Sequence to Sequence ECG Cardiac Rhythm Classification using Convolutional Recurrent Neural Networks

Teeranan Pokaprakarn, Rebecca R. Kitzmiller, J. Randall Moorman, Doug E. Lake, Ashok K. Krishnamurthy, and Michael R. Kosorok

***Abstract*—This paper proposes a novel deep learning architecture involving combinations of Convolutional Neural Networks (CNN) layers and Recurrent neural networks (RNN) layers that can be used to perform segmentation and classification of 5 cardiac rhythms based on ECG recordings. The algorithm is developed in a sequence to sequence setting where the input is a sequence of five second ECG signal sliding windows and the output is a sequence of cardiac rhythm labels. The novel architecture processes as input both the spectrograms of the ECG signal as well as the heartbeats' signal waveform. Additionally, we are able to train the model in the presence of label noise. The model's performance and generalizability is verified on an external database different from the one we used to train. Experimental result shows this approach can achieve an average F1 scores of 0.89 (averaged across 5 classes). The proposed model also achieves comparable classification performance to existing state-of-the-art approach with considerably less number of training parameters.**

***Index Terms*—Arrhythmia, Atrial Fibrillation (AFIB), Cardiac Rhythm, Deep Learning, Electrocardiogram (ECG), Feature extraction, Long short-term memory (LSTM), Neural Networks, Recurrent Neural Networks (RNN)**

I. Introduction

An electrocardiogram (ECG) contains information on the electrical signal activities of the heart. It can be used to produce various measures such as heart rate as well as to diagnose a condition like arrhythmia. Discerning an arrhythmia such as Atrial Fibrillation (AF) from other common cardiac rhythms with continuous monitoring is of high value because, if left undetected, those with silent AF (estimated to be about 1/3 of AF patients) experience the same risk for stroke as symptomatic AF patients [1] and AF is associated with an estimated 5-fold increase in risk of stroke [2]. Since AF is implicated in over 454,000 hospital admissions [2] and 158,000 deaths each year [3], improved automated methods to passively identify AF may significantly reduce associated morbidity and mortality for those unaware they are at risk. Diagnosis often depends on patients' first recognizing symptoms (e.g., elevated heart rate, palpitations, fatigue) [4] and seeking care where careful clinical assessment and expert cardiologist review of electrocardiogram (ECG) data remains the goal standard to confirmed diagnosis and select intervention [5]. Although the American Heart Association treatment guidelines fail to mandate a minimum threshold of AF burden (i.e, number of minutes in 24 hours) that require treatment [5] recent studies note significant increase in stroke risk as AF burden increases (i.e., 2.5 hours in 24 hours) (see [6]–[8]) suggesting the importance of early identification methods.

Machine learning based detection methods, where a model is trained using an existing database with annotated ECG records to learn the task of classifying arrhythmias, have had a long history of research mostly on the task of heartbeat classification. Deep learning based models has also emerged as successful in this task. [9]–[13] Most approaches focus on the beat-level classification problem where each beat is classified into various types rather than on the classification of cardiac rhythms. These algorithms are trained mostly using the publicly available MIT-BIH arrhythmia database [14], whose small size and data imbalance in terms of cardiac rhythm classes may have hindered the ability to train machine learning based model for cardiac rhythm detection. The detection of cardiac rhythms especially Atrial Fibrillation(AF) using machine learning methods has also gained recent interests. These includes deep learning approaches such as [15], [16] for AF detection as well as the deep learning model in [17] which can classify 12 rhythm classes. The detection of cardiac

This work was supported in part by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR002489 and Grant Award Number KL2TR001109 as well as by the National Cancer Institute through Grant Award Number P01 CA142538.

Teeranan Pokaprakarn and Michael R. Kosorok are with the Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27516 USA (e-mail: teeranan@email.unc.edu (corresponding author), kosorok@bios.unc.edu)

Rebecca R. Kitzmiller is with the School of Nursing, University of North Carolina, Chapel Hill, NC 27516 USA (e-mail:kitzm002@email.unc.edu)

J. Randall Moorman is with the Cardiology Division, Department of Internal Medicine, School of Medicine, University of Virginia, Charlottesville, VA 22903 USA and with AMP3D, Advanced Medical Predictive Devices, Diagnostics, and Displays, Inc, Charlottesville, VA 22902 USA. Conflict Statement: J. Randall Moorman owns stock in Medical Predictive Science Corporation and Advanced Medical Predictive Devices, Diagnostics, and Displays (e-mail:rm3h@virginia.edu)

Doug E. Lake is with the Department of Medicine, Cardiovascular Medicine, University of Virginia, Charlottesville, VA 22903 USA (e-mail: dlake@virginia.edu)

Ashok K. Krishnamurthy is with the Renaissance Computing Institute (RENCI) and the Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599 USA (e-mail: ashok@renci.org)

rhythms such as Atrial Fibrillation(AF) where segments in a ECG recording are given rhythm labels is the focus of this work. Each cardiac rhythm has distinct characteristics. For instance, AF is characterized by fast irregular rhythm and the absence of P waves in ECG.

The contributions of the current paper are as follow. First, we propose a novel architecture depicted in Fig. 1 that incorporates each heartbeat's waveform morphology and the spectrogram-based pattern of heartbeats over time. This is specifically designed for the cardiac rhythm detection task that requires both beat-level information as well as the temporal pattern of the heart beats and their relationships. Second, the proposed approach works in a sequence to sequence setting where the input is a sequence of signal segments and the output is a sequence of cardiac rhythm labels. This allows both segmentation and classification of cardiac rhythm within an ECG recording of varying lengths. On the clinical side, this means that the burden of AF, i.e., the frequency and duration of the episode can be measured, and therefore help inform clinical decision making [18]. Third, we demonstrates the ability to train the model robustly in the presence of noisy labels (i.e., labels are not ground-truth annotations) by using an existing Holter monitor database with labels supplied by the manufacturer's automatic classifier. This is inspired by previous works outside of the healthcare domain that uses large datasets with noisy annotations instead of laboriously collected clean annotations to train neural networks [19]–[21]. Finally, we assess model performance on samples from an external database with ground-truth annotations to verify that the model generalizes beyond the database we use to train the model.

Related works with details on other existing approaches to arrhythmia detection are presented in Section II. Section III describes our proposed methodology. The experimental setup is given in Section IV, and the details on how we train the model is given in Section V. Finally, Section VI, VII, VIII contains the results,discussion, and conclusion respectively.

II. Related Works

Much of the previous work in arrhythmia detection that uses deep learning approaches (for review, see [22], [23]) has focused on the heartbeat recognition/classification problem. Other proposed machine learning based approaches include Support Vector Machine based systems [24], [25]. These approaches rely on hand-crafted features, unlike deep learning approaches, where feature extraction is learned by the model. Successful deep learning approaches include, for instance, using artificial neural networks with a Kohonen layer [9], a 1-D convolutional neural network [10], [11], convolutional denoising autoencoders [26], and an LSTM-based algorithm [12]. A sequence to sequence model has also been proposed at the beat-level classification [13]. Similar to others, a part of our work includes extracting each heartbeat's features from ECGs using Convolutional Neural Networks (CNNs). The focus of our work is on the task of cardiac rhythm detection (not on beat level classification), a task that requires incorporating these features with the timing and the context of the surrounding beats in order to detect the rhythm class.

Because Atrial Fibrillation is the most common form of cardiac arrhythmia, the detection of this rhythm has been an active area of research. Some studies use RR interval analysis for this task. These include studying sample entropy measures [27] as well as scatterplots of RR intervals versus change of RR intervals [28]. These detection methods rely only on the temporal characteristics of cardiac rhythms. In recent years, deep learning approaches have also gained popularity. For example CNN [15], as well as LSTM [16], [29], have been used for the classification task. More recently, the 2017 PhysioNet Challenge [30] resulted in many methods successful in classifying short (9s - 60s) single-lead ECG recordings to 4 classes: 1) Normal sinus rhythm, 2) AF, 3) Other rhythm, or 4) Noise. Some of the top performing methods include using a combination of expert generated features and features from neural networks [31]; Recurrent Neural Networks (RNN) [32], [33]; Convolutional Recurrent Neural Networks (CRNN) [34]; and Random Forests based on engineered features [35]. One limitation of the aforementioned methods and the PhysioNet Challenge methods is that the algorithms only produce as output a single label for each input ECG recording. Unlike real-world ECG signals, the 2017 PhysioNet Challenge data were segmented to contain only one rhythm class. Our approach allows the input to contain more than one rhythm class and performs segmentation as well as classification. One way to perform segmentation and classification at the same time is for the output of the algorithm to be a sequence of labels. This is proposed in [17] where a 34-layers CNN takes as input a 30 s record and outputs a sequence of 23 rhythm labels. The deep architecture allows the model to reduce the 30 s 200 Hz record to a sequence of output labels. We take a different approach allowing the input to be a sequence of sliding (much shorter) windows of signal that then generate a sequence of rhythm label outputs. This allows the Convolutional Neural Network part of our architecture to have fewer layers while Recurrent Neural Network layers are used to process the signal windows sequentially and generate a sequence of labels as output. This approach also allows the input recording to be of varying length, unlike, the CNN in [17] that requires a fixed length input.

III. Proposed Methodology

The high-level overview of the proposed methodology is illustrated in Fig. 1. The algorithm takes as input a sequence of 5 seconds ECG signal segments and outputs a sequence of ECG arrhythmia rhythm labels. 5 second ECG signal segments are fed into two parts of the model to generate two different sets of features that represent the segment: 1. heartbeat waveform features and 2. spectrogram features. This architecture is uniquely designed to synthesize two sources of information from ECG signals specifically for the rhythm classification task: morphology of each heart beat as well as the pattern of heart beats underlying the rhythm over a longer period of time.

Note that the parameter values given in this section are in the context of ECG recordings with a sampling rate of 200 Hz.

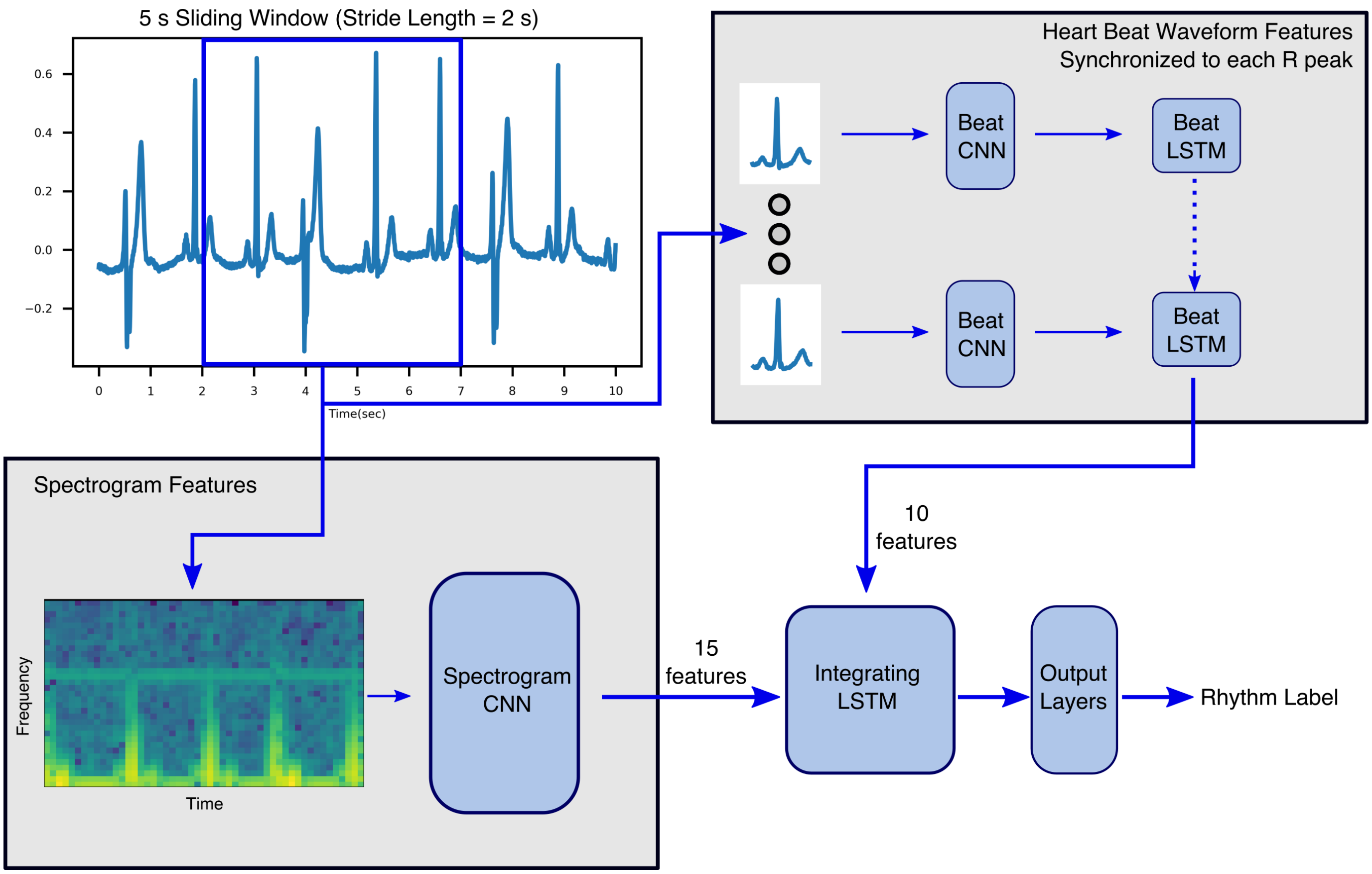

Fig. 1. Overview diagram of the proposed model

A. Pre-processing

Only minimal pre-processing transformation is applied to the raw ECG signal. For each segment of ECG signal, the signal is demeaned and rescaled so that the range is 1 as described by the function

$$f(x) = \frac{x - \bar{x}}{\max(x) - \min(x)},$$

where x is a vector of the raw values of a segment ECG signal.

In addition, the two pre-processing steps described below are applied for each of the two main feature sets of the model.

1) R-peak detection: A QRS complex detection algorithm is executed to detect the QRS complex so that the R peaks can be located in the ECG signal segment. There are many accurate and reliable algorithms available for this task (see [36] for comparison). Tools, for example [37], that analyze ECG signals often start with this QRS detection step in order to extract useful information such as the RR interval time series. For our method, we need the R peak locations to extract the beat waveform features of the signal values surrounding these locations.

2) Spectrogram: A spectrogram is often used in end-to-end automatic speech recognition systems (ASRs), for instance in [38]–[41], to generate numerical input features for neural networks from raw audio signal. To produce a spectrogram, a Short-Time Fourier Transform (STFT) of the ECG signal is first computed with the following parameter: size of FFT window = 64 samples with Hann window function, stride length = 20 samples. The power spectrum is then computed and transformed to the decibel scale.

B. Heartbeat Waveform Features

The objective of this feature set is to extract information about the waveform of each heartbeat from a given segment of ECG signal. The morphology around the QRS complex helps distinguish between types of beats, for instance, determining a premature ventricular contraction beat from a normal sinus beat. Using the locations of detected R peaks, we extract 0.7 seconds of signal surrounding each location (0.3 seconds before and 0.4 seconds after). This means that for each segment of ECG signal, we have a sequence of 0.7 second signal chunks (one for each heartbeat). Note that different segments of ECG signal will have different numbers of beats and therefore the sequences may be of different lengths.

Each 0.7 second signal chunk is passed through a 4-layer Convolutional Neural Network (CNN) to be encoded into 10 features. Each convolutional layer is followed by a Rectified Linear Unit (ReLU) activation function [42] and Batch Normalization (BN) [43]. Downsampling is done directly by each convolutional layer with a stride of 2. Moreover, the number of filters is also halved at every convolutional layer. The architecture of this CNN is shown in Fig. 2.

The 10 features output from the Beat CNN is then fed into a unidirectional Long Short-term Memory (LSTM) [44] with 1 hidden layer and 10 features in the hidden states.

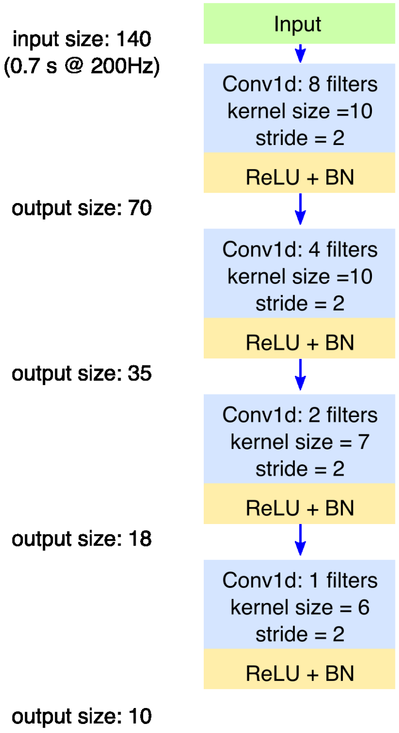

Fig. 2. Diagram of the Convolutional Layers used to process each heart beat's signal

To obtain a fixed-dimensional feature set for the varying length sequence of beats, we adopt a many-to-one sequence modelling structure. Therefore, only the 10-features output from the last beat's LSTM is passed to the integrating LSTM (described in Section III-D).

C. Spectrogram Features

A two-dimensional spectrogram with time on the x-axis and frequency on the y-axis is fed into another Convolutional Neural Network (CNN) to produce 15 features. Two-dimensional convolution is performed at each layer followed by a Rectified Linear Unit (ReLU) activation function [42] and Batch Normalization (BN) [43]. Downsampling is then performed by maxpooling. This architecture is illustrated in Fig. 3. This CNN extracts the local features of the spectrogram of a 5 seconds ECG signal segment.

D. Integrating LSTM and Output Layers

Feature sets from section III-B (15 features) and section III-C (10 features) are then combined to form the local representation of the 5 seconds ECG signal segment. This combined set of features is then fed into a 2-layer bidirectional LSTM with 10 hidden units. The bidirectional LSTM serves as an aggregator of information across the surrounding ECG signal segments so that both the local features of the ECG signal segment as well as the surrounding context of that segment can be summarized and fed as input to the output layers. Note that we adopt the many-to-many RNN sequence structure so that a label is generated for each 5s window. Finally, 20 features (10 from each direction) are passed to the 2 output layers. The first output layer has 10 hidden units with ReLU activation function [42], and in the last layer the number of hidden units equals the number of ECG rhythm classes with the softmax functions as activation at the output nodes.

IV. Experimental Setup

The ECG cardiac rhythm detection model with the proposed architecture described above is trained with a database of Holter Recording and its performance is evaluated on an external database.

A. Datasets

1) UVA Holter Recordings: Training a neural network from scratch requires a large dataset. We first use a subset of a large dataset which contains 24 hours of Holter recordings sampled at 200 Hz collected at University of Virginia (UVA) Heart Station and used in previous studies on heart rate dynamics [45], [46]. We restrict the dataset to patients of age 25 years or older. The age restriction is chosen to resemble patients in the MIT-BIH Arrhythmia Database [14] which is the external database we use to test the model. The total number of patients in the subset we used is 1928. The included patients are primarily female (54.41%) and range in age from 25 to 91 (mean=59). Twenty percent (n= 386) of these patients are then reserved for the test set. The remaining 1542 patients are used in the training phase. Since most of the sample sections contain only Normal Sinus Rhythm (NSR), we only include 10 percent of these NSR sections for training. However, for the test set, all the sections are used. ECG signal sections from the 1542 patients reserved for training are then split 80 percent for the training set and 20 percent for the validation set. The original data collection [45] was carried under the University of Virginia's Institutional Review Board; however, this analysis is performed on de-identified Holter recordings.

2) MIT-BIH Arrhythmia Database: To test how well the model generalizes, an external database: MIT-BIH Arrhythmia Database [14] is used as a test set. This database contains 48 half-hour two-channel ambulatory ECG recordings. The recording is sampled at 360 Hz. To be consistent with the UVA Philips Holter recordings, we resample these recordings to 200 Hz. Moreover, only lead II is used in this experiment. The annotations in this database were verified by two cardiologists with discrepancies resolved by consensus [14]. For the experiment in this paper, we consider these annotations ground-truth.

The ECG signal from both datasets are then set up according to the formulation described below in section IV-B. Sections of the ECG are also filtered out if their Signal Quality Index (SQI) is below 0.7 as measured by the PhysioNet Cardiovascular Signal Toolbox [37]. In addition, based on data availability, only five rhythm classes are used in both training and testing of this experiment: 1. Normal Sinus Rhythm (NSR), 2. Atrial Fibrillation (AF), 3. Supraventricular Tachyarrhythmia (SVTA), 4. Ventricular Bigeminy (B), 5. Ventricular Trigeminy (T). Sections of ECG recordings with rhythm labels other than these 5 classes are excluded.

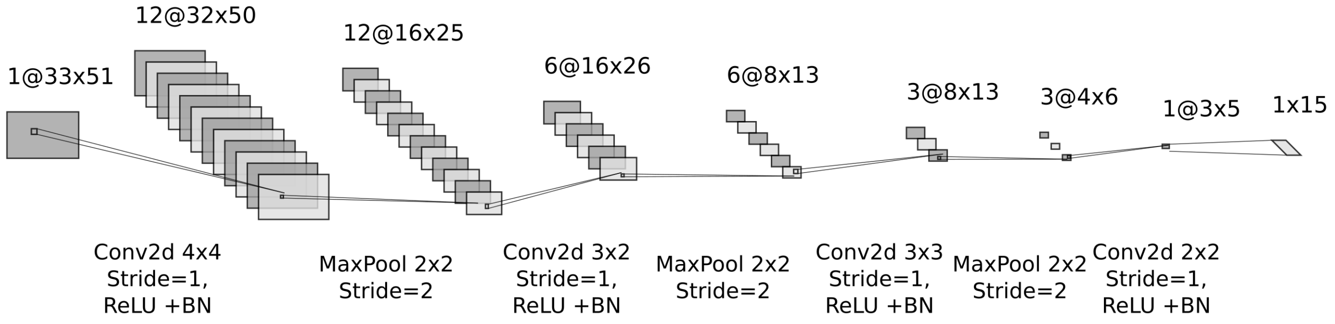
1@33x51
12@32x50
12@16x25
6@16x26
6@8x13
3@8x13
3@4x6
1@3x5
1x15
Conv2d 4x4 Stride=1, ReLU +BN
MaxPool 2x2 Stride=2
Conv2d 3x2 Stride=1, ReLU +BN
MaxPool 2x2 Stride=2
Conv2d 3x3 Stride=1, ReLU +BN
MaxPool 2x2 Stride=2
Conv2d 2x2 Stride=1, ReLU +BN

Fig. 3. Diagram of the convolutional layers used to process a spectrogram of a 5s ECG signal segment

B. Problem Formulation

We formulate arrhythmia detection as a sequence-to-sequence learning task since this allows ECG recordings to be both segmented and classified to various rhythms. The input to our algorithm is a sequence of ECG signal segments and the output is a sequence of ECG cardiac rhythm labels. In our experiment, a sliding window (length: 5s, stride: 2s) is used to generate a sequence of ECG signal segments within an ECG section of length 123s (about 2 min). Each 5s window is then assigned to one rhythm label and the algorithm produces a label output every 2 seconds. Therefore, in this setup, a 123s ECG signal section contains 60 rhythm labels.

During training, the labels and the timing of rhythm changes from an automatic classifier from Philips Holter Software is used as supervision to train the neural network. As such, we implement learning paradigms previously proposed to train the network in the presence of noisy labels, i.e., some proportion of the labels for the supervised learning are misclassified. The training with noisy labels is described in section V-B.

V. Training Procedure

We train the neural networks from scratch in mini batches of size 50 samples with the oversampling scheme described in Section V-A. Each epoch contains 100,000 samples drawn from our training set. We also implement two recently proposed learning paradigms: Co-teaching [47] and PENCIL [48] described below during training to account for the fact that we are training the model in the presence of label noise. The number of epochs is 20 for both the model trained under the Co-teaching paradigm and the model trained under the PENCIL paradigm.

A. Dataset Imbalance

The number of ECG signal sample sections of length 123s used in the training set is 174,956 samples ($n = 174,956$). Although this is a relatively large training dataset, the dataset is imbalanced in terms of the rhythm classes as shown in Table I (Note that each sample can contain more than one rhythm classes). To address this imbalance, during training we sample these sections into each mini batch with different sampling weights assigned according to which rhythm class the sample contains. The last column of Table I shows the value of the weight w_i for each class. Since each ECG signal sample section can contain more than one rhythm class, the maximum of the weights w_i assigned for each class in that section is used. Each mini batch is a weighted random sample of size 50 drawn with replacement from the training set of size n where the sampling probability is $w_i / \sum_{j=1}^{n} w_j$ for each sample.

TABLE I
Class Imbalance

Class	Number of samples containing the rhythm	w_i
NSR	93,997	1
AF	80,757	1
SVTA	20,418	4
B	13,649	5
T	15,699	5

B. Noisy Label

Because our model is trained under the supervision of labels obtained from an automatic classifier, we assume that there is a certain level of noise in the label. It is known that deep neural network has the capacity to overfit random labels [49]. Therefore, the presence of noisy labels can negatively affect the training of the model and may result in poor generalization. To address this issue, we implemented two previously proposed learning methods described below to ensure robustness during the training of our model.

1) Co-teaching [47]: Co-teaching is a learning paradigm that utilizes two peer networks to select instances for each other to train. This sample selection is an attempt to try to select clean instances to train on so that the training data that each network uses is less noisy in terms of the label. Thus, large loss instances aren't used when updating the parameter of the network. In our experiment, $R(T)$ which controls the proportion of instances to keep during training for each epoch is set to $R(T) = 1 - 0.1 \cdot \min(\frac{T}{10}, 1)$ where $T = 1, ..., 20$ is the epoch number.

2) PENCIL [48]: Probabilistic end-to-end noise correction (PENCIL) is another framework that deals with noisy labels in a different way. In the PENCIL framework, noisy labels are used to initialize a label distribution which is then allowed to be updated during training. The proposed loss function in

PENCIL is comprised of 3 components: classification loss ($\mathcal{L}_c$), compatibility loss ($\mathcal{L}_o$), and entropy loss ($\mathcal{L}_e$). There are two hyperparameters in this loss function, α which controls the weight of $\mathcal{L}_o$ and β which controls weight of $\mathcal{L}_e$. In our experiment, we set α to 0.1 and β to 0.4. Additionally, $\lambda = 10000$ is used to update the label distribution. The number of epochs for the three steps are 3 epochs (backbone learning), 10 epochs (PENCIL learning), 7 epochs (final fine-tuning). Following examples in [48], we also halve the learning rate at the beginning of the second and third step.

C. Optimization and Hyperparameters

Optimization of all the weights in our proposed neural network is done by the Adam [50] optimizer with default parameters ($\beta_1 = 0.9, \beta_2 = 0.999$) as set by Pytorch. Various initial learning rates (0.001, 0.01, and 0.1) were tried and the learning rate of 0.01 works best based on the validation set. We also choose a relatively small batch size of 50 as findings from [51] suggest that smaller batch sizes tend to lead to better generalization.

D. Evaluation of Classification Performance

To evaluate the performance of the proposed algorithm, four measures commonly used to evaluate classification algorithm are reported: sensitivity (SEN; also known as recall), positive predictive value (PPV; also known as precision), specificity (SPEC), and F1 score which is the harmonic mean of SEN and PPV. The performance is evaluated at the ”sequence-level” (as in [17]) where a sequence of predicted rhythm label outputs is compared against a sequence of rhythm labels from a reference annotation. In our setup, the sequence of labels is generated every two seconds based on a 5s sliding window of ECG signal.

VI. Results

First, we test both the model trained under the Co-teaching and PENCIL paradigm on the UVA Holter recordings test set we reserved. The test set consists of 195,849 sample sections. Note that since the annotations come from the Philips Holter Software, the results shown in Table II and III are a comparison of our trained model labels and the labels generated by the system we used to supervise the training. For comparison with ground-truths annotation, we rely on the samples from the MIT-BIH Arrhythmia Database.

Second, samples from the MIT-BIH Arrhythmia Database are used to test both models. After restricting to just 5 classes, the test set size from the MIT-BIH arrhythmia database consists of 386 sample sections. There are 60 labels in each section based on the experimental setup described in section IV-B and, therefore, 23,160 labels in total in the test set. The overall test accuracy is 97.60 percent for Model A (Co-teaching) and 97.43 percent for Model B (PENCIL). Because of the class imbalance, we examine the models' performance for each class as shown in Table IV and Table V. This result shows that both models perform similarly well and that the models trained can generalize beyond the database that was used for training.

TABLE II
UVA Holter Results for Model A: Co-teaching

Class	SPEC (%)	SEN (%)	PPV (%)	F1 score (%)
NSR	96.39	97.91	99.48	98.69
AF	98.11	95.32	85.25	90.00
SVTA	99.77	90.64	81.00	85.54
B	99.88	89.86	81.56	85.51
T	99.95	87.59	90.24	88.89

TABLE III
UVA Holter Results for Model B: PENCIL

Class	SPEC (%)	SEN (%)	PPV (%)	F1 score (%)
NSR	96.43	97.68	99.48	98.58
AF	97.88	95.78	83.80	89.39
SVTA	99.81	92.36	84.08	88.02
B	99.91	87.77	84.23	85.96
T	99.96	88.16	90.75	89.44

TABLE IV
MIT-BIH Results for Model A: Co-teaching

Class	SPEC (%)	SEN (%)	PPV (%)	F1 score (%)
NSR	95.85	98.01	99.24	98.62
AF	99.13	98.83	92.94	95.79
SVTA	99.98	70.37	88.37	78.35
B	99.20	91.40	78.09	84.22
T	99.81	85.34	89.20	87.22

TABLE V
MIT-BIH Results for Model B: PENCIL

Class	SPEC (%)	SEN (%)	PPV (%)	F1 score (%)
NSR	95.12	97.92	99.10	98.51
AF	98.98	99.62	91.88	95.59
SVTA	99.95	70.37	77.55	73.79
B	99.27	85.53	78.35	81.78
T	99.85	85.34	91.49	88.31

Next, we compare the two models trained under the two learning paradigms for noisy labels described in Section V.B against a baseline model trained with the same hyperparameters but without the implementation of those learning methods. The baseline model was trained with early stopping. When the average F1 score across 5 classes on the validation set stopped improving for 5 epochs, the training ended and we save the best model in terms of the average F1 score as evaluated on the validation set. The performance of the models in the MIT-BIH test set is reported in Table VI and the comparison shows that the two learning methods that account for noisy labels have helped the training of our proposed model.

For comparison against existing approaches, we cannot directly compare the sequence of outputs from our model which consists of one label for each signal interval and do segmentation for samples with more than one class against other works. Therefore, we slightly modified our model to perform only one classification for each recording by replacing the bidirectional LSTM with 10 hidden-units with a unidirec-

TABLE VI
Comparison against baseline model without noisy label learning paradigms

Model	NSR	AF	SVTA	B	T
Baseline	96.51	83.57	76.36	76.74	80.92
Model A: Co-teaching	98.62	95.79	78.35	84.22	87.22
Model B: PENCIL	98.51	95.59	73.79	81.78	88.31

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3098662, IEEE Journal of Biomedical and Health Informatics

TABLE VII
COMPARISON AGAINST RECENT WORK USING F1 SCORES(%) ESTIMATED BY 5-FOLDS CROSS VALIDATION

Model	NSR	AF	Other	Noisy	Overall
Proposed†	87.3	71.4	67.8	55.6	75.5
Zihlmann et al.† [34]	87.4	69.9	66.5	54.9	74.6
Proposed*	88.2	74.6	69.5	58.1	77.4
Zihlmann et al.* [34]	88.8	76.4	72.6	64.5	79.2

F1 score averaged across NSR,AF,Other is reported as Overall following [30]
†denotes training without any data augmentation
* denotes training with data augmentation described in [34]

tional LSTM with 20 hidden-units as the integrating LSTM. The many-to-one RNN structure is adopted in this case instead to perform only one classification from the entire sample. We evaluate our proposed CRNN architecture against another CRNN [34] which achieved second best average F1 score (0.82 vs the highest 0.83) in the PhysioNet/CinC Challenge 2017(for details, see [30]). Note also that the method in [34] is the top ranked end-to-end deep learning approach that does not rely on any feature engineering. Table VII shows the comparison reporting the F1 scores estimated using the same 5-folds cross validation experiment in [34] to ensure fair comparison. We believe the proposed method achieves comparable results with much fewer number of trainable parameters in the model (∼10,000 vs ∼5 millions parameters). This suggests that our model may be easier to train as well as potentially less prone to overfitting to the training data which can lead to poor generalizability.

VII. DISCUSSION

The model complexity of our proposed model is relatively lightweight with 8,042 parameters (864 in spectrogram CNN; 513 in Beat CNN; 880 in Beat LSTM; 5,520 in Integrating LSTM; and 265 in Output Layers).

Next, an ablation study is conducted to examine the value of the two blocks in the proposed models which generate two features sets: 1. Spectrogram Features 2. Heartbeat waveform features from a sequence of beats. Table VIII shows the experimental results of removing each of the two feature sets from the model while keeping the rest of the training procedure constant. All three models in Table VIII are trained using the Co-teaching learning schemes and the test dataset is the MIT-BIH Arrhythmia Dataset. ”Beat Sequence Only” model removes the spectrogram features block from the model while the ”Spectrogram Only” model removes the heartbeat waveform features. Compared to the proposed model, we see that there is a degradation in model performance across the 5 rhythm classes for the ”Beat Sequence Only” model which relies only on the sequences of beat waveform morphology. This suggests that the 5s window spectrogram contains useful additional information for classifying these cardiac rhythms. With regards to the ”Spectrogram Only” model, the F1 scores are lower only for AF and T. This suggests that explicitly extracting heartbeat waveform as a sequence can help in detecting for instance trigeminy where every third heartbeat is a premature ventricular contraction (PVC).

Lastly, we examine some common mistakes made by the model. The confusion matrix for Model A (Co-Teaching) using ground-truth annotations from the MIT-BIH Arrhythmia Database is shown in Fig. 4. The most common mistake is that Supraventricular Tachyarrhythmia (SVTA) is mistaken as Normal Sinus Rhythm (NSR). The model also sometimes misses very short segments of SVTA. Less common mistakes include slightly shifted locations of the onsets and offsets of Bigeminy (B) and Trigeminy (T) and, therefore, some B and T labels are still labelled as NSR by the model.

TABLE VIII
F1 SCORES (%) AFTER REMOVING COMPONENTS IN THE MODEL

Model	NSR	AF	SVTA	B	T
Proposed Model	98.62	95.79	78.35	84.22	87.22
Beat Sequence Only	95.87	79.27	48.64	70.81	64.45
Spectrogram Only	97.17	91.49	79.61	85.85	43.28

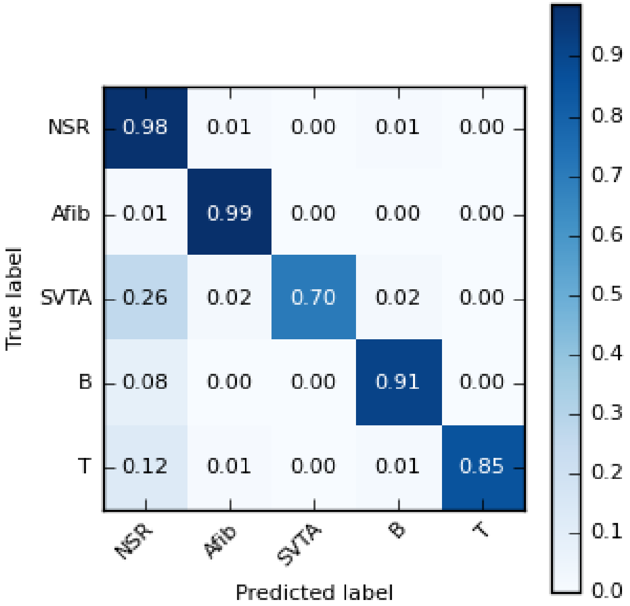

Fig. 4. Normalized Confusion Matrix for Model A: Co-teaching

VIII. CONCLUSION

To conclude, we present in this work a novel architecture that uniquely combines the morphology of each heart beat and the overall pattern of heart beats over a period of time. While other previous deep learning approaches have mostly focused on beat classifications or AF classification, our approach demonstrates the ability to differentiate various cardiac rhythms and the sequence to sequence learning setting allows our model to do both segmentation and classification of rhythms from variable-length ECG recordings. This can aid cardiologists in locating a specific rhythm within a long recording to confirm a heart arrhythmia and estimate its burden. We also demonstrate the ability to train this model in the presence of noisy labels. The use of existing databases to train deep learning models even with some noise in the labels can be helpful especially in healthcare settings where large datasets can be hard or expensive to obtain. The training with noisy labels is in contrast to other work in arrhythmia detection which often relies on databases with expert annotated labels such as the MIT-BIH arrhythmia database [14] and the much larger database used in [17]. We also show that the performance of the learned model generalizes beyond the database we used to train.

One area to expand in this research is to include other rhythms such as Atrial Flutter. Atrial Flutter is not in our analysis mainly due to data availability. However, the distinction between Atrial Flutter and Atrial Fibrillation can be challenging due to their similarities [27], [52] and often result in mislabeling by current automated systems. Further work to discern between these arrhythmias has important clinical treatment implications as most interventions come with patient risk of harm and side effect [5]. Additionally, since the experiment in this work was done only after filtering out sections of the ECG with low signal quality index, another issue to explore is to test the robustness of the method to detecting cardiac rhythms under different levels of noise and presence of artefacts in the ECG signal due to various sources such as the movement of patients.

Wearable devices hold promises for continuous health monitoring including heart activity [12], [53]–[55]. Recent innovations in wearable technology and their increasing acceptance may provide opportunities to better advise patients of potential lethal arrhythmias, such as AF [56], [57]. Our approach may lead to improved effectiveness of personal wearable devices as well as electrocardiogram (ECG) monitoring in intensive care units where AF may go undetected, contributing to increased length of ICU stay and in-hospital mortality [58]. Improvement in monitoring and diagnostics tools for heart health would have a significant public health benefit as cardiovascular disease burden remains a significant global issue.

References

[1] A. J. Camm, G. Corbucci, and L. Padeletti, "Usefulness of continuous electrocardiographic monitoring for atrial fibrillation," *American Journal of Cardiology*, vol. 110, no. 2, pp. 270–276, 2012.

[2] E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, S. R. Das, F. N. Delling, L. Djousse, M. S. V. Elkind, J. F. Ferguson, M. Fornage, L. C. Jordan, S. S. Khan, B. M. Kissela, K. L. Knutson, T. W. Kwan, D. T. Lackland, T. T. Lewis, J. H. Lichtman, C. T. Longenecker, M. S. Loop, P. L. Lutsey, S. S. Martin, K. Matsushita, A. E. Moran, M. E. Mussolino, M. O'Flaherty, A. Pandey, A. M. Perak, W. D. Rosamond, G. A. Roth, U. K. A. Sampson, G. M. Satou, E. B. Schroeder, S. H. Shah, N. L. Spartano, A. Stokes, D. L. Tirschwell, C. W. Tsao, M. P. Turakhia, L. B. VanWagner, J. T. Wilkins, S. S. Wong, S. S. Virani, and N. null, "Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.

[3] CDC, "Heart Disease: Atrial Fibrillation. Centers for Disease Control and Prevention," 2020. [Online]. Available: https://www.cdc.gov/heartdisease/atrial_fibrillation.htm

[4] J. A. Reiffel, "When Silence Isn't Golden: The Case of "Silent" Atrial Fibrillation." *The Journal of innovations in cardiac rhythm management*, vol. 8, no. 11, pp. 2886–2893, nov 2017.

[5] C. T. January, L. S. Wann, H. Calkins, L. Y. Chen, J. E. Cigarroa, J. C. Cleveland, P. T. Ellinor, M. D. Ezekowitz, M. E. Field, K. L. Furie, P. A. Heidenreich, K. T. Murray, J. B. Shea, C. M. Tracy, and C. W. Yancy, "2019 AHA/ACC/HRS Focused Update of the 2014 AHA/ACC/HRS Guideline for the Management of Patients With Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart R," *Circulation*, vol. 140, no. 2, pp. e125–e151, 2019.

[6] A. S. Go, K. Reynolds, J. Yang, N. Gupta, J. Lenane, S. H. Sung, T. N. Harrison, T. I. Liu, and M. D. Solomon, "Association of Burden of Atrial Fibrillation With Risk of Ischemic Stroke in Adults With Paroxysmal Atrial Fibrillation: The KP-RHYTHM Study," *JAMA Cardiology*, vol. 3, no. 7, pp. 601–608, 2018. [Online]. Available: https://doi.org/10.1001/jamacardio.2018.1176

[7] R. M. Kaplan, J. Koehler, P. D. Ziegler, S. Sarkar, S. Zweibel, and R. S. Passman, "Stroke Risk as a Function of Atrial Fibrillation Duration and CHA2DS2-VASc Score," *Circulation*, vol. 140, no. 20, pp. 1639–1646, 2019.

[8] I. C. Van Gelder, J. S. Healey, H. J. G. M. Crijns, J. Wang, S. H. Hohnloser, M. R. Gold, A. Capucci, C.-P. Lau, C. A. Morillo, A. H. Hobbelt, M. Rienstra, and S. J. Connolly, "Duration of device-detected subclinical atrial fibrillation and occurrence of stroke in ASSERT." *European heart journal*, vol. 38, no. 17, pp. 1339–1344, may 2017.

[9] S. L. Melo, L. P. Calôba, and J. Nadal, "Arrhythmia analysis using artificial neural network and decimated electrocardiographic data," *Computers in Cardiology*, pp. 73–76, 2000.

[10] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. S. Tan, "A deep convolutional neural network model to classify heartbeats," *Computers in Biology and Medicine*, vol. 89, no. July, pp. 389–396, 2017.

[11] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, mar 2016.

[12] S. Saadatnejad, M. Oveisi, and M. Hashemi, "LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable Devices," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2019.

[13] S. Mousavi and F. Afghah, "Inter- and Intra- Patient ECG Heartbeat Classification for Arrhythmia Detection: A Sequence to Sequence Deep Learning Approach," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May, no. 1657260, pp. 1308–1312, 2019.

[14] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia Database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, may 2001.

[15] Y. Xia, N. Wulan, K. Wang, and H. Zhang, "Detecting atrial fibrillation by deep convolutional neural networks." *Computers in biology and medicine*, vol. 93, pp. 84–92, feb 2018.

[16] Y. Chang, S. Wu, L. Tseng, H. Chao, and C. Ko, "AF Detection by Exploiting the Spectral and Temporal Characteristics of ECG Signals With the LSTM Model," in *2018 Computing in Cardiology Conference (CinC)*, vol. 45, 2018, pp. 1–4.

[17] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[18] E. N. Prystowsky, "Management of atrial fibrillation: therapeutic options and clinical decisions." *The American journal of cardiology*, vol. 85, no. 10A, pp. 3D–11D, may 2000.

[19] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning Object Categories From Internet Image Searches," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1453–1466, aug 2010.

[20] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang, "Learning from massive noisy labeled data for image classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2015, pp. 2691–2699.

[21] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei, "The Unreasonable Effectiveness of Noisy Data for Fine-Grained Recognition," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 301–320.

[22] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi, "A review on deep learning methods for ECG arrhythmia classification," *Expert Systems with Applications: X*, vol. 7, p. 100033, 2020. [Online]. Available: https://doi.org/10.1016/j.eswax.2020.100033

[23] S. M. P. Dinakarrao, A. Jantsch, and M. Shafique, "Computer-Aided Arrhythmia Diagnosis with Bio-Signal Processing: A Survey of Trends and Techniques," *ACM Comput. Surv.*, vol. 52, no. 2, mar 2019. [Online]. Available: https://doi.org/10.1145/3297711

[24] S. Osowski, L. T. Hoai, and T. Markiewicz, "Support Vector Machine-Based Expert System for Reliable Heartbeat Recognition," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 4, pp. 582–589, 2004.

[25] C. Ye, M. T. Coimbra, and B. V. Vijaya Kumar, "Arrhythmia detection and classification using morphological and dynamic features of ECG

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3098662, IEEE Journal of Biomedical and Health Informatics

signals," *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, pp. 1918–1921, 2010.

[26] K. Ochiai, S. Takahashi, and Y. Fukazawa, "Arrhythmia Detection from 2-lead ECG using Convolutional Denoising Autoencoders," in *KDD'18 Deep Learning Day, London, UK*, 2018.

[27] D. E. Lake and J. R. Moorman, "Accurate estimation of entropy in very short physiological time series: the problem of atrial fibrillation detection in implanted ventricular devices," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 300, no. 1, pp. H319–H325, 2011.

[28] J. Lian, L. Wang, and D. Muessig, "A simple method to detect atrial fibrillation using RR intervals," *American Journal of Cardiology*, vol. 107, no. 10, pp. 1494–1497, 2011.

[29] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," *Expert Systems with Applications*, vol. 115, pp. 465–473, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418305190

[30] G. D. Clifford, C. Liu, B. Moody, L. H. Lehman, I. Silva, Q. Li, A. E. Johnson, and R. G. Mark, "AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017," *Computing in Cardiology*, vol. 44, pp. 1–4, 2017.

[31] S. Hong, M. Wu, Y. Zhou, Q. Wang, J. Shang, H. Li, and J. Xie, "ENCASE: An ENsemble ClASsifiEr for ECG classification using expert features and deep neural networks," *Computing in Cardiology*, vol. 44, pp. 1–4, 2017.

[32] P. Schwab, G. C. Scebba, J. Zhang, M. Delai, and W. Karlen, "Beat by beat: Classifying cardiac arrhythmias with recurrent neural networks," *Computing in Cardiology*, vol. 44, pp. 1–4, 2017.

[33] T. Teijeiro, C. A. García, D. Castro, and P. Félix, "Arrhythmia classification from the abductive interpretation of short single-lead ECG records," *Computing in Cardiology*, vol. 44, pp. 1–4, 2017.

[34] M. Zihlmann, D. Perekrestenko, and M. Tschannen, "Convolutional recurrent neural networks for electrocardiogram classification," in *2017 Computing in Cardiology (CinC)*, 2017, pp. 1–4.

[35] M. Zabihi, A. B. Rad, A. K. Katsaggelos, S. Kiranyaz, S. Narkilahti, and M. Gabbouj, "Detection of atrial fibrillation in ECG hand-held devices using a random forest classifier," *Computing in Cardiology*, vol. 44, pp. 1–4, 2017.

[36] M. Llamedo and J. P. Martínez, "QRS detectors performance comparison in public databases," *Computing in Cardiology*, vol. 41, no. January, pp. 357–360, 2014.

[37] A. N. Vest, G. D. Poian, Q. Li, C. Liu, S. Nemati, A. J. Shah, and G. D. Clifford, "An open source benchmarked toolbox for cardiovascular waveform and interval analysis," *Physiological Measurement*, vol. 39, no. 10, p. 105004, oct 2018.

[38] L. Deng, M. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. Hinton, "Binary Coding of Speech Spectrograms Using a Deep Auto-encoder," *Interspeech*, no. September, pp. 1692–1695, 2010.

[39] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. Laurent, Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, pp. 410–414, 2016.

[40] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pp. 5884–5888, 2018.

[41] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural Speech Synthesis with Transformer Network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6706–6713, 2019.

[42] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *ICML*, 2010, pp. 807–814.

[43] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 2015, pp. 448–456.

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[45] T. J. Moss, D. E. Lake, and J. R. Moorman, "Local dynamics of heart rate: Detection and prognostic implications," *Physiological Measurement*, vol. 35, no. 10, pp. 1929–1942, 2014.

[46] M. Carrara, L. Carozzi, T. J. Moss, M. De Pasquale, S. Cerutti, M. Ferrario, D. E. Lake, and J. R. Moorman, "Heart rate dynamics distinguish among atrial fibrillation, normal sinus rhythm and sinus rhythm with frequent ectopy," *Physiological Measurement*, vol. 36, no. 9, pp. 1873–1888, 2015.

[47] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. W. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in Neural Information Processing Systems*, vol. 2018-Decem, no. NeurIPS, pp. 8527–8537, 2018.

[48] K. Yi and J. Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 7010–7018, 2019.

[49] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *5th International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://arxiv.org/abs/1611.03530

[50] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, {ICLR} 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[51] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, "On large-batch training for deep learning: Generalization gap and sharp minima," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–16, 2017.

[52] V. Lee, G. Xu, V. Liu, P. Farrehi, and J. Borjigin, "Accurate detection of atrial fibrillation and atrial flutter using the electrocardiomatrix technique," *Journal of Electrocardiology*, vol. 51, no. 6, pp. S121–S125, 2018.

[53] J. M. Bote, J. Recas, F. Rincón, D. Atienza, and R. Hermida, "A Modular Low-Complexity ECG Delineation Algorithm for Real-Time Embedded Systems," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 429–441, mar 2018.

[54] X. Wang, Q. Gui, B. Liu, Z. Jin, and Y. Chen, "Enabling Smart Personalized Healthcare: A Hybrid Mobile-Cloud Approach for ECG Telemonitoring," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 739–745, may 2014.

[55] S. P. Shashikumar, A. J. Shah, Q. Li, G. D. Clifford, and S. Nemati, "A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology," in *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, feb 2017, pp. 141–144.

[56] M. P. Turakhia, M. Desai, H. Hedlin, A. Rajmane, N. Talati, T. Ferris, S. Desai, D. Nag, M. Patel, P. Kowey, J. S. Rumsfeld, A. M. Russo, M. T. Hills, C. B. Granger, K. W. Mahaffey, and M. V. Perez, "Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study," *American Heart Journal*, vol. 207, pp. 66–75, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002870318302710

[57] J. M. Bumgarner, C. T. Lambert, A. A. Hussein, D. J. Cantillon, B. Baranowski, K. Wolski, B. D. Lindsay, O. M. Wazni, and K. G. Tarakji, "Smartwatch Algorithm for Automated Detection of Atrial Fibrillation." *Journal of the American College of Cardiology*, vol. 71, no. 21, pp. 2381–2388, may 2018.

[58] T. J. Moss, J. F. Calland, K. B. Enfield, D. C. Gomez-Manjarres, C. Ruminski, J. P. Dimarco, D. E. Lake, and J. R. Moorman, "New-onset atrial fibrillation in the critically III," *Critical Care Medicine*, vol. 45, no. 5, pp. 790–797, 2017.