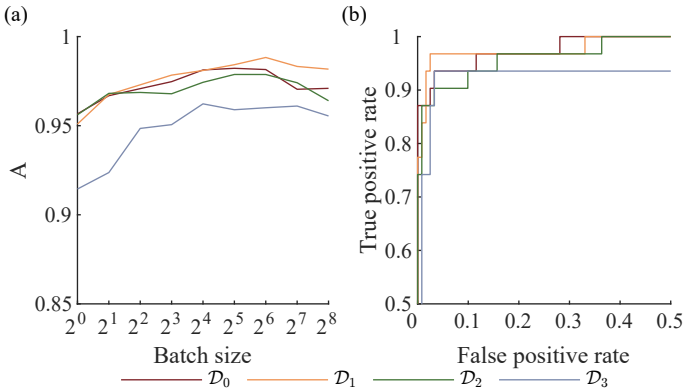

Fig. 3: (a) Area under the ROC (A) as a function of batch size using CNN models on the validation dataset. (b) ROCs for different models using a batch size of 32.

B. GBDT hyperparameter selection

Based on systematic experimentation on the training dataset, the following hyperparameter values were found adequate and therefore used for performance evaluation: The learning rate, controlling the step size in the optimization process, is set to 0.01; the maximum number of decision splits, controlling the maximum number of decision splits per tree, is set to 6; the minimum samples split, determining the minimum number of samples required to split an internal node, is set to 2; and the minimum samples per leaf, determining the minimum number of samples that must be present in a leaf node, is set to 20. The Gini impurity, measuring the likelihood of incorrectly classifying a randomly chosen element based on the label distribution in the node, is used as a split criterion.

The number of trees in the boosting process is an important hyperparameter. Increasing the number of decision trees often leads to better performance, as more trees can refine the model's decision. However, too many trees can result in overfitting. Using the validation dataset, Fig. 4(a) shows A as a function of the number of decision trees for the four models. While A increases with an increasing number of decision trees, adding more decision trees contributes negligibly to performance above a certain point. Therefore, the number of trees is set to 150 for all models.

Figure 4(b) presents the ROC for each model. The detection thresholds are as follows: 0.52 for $\mathcal{G}_0$, 0.33 for $\mathcal{G}_1$, 0.49 for $\mathcal{G}_2$, and 0.41 for $\mathcal{G}_3$. The best performance is obtained using $\mathcal{G}_0$, with A = 0.97, Se = 0.90, Sp = 0.92. The models $\mathcal{G}_2$

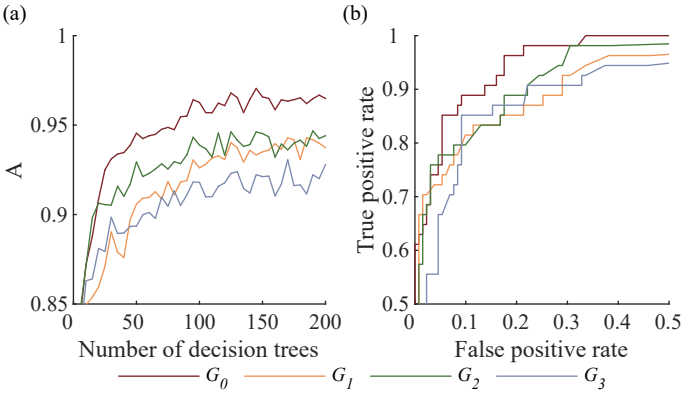

Fig. 4: (a) Area under the ROC (A) as a function of the number of decision trees using the GBDT models on the validation dataset. (b) ROCs for different models using 150 decision trees.

and $\mathcal{G}_1$ exhibit similar performance, resulting in A = 0.95, Se = 0.90, Sp = 0.90, and A = 0.94, Se = 0.87, Sp = 0.91, respectively. The model $\mathcal{G}_3$ perform slightly worse, yielding A = 0.92, Se = 0.87, Sp = 0.89, respectively.

V. Results

A. Feature distribution among datasets

Figure 5 shows the feature distributions of the training, validation, and test datasets. For most features and leads, the distribution shapes are similar ($p > 0.05$). An exception is the J point in leads I and V3-LA of the test dataset, which exhibits slightly lower values compared to the training and validation datasets.

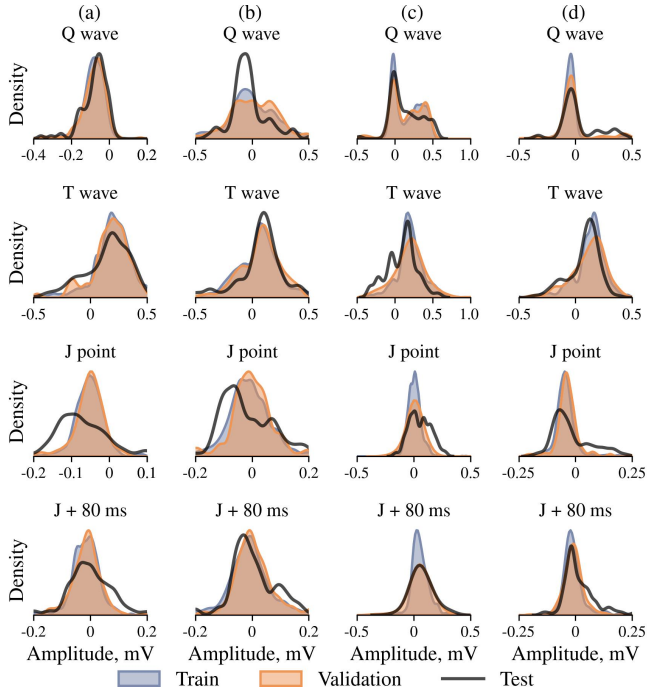

Fig. 5: Feature distributions on the training, validation, and test datasets for: (a) lead I, (b) lead V3-LA, (c) lead V5-LA, and (d) lead A-LA.

B. Performance on the test dataset

The best AMI detection performance on the test dataset is, as expected, obtained using $\mathcal{G}_0$ and $\mathcal{D}_0$ which both involve four leads, see Fig. 6. Of these two models, $\mathcal{D}_0$ performs slightly better than $\mathcal{G}_0$, with Se of 0.77 and Sp of 0.75 compared to 0.77 and 0.72, respectively. However, a more substantial drop in performance is observed for the CNN models compared to the GBDT models. Relative to the validation dataset, Se drops by 19.2%, 26.1%, 24.3%, 26.8% for $\mathcal{D}_0$, $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$ and by 15.5%, 19.2%, 21.2%, 18.2% for $\mathcal{G}_0$, $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$. The drop in Sp is similar – 21.4%, 28.6%, 24.7%, 23.4%, for $\mathcal{D}_0$, $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$, respectively, and 21.3%, 17.6%, 19.4%, 22.5%, for $\mathcal{G}_0$, $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$, respectively. However, after excluding patients with other CVDs, the drop in Sp becomes much less pronounced – 1.7%, 14.5%, 8.8%, 7.2% for $\mathcal{D}_0$, $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, and 1.3%, -1.0%, 0.1%, 7.2% for $\mathcal{G}_0$, $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$.

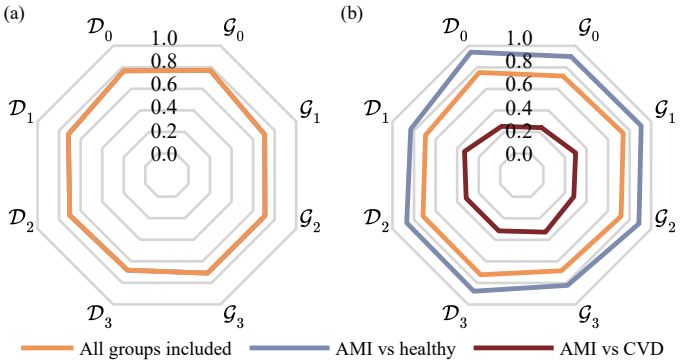

Fig. 6: (a) Sensitivity and (b) specificity of the CNN and GDBT models obtained on the test dataset. Note that sensitivity is the same across all three subsets, since they all include the same AMI patients.

C. Impact of wECG recording duration

To evaluate the impact of a shorter wECG recording duration on performance, sensitivity and specificity was also determined when using only the first 30 s instead of the entire 1-min recording. As indicated by Table II, the performance is largely unchanged across most models, except for $\mathcal{G}_0$, $\mathcal{G}_0$, $\mathcal{G}_2$, and $\mathcal{G}_3$, whose performance improved. On the other hand, 11 wECGs (six from the AMI group, two from the other CVDs group, and three from the healthy group) were excluded because they did not meet the criterion of having at least 10 representative beats for averaging. In addition, fewer representative beats reduce the validity of the average beat which in turn contributes to the slight changes in performance.

D. Detection of infarcts from different locations

Figure 7 shows the sensitivity of the different models when detecting AMI for different infarct locations. Both types of models successfully detects infarcts in lateral and anterior locations, with Se ranging from 0.72 to 0.93 for CNN models and from 0.85 to 0.98 for GBDT models. While CNN models are relatively sensitive to inferior infarcts, GBDT models perform poorly, with the exception of $\mathcal{G}_0$.

TABLE II: Change in sensitivity and specificity when using the first 30 s of the wECG instead of the entire 1-min wECG.

Model	Se	Sp
$\mathcal{D}_0$	-0.01	0.03
$\mathcal{D}_1$	-0.00	0.01
$\mathcal{D}_2$	0.03	0.00
$\mathcal{D}_3$	-0.03	0.05
$\mathcal{G}_0$	0.08	0.05
$\mathcal{G}_1$	0.00	0.01
$\mathcal{G}_2$	0.10	0.06
$\mathcal{G}_3$	0.07	0.03

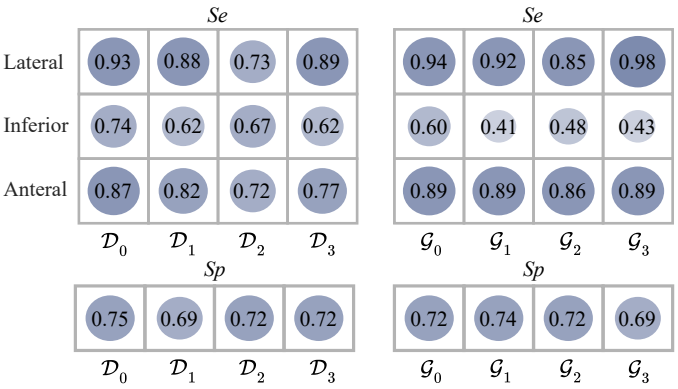

Fig. 7: Sensitivity of different models when detecting AMI across different infarct locations.

E. Detection in wECGs obtained with/without assistance

Table III presents the agreement between decisions when the wECG is acquired with and without technician assistance. Discrepancies in detection occurred in 19.5% (23/118) of the wECGs for $\mathcal{G}_3$ and in 17.8% (21/118) for $\mathcal{D}_3$, with corresponding κ values equal to 0.64 and 0.61, respectively. This difference can be attributed to inaccurate touch sites, leading to changes in wECG morphology.

TABLE III: Agreement between decisions when the wECG is acquired with and without technician assistance. The superscript "w" refers to the model using the wECG acquired without technician assistance.

	AMI	Non AMI
$\mathcal{D}_3$	$\mathcal{D}_3^w$	
AMI	44	9
Non AMI	12	53
$\mathcal{G}_3$	$\mathcal{G}_3^w$	
AMI	43	13
Non AMI	10	52

F. Explanations

Figure 8 shows how the initialization of $\mathcal{D}_0$ and $\mathcal{G}_0$ influence the explanations on the test dataset. The results indicate that the explanations of $\mathcal{D}_0$ are highly dependent on initialization, suggesting that CNN optimization does not always result in the same intervals.

The decision is most often influenced by the QRS complex (median count: 64), followed by the T wave (median: 19) and the ST segment (median: 13). In contrast, using $\mathcal{G}_0$, the J point emerged as contributing the most (median: 51), followed by the T wave amplitude (median: 34).

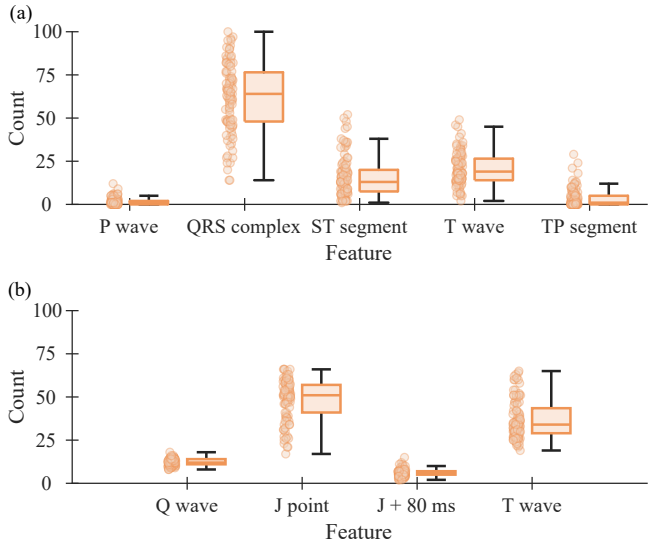

Fig. 8: Impact of the initialization of (a) $\mathcal{D}_0$ on the most important interval and (b) $\mathcal{G}_0$ on the most important feature. Each circle represents the number of times a specific interval or feature was identified as the most influential in the model's decision across the entire test dataset for a single initialization.

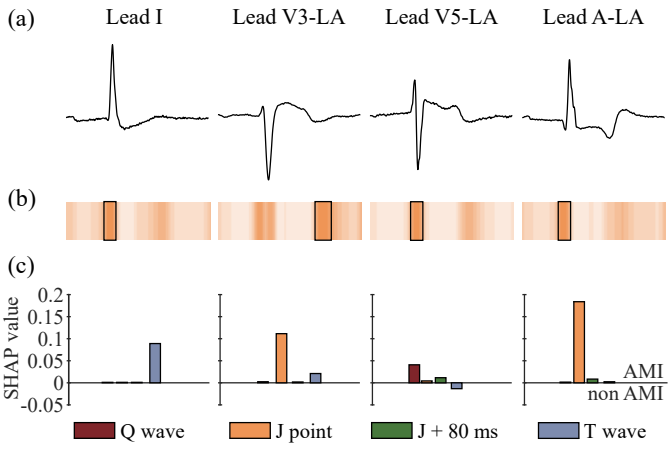

Fig. 9: Explanation for a patient with AMI: (a) wECG, (b) activation map generated by Grad-CAM to explain CNN detection, where the rectangle highlights the beat interval contributing the most to the decision for that specific lead, and (c) SHAP values explaining GBDT detection.

Figure 9 shows an example of explanations for a patient with AMI using $\mathcal{G}_0$ and $\mathcal{D}_0$. Interestingly, using $\mathcal{D}_0$, the intervals that contributed the most to the decision (e.g., the end of the T wave for lead V3-LA and the QRS complex for lead V5-LA) do not correspond to the intervals typically analyzed in clinical practice for AMI detection.

To shed further light on the agreement between explanations of $\mathcal{G}_0$ and $\mathcal{D}_0$, Fig. 10 shows a confusion matrix for patients with AMI. Only 20.2% (34/168) of the intervals identified by

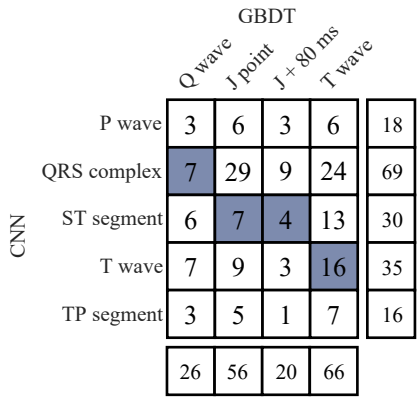

Fig. 10: Agreement between GBDT and CNN explanations combined for all four models in AMI patients. Highlighted cells indicate explanatory GBDT features corresponding to explanatory CNN intervals.

the CNN as contributing the most to the decision corresponding to the respective GBDT features. This discrepancy is primarily due to the QRS complex, identified as the most important interval for the CNN models in 41.1% (69/168) of all intervals.

VI. Discussion

The main finding of the study is that the GBDT models, which rely on clinically meaningful features, are more robust than the CNN models when transitioning from the training/validation dataset to the test dataset. This can be attributed to the capability of the CNN to learn dataset-specific features; this finding applies considerably less to the GBDT models. As a result, while the CNN models performed well on the validation dataset, the performance dropped more substantially when analyzing wECGs of the test dataset than did that of the GBDT models. Moreover, the explanations of the GBDT models are less sensitive to initialization, see Fig. 8. The finding that the QRS complex, the TP interval, and the P wave were sometimes identified as the most influential to the decision is consistent with previous research [30], [31], which reported that the intervals contributing the most to CNN decisions often do not agree with intervals conventionally used for infarction detection.

The results indicate that the CNN models perform slightly better on the test dataset than the GBDT models, see Fig. 6. However, it is important to note that decision trees are better suited for small datasets, while CNNs often perform better using large training datasets. Consequently, final conclusions regarding model performance should not be drawn from this study alone.

A. Machine learning for detection of myocardial infarction

While the majority of studies have used the standard 12-lead ECG as input [44], reduced-lead configurations have also attracted research interest. For example, a deep learning model was developed to diagnose infarction based on six limb leads [45]. However, the model exhibited poor performance due to the absence of information from the precordial leads. The performance improved when the model input was augmented with synthesized precordial leads, resulting in performance comparable to that using the standard 12-lead ECG. Other studies combined limb and precordial leads, such as a CNN trained on four leads, of which three were from precordial sites [17], [46]. Efforts have also been made to use single-lead input. For example, lead I was used as input to a long short-term memory network [47], whereas lead II was used as input to a k-nearest neighbors classifier, a support vector machine [48], and a long short-term memory network [49]. Leads I and II were explored in various CNN architectures [22], [23], [47], [47], [50]. However, the clinical utility of a single lead remains unclear as anterior and septal AMI is likely not reflected in leads I and II.

Current research on AMI detection using wrist-worn devices is still in an early stage and relies primarily on case studies and manual interpretation, e.g., [9], [51], [52]. Automated AMI detection was partially addressed in a study that simulated sequential ECG acquisition using asynchronous leads from a standard 12-lead ECG [53]. The study assumed that precordial leads obtained from a smartwatch between the precordial site and the right index finger are equivalent to those in a standard 12-lead ECG. However, this assumption is an approximation, as acquiring precordial leads requires a central terminal obtained by connecting the three limb electrodes, which is not feasible with smartwatches. Using a separate dataset for testing, the sensitivity of AMI detection ranged from 0.59 for two input leads to 0.68 for four input leads at a fixed specificity of 0.87 [53].

Previous studies have applied CNN and GBDT models with architectures similar to those used in this study. For example, individual beats from the 12-lead ECG were used as input to a CNN [31], [54]. The present study uses average beats rather than individual beats, as the lower quality of the wECG makes beat-level analysis more susceptible to noise. A modified random forest classifier based on morphological ECG features is comparable to the use of GBDT in this study [15]; however, the focus here is on features commonly analyzed in clinical practice to facilitate interpretation. Given the substantial differences in number of leads, electrode configurations, and signal quality of the wECG compared to those of the standard 12-lead ECG, the models were specifically designed and trained to learn the characteristics of the wECG.

B. Implications of dataset shift

Machine learning models for infarction detection are typically evaluated using cross-validation [23]. However, this approach can result in overoptimistic performance estimates [55], [56], as cross-validation within a single dataset does not expose the model to the variability encountered in real-world scenarios. Since self-acquired wECGs are often influenced by varying levels of noise and artifacts and exhibit morphological changes due to inaccuracies in touch site, we used separate datasets for training/validation and testing to better reflect real-world conditions, even at the cost of lower performance.

The present study shows that the drop in performance when transitioning from the training/validation to test dataset varies

across models and model types, suggesting that the extent of performance drop depends on both model architecture and features used. The more pronounced drop in performance of the CNN models relative to that of the GBDT models can be attributed to a covariate dataset shift [57]. Machine learning models perform best under the assumption that the training and test datasets have the same distribution. However, this assumption rarely holds in real-world scenarios, where discrepancies between the two distributions are inevitable.

Another reason for the performance drop on the test dataset may be differences in the timing of ECG acquisition relative to the onset of AMI. While this is not specified for the PTB-XL dataset [33], the wECG was acquired within 24 hours following percutaneous coronary intervention, a period during which ECG features are likely to change rather than remain stable. For example, ST segment recovery can begin as early as 30 min after the intervention [58]. Within two days post intervention, biphasic T wave changes in leads V2, V3, and V4 are commonly observed in patients with anterior STEMI [59]. Moreover, the development of Q waves may begin within minutes to hours after myocardial injury [40]. Preferably, the wECG should be acquired prior to intervention to more accurately reflect the real-world scenario in which a wrist-worn device is supposed to be used for early detection.

The difference in the residual noise level of the wECGs between the training/validation and test datasets may also have influenced detection performance. The noise level after bandpass filtering typically falls within the range of 50–80 μV RMS. With about 50 representative beats available per minute, averaging reduces the noise level to 7–11 μV RMS. In contrast, the noise level of the ECGs from the PTB-XL database is generally less than 20 μV RMS, however, the shorter recording duration allow averaging of about 10 beats, reducing the noise level to less than 7 μV RMS.

Techniques for handling dataset shifts, such as adjusting the training loss function using importance weights, reweighting the input data to match the test distribution, or invoking a transfer learning approach, may improve the performance on a test dataset [57]. In real-world scenarios, the issue still remains as the performance drop is likely due to temporal shifts [60].

C. Effect of infarct location on detection sensitivity

The best performance obtained using $\mathcal{G}_0$ and $\mathcal{D}_0$ suggests that the combination of leads I, V3-LA, V5-LA, and A-LA is sufficient for detecting AMI of different infarct locations. However, all models, particularly those based on GBDT, perform poorly in detecting inferior infarcts, which is unexpected given that such infarcts typically manifest in leads II, III, and aVF. Therefore, $\mathcal{D}_3$ and $\mathcal{G}_3$, both including the abdomen touch site, are expected to perform better. Another unexpected finding is that $\mathcal{D}_2$, including the V5 touch site, misses 27% of the lateral infarcts despite their typical manifestation in leads I, aVL, V5, and V6. In contrast, $\mathcal{G}_2$ detects lateral AMI with a sensitivity of 0.85. Why $\mathcal{G}_2$ and $\mathcal{D}_2$, despite using the same wECG as an input, behave differently remains to be answered. The ability of the models to detect rare isolated septal infarctions, primarily presenting changes in leads V1 and V2, also needs further investigation.

D. Feature selection

Most research on AMI detection has not accounted for the time elapsed since the infarction, often mixing ECGs from AMI with those from past infarction. However, this distinction is essential because beat morphology during AMI evolves over time, meaning that the relevance of a certain feature depends on when the ECG is acquired.

Within a few hours, ST segment elevation becomes noticeable, i.e., the hallmark of ST-elevation AMI. Therefore, the amplitudes at the J point and J + 80 ms were selected to describe the initial and ending parts of the ST segment. As the infarction progresses, the T wave may become inverted or flattened, making T wave amplitude an important feature as well. Meanwhile, an increase in Q wave amplitude indicates that a large area of the myocardium has suffered damage [61].

A peaked T wave is yet another feature that may appear shortly after the onset of symptoms and has therefore been suggested as an early indicator of AMI, with recognition in clinical practice guidelines [40]. However, these changes typically last only up to 30 minutes, making them less reliable for diagnosing AMI [62]. In the present study, patients were included after having undergone percutaneous coronary intervention, and, therefore, several hours had passed between the onset of AMI and the acquisition of the wECG, thereby reducing the relevance of the peaked T wave.

QRS complex-based features, such as duration, fragmentation, and angle, have been associated with STEMI [63]. To quantify such changes, QRS scores have been developed and shown to be effective in assessing the severity of myocardial injury [64]. Features derived from the high-frequency QRS, with a bandwidth from 150 Hz to 250 Hz, have been used to diagnose AMI [65], and has proven valuable in detecting acute coronary artery occlusion [66] and evaluating patients with chest pain [67]. It remains to be investigated whether CNN-based models have learned to capture QRS changes while assigning less weight to conventional features, e.g., the ST-segment, leading to that the QRS complex becomes the most important interval.

E. Electrode touch site

The proposed approach depends on accurately touching specific body sites; however, this requires attention as electrode placement errors are unavoidable even in clinical practice. Studies have shown significant inter-rater variability in electrode positioning, especially when ECGs are obtained by less qualified staff [68]. For example, average displacements of 13.5 mm vertically and 16.5 mm horizontally have been reported when the ECG was acquired by nurses in emergency settings, with greater inconsistency observed in chest leads and among females [69]. Such inaccuracies can distort ECG morphology, potentially mimicking or obscuring the diagnosis of conditions like AMI [68]. Morphological changes become especially pronounced when the precordial lead displacement exceeds 2 cm [70]. Lead V2 is the most sensitive to misplacement, followed by leads V3, V1, and V4, whereas leads V5 and V6 are less influenced, mainly manifested by a reduction in signal amplitude rather than by alterations in morphology [70].

To evaluate the impact of touch site errors on detection performance, participants were asked to touch the abdominal site without technician assistance. The outcome remained unchanged in 82.2% and 80.5% when using $\mathcal{D}_3$ and $\mathcal{G}_3$ respectively (see Table III). Since this experiment was conducted after the site had been previously touched with technician assistance, it is likely that participants learned the correct position through demonstration. Consequently, correctly recalling the correct touch site in a real-world setting may be more challenging than in a controlled environment, potentially leading to more incorrect detections. An alternative touch site that may be easier to replicate is the lower sternum at the level of the fifth intercostal space, corresponding to the position of electrode E in the EASI system.

F. Limitations

A limitation of the present study is that the models were trained using wECGs converted from the 12-lead ECG. Better performance is likely to be achieved with training data collected from a wrist-worn device. The differences between wECGs in the training and test datasets may be explained by the electrode placement, which was approximately 1–2 cm away from the conventional V3 and V5 sites, due to the electrodes attached for acquiring the standard 12-lead ECG.

Another limitation is the demographic imbalance between the AMI and non-AMI groups used for model training and testing. Since the use of a CNN has been found useful for estimating age from the ECG [71], age-related changes, such as a subtle decline in QRS amplitude [72], may have introduced confounding effects. The imbalance in sex distribution between the groups may represent another source of bias, as sex-related differences in the ECG, such as a steeper slope of the ST segment observed in males [72], could have influenced model performance as well.

The features were extracted from the average beat using a wave delineator whose performance drops in conditions that cause major changes in the ECG, such as left or right bundle branch block or AMI. In such conditions, the widened and/or distorted QRS complexes can lead to inaccuracies in localizing the Q wave and the J point. Since incorrect determination of the J point causes the J + 80 ms point to be incorrect, these inaccuracies may have contributed to a greater deviation in feature values among patient groups with AMI and other CVDs compared to the healthy group, which in turn may have improved the performance of the GBDT models.

VII. Conclusions

The present study shows that the machine learning models explored for AMI detection offer acceptable performance, despite using only two to four ECG leads acquired with a wrist-worn device. The CNN models perform slightly better than the GBDT models. However, caution is warranted as the explanations of the CNN-based decisions rarely agree with the ECG intervals typically used in clinical practice.

VIII. Acknowledgments

This work was supported by the Research Council of Lithuania under Agreement S-MIP-23-132.

IX. Conflicts of Interest

References

[1] "Preventable and treatable mortality statistics," *Eurostat Statistics Explained*, 2024.
[2] S. S. Martin *et al.*, "2024 heart disease and stroke statistics: A report of US and global data from the American Heart Association," *Circulation*, vol. 149, no. 8, pp. e347–e913, 2024.
[3] G. De Luca *et al.*, "Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction: Every minute of delay counts," *Circulation*, vol. 109, no. 10, pp. 1223–1225, 2004.
[4] S. T. Vernon *et al.*, "ST-segment–elevation myocardial infarction (STEMI) patients without standard modifiable cardiovascular risk factors—how common are they, and what are their outcomes?" *J. Am. Heart Assoc.*, vol. 8, no. 21, p. e013296, 2019.
[5] M. P. Turakhia, "Addressing the new last-mile problem in health care with home-based complex diagnostics," *JAMA Cardiol.*, vol. 5, no. 10, pp. 1180–1181, 2020.
[6] A. Aldujeli *et al.*, "Delays in presentation in patients with acute myocardial infarction during the COVID-19 pandemic," *Cardiol. Res.*, vol. 11, no. 6, p. 386, 2020.
[7] J. B. Muhlestein *et al.*, "Feasibility of combining serial smartphone single-lead electrocardiograms for the diagnosis of ST-elevation myocardial infarction," *Am. Heart J.*, vol. 221, pp. 125–135, 2020.
[8] A. Samol *et al.*, "Single-lead ECG recordings including Einthoven and Wilson leads by a smartwatch: A new era of patient directed early ECG differential diagnosis of cardiac diseases?" *Sensors*, vol. 19, no. 20, p. 4377, 2019.
[9] C. A. M. Spaccarotella *et al.*, "Multichannel electrocardiograms obtained by a smartwatch for the diagnosis of ST-segment changes," *JAMA Cardiol.*, vol. 5, no. 10, pp. 1176–1180, 2020.
[10] T. Lindow and O. Pahlm, "Smartphone 12-lead ECG—exciting but must be handled with care," *Am. Heart J.*, vol. 226, pp. 267–268, 2020.
[11] D. Duncker *et al.*, "Smart wearables for cardiac monitoring—real-world use beyond atrial fibrillation," *Sensors*, vol. 21, no. 7, p. 2539, 2021.
[12] J. Bacevičius *et al.*, "Six-lead electrocardiography compared to single-lead electrocardiography and photoplethysmography of a wrist-worn device for atrial fibrillation detection controlled by premature atrial or ventricular contractions: Six is smarter than one," *Front. Cardiovasc. Med.*, vol. 10, p. 1160242, 2023.
[13] M. Arif, I. A. Malagore, and F. A. Afsar, "Detection and localization of myocardial infarction using k-nearest neighbor classifier," *J. Med. Syst.*, vol. 36, pp. 279–289, 2012.
[14] L. Sharma, R. Tripathy, and S. Dandapat, "Multiscale energy and eigenspace approach to detection and localization of myocardial infarction," *IEEE J. Biomed. Health Inform.*, vol. 62, no. 7, pp. 1827–1837, 2015.
[15] C. Liang *et al.*, "An interpretable ensemble trees method with joint analysis of static and dynamic features for myocardial infarction detection," *Physiol. Meas.*, vol. 45, no. 8, p. 085006, 2024.
[16] U. R. Acharya *et al.*, "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals," *Inf. Sci.*, vol. 415, pp. 190–198, 2017.
[17] W. Liu *et al.*, "Real-time multilead convolutional neural network for myocardial infarction detection," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1434–1444, 2017.
[18] ——, "Multiple-feature-branch convolutional neural network for myocardial infarction diagnosis using electrocardiogram," *Biomed. Signal Process. Control*, vol. 45, pp. 22–32, 2018.
[19] N. Liu *et al.*, "A simple and effective method for detecting myocardial infarction based on deep convolutional neural network," *J. Med. Imaging Health Inform.*, vol. 8, no. 7, pp. 1508–1512, 2018.
[20] U. B. Baloglu *et al.*, "Classification of myocardial infarction with multi-lead ECG signals and deep CNN," *Pattern Recognit. Lett.*, vol. 122, pp. 23–30, 2019.
[21] W. Liu *et al.*, "MFB-CBRNN: A hybrid network for MI detection using 12-lead ECGs," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 503–514, 2019.
[22] K. Feng *et al.*, "Myocardial infarction classification based on convolutional neural network and recurrent neural network," *Appl. Sci.*, vol. 9, no. 9, p. 1879, 2019.
[23] V. Jahmunah *et al.*, "Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals," *Comput. Biol. Med.*, vol. 134, p. 104457, 2021.

[24] L. Fu *et al.*, "Hybrid network with attention mechanism for detection and location of myocardial infarction based on 12-lead electrocardiogram signals," *Sensors*, vol. 20, no. 4, p. 1020, 2020.

[25] K. Jafarian *et al.*, "Automating detection and localization of myocardial infarction using shallow and end-to-end deep neural networks," *Appl. Soft Comput.*, vol. 93, p. 106383, 2020.

[26] J. Zhang *et al.*, "Automated detection and localization of myocardial infarction with staked sparse autoencoder and treebagger," *IEEE Access*, vol. 7, pp. 70 634–70 642, 2019.

[27] C. Han and L. Shi, "ML–ResNet: a novel network to detect and locate myocardial infarction using 12 leads ECG," *Comput. Methods Programs Biomed.*, vol. 185, p. 105138, 2020.

[28] J. Qu *et al.*, "An interpretable shapelets-based method for myocardial infarction detection using dynamic learning and deep learning," *Physiol. Meas.*, vol. 45, no. 3, p. 035001, 2024.

[29] W. Zhang *et al.*, "Interpretable detection and location of myocardial infarction based on ventricular fusion rule features," *J. Healthc. Eng.*, vol. 2021, no. 1, p. 4123471, 2021.

[30] M. Bodini, M. W. Rivolta, and R. Sassi, "Interpretability analysis of machine learning algorithms in the detection of ST-elevation myocardial infarction," in *Proc. Comput. Cardiol.*, 2020, pp. 1–4.

[31] V. Jahmunah *et al.*, "Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals," *Comput. Biol. Med.*, vol. 146, p. 105550, 2022.

[32] K. Jančiulevičiūtė *et al.*, "wECGdb: An ECG database acquired using a wrist-worn device from patients with acute myocardial infarction and controls (version v1)," 2025, [Data set]. [Online]. Available: https://doi.org/10.5281/zenodo.15235775

[33] P. Wagner *et al.*, "PTB-XL, a large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, pp. 1–15, 2020.

[34] K. Jančiulevičiūtė *et al.*, "An echo state network for synthesizing the standard 12-lead ECG from a two-lead ECG obtained from a single touch of a wrist-worn device," *Biomed. Signal Process. Control.*, vol. 109, p. 108008, 2025.

[35] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter," *Physiol. Meas.*, vol. 29, no. 1, pp. 15–32, 2008.

[36] J. Behar *et al.*, "ECG signal quality during arrhythmia and its application to false alarm reduction," *IEEE J. Biomed. Health Inform.*, vol. 60, no. 6, pp. 1660–1666, 2013.

[37] N. Pilia *et al.*, "ECGdeli—an open source ECG delineation toolbox for MATLAB," *SoftwareX*, vol. 13, p. 100639, 2021.

[38] M. Butkuvienė *et al.*, "Considerations on performance evaluation of atrial fibrillation detectors," *IEEE J. Biomed. Health Inform.*, vol. 68, no. 11, pp. 3250–3260, 2021.

[39] R. R. Selvaraju *et al.*, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[40] K. Thygesen *et al.*, "Fourth universal definition of myocardial infarction (2018)," *Circulation*, vol. 138, no. 20, pp. e618–e651, 2018.

[41] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural. Inf. Process. Syst.*, vol. 30, 2017.

[42] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.

[43] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014.

[44] P. Xiong, S. M.-Y. Lee, and G. Chan, "Deep learning for detecting and locating myocardial infarction by electrocardiogram: a literature review," *Front. Cardiovasc. Med.*, vol. 9, p. 860032, 2022.

[45] Y. Cho *et al.*, "Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography," *Sci. Rep.*, vol. 10, no. 1, p. 20495, 2020.

[46] Y. Cao *et al.*, "ML-Net: multi-channel lightweight network for detecting myocardial infarction," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 10, pp. 3721–3731, 2021.

[47] N. Strodthoff and C. Strodthoff, "Detecting and interpreting myocardial infarction using fully convolutional neural networks," *Physiol. Meas.*, vol. 40, no. 1, p. 015001, 2019.

[48] B. Fatimah *et al.*, "Efficient detection of myocardial infarction from single lead ECG signal," *Biomed. Signal Process. Control*, vol. 68, p. 102678, 2021.

[49] H. Martin *et al.*, "Real-time frequency-independent single-lead and single-beat myocardial infarction detection," *Artif. Intell. Med.*, vol. 121, p. 102179, 2021.

[50] H. M. Rai *et al.*, "Myocardial infarction detection using deep learning and ensemble technique from ECG signals," in *Proc. Second Int. Conf. Comput. Commun. Cyber Secur.: IC4S 2020*. Springer, 2021, pp. 717–730.

[51] K. Stark *et al.*, "Watch out for ST-elevation myocardial infarction: a case report of ST-elevation in single-lead electrocardiogram tracing of a smartwatch," *Eur. Heart J. Case Rep.*, vol. 4, no. 6, pp. 1–4, 2020.

[52] K. Li *et al.*, "Using the Apple watch to record multiple-lead electrocardiograms in detecting myocardial infarction: where are we now?" *Tex. Heart Inst. J.*, vol. 49, no. 4, p. e227845, 2022.

[53] C. Han *et al.*, "Automated detection of acute myocardial infarction using asynchronous electrocardiogram signals—preview of implementing artificial intelligence with multichannel electrocardiographs obtained from smartwatches: retrospective study," *J. Med. Internet Res.*, vol. 23, no. 9, p. e31129, 2021.

[54] J.-Z. Jian *et al.*, "Detection of myocardial infarction using ECG and multi-scale feature concatenate," *Sensors*, vol. 21, no. 5, p. 1906, 2021.

[55] A. Isaksson *et al.*, "Cross-validation and bootstrapping are unreliable in small sample classification," *Pattern Recognit. Lett.*, vol. 29, no. 14, pp. 1960–1965, 2008.

[56] S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: What does it estimate and how well does it do it?" *J. Am. Stat. Assoc.*, vol. 119, no. 546, pp. 1434–1445, 2024.

[57] G. D. Y *et al.*, "Covariate shift: A review and analysis on classifiers," in *2019 Glob. Conf. Adv.Technol.*, 2019, pp. 1–6.

[58] C. E. Buller *et al.*, "ST-segment recovery and outcome after primary percutaneous coronary intervention for ST-elevation myocardial infarction: insights from the Assessment of Pexelizumab in Acute Myocardial Infarction (APEX-AMI) trial," *Circulation*, vol. 118, no. 13, pp. 1335–1346, 2008.

[59] R. Delewi *et al.*, "Pathological Q waves in myocardial infarction in patients treated by primary PCI," *JACC Cardiovasc. Imaging*, vol. 6, no. 3, pp. 324–331, 2013.

[60] L. L. Guo *et al.*, "Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine," *Appl. Clin. Inform.*, vol. 12, no. 04, pp. 808–815, 2021.

[61] R. Delewi *et al.*, "Pathological Q waves in myocardial infarction in patients treated by primary PCI," *JACC Cardiovasc. Imaging*, vol. 6, no. 3, pp. 324–331, 2013.

[62] L. Koechlin *et al.*, "Hyperacute T wave in the early diagnosis of acute myocardial infarction," *Ann. Emerg. Med.*, vol. 82, no. 2, pp. 194–202, 2023.

[63] N. Yang *et al.*, "What can we find in QRS in patients with ST-segment-elevation myocardial infarction?" *J. Electrocardiol.*, vol. 75, pp. 52–59, 2022.

[64] E. P. Bounous Jr *et al.*, "Prognostic value of the simplified Selvester QRS score in patients with coronary artery disease," *J. Am. Coll. Cardiol.*, vol. 11, no. 1, pp. 35–41, 1988.

[65] G. Amit *et al.*, "High-frequency QRS analysis in patients with acute myocardial infarction: A preliminary study," *Ann. Noninvasive. Electrocardiol.*, vol. 18, no. 2, pp. 149–156, 2013.

[66] J. Pettersson *et al.*, "Changes in high-frequency QRS components are more sensitive than ST-segment deviation for detecting acute coronary artery occlusion," *J. Am. Coll. Cardiol.*, vol. 36, no. 6, pp. 1827–1834, 2000.

[67] O. Galante *et al.*, "High-frequency QRS analysis in the evaluation of chest pain in the emergency department," *J. Electrocardiol.*, vol. 50, no. 4, pp. 457–465, 2017.

[68] K. Khunti, "Accurate interpretation of the 12-lead ECG electrode placement: A systematic review," *Health Educ. J.*, vol. 73, no. 5, pp. 610–623, 2014.

[69] K. McCann *et al.*, "Accuracy of ECG electrode placement by emergency department clinicians," *Emerg. Med. Australas*, vol. 19, no. 5, pp. 442–448, 2007.

[70] M. Kania *et al.*, "The effect of precordial lead displacement on ECG morphology," *Med. Biol. Eng. Comput.*, vol. 52, pp. 109–119, 2014.

[71] Z. I. Attia *et al.*, "Age and sex estimation using artificial intelligence from standard 12-lead ECGs," *Circ. Arrhythm. Electrophysiol.*, vol. 12, no. 9, p. e007284, 2019.

[72] L. S. Green *et al.*, "Effects of age, sex, and body habitus on QRS and ST-T potential maps of 1100 normal subjects," *Circulation*, vol. 71, no. 2, pp. 244–253, 1985.