Enhancing convolutional neural network predictions of electrocardiograms with left ventricular dysfunction using a novel sub-waveform representation

Hossein Honarvar, PhD,*[1] Chirag Agarwal, PhD,† Sulaiman Somani, MD,*
Akhil Vaid, MD,* Joshua Lampert, MD,‡ Tingyi Wanyan, PhD,*§ Vivek Y. Reddy, MD,‡
Girish N. Nadkarni, MD,*‖¶ Riccardo Miotto, PhD,* Marinka Zitnik, PhD,†
Fei Wang, PhD,# Benjamin S. Glicksberg, PhD***[1]

*From the *Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, New York, †Department of Biomedical Informatics, Harvard University, Boston, Massachusetts, ‡Helmsley Center for Cardiac Electrophysiology, Mount Sinai Hospital, New York, New York, §School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana, ‖Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, ¶The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, #Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, and **Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York.*

BACKGROUND Electrocardiogram (ECG) deep learning (DL) has promise to improve the outcomes of patients with cardiovascular abnormalities. In ECG DL, researchers often use convolutional neural networks (CNNs) and traditionally use the full duration of raw ECG waveforms that create redundancies in feature learning and result in inaccurate predictions with large uncertainties.

OBJECTIVE For enhancing these predictions, we introduced a sub-waveform representation that leverages the rhythmic pattern of ECG waveforms (data-centric approach) rather than changing the CNN architecture (model-centric approach).

RESULTS We applied the proposed representation to a population with 92,446 patients to identify left ventricular dysfunction. We found that the sub-waveform representation increases the performance metrics compared to the full-waveform representation. We observed a 2% increase for area under the receiver operating characteristic curve and 10% increase for area under the precision-recall curve. We also carefully examined three reliability components of explainability, interpretability, and fairness. We provided an explanation for enhancements obtained by heartbeat alignment mechanism. By developing a new scoring system, we interpreted the clinical relevance of ECG features and showed that sub-waveform representation further pushes the scores towards clinical predictions. Finally, we showed that the new representation significantly reduces prediction uncertainties within subgroups that contributes to individual fairness.

CONCLUSION We expect that this added control over the granularity of ECG data will improve the DL modeling for new artificial intelligence technologies in the cardiovascular space.

KEYWORDS Deep learning, Cardiology, Electrocardiograms, Sub-waveform representation, Machine Learning

(Cardiovascular Digital Health Journal 2022;3:220–231) © 2022 Heart Rhythm Society. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Electrocardiography is a common technique for recording electrical activity of the heart over a time period and is used as a noninvasive, first-line, and inexpensive diagnostic tool.[1] This process generates electrocardiogram (ECG) data that provide physiological and structural information about the heart and is normally used for diagnosing cardiac-related diseases. Each ECG is obtained by placing several electrodes on the skin in different parts of the body. The arrangement of these electrodes with respect to each other

[1] Co-corresponding authors. **Address reprint requests and correspondence:** Dr Benjamin S. Glicksberg, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029. E-mail address: benjamin.glicksberg@mssm.edu.

https://doi.org/10.1016/j.cvdhj.2022.07.074

KEY FINDINGS

- The sub-waveform representation of ECGs improved the precision and reduced the uncertainties of the CNN predictions.
- The developed scoring system directly quantified the important ECG features in the waveforms and facilitated the interpretation of the CNN predictions.
- The sub-waveform representation of ECGs contributed to minimizing the disparities of CNN predictions within the protected subgroups.

is called a "lead." Each lead records a time-dependent waveform (eg, 10-second duration), with the magnitude being the voltage difference between the corresponding electrodes. A conventional configuration of ECGs is 12-lead, with 3 limb leads (I, II, III), 3 augmented limb leads (aVF, aVL, aVR), and 6 precordial leads (V_1, V_2, V_3, V_4, V_5, V_6).

To accelerate disease diagnosis and management, extracting clinical information from ECG waveforms using artificial intelligence (AI) algorithms has seen a resurgence recently, mainly owing to large datasets, powerful computers, and new methodological developments.[2–4] One of the main promises of AI is finding hidden patterns in the data that are invisible to human experts but visible to intelligent algorithms.[5,6] In the past, conventional statistical and machine learning methods such as support vector machines, logistic regression, random forests, and gradient boosting were used for diagnostic predictions from ECGs.[7] But these methods normally require features that are manually provided by domain experts or signal processing techniques such as Fourier transform. This ad hoc feature extraction is cumbersome and expensive in terms of time and human efforts because of the unstructured nature of ECGs, scale, and the possibility of inaccurate predictions or learning irrelevant artifacts. To overcome these issues, deep learning (DL) has emerged as a powerful analytics tool that automates feature extraction for large-scale unstructured data types such as text, imaging, and, more recently, ECG waveforms,[2,5,6,8] to decode complex patterns in data and to provide valuable clinical insights.

In the literature, there is a body of work on changing the architecture of neural network (NN) that is the typical core of DL in order to improve the predictions from ECG waveforms.[2] Two widely used architectures are convolutional NN (CNN) and recurrent NN.[9–11] There have also been efforts on combining CNN and recurrent NN to further enhance the predictions.[12] Despite the important role of NN in processing the data and finding useful patterns, major enhancements in learning can be achieved by manipulations at the data level while the NN is kept fixed. For many applications, the boost in performance from data preparation, labeling, preprocessing, and representation might outweigh searching for the optimal NN architecture. With emerging DL applications for ECG data type, the need for data-level manipulations has become more important than ever, since each type has its own complexities and therefore needs customized recipes. In the literature, there have been several studies that investigate the impact of transforming ECG waveforms on predictions. One common strategy is augmenting waveforms by perturbing the waveforms locally or creating slices from the original waveform. For example, the window warping method perturbs the waveforms by dilating or squeezing a small region of the waveform.[13] Another popular augmentation technique is window slicing, which creates random slices of the waveforms with the same label as the original full waveform by sliding windows of the same size.[14] To alleviate the potential problems caused by random slicing, a concatenate and resampling method has been proposed that slices the waveform according to the peaks and then concatenates the slices to reconstruct the length of the original waveform.[15] In addition to augmentation methods, the grid-like structural representation of waveforms for NN has been examined in the form of a 1-dimensional (1D) waveform with 8 leads as channels or a 2D image with time being the width and leads forming the height of the image.[9,10,16] Also, representing ECG waveforms as spectrograms using wavelet or Fourier transform techniques has been studied.[17] However, to date, finding the optimal and reliable representation at the waveform level has not yet been explored. The current state-of-the-art research in DL of ECG often uses the raw waveforms obtained from the device, which we refer to as full-waveform representation, shown in Figure 1A.

In this work, we introduce a novel sub-waveform representation that extends the full-waveform representation of ECGs to improve the DL predictions. We apply this methodology to identify left ventricular dysfunction (LVD), which is present in 1.4%–2.2% of the population and 9% among the elderly.[18] Diagnosing this dysfunction is important to prevent complications such as heart failure and to reduce mortality risk. Once diagnosed, treatment strategies and device implantation are normally effective. A primary tool for diagnosing heart failure, which might be due to LVD, is B-type natriuretic peptide (BNP) levels, which require invasive blood draws. BNP levels can be falsely low in obese patients and falsely high in patients taking certain drugs, such as angiotensin receptor–neprilysin inhibitors (eg, sacubitril-valsartan), as these drugs reduce the clearance of BNP,[19] LVD is normally quantified through measuring the left ventricular ejection fraction (LVEF) from manual inspection of echocardiograms (echos), which generate ultrasound videos of the heart.[20,21] The manual processing of echos can result in inaccurate predictions with large uncertainties.[22,23] The LVEF prediction from echo videos has been automated using DL to accelerate the process and improve the accuracy of the predictions.[24] Recently, DL has been used to predict the LVEF from ECG waveforms with the echos as ground truth.[25] In this work, we focus our analysis on this later development, since we are interested in exploring the ECG data

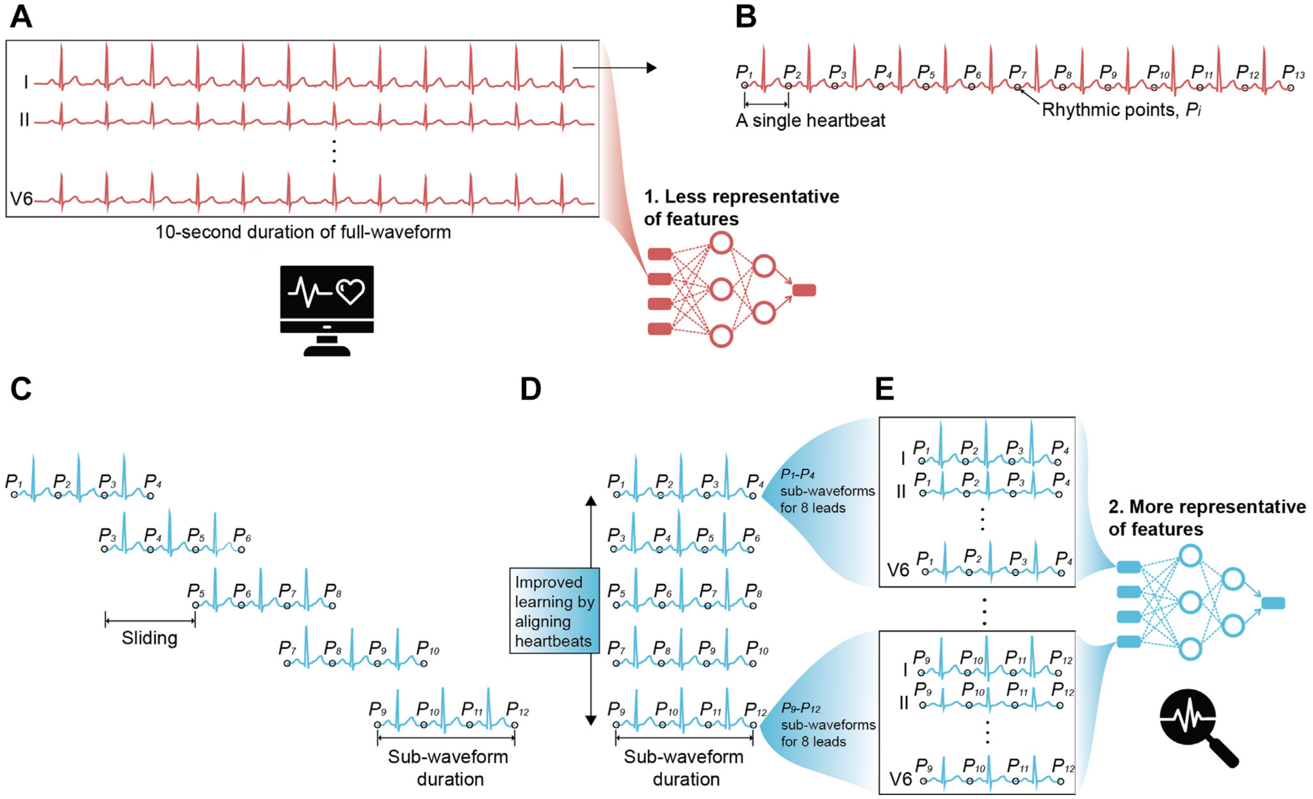
A
I
II
V6
10-second duration of full-waveform
B
P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13
A single heartbeat
Rhythmic points, Pi
1. Less representative of features
C
Sliding
Sub-waveform duration
D
Improved learning by aligning heartbeats
Sub-waveform duration
E
P1-P4 sub-waveforms for 8 leads
P9-P12 sub-waveforms for 8 leads
2. More representative of features

Figure 1 Electrocardiography (ECG) waveform representations. **A:** Full-waveform representation that directly comes from an ECG device. **B–D:** Rhythmic discretization (**B**) and sliding of the lead I full waveform (**C**) to create the lead I sub-waveform representation (**D**). **E:** The same discretization and sliding as lead I is applied to other 7 leads to create sub-waveform representation for all 8 leads.

and a clinical randomized trial has recently been conducted to assess the efficacy of this ECG-DL tool.[26] Therefore, outcomes (ground truth labels for DL models) are LVEF values that are extracted from echo reports.[20,21]

For LVD population, we show that sub-waveform representation increases the predictive performance. In addition, we explain the underlying mechanism for the improvements gained in DL models using the proposed representation. We develop an interpretation framework for quantifying the importance scores of ECG features and investigate the differences between the interpretation of 2 representations. Finally, we investigate the impact of sub-waveform representation on the disparities in different subgroups.

Materials and Methods

Data source

We extracted data from 92,446 patients from 5 hospitals in the Mount Sinai Health System, serving a diverse and urban population in New York City. Our study was approved by the Mount Sinai Institutional Review Board. The characteristic of this dataset is shown in Table 1. Our raw ECGs were stored in XML files that have patient and test demographics, diagnostic information, and waveform details. Each waveform is recorded at 500 Hz for a duration of 10 seconds (total of 5000 data points) for 12 leads. For DL modeling, the 12-lead ECG is reduced to 8 leads because leads III, aVF, aVL, and aVR can be derived from the linear combination of leads I and II.

Data preprocessing

Before feeding the waveforms into NN, we performed several preprocessing steps, as follows. To remove baseline drift that may stem from baseline respiration or lead migration, we used a median filter (width of 1 second or 500 Hz). We removed ECGs flagged with a “Poor Diagnostic Code” from the confirmed reading. We standardized waveforms to have zero-mean and unit-variance.

Outcome

We converted the continuous LVEF values to binary values. As suggested,[25] we chose LVEF ≤35% to be the positive LVD (low LVEF) and LVEF >35% to be the negative LVD. In terms of linking ECGs to the echo report, we linked each ECG to the nearest-date echo report. We only have 1 ECG per patient; therefore, all our ECG-echo pairs are unique.

ECG sub-waveform representation

The ECG waveforms are a measure of hemodynamics of the heart.[27] Like other physical waves, which are created by exciting a medium, ECG waveforms are created by exciting the heart through hemodynamics. Therefore, ECG waveforms should intuitively inherit the fundamentals of wave physics. From the physics point of view, waves in nature generally have 2 components in terms of morphology: coherent (ordered) and incoherent (disordered).[28] The

Table 1 Dataset for identifying left ventricular dysfunction (LVD) by predicting left ventricular ejection fraction (LVEF) from ECG waveforms of 92,446 unique patients from 5 hospitals at the Mount Sinai Health System

Dataset characteristics	Training	Development	Test
No. patients	73,956	9246	9244
Percent positive	7.6	7.6	7.6
Race (%)			
White	41.2	41.6	40.9
Black	25.1	25.0	25.4
Asian	6.6	6.6	6.4
Other	27.1	26.8	27.3
Ethnicity (%)			
Hispanic	18.4	17.9	17.4
Non-Hispanic	53.1	53.4	53.9
Other	28.5	28.6	28.7
Sex (%)			
Male	54.9	54.7	53.9
Female	45.1	45.2	46.0
Age (mean ± SD)	63.4 ± 15.1	63.2 ±14.8	63.5 ± 15.1

ordered component of waves normally provides opportunities for controlling the behavior of waves in a desirable manner. For example, this controllability property has resulted in developing new technologies to efficiently control light and heat waves in materials.[29–31] From the clinical perspective, ECG waveforms do indeed have the ordered and disordered components, and they are called rhythmic and arrhythmic ECGs.[32] As is the characteristic of ordered waves, the rhythmic ECG waveforms offer opportunities for engineering the granularity of the data to enhance learning. This physics-based inspiration forms the basis of this work and has motivated us to create the ECG sub-waveform representation.

The algorithmic development of proposed sub-waveform representation is illustrated in Figure1B–1E. The main mechanism underlying the improvements gained by sub-waveform representation is *heartbeat alignment,* which will be explained in detail throughout this work. It is important to emphasize that our data-centric approach is fundamentally different from previous model-centric approaches because we are only transforming ECG waveforms rather than changing the NN architecture. This transformation introduces a new space for learning optimal features and can potentially be used for any architecture and any task. In addition, it is worth clarifying that our approach is different from slicing-based augmentation techniques mentioned earlier. The major difference is that our approach uses a rhythmic discretization based on a reference rhythmic ECG. This important property is the core of the proposed approach because it allows aligning heartbeats for rhythmic ECGs with minimal impact on arrhythmic ECGs and improves the optimality and reliability of ECG DL predictions. Augmentation is not our main goal, although the sub-waveform representation can still benefit from augmentation effects. Another important difference is that most of the previous studies on augmentation used either the full or near-full length of the original full waveform. However, our goal is to find an optimal representation by creating sub-waveforms with resolution on the order of heartbeats. In fact, for these reasons, we call it a sub-waveform representation.

For numerical implementation of our sub-waveform representation, we use a reference rhythmic ECG that has 13 heartbeats and has a heart rate of 78 beats per minute (bpm), which is around the average heart rate of adults.[33] Our reference heartbeat is 0.74 seconds and we use this heartbeat to discretize all ECG full waveforms, including rhythmic and arrhythmic ECGs, and to ensure the same input length for all ECGs, as required for DL models. Each heartbeat is the smallest unit of a waveform. In Figure 1B, we visualize this discretization for lead I. The rhythmic discretization allows us to extract sub-waveforms at rhythmic points with a duration that is a multiple of the heartbeat. To diversify the sub-waveforms, we introduce a "sliding" parameter that controls where the sub-waveform should originate and how many sub-waveforms are created depending on the duration of sub-waveforms. In Figure 1C, we illustrate how a sub-waveform with 3-heartbeat duration and 2-heartbeat sliding results in 5 sub-waveforms that are generated from the original full waveform. Once the sub-waveforms are created, the new representation for lead I is formed, as shown in Figure 1D. By applying the same procedure used for lead I for all leads, the final sub-waveform representation for 8 leads is derived (shown in Figure 1E). In our Supplemental Material, we provide the pseudocode for our algorithm.

Deep learning setup and evaluation

We used a deep neural network that has a similar structure as in reference 9. This deep neural network takes preprocessed waveforms as inputs and output a binary prediction. Each ECG is represented as a 1D waveform with 8 leads as channels. Our network has 26 layers and takes advantage of residual connections to make the optimization more effective by avoiding exploding or vanishing gradients.[34] There are 3 convolutional layers followed by 11 residual blocks (2 convolutional layers per block) and 1 dense layer. After each convolutional layer, we apply a batch normalization to improve learning and a rectified linear unit to activate the nonlinearities in the network.[35] We use dropout regularizer to minimize overfitting in training.[36] The final layer is a fully connected layer followed by a sigmoid function. For training, we randomly initialized weights.[37] We used the Adam optimizer for backpropagation using initial learning rate of 0.005 and default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$.[38] A scheduler is used to decay the learning rate by a factor of 0.1 if a metric is not improved for a few epochs.

We used a development set for evaluating our model during training and we used the holdout test set for reporting the final results. We trained DL models to classify LVEF severity using ECGs of 73,956 patients and used ECGs of 9244 patients for development. We evaluated the performance using 2 metrics: area under receiving operating characteristic (AUROC) and area under precision-recall curve (AUPRC).

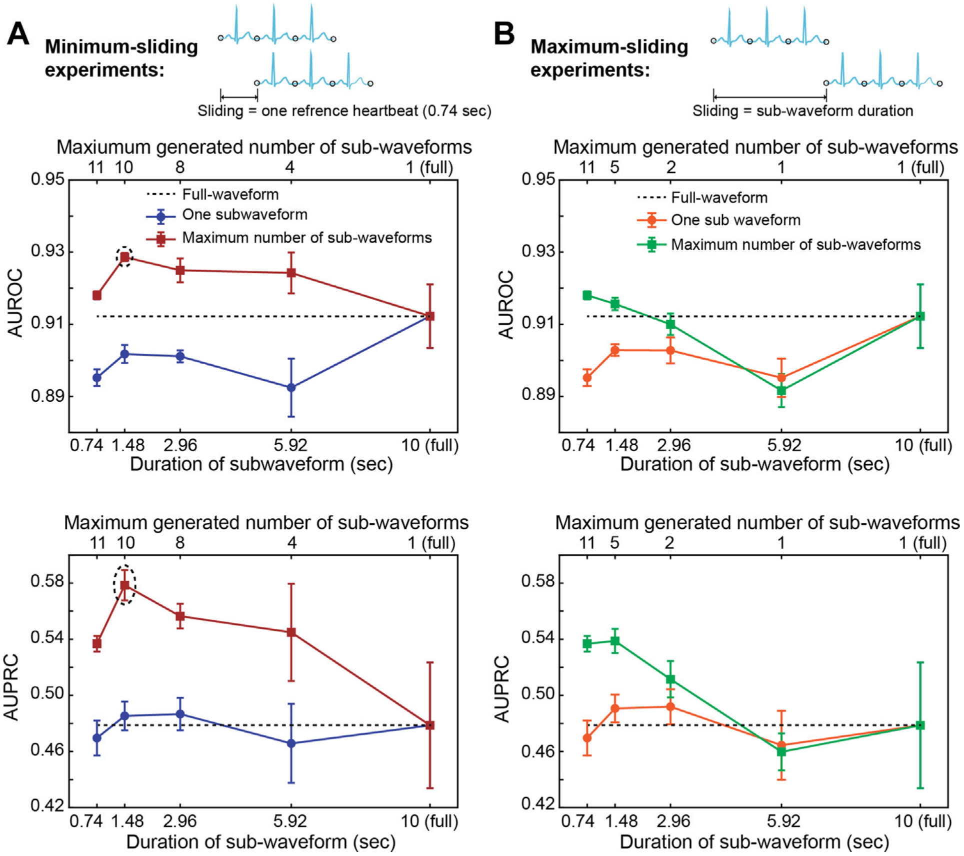

Figure 2 Systematic experiments for evaluating performance of sub-waveform with respect to full-waveform representation (baseline) for left ventricular dysfunction (LVD) case study for 9244 patients in holdout test set. **A:** Minimum sliding is used and is equal to 0.74 seconds. **B:** Maximum sliding is used and is equal to sub-waveform duration. For both experiments in panels A and B, we examine 2 sets of a number of sub-waveforms: 1 sub-waveform and maximum number of sub-waveforms. The optimal sub-waveform is highlighted by dashed circle and has a duration of 1.48 seconds with 10 sub-waveforms. Our sub-waveform representation provides more accurate predictions with smaller uncertainties.

We report the performance for 9244 patients in a holdout test set. For quantitative evaluation of our models, we calculated 2 curves: receiver operating characteristic (ROC) and precision recall curve (PRC). We then reported 2 metrics calculated from these 2 curves: Area Under ROC (AUROC) and Area Under PRC (AUROC), as shown in Figure 2. For each final prediction, we averaged the predictions of 5 independent DL runs with randomly initiated weights and for bootstrapping, we extracted 1 randomly selected sub-waveform for each run. Reported uncertainty for each prediction is the standard deviation of the 5 bootstrapped predictions.

Experimental design and framework

As the baseline, we predicted the LVEF for our 9244 patients in a holdout test set using full-waveform representation using a 1D CNN model. For sub-waveform representation, we performed a set of systematic experiments to understand how it performs compared to the full-waveform case.

In Figure 2A and 2B, we show 2 sets of experiments that are controlled by the sliding parameter (a multiple of reference heartbeat). We set the sliding to be (1) minimum, which equals to 1 reference heartbeat, and (2) maximum, which equals to sub-waveform length—a sliding greater than sub-waveform length results in losing information from the original full waveform. For each sliding, we varied the sub-waveform duration from 0.74 to 5.92 seconds. For each sliding and duration, we predicted the performance for 1 randomly selected sub-waveform and the maximum generated number of sub-waveforms.

Explanation framework

To explain our DL results, we developed an explanation framework to directly show how aligning heartbeats improves the performance. In ECG data, heart rate is predicted by measuring the time between heartbeats. Since each ECG captures several heartbeats (13 in the reference full-waveform representation), we averaged the predicted heart rates across all the heartbeats. Using heart rate variability (quantified by coefficient of variation, CV, for each predicted heart rate), we divided our ECG data into 2 groups: rhythmic (CV $\leq$0.01) and arrhythmic (CV $>$0.01). The statistical plots (heart rate for each patient and count vs heart rate) for 2 two groups are shown in Figure 3A. We note that our reference

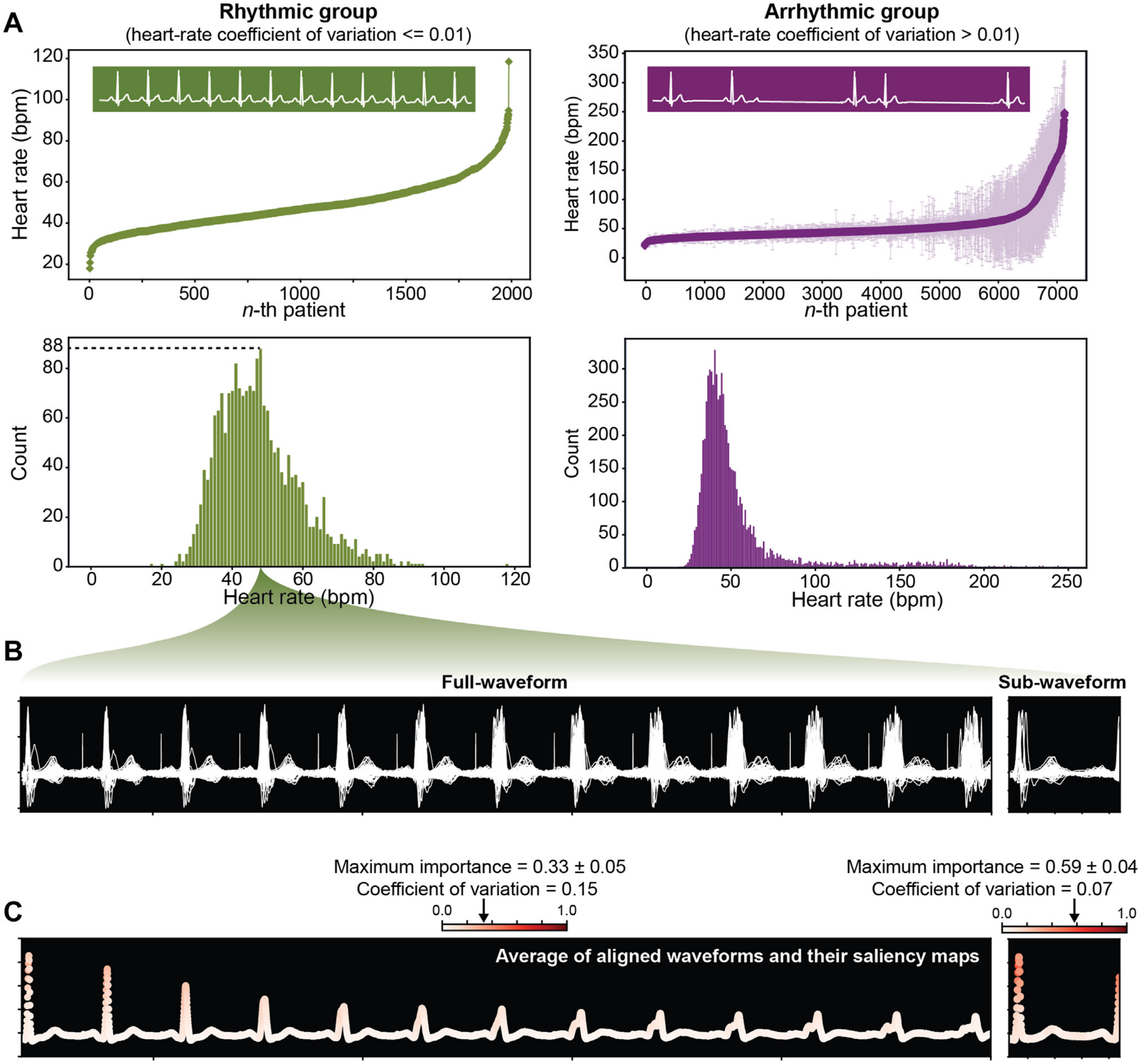

Figure 3 Explanation for enhanced learning by sub-waveform representation for 9244 patients in holdout test set. **A:** Full-waveform heart rate for each patient (top) and heart rate histogram (bottom) for rhythmic and arrhythmic groups. Heart rate coefficient of variation (CV) is set to 0.01 to define the 2 groups. The error bars come from averaging across different heartbeats of full-waveform representation. **B:** Eighty-eight lead I waveforms are taken from rhythmic group corresponding to heart rate with maximum count and waveforms are left aligned for each representation. **C:** The aligned waveforms and their saliency maps are averaged. The absolute values of a saliency map are min-max normalized for each waveform. As predicted by higher averaged maximum importance and lower CV, features are more aligned with less uncertainty in sub-waveform representation.

heart rate of 78 bpm falls within the 95th percentile of rhythmic and arrhythmic ECGs.

To understand the impact of heartbeat alignment on performance, we focused on the rhythmic group because arrhythmic ECGs cannot be aligned owing to the random distribution of heartbeats across the full waveform. Thus we statistically investigated the learning for a set of rhythmic ECGs with maximum count that corresponds to 88 ECGs with a heart rate of 48-bpm. We focused our analysis on lead I, as that is also used in wearable devices,[39] but a similar analysis is extensible to other leads. We left-padded the lead I waveforms by matching the first QRS complexes and then right-padded the waveforms to match the length of the left-padded waveforms. The aligned waveforms are shown in Figure 3B for both full-waveform and sub-waveform representations. For each of the 88 waveforms, we calculated a saliency map using the DeepLIFT approach, which shows the gradient of predicted outcome with respect to the changes in input ECG.[40] For a binary outcome, the positive and negative values of a saliency map show the contribution of features to positive and negative outcomes, respectively. Because we are interested in exploring the impact of alignment on all features, we kept both positive

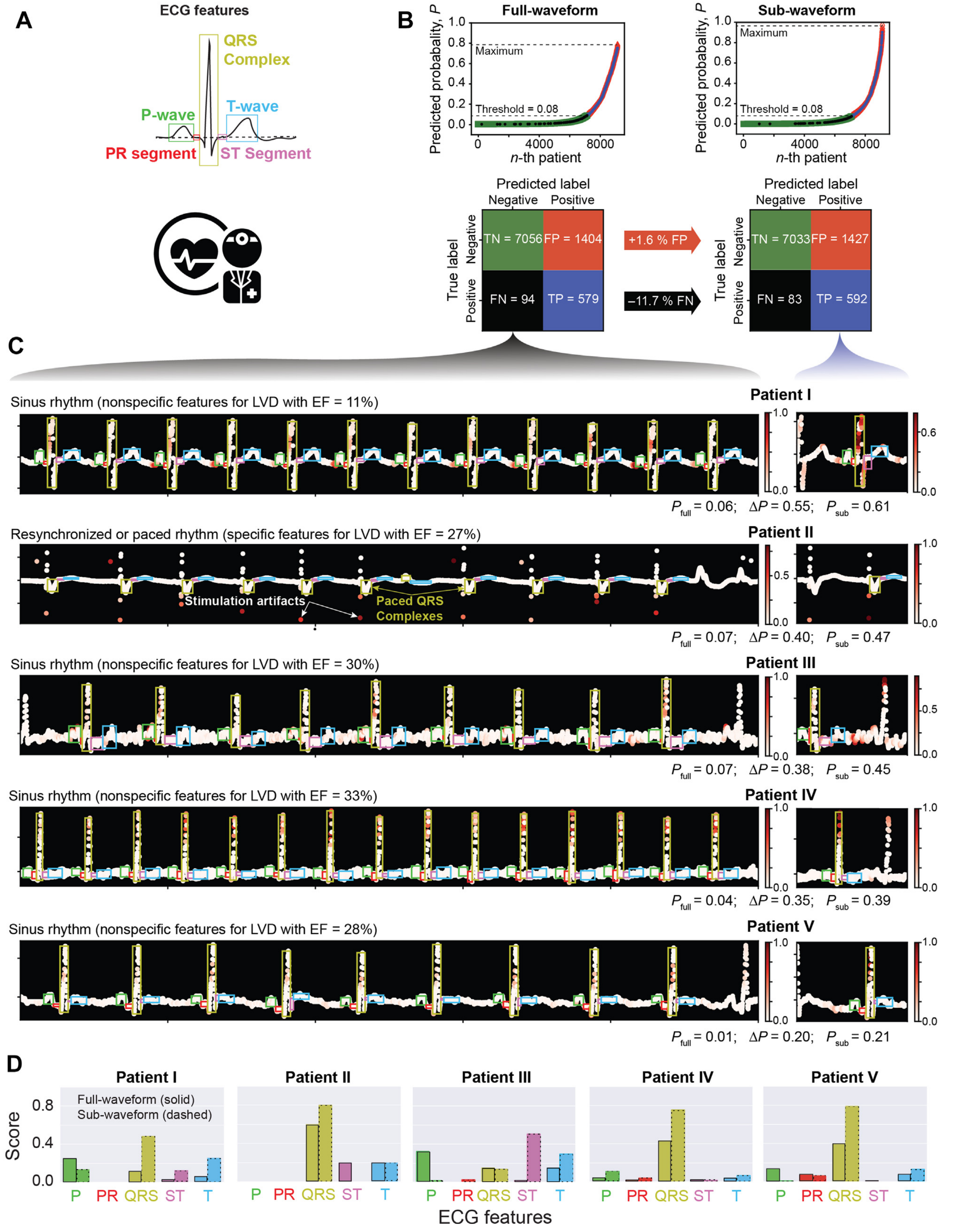

Figure 4 Interpretation of electrocardiographic (ECG) features for full-waveform and sub-waveform representations. **A:** ECG features: P wave, PR segment, QRS complex, ST segment, and T wave. The bounding boxes show the extents of each feature. **B:** Predicted probability for each patient and corresponding confusion matrix. **C:** Positive saliency maps for top 5 patients with positive outcome that are classified as false-negative by full-waveform representation and turned into true positive by sub-waveform. Top-5 shows the ranking with respect to the difference in probabilities of sub-waveform and full-waveform representations. **D:** For each patient, P, PR, QRS, ST, and T importance scores are calculated and shown as a bar chart for both representations.

and negative portions of the saliency map and used the min-max normalization of the absolute values of the saliency map. We then averaged aligned waveforms and their saliency maps for both representations, as shown in Figure 3C. We only included the nonzero (ie, nonpadded) overlapping parts of all the waveforms for averaging across samples. In the Supplemental Material, we provide the algorithmic steps for our explanation framework.

Interpretation framework

The black-box nature of DL precludes understanding what features in ECG waveforms are meaningful and important for a particular task.[41–43] Despite the high performance that DL models for ECGs can achieve in a variety of tasks, interpreting the most relevant features is still a challenge, thus limiting their use in clinical workflows.[5,6,16] Therefore, identifying clinically relevant features is crucial for tailoring further workup and treatment strategies to optimize efficacy in improving symptoms and clinical outcomes.[44] As shown in Figure 4A, any waveform is characterized by 5 ECG features: P wave, PR segment, QRS complex, ST segment, and T wave.[1] These features provide significant clinical information for assessing ECGs.

We developed an algorithm to interpret the clinical relevance of important ECG features for DL predictions. For each patient, we calculated the positive saliency maps for both representations (Figure 4C), as detailed in Supplemental Material. We highlight that we focused only on the positive portion of a saliency map that contributes to the positive outcome, since we are interested in patients with positive outcome, unlike an explanation saliency map, for which we considered both positive and negative portions.

Owing to the higher predictive performance of sub-waveform representation, we are interested in interpreting the ECGs of patients that are classified as false-negative (FN) in full waveform and turn into true positive (TP) using sub-waveform, because this allows us to better understand the clinical relevance of our predictions and how sub-waveform representation drives the decisions from FN to TP. Therefore, we ranked the ECGs based on the difference in DL predicted probabilities of sub-waveform TP and full-waveform FN.

We focused on the top 5 ranked ECGs and our cardiac electrophysiology team annotated the 5 ECG features using the full waveform and across all heartbeats and leads (see Supplemental Material). For sub-waveform, we used the same exact annotations as full waveform to provide fair comparisons. For each ECG feature, we calculated a normalized score using the saliency map and annotations, which is also described in the Supplemental Material. Each score shows the contribution of the ECG feature to the predicted probability—the higher the score, the higher the contribution to predicted probability.

Subgroup analysis

Another concern for DL in healthcare is the potential disparity that the predictive capabilities of these models are not fair to all subpopulations, which negatively impact certain subgroups in society.[45] To mitigate these biases, fairness researchers have proposed algorithmic techniques to minimize disparities in predictions (1) across subgroups (called “group fairness”) and (2) within a subgroup (called “individual fairness”).[46] In our study, a potential source of bias for predictive performance may arise from higher number of rhythmic vs arrhythmic ECGs in each subgroup,[47] which can be mitigated by optimizing the waveform representation. To investigate the impact of waveform representation on subgroup disparities, we quantified the prevalence of arrhythmia and performance metrics for 14 racial, ethnic, and sex subgroups in our holdout test.

Results

Predictive performance for identification of LVD

The performance metrics for full waveform (baseline) are shown in Figure 2 and highlighted by dashed horizontal lines. Our full-waveform predictions are close to the results in reference 25 despite using different datasets and model architectures. We found that representations with maximum number of sub-waveforms outperforms 1 sub-waveform for both sliding experiments because of the alignment effect, as explained below. In addition, maximum sliding (Figure 2B) has a lower performance than minimum sliding (Figure 2A) because a smaller number of sub-waveforms are generated, which weakens the alignment. We found the optimal performance to be when the duration is 1.48 seconds (2 reference heartbeats with 10 sub-waveforms), as highlighted by dashed circles. In addition to changes in performance, another important observation is changes in uncertainties. We observed that sub-waveform predictions have smaller uncertainties compared to full-waveform predictions. To better understand the variabilities in a prediction, we used the coefficient of variation (CV)—the ratio of standard deviation (uncertainty) to mean. In particular, the full-waveform CV is 0.010 for AUROC and 0.094 for AUPRC and these values for the optimal sub-waveform are reduced to 0.001 and 0.017.

Explanation for underlying mechanism of DL improvements

We observed that the averaged maximum importance is increased from 0.33 to 0.59 and the averaged CV is decreased from 0.15 to 0.07 by changing the full-waveform to optimal sub-waveform representation, as illustrated in Figure 3C. This observation is remarkable, as it directly shows the enhancement owing to alignment of rhythmic sub-waveforms. In the full-waveform representation of rhythmic ECGs, the repeated heartbeats that are similar in shape introduce redundancies in the learned weights and sub-waveform representation improves this deterioration in learning by aligning heartbeats. In particular, as shown in the Supplemental Material, we found that full waveform causes significant overfitting in sallower NN while sub-waveform is stable in terms of overfitting with respect to changes in depth of NN.

Even though the main improvements arise from rhythmic ECGs, we emphasize that sub-waveform representation still captures the features of arrhythmic ECGs, as supported by our performance results. This is because the optimal case's duration corresponds to the maximum heart rate of 156 bpm, which falls within the 99th percentile of arrhythmic ECGs.

Interpretation of ECG features

For interpretation analysis, we focused on the full-waveform model with AUROC = 0.918, AUPRC = 0.473 and the optimal sub-waveform model with AUROC = 0.926 and AUPRC = 0.578. The predicted probabilities and confusion matrices for both representations are shown in Figure 4B. In fact, there were 35 patients with full-waveform FN converted to sub-waveform TP, compared to 22 patients that were full-waveform TP and sub-waveform FN. Following our ranking strategy, we focused on the top 5 patients, with the difference in probabilities ranging from 0.55 to 0.20.

Using sub-waveform representation, we observed a 1.6% increase in FP and an 11.7% decrease in FN (Figure 4B). It is important to highlight the decrease in FN, as the class imbalance for the 2 representations is the same and the percent positive is 0.076. For our holdout test set, there are only 703 positive patients and 8543 negative patients. Especially for emerging AI technologies in healthcare, it is extremely important (1) not to miss any positive patient (ie, low FN) and (2) not to have many false alarms that can impose additional costs and alarm fatigue (ie, low FP).[48] We observed that out of the top 5 ranked patients, the ECGs for patients I and III–V are clearly rhythmic. But the rhythmic pattern for patient II is less clear. To better understand the characteristics of these ECGs, our cardiac electrophysiology experts confirmed that patients I and III–V have overall sinus rhythms[49] and patient II has a vector of pacing that is consistent with cardiac resynchronization therapy (CRT), which suggests that the patient likely has underlying LVD, as CRT is normally used in patients with heart failure and left bundle branch block to restore left and right ventricular synchrony or in patients with LVD who have a high burden of pacing.[50] This clinical validation is important to reconfirm that sub-waveform mainly impacts the learning of rhythmic ECGs, as per the underlying hypotheses.

In Figure 4D, we observed that sub-waveform representation transforms the important ECG features in all patients. To better understand the meaning of the scores and connect our DL predictions of important features with cardiologists' predictions, we asked our cardiac electrophysiology team to determine the specific and nonspecific LVD-related features. They identified only patient II as the person with specific features for LVD, mainly owing to the paced vector of QRS complexes being consistent with CRT pacing. We then investigated our saliency maps (either of full waveform or sub-waveform) to search for this feature. We observed that saliency maps visually highlight the stimulation artifact as the most important feature, but this is clinically not meaningful. We noted that the ECG of patient II does not have a P wave and PR segment. Also, determining the extents of QRS complex, ST segment, and T wave is nontrivial for a nonexpert. We observed that the importance scores can predict the paced QRS complex as the most important. In full-waveform representation, the QRS complex has a score of

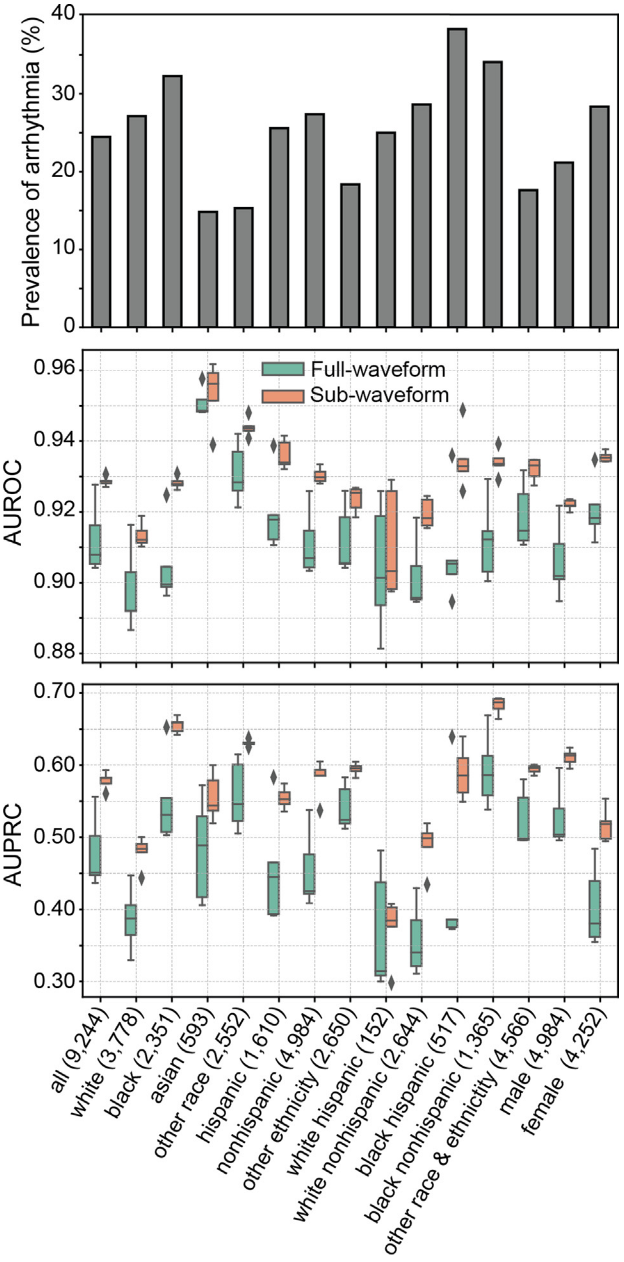

Figure 5 Impact of electrocardiographic waveform representation on 14 subgroups in holdout test set. Top plot shows prevalence of arrhythmia in each subgroup. Area under receiver operating characteristic (AUROC) and area under precision-recall curve (AUPRC) for full-waveform (green) and sub-waveform (orange) representations are shown in middle and bottom plots. The number of patients in each subgroup is shown in the parenthesis. Less prevalence of arrhythmia in a subgroup may induce redundancies owing to higher number of rhythmic full waveforms that cause higher deep learning prediction uncertainties. Sub-waveform representation provides individual fairness by reducing these disparities within a subgroup.

0.6 and it changes to 0.8 in the sub-waveform case by suppressing the ST-segment score of 0.2. This clinically meaningful enhancement is the leading cause for pushing the predicted probability of LVD from 0.07 to 0.47. In terms of nonspecific features, they noted the widened and notched QRS complex in patient I, as captured by our scoring, to be abnormal depolarization and nonspecific for LVD. They confirmed ST depression in patient III as a nonspecific feature for LVD in association with myocardial ischemia and our scoring system also predicted that important feature switches from P wave, with a score of 0.32, to ST segment, with a score of 0.50.

Impact of ECG waveform representation on subgroups

In Figure 5, the prevalence of arrhythmia for 14 racial, ethnic, and sex subgroups in our holdout test set is shown. We have predicted the AUROC and AUPRC for these subgroups and the results are shown in Figure 5. For both representations, we observed disparities across subgroups that can be mitigated using group fairness techniques; this is not the focus of this work. Relevant to the current work, we focus on individual fairness, since we observed that for each subgroup, DL uncertainty using the sub-waveform representation is much smaller than using the full-waveform representation. This implies that sub-waveform representation can help to mitigate the disparities in DL predictions within a subgroup with a lower prevalence of arrhythmia (ie, higher number of rhythmic ECGs).

Discussion

This work proposes a new sub-waveform representation of ECGs to enhance the DL predictions. This extension to traditional full-waveform representation shows that rearranging the waveforms is an added control that provides opportunities for enhancing DL modeling capabilities. We clarify that the sub-waveform representation is solely a form of input-data transformation and to investigate its impact on learning, we intentionally kept all parameters of NN fixed.[51] Therefore, any gain is mainly owing to the sub-waveform representation.

We systematically investigated the performance of the proposed representation for identifying LVD and observed improvements in performance metrics and, importantly, reductions in uncertainties. As uncertainty is a source of bias[52] and we observed variations in prevalence of arrhythmia across subgroups, we investigated the impact of data representation on subgroups. Because predictions for a subgroup with low prevalence of arrhythmia (ie, higher number of rhythmic ECGs) can be biased owing to full-waveform redundancies, they can thus benefit from sub-waveform representation owing to significantly reduced uncertainties for rhythmic ECGs. We emphasize that the change in uncertainty is a direct consequence of ECG waveform representations and their impact on DL optimization stability. Our subgroup analysis shows the importance of the proposed representation for individual fairness, which primarily aims at providing homogenous predictions for individuals within the same subgroup.[53] Indeed, we showed that waveform representation directly controls the fluctuations in DL predictions and thus can be used as a bias mitigation tool. To the best of our knowledge, individual fairness has not been investigated in this context, which can help to achieve the ultimate goal of operationalizing fairness for new ECG-AI systems in the real world.[54]

We explained how assigning different weights to similarly shaped heartbeats in the full-waveform of each lead impedes the NN from finding the optimal features across similar ECG examples. This creates redundancies that cause overfitting and deteriorate the learning. However, by using sub-waveform representation, assigning different weights to similar heartbeats is less likely because of the reduced number of heartbeats and increased number of aligned sub-waveforms, which can also be treated as new training examples per patient. Our explanation analysis provided the evidence for our hypothesis on improved learning by aligning heartbeats in the sub-waveform representation, which provides better localization of important features with less uncertainty.

To provide clinical interpretation of our predictions, we developed a novel scoring system for quantifying the DL importance of ECG features. Although visualizing saliency maps that highlight only the important ECG data points have been used before extensively, to the best of our knowledge this is the first attempt in developing such a scoring system for ECG-DL modeling. As we showed throughout this work, we need to investigate the importance of a collection of data points that represent an ECG feature rather than only showing the importance of data points. Our interpretation framework showed that sub-waveform representation generally performs better at localizing and highlighting important features that result in a higher prediction probability. To verify the predicted scores, we applied our framework to a paced ECG with specific features of LVD, for which, determining the importance of ECG features by visual examination of saliency maps was difficult. The importance scores were qualitatively in high agreement with the predictions of our cardiac electrophysiology team, connecting which would have been very difficult without quantifying importance scores. Other than this clinical validation, we also showed how our interpretation framework assists clinicians in cases for which the ECG features are nonspecific. For example, our scoring system was useful to assist the clinicians in determining the relative importance of otherwise nonspecific findings such as widened QRS complex or ST-segment depression.

In summary, we introduced a novel sub-waveform representation to enhance the DL modeling of ECG waveforms. We showed that the proposed representation can perform better than the traditional full-waveform representation for the identification of LVD. We explained the underlying mechanism for DL improvements gained by aligning sub-waveforms. We provided an interpretation framework and validated this framework against cardiologists' predictions. Finally, we showed the advantages of developed representation in mitigating uncertainties within subgroups.

Limitations

We highlight several limitations of our study. Sub-waveform representation can be ineffective for ECGs that have strong temporal dependencies. Also, we do not have an evaluation of the impact of representation on other outcomes, data sources, and architectures. In addition, we have interpreted our results only for 5 patients.

For future work, it is important to explore the potential benefits of the proposed representation on a variety of outcomes. Another path forward is exploring the multiscale potential of full-waveform and sub-waveform representations to gain further improvements by combining representations at different scales (ie, waveform durations). For example, one could train a model using full-waveform representations and fuse the predictions with the model(s) trained using 1 or more sub-waveform representations. This can potentially provide even further improvements. In addition, expanding and validating the interpretation framework at a larger scale for both representations is of great significance for clinical workflow because a rigorous scoring system could be used as a tool for (1) confirming clinical understanding of DL predictions when ECG features are specific and (2) informing clinicians, especially when features are nonspecific and cannot be easily determined by clinicians. In addition, more work is needed to explore the disparities arising from waveform representation and its relation to prevalence of arrhythmia for providing fairer predictions.

Data availability

The ECG data and labels from echo reports are not publicly available due to HIPAA privacy rules.

Code availability

For reproducibility, all the codes used to generate the results in this work are publicly available at https://github.com/Glicksberg-Lab/ECG_Representation. These codes include the ECG-DL trainer, evaluator, explainer, interpreter, and visualizer. Further details are provided in the GitHub link.

Acknowledgments

We acknowledge Jay Havaldar, Mark Shervy, and Manbir Singh for IT support. We thank Eddye Golden, Shelly Kaur, and Yovanna Roa for administrative and project management support.

Funding Support

This study was supported by the National Center for Advancing Translational Sciences, National Institutes of Health (NIH) U54 TR001433-05. F.W. would also like to acknowledge the support from National Science Foundation (NSF 1750326 and NIH RF1AG072449).

Disclosures

B.S.G. has received consulting fees from Anthem AI and consulting and advisory fees from Prometheus Biosciences. G.N.N. has received consulting fees from AstraZeneca, Reata, BioVie, Siemens Healthineers and GLG Consulting; grant funding from Goldfinch Bio and Renalytix; financial compensation as a scientific board member and adviser to Renalytix; owns equity in Renalytix and Pensieve Health as a cofounder and is on the advisory board of Neurona Health. The other authors declare no competing interests.

Authorship

All authors attest they meet the current ICMJE criteria for authorship.

Patient Consent

All clinical data were de-identified and written informed consent was waived.

Ethics Statement

This study has been approved by the institutional review board at the Icahn School of Medicine at Mount Sinai.

Appendix

Supplementary data

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.cvdhj.2022.07.074

References

1. Goldberger AL, Goldberger ZD, Shvilkin A. Clinical electrocardiography: a simplified approach e-book. Elsevier Health Sciences; 2017.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–444.
3. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719–731.
4. Wagner P, Strodthoff N, Bousseljot RD, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data 2020;7:1–15.
5. Somani S, Russak AJ, Richter F, et al. Deep learning and the electrocardiogram: review of the current state-of-the-art. Europace 2021;23:1179–1191.
6. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol 2021;1–14.
7. Minchole A, Camps J, Lyon A, Rodríguez B. Machine learning in the electrocardiogram. J Electrocardiol 2019;57:S61–S64.
8. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19:1236–1246.
9. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25:65–69.
10. Wołk K, Wołk A. Early and remote detection of possible heartbeat problems with convolutional neural networks and multipart interactive training. IEEE Access 2019;7:145921–145927.
11. Chauhan S, Vig L, Ahmad S. ECG anomaly class identification using LSTM and error profile modeling. Comput Biol Med 2019;109:14–21.
12. He R, Liu Y, Wang K, et al. Automatic cardiac arrhythmia classification using combination of deep residual network and bidirectional LSTM. IEEE Access 2019;7:102119–102135.
13. Le Guennec A, Malinowski S, Tavenard R. Data augmentation for time series classification using convolutional neural networks, ECML/PKDD Workshop on

Advanced Analytics and Learning on Temporal Data. Italy: Riva Del Garda; 2016, https://core.ac.uk/download/pdf/48148906.pdf. Accessed September 13, 2022.

14. Cui Z, Chen W, Chen Y. Multi-scale convolutional neural networks for time series classification. arXiv preprint arXiv:1603.06995, 2016.
15. Cao P, Li X, Mao K, et al. A novel data augmentation method to enhance deep neural networks for detection of atrial fibrillation. Biomed Signal Process Control 2020;56:101675.
16. Hong S, Zhou Y, Shang J, Xiao C, Sun J. Opportunities and challenges of deep learning methods for electrocardiogram data: a systematic review. Comput Biol Med 2020;103801.
17. Huang J, Chen B, Yao B, He W. ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network. IEEE Access 2019; 7:92871–92880.
18. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019;25:70–74.
19. Leong DP, McMurray JJ, Joseph PG, Yusuf S. From ACE inhibitors/ARBs to ARNIs in coronary artery disease and heart failure (Part 2/5). J Am Coll Cardiol 2019;74:683–698.
20. Yamani H, Cai Q, Ahmad M. Three-dimensional echocardiography in evaluation of left ventricular indices. Echocardiography 2012;29:66–75.
21. Quiñones MA, Waggoner AD, Reduto LA, et al. A new, simplified and accurate method for determining ejection fraction with two-dimensional echocardiography. Circulation 1981;64:744–753.
22. Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. Eur Heart J Cardiovasc Imaging 2015;16:233–271.
23. Farsalinos KE, Daraban AM, Ünlü S, Thomas JD, Badano LP, Voigt JU. Head-to-head comparison of global longitudinal strain measurements among nine different vendors: the EACVI/ASE Inter-Vendor Comparison Study. J Am Soc Echocardiogr 2015;28:1171–1181.e2.
24. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580:252–256.
25. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat Med 2019;25:70–74.
26. Yao X, Rushlow DR, Inselman JW, et al. Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med 2021;27:815–819.
27. Mittal R, Seo JH, Vedula V, et al. Computational modeling of cardiac hemodynamics: current status and future outlook. J Comput Phys 2016;305:1065–1082.
28. Streltsov A, Adesso G, Plenio MB. Colloquium: quantum coherence as a resource. Rev Mod Phys 2017;89:041003.
29. Popmintchev T, Chen MC, Popmintchev D, et al. Bright coherent ultrahigh harmonics in the keV x-ray regime from mid-infrared femtosecond lasers. Science 2012;336:1287–1291.
30. Hussein MI, Tsai CN, Honarvar H. Thermal conductivity reduction in a nanophononic metamaterial versus a nanophononic crystal: a review and comparative analysis. Adv Funct Mater 2020;30:1906718.
31. Honarvar H, Hussein MI. Two orders of magnitude reduction in silicon membrane thermal conductivity by resonance hybridizations. Phys Rev B 2018; 97:195413.
32. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet 2019; 394:861–867.
33. Coumel P, Maison-Blanche P, Catuli D. Heart rate and heart rate variability in normal young adults. J Cardiovasc Electrophysiol 1994;5:899–911.
34. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: European Conference on Computer Vision. Springer; 2016.
35. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR; 2015.
36. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014; 15:1929–1958.
37. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision; 2015.
38. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
39. Spaccarotella CAM, Polimeni A, Migliarino S, et al. Multichannel electrocardiograms obtained by a Smartwatch for the diagnosis of ST-segment changes. JAMA Cardiol 2020;5:1176–1180.
40. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: International Conference on Machine Learning. PMLR; 2017.
41. Wang F, Casalino LP, Khullar D. Deep learning in medicine—promise, progress, and challenges. JAMA Intern Med 2019;179:293–294.
42. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? Ann Intern Med 2020;172:59–60.
43. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–215.
44. Lampert J, Miller M, Halperin JL, et al. Prognostic value of electrocardiographic QRS diminution in patients with COVID-19. J Am Coll Cardiol 2021; 77:2258–2259.
45. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci 2020;4:123–144.
46. Bellamy RK, Dey K, Hind M, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Devel 2019;63. 4:1–4:15.
47. Dewland TA, Olgin JE, Vittinghoff E, Marcus GM. Incident atrial fibrillation among Asians, Hispanics, blacks, and whites. Circulation 2013;128:2470–2477.
48. Babic B, Gerke S, Evgeniou T, Cohen IG. Direct-to-consumer medical machine learning and artificial intelligence applications. Nat Mach Intell 2021; 3:283–287.
49. Corley SD, Epstein AE, DiMarco JP, et al. Relationships between sinus rhythm, treatment, and survival in the Atrial Fibrillation Follow-Up Investigation of Rhythm Management (AFFIRM) Study. Circulation 2004;109:1509–1513.
50. Abraham WT, Hayes DL. Cardiac resynchronization therapy for heart failure. Circulation 2003;108:2596–2603.
51. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3), 2021, pp.107-115.
52. Lachish S, Murray KA. The certainty of uncertainty: potential sources of bias and imprecision in disease ecology studies. Front Vet Sci 2018;5:90.
53. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: International Conference on Machine Learning. PMLR; 2013.
54. Gichoya JW, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ Health Care Inform 2021;28:e100289.