MINA: Multilevel Knowledge-Guided Attention for Modeling Electrocardiography Signals

Shenda Hong1,2,5, Cao Xiao3, Tengfei Ma4, Hongyan Li1,2 and Jimeng Sun5
1School of Electronics Engineering and Computer Science, Peking University, China
2Key Laboratory of Machine Perception (Ministry of Education), Peking University, China
3Analytics Center of Excellence, IQVIA, USA
4IBM Research, USA
5Department of Computational Science and Engineering, Georgia Institute of Technology, USA

Abstract

Electrocardiography (ECG) signals are commonly used to diagnose various cardiac abnormalities. Recently, deep learning models showed initial success on modeling ECG data, however they are mostly black-box, thus lack interpretability needed for clinical usage. In this work, we propose Multilevel kNowledge-guided Attention networks (MINA) that predict heart diseases from ECG signals with intuitive explanation aligned with medical knowledge. By extracting multilevel (beat-, rhythm- and frequency-level) domain knowledge features separately, MINA combines the medical knowledge and ECG data via a multilevel attention model, making the learned models highly interpretable. Our experiments showed MINA achieved PR-AUC 0.9436 (outperforming the best baseline by 5.51%) in real world ECG dataset. Finally, MINA also demonstrated robust performance and strong interpretability against signal distortion and noise contamination.

1 Introduction

Heart diseases are among the leading causes of death of the world [Benjamin et al., 2018]. The routine monitoring of physiological signals is deemed important in heart disease prevention. Among existing monitoring technologies, electrocardiography (ECG) is a commonly used non-invasive and convenient diagnostic tool that records physiological activities of heart over a period of time. Deciphering ECG signals can help detect many heart diseases such as atrial fibrillation (AF), myocardial infarction (MI), and heart failure (HF) [Turakhia, 2018; Yanowitz, 2012].

An example of real world ECG signal is shown in Figure 1. ECG signals from cases and controls of heart diseases show different patterns at 1) beat level, 2) rhythm level, and 3) frequency level, each representing different anomalous activities of the heart. For example, beat morphology such as P wave (atrial depolarization) and QRS complex (ventricular depolarization) can reflect conditions related to heart electric conduction. Rhythm level patterns capture rhythm features across beats and reflect cardiac arrhythmia conditions (abnormal heart rhythms). Frequency level is about frequency variations and sheds light on the diagnosis of ventricular flutter and ventricular fibrillation. Learning these patterns to support diagnoses has been an important research area in ECG analysis [Roopa and Harish, 2017; Lake and Moorman, 2010; Linker, 2016; Tateno and Glass, 2001].

In real clinical settings, in addition to the demand of an accurate classification, the interpretability of the results is equally important [Tsai et al., 2003]. Cardiologists need to provide both diagnosis and detailed explanations to support diagnosis [EC57, 2012]. Also, many heart diseases do not pose abnormal ECG diagram constantly [Benjamin et al., 2018; Yanowitz, 2012], especially during the early stage of the diseases. Therefore, interpretability of the results, particularly highlighting diagnosis related parts of the data, is crucial for early diagnosis and better clinical decisions.

Traditional machine learning methods either learn time domain patterns including beat level [Ladavich and Ghoaraani, 2015; Prirerfellner et al., 2014] and rhythm level [Huang et al., 2011], or extract frequency patterns using signal processing techniques such as discrete wavelet transform [García et al., 2016]. However, time domain approaches are easily affected by noise or signal distortion [Rodríguez et al., 2015]; while frequency domain methods cannot model rare events or some temporal dynamics that occur in time domain. Besides, they all require laborious feature engineering, and their performance also relies on the quality of the constructed features.

Recently, deep learning models showed initial success in modeling ECG data. Convolutional neural networks (CNN) were used to learn beat level patterns [Kiranyaz et al., 2016; Rajpurkar et al., 2017; Hannun et al., 2019]. Recurrent neural networks (RNN) are suitable for capturing rhythm fea-
tatures [Schwab et al., 2017; Hong et al., 2017; Zihlmann et al., 2017]. Moreover, attention mechanism is employed to extract interpretable rhythm features [Schwab et al., 2017]. Despite their progress, these models were either black-box or only highlighted one aspect of patterns (such as rhythm features as in [Schwab et al., 2017]), thus lack the comprehensive interpretability of the results for real clinical usage.

In this work, we propose Multilevel kNowledge-guide Attention model (MINA) to learn and integrate different levels of features from ECG which are aligned with clinical knowledge. For each level MINA extracts level-specific domain knowledge features and uses them to guide the attention, including beat morphology knowledge that guides attentive CNN and rhythm knowledge that guides attentional RNN. MINA also performs attention fusion across time- and frequency domains. We proposed new evaluation approaches by interfering ECG signals with noise and signal distortion. We evaluated interpretability and robustness of the model by tracking intermediate reactions across layers from multilevel attentions to the final predictions.

Experimental results show MINA can correctly identify critical beat location, significant rhythm variation, important frequency component and remain robust in prediction under signal distortion or noise contamination. Tested on the atrial fibrillation prediction, MINA achieved PR-AUC 0.9436 (outperforming the best baseline by 5.51%). Finally, MINA also showed strong result interpretability and more robust performance than baselines.

2 Related Work

Traditional methods include time domain methods such as beat level methods [Ladavich and Ghoraani, 2015; Pürerfellner et al., 2014] and rhythm level ones [Tateno and Glass, 2001; Huang et al., 2011; Oster and Clifford, 2015], both depending on segmentation by detecting QRS complex. However, time domain methods rely on the accuracy of QRS detection, thus are easily affected by noise or signal distortion. Frequency domain approaches, on the other hand, cannot model rare events and other time-domain patterns and thus lack interpretability. Moreover, both types of features are subjective.

Recently, deep neural networks (DNNs) have been used in ECG diagnosis [Kiranyaz et al., 2016; Rajpurkar et al., 2017; Hannun et al., 2019; Zihlmann et al., 2017; Hong et al., 2017; Schwab et al., 2017]. Many of them have demonstrated state-of-the-art performance due to their ability in extracting effective features [Rajpurkar et al., 2017; Hong et al., 2017]. Some of them build an end-to-end classifier [Kiranyaz et al., 2016; Rajpurkar et al., 2017; Zihlmann et al., 2017], others build a mixture model which combines traditional feature engineering methods and deep models [Hong et al., 2017; Schwab et al., 2017; Hong et al., 2019]. However, existing deep models are insufficient in three aspects. First, they neglect the characteristics of ECG signals when design model architecture, namely, beat morphological, rhythm variations. Second, they only analyze ECG signals in time domain. Last, they are “black-box” and thus not interpretable. In real world medical applications, interpretability is critical for clinicians to accept machine recommendations and implement intervention.

Figure 2: MINA takes raw ECG signals as input and outputs probabilities of disease onset. MINA used knowledge-guided attention to learn informative beat-, rhythm-, and frequency level patterns, and then performs attentive signal fusion for improved prediction.

3 Method

In this section, we will introduce the model design of MINA. Section 3.1 provides an overview and introduces all notations. Section 3.2 describes the basic framework, including each layer of MINA. Section 3.3 proposes our new attention mechanism which is integrated in MINA. Section 3.4 describes how we evaluate interpretability and robustness. Figure 2 depicts the architecture of MINA.

3.1 Overview of MINA

Here we briefly describe the framework and introduce notations used in this paper. Assume we are given a single lead ECG signal \( x \in \mathbb{R}^n \) and use it to predict class probability. We firstly transform it into multi-channel signals with \( F \) channels across different frequency bands where \( i \)th signal is denoted as \( x^{(i)} \in \mathbb{R}^n \). We then split each \( x^{(i)} \) into \( M \) segments \( s^{(k)} \). Next we apply CNN and RNN consecutively on \( s^{(k)} \) to obtain beat level attention \( d^{(k)} \), \( 1 \leq k \leq M \) and rhythm level attention \( c^{(i)} \). This follows by a fully connected layer that transforms \( c^{(i)} \) into \( q^{(i)} \). We then take weighted average to integrate \( Q = [q^{(1)}, \ldots, q^{(F)}] \) across all channels to output frequency attention \( d \), which will be used in prediction. To improve model accuracy and interpretability, we propose a knowledge guided attention to learn attention vectors from beat-, rhythm-, and frequency levels, denoted as \( \alpha, \beta, \gamma \) respectively. More details will be described in Section 3.2. The notations are summarized in Table 1. Detailed configurations of MINA are shown in the Experiments section.

3.2 Description of MINA

Signal Transformation and Segmentation

In order to utilize the frequency-domain information, we employ an efficient strategy by decomposing original ECG signals into different frequency bands (where each band is regarded as a channel). Then we can concurrently model signals of each channel.

Specifically, we propose a new time-frequency transformation layer to transform a single lead ECG signal into multi-channel ones. Here we use Finite Impulse Response (FIR)
handpass filter [Oppenheim et al., 1996] to transform single lead ECG signal \( x \) into \( F \) multi-channel ECG signals \( X = [x^{(1)}, x^{(2)}, \ldots, x^{(F)}] \).

Then for each channel, we split \( x^{(i)} \in \mathbb{R}^n \) into a sequence of \( M \) equal length segments. Unlike previous deep models [Schwab et al., 2017; Kiranyaz et al., 2016] that perform segmentation using QRS complex detection, which is easily affected by signal quality, we simply use sliding window segmentation. By cutting each \( i \)th segment is indexed by \((i - 1) \times T \) and \( i \times T - 1 \), we receive \( M \) equal length segments \( s \in \mathbb{R}^T \) (without the loss of generality, we assume that \( n = M \times k \), otherwise we can cut off last part which is shorter than \( T \)). In general, segment length \( T \) needs to be shorter than the length of one heart beat, so that we can extract patterns in beat level. Detailed configurations can be found in Implementation Details section.

### Beat Level Attentive Convolutional Layer

For beat level patterns, we mainly consider the abnormal rhythm variation. To capture them from beat sequences, RNNs are a natural choice due to their abilities to learn on data with temporal dependencies. Again to improve interpretability and accuracies, we use knowledge guided attention with rhythm knowledge.

Specifically, we use a bidirectional Long Short-Term Memory network [Schuster and Paliwal, 1997] (Bi-LSTM) to get rhythm level annotations of segments. The bidirectional LSTM is denoted here as \( h^{(k)} = \text{BiLSTM}(o^{(1)}, \ldots, o^{(k)}) \). We concatenate the forward and backward outputs of Bi-LSTM and receive the rhythm level feature \( H \in \mathbb{R}^{J \times M} \), \( H = [h^{(1)}, \ldots, h^{(M)}] \). Here we use knowledge-guided attention to output the rhythm level attention \( c = \sum_{k=1}^{M} \beta_k h^{(k)} \), where \( \beta_k \) represents the weight of \( k \)th rhythm level hidden state \( h^{(k)} \).

### Fusion and Prediction

At the beginning we decompose ECG signals into multiples channels (i.e., frequency bands) and learn rhythm level features \( \{c^{(i)}\} \) from each channel. Now we will perform attention fusion across all channels to have a more comprehensive view about the signal.

We first perform fully connected transformation: \( \mathbf{Q} = W_c^T \mathbf{C} + b_c \), where \( \mathbf{C} \in \mathbb{R}^{J\times F} \), \( \mathbf{C} = [c^{(1)}, \ldots, c^{(F)}] \), \( W_c \in \mathbb{R}^{J\times H} \), \( b_c \in \mathbb{R}^H \) and \( \mathbf{Q} \in \mathbb{R}^{H\times F} \). Then, since the importance of these channels may not be homogeneous, we will take weighted average of \( q^{(i)} \) to calculate frequency level attention \( d = \sum_{i=1}^{F} \gamma_i q^{(i)} \) where \( \gamma_i \) is the weight of \( q^{(i)} \), \( Q = [q^{(1)}, \ldots, q^{(F)}] \). We use frequency knowledge, signals with greater energy are more informative, to determine the weight \( \gamma \). Here we use power spectral density to measure energy.

Last, given integrated features \( d \) we make prediction using \( p = \text{softmax}(W_d \mathbf{b} + d) \), where \( W \in \mathbb{R}^{H \times C} \), \( \mathbf{b} \in \mathbb{R}^H \) and \( \mathbf{Q} \in \mathbb{R}^{H\times F} \). Bullis the class entropy loss \( CE(p) = - \sum_{z=1}^{C} \mathbb{I}(z = 1) \mathbf{w}_z \log p_z \), where \( C \) is the number of classes, \( z \) is the ground truth, \( \mathbf{w} \) is the weight vector with the same shape as \( z \), \( \mathbb{I} \) is the indication function. \( \mathbf{w} \) is adjusted to handle with class imbalance problem which is common in medical area.

### 3.3 Knowledge Guided Attention of MINA

We now describe how to compute multilevel attention weights \( \alpha, \beta, \gamma \). The attention mechanism can be regarded as a two-layer neural network: the 1st fully connected layer calculates the scores for computing weights; the 2nd fully connected layer computes the weights with via softmax activation.

In the first layer, the scores are computed based on the following features. (1) Multilevel outputs \( L \in \mathbb{R}^{K \times N} \), \( H \in \mathbb{R}^{J \times M} \). (2) Domain knowledge features including beat level \( K_\alpha \in \mathbb{R}^{E \times N} \), rhythm

<table>
<thead>
<tr>
<th>Table 1: Notations for MINA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notation</td>
</tr>
<tr>
<td>( C, P )</td>
</tr>
<tr>
<td>( n, M, T )</td>
</tr>
<tr>
<td>( x \in \mathbb{R}^n )</td>
</tr>
<tr>
<td>( X \in \mathbb{R}^{R \times N} ), ( x^{(i)} \in \mathbb{R}^n )</td>
</tr>
<tr>
<td>( s \in \mathbb{R}^T )</td>
</tr>
<tr>
<td>( S \in \mathbb{R}^{M \times T} ), ( k \in \mathbb{R}^T )</td>
</tr>
<tr>
<td>( L_i \in \mathbb{R}^{K \times M} )</td>
</tr>
<tr>
<td>( L^{(i)} \in \mathbb{R}^M )</td>
</tr>
<tr>
<td>( a \in \mathbb{R}^R )</td>
</tr>
<tr>
<td>( O \in \mathbb{R}^{R \times M} ), ( o^{(i)} \in \mathbb{R}^N )</td>
</tr>
<tr>
<td>( H \in \mathbb{R}^{J \times M} ), ( h^{(k)} \in \mathbb{R}^J )</td>
</tr>
<tr>
<td>( c \in \mathbb{R}^{J} )</td>
</tr>
<tr>
<td>( C \in \mathbb{R}^{J \times F} ), ( c^{(i)} \in \mathbb{R}^J )</td>
</tr>
<tr>
<td>( W_c \in \mathbb{R}^{H \times M} ), ( b_c \in \mathbb{R}^H )</td>
</tr>
<tr>
<td>( Q \in \mathbb{R}^{H \times F} ), ( q^{(i)} \in \mathbb{R}^H )</td>
</tr>
<tr>
<td>( d \in \mathbb{R}^{F} )</td>
</tr>
<tr>
<td>( W \in \mathbb{R}^{R \times M} ), ( b \in \mathbb{R}^{F} )</td>
</tr>
<tr>
<td>( p, p_i )</td>
</tr>
<tr>
<td>( w, w_i )</td>
</tr>
<tr>
<td>( z, z_i )</td>
</tr>
<tr>
<td>( \alpha, \alpha_i \in \mathbb{R} \cup \mathbb{N} )</td>
</tr>
<tr>
<td>( \alpha_i )</td>
</tr>
<tr>
<td>( \beta, \beta_i \in \mathbb{R} \cup \mathbb{N} )</td>
</tr>
<tr>
<td>( \gamma \in \mathbb{R} )</td>
</tr>
<tr>
<td>( K_\alpha \in \mathbb{R}^{K \times M} )</td>
</tr>
</tbody>
</table>

| Rhythm Level Attentive Recurrent Layer |

For rhythm level patterns, we mainly consider the abnormal rhythm variation. To capture them from beat sequences, RNNs are a natural choice due to their abilities to learn on data with temporal dependencies. Again to improve interpretability and accuracies, we use knowledge guided attention with rhythm knowledge.

Specifically, we use a bidirectional Long Short-Term Memory network [Schuster and Paliwal, 1997] (Bi-LSTM) to get rhythm level annotations of segments. The bidirectional LSTM is denoted here as \( h^{(k)} = \text{BiLSTM}(o^{(1)}, \ldots, o^{(k)}) \). We concatenate the forward and backward outputs of Bi-LSTM and receive the rhythm level feature \( H \in \mathbb{R}^{J \times M} \), \( H = [h^{(1)}, \ldots, h^{(M)}] \). Here we use knowledge-guided attention to output the rhythm level attention \( c = \sum_{k=1}^{M} \beta_k h^{(k)} \), where \( \beta_k \) represents the weight of \( k \)th rhythm level hidden state \( h^{(k)} \).
level $K_\beta \in \mathbb{R}^{E_2 \times M}$, and frequency level $K_\gamma \in \mathbb{R}^{E_\gamma \times F}$. Concretely, three levels of domain knowledge features can be represented as below.

- **Beat Level $K_\alpha$**: For beat level knowledge we mainly consider the abnormal wave shapes or sharply changed points such as QRS complex [Kashani and Barold, 2005]. To represent it we compute first-order difference $\Delta$ and a convolutional operation $\text{Conv}_\alpha$ on each segment $s$ to extract the beat level knowledge feature $K_\alpha = \text{Conv}_\alpha(\Delta(s))$, $\Delta(s)_i = s_i - s_{i-1}$ and $\Delta(s)_0 = s_0$. $s_i$ is the $i$th value in $s$. Detailed configurations of $\text{Conv}_\alpha$ are introduced in Implementation Details section.

- **Rhythm Level $K_\beta$**: Attention weights $\beta$ focus on rhythm level variation, such as severe fluctuation in ventricular fibrillation disease [Yanowitz, 2012]. To characterize it we compute standard deviation on each segment in $S$ to extract the rhythm level knowledge feature vector $K_\beta = \mathcal{D}(S)$, where $\mathcal{D}$ calculate standard deviation of each $s$ in $S$. $\mathcal{D}(s) = \frac{1}{n} \sum (s_i - \bar{s})^2$ to generate prediction $\text{MINA}$.

- **Frequency Level $K_\gamma$**: On frequency level, signals with greater energy contain more information and thus need more attention [Yanowitz, 2012]. So we use power spectral density (PSD), a popular measure of energy, to extract the frequency level knowledge feature vector $K_\gamma = \mathcal{F}(X)$, where $\mathcal{F}$ calculate PSD [Oppenheim et al., 1996] using a periodogram of each $x^{(i)}$ in $X$.

Then, we concatenate model outputs and knowledge features to compute scores and attention weights.

$$
\alpha = \text{softmax}(V_\alpha^T(W_\alpha^T [L_{K_\alpha} \odot b_\alpha]))
$$

$$
\beta = \text{softmax}(V_\beta^T(W_\beta^T [H_{K_\beta} \odot b_\beta]))
$$

$$
\gamma = \text{softmax}(V_\gamma^T(W_\gamma^T [Q_{K_\gamma} \odot b_\gamma]))
$$

where, $W_\alpha \in \mathbb{R}^{(K+E_\alpha) \times D_\alpha}$, $W_\beta \in \mathbb{R}^{(J+E_\beta) \times D_\beta}$, $W_\gamma \in \mathbb{R}^{(H+E_\gamma) \times D_\gamma}$, $b_\alpha \in \mathbb{R}^{D_\alpha}$, $b_\beta \in \mathbb{R}^{D_\beta}$, $b_\gamma \in \mathbb{R}^{D_\gamma}$ represent weights and biases in the first layer. $V_\alpha \in \mathbb{R}^{D_\alpha \times 1}$, $V_\beta \in \mathbb{R}^{D_\beta \times 1}$, $V_\gamma \in \mathbb{R}^{D_\gamma \times 1}$ represent weights in the second layer. $\odot$ is addition with broadcasting.

### 3.4 Method for Evaluating Interpretability and Robustness

To evaluate the interpretability and robustness of MINA, we perturb the signals and observe attention weights and prediction results. The evaluation method is illustrated in Figure 3.

Concretely, we add signal distortion (low frequency interferer) or noise (high frequency interferer) to the original ECG signal $x$ and get $x'$, here we choose baseline signal distortion and white noise. For the perturbed signals $x'$, we applied MINA to generate prediction $p'$ and output multilevel attention weights $\alpha', \beta', \gamma'$. We compare them with the original results $p$ and $\alpha$, $\beta$, $\gamma$ from unperturbed data.

To evaluate the interpretability of MINA, we visually check whether attention weights are in line with medical evidences. For beat level attention weights of $M$ segments

![Figure 3: Analysis of multi-level attention change (Orange) and final prediction change (Blue).](image)

### Table 2: Data profile of PhysioNet Challenge 2017 dataset

<table>
<thead>
<tr>
<th>Type</th>
<th># recording</th>
<th># of points</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF</td>
<td>738</td>
<td>9631</td>
</tr>
<tr>
<td>non-AF</td>
<td>7790</td>
<td>9760</td>
</tr>
</tbody>
</table>

We evaluate the robustness of MINA based on the two tasks: (1) we visually compare whether the new attention weights after perturbation are still in line with medical evidences, using the same way above, (2) we gradually change the interfered amplitude and evaluate the overall performance changes. The more robust model will be less impacted. Moreover, these results can also be used to evaluate interpretability, since interpretable model can highlight meaningful information, while also suppress unrelated parts.

### 4 Experiments

In this section, we first describe the dataset used for the experiments, followed by the description of the baseline models. Then we discuss the model performance.

#### 4.1 Source of Data

We conducted all experiments using real world ECG data from PhysioNet Challenge 2017 databases [Clifford et al., 2017]. The dataset contains 8,528 de-identified ECG recordings lasting from 9s to just over 60s and sampled at 300Hz by the AliveCor device, 738 from AF patients and 7790 from controls as predefined by the challenge. We first divided the data into a training set (75%), a validation set (10%) and a test set (15%) to train and evaluate in all tasks. Then, we preprocess them to get equal length data, where $n = 3000$. The summary statistics of the data is described in Table 2. In this study, the objective is to discriminate records of AF patients from those of controls.

5891
4.2 Baseline Models
We will compare MINA with the following models:

- **Expert**: A combination of extracted features used in AF diagnosis including: rhythm features like sample entropy on QRS interval [Lake and Moorman, 2010]; cumulative distribution functions [Tateno and Glass, 2001]; thresholding on the median absolute deviation (MAD) of RR intervals [Linker, 2016]; heart rate variability in Poincare plot [Park et al., 2009]; morphological features like location, area, duration, interval, amplitude and slope of related P wave, QRS complex, ST segment and T wave; frequency features like frequency band power. We used QRS segmentation method in [Pan and Tompkins, 1985] and trained an LR classifier using these features. Then, we build both logistic regression (ExpertLR) and random forest (ExpertRF) on above extracted features.

- **CNN**: Convolutional layers are performed on ECG segments with shared weights. We use global average pooling to combine features, and fully connect (FC) layer and softmax layer for prediction. The model architecture is modified based on [Kiranyaz et al., 2016] to handle ECG segments. The hyper-parameters in CNN, FC and softmax are the same as MINA to match the model complexity.

- **CRNN**: We used shared weights convolutional layers on ECG segments, and replaced the global average pooling with bi-directional LSTM. Then FC and softmax are applied to the top hidden layer. The architecture is modified based on [Zihlmann et al., 2017], but only keep one convolutional layer. Other hyper-parameters in CNN, RNN, FC and softmax are the same as MINA.

- **ACRNN**: Based on CRNN, with additional beat level attentions and rhythm level attentions. Other hyper-parameters are the same as MINA.

4.3 Implementation Details
In convolutional layers of CNN, CRNN, ACRNN and MINA, we use one layer for each model. The number of filters is set to 64, the filter size is set to 32 and strider is set to 2. Pooling is replaced by attention mechanism. Conv of $K_f$ has one filter with size set to 32, the strider is also 2. In recurrent layers of CRNN, ACRNN and MINA, we also use one single layer for each model, the number of hidden units in each LSTM is set to 32. The dropout rate in the fully connected prediction layer is set to 0.5. In sliding window segmentation, we use non-overlapping stride with $T = 50$. Deep models are trained with the mini-batch of 128 samples for 50 iterations, which was a sufficient number of iterations for achieving the best performance for the classification task. The final model was selected using early stopping criteria on validation set. We then tested each model for 5 times using different random seeds, and report their mean values with standard deviation. All models were implemented in PyTorch version 0.3.1, and trained with a system equipped with 64GB RAM, 12 Intel Core i7-6850K 3.60GHz CPUs and Nvidia GeForce GTX 1080. All models were optimized using Adam [Kingma and Ba, 2014], with the learning rate set to 0.003. Our code is publicly available at https://github.com/hsd1503/MINA.

Table 3: Performance Comparison on AF Prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>ROC-AUC</th>
<th>PR-AUC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExpertLR</td>
<td>0.9350 ± 0.0000</td>
<td>0.8730 ± 0.0000</td>
<td>0.8023 ± 0.0000</td>
</tr>
<tr>
<td>ExpertRF</td>
<td>0.9394 ± 0.0000</td>
<td>0.8816 ± 0.0000</td>
<td>0.8180 ± 0.0000</td>
</tr>
<tr>
<td>CNN</td>
<td>0.8711 ± 0.0036</td>
<td>0.8669 ± 0.0068</td>
<td>0.7914 ± 0.0090</td>
</tr>
<tr>
<td>CRNN</td>
<td>0.9040 ± 0.0115</td>
<td>0.8943 ± 0.0111</td>
<td>0.8262 ± 0.0215</td>
</tr>
<tr>
<td>ACRNN</td>
<td>0.9072 ± 0.0047</td>
<td>0.8935 ± 0.0087</td>
<td>0.8248 ± 0.0229</td>
</tr>
<tr>
<td>MINA</td>
<td>0.9488 ± 0.0081</td>
<td>0.9436 ± 0.0082</td>
<td>0.8342 ± 0.0352</td>
</tr>
</tbody>
</table>

Figure 4: An ECG signal of AF patient (left top), MINA learns beat level attention which points to the position of significant QRS complexes and abnormal P waves. Rhythm level attention shows the abnormal RR interval. The frequency channel with highest attention correspond to the frequency bands where QRS complex is dominant.

4.4 Performance Comparison
Performance was measured by the Area under the Receiver Operating Characteristic (ROC-AUC), Area under the Precision-Recall Curve (PR-AUC) and the F1 score. The PR-AUC is considered a better measure for imbalanced data like ours [Davis and Goadrich, 2006]. Table 3 shows MINA outperforms all baselines, and shows 5.51% higher PR-AUC than the second best models.

5 Interpretablity and Robustness Analysis
5.1 MINA Automatically Extracts Clinically Meaningful Patterns
When reading an ECG record (upper left in Figure 4), cardiologists will make AF diagnosis based on following clinical evidences: 1) the absence of P wave: a small upward wave before QRS complex; 2) the irregular RR interval such as the much wider one between the 4th and the 5th QRS complex. MINA learns these patterns automatically via beat-, rhythm- and frequency level attention weights. From Figure 4, the beat level attentions point to where QRS complex or absent P waves occur. The rhythm level attentions indicate the location of abnormal RR interval, which precisely matches the clinical evidence. Besides, from the frequency level attentions, we notice channel 10Hz-50Hz receives the highest attention weight so MINA pays more attention to it. In fact, QRS complex, the most significant clinical evidence in ECG diagnosis, is known to be dominant in 10Hz-50Hz [Tateno and Glass, 2001; Linker, 2016; Lake and Moorman, 2010].

5.2 MINA Remains Interpretable and Robust Against Baseline Signal Distortion
The baseline wander distortion is a low frequency noise with slow but large changes of the signal offset. It is a common
issue that drops ECG analysis performance. In this experiment, we mimic the real world setting by distorting data and observe whether MINA can still make robust and interpretable predictions.

For the experiment we interfered the signal in Figure 4 with baseline wander distortion. The interfered signal is plotted in Figure 5(a). From the original frequency attention in Figure 4, it is easy to see Channel 1 (<0.5Hz) has the lowest weights, while Channel 2 (0.5Hz-50Hz) weights much higher. Thus Channel 1 can be interpreted as baseline component while Channel 2 as clean signal component. MINA pays more attention to Channel 2 than Channel 1. After signal distortion, the importance of both channels remain the same, which is also reflected from their beat level and rhythm level attentions. Channel 1 shows no significant patterns, but the more informative Channel 2 have similar beat- and rhythm level patterns as unperturbed data, which indicates the interpretability of MINA will be less impacted by data distortion.

To evaluate model robustness, we compare the performance change along the increase of distortion amplitude on the entire test set. As shown in Figure 5(d), MINA still has much lower performance drop even after distortion by large amplitude. While all baselines start to have large performance drop even with little distortion. This is mainly thanks to frequency attention fusion. In training process, the model already identified Channel 1 a baseline signal. Thus baseline distortion will have less impact on important signals in clean signal channel. Since baseline signal distortion occurs in real clinical setting, MINA will provide more accurate prediction in these scenarios.

5.3 **MINA Remains Interpretable and Robust in the Presence of Noise**

The high frequency noise contamination is another common issues. For this experiment, we perturbed the signal in Figure 4 with white noise. The perturbed signal is in Figure 6(a). Similar to last experiment, from original frequency attentions we know Channel 3 (>50Hz) has lower weights. It is a channel known for high noise. While Channel 2 (0.5Hz-50Hz) weights much higher and is known as a clean signal channel.

After noise contamination, the noise impacts more to the noise Channel which is less important in the prediction of MINA, but the more informative Channel 2 have similar beat- and rhythm level patterns as unperturbed data, which indicates the interpretability of MINA will be less impacted by noise contamination. In Figure 6(d), we compare the PR-AUC change along the increase of noise amplitude on the entire test set. MINA is less impacted by noise than other methods, demonstrating more robust performance in the presence of noise thanks to frequency attention fusion.

6 Conclusion and Future Work

In this paper, we propose MINA, a deep multilevel knowledge-guided attention networks that interpretatively predict heart disease from ECG signals. MINA outperformed baselines in heart disease prediction task. Experimental results also showed robustness and strong interpretability against signal distortion and noise contamination. In future, we can extend to a broad range of disease where ECG signals can be treated as additional information in the diagnosis, on top of other health data such as electronic health records. Then we will need to investigate interpretable prediction based on multimodal data, which is a possibly rewarding avenue of future research.

Acknowledgements

This work was supported by the National Science Foundation, award IIS-1418511, CCF-1533768 and IIS-1838042, the National Institute of Health award 1R01MD011682-01 and R56HL138415. We also thanks valuable discussions with Li Jiang from BOE.

References

