Detection of Paroxysmal Atrial Fibrillation using Attention-based Bidirectional Recurrent Neural Networks

Supreeth P. Shashikumar
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA
supreeth@gatech.edu

Amit J. Shah
Department of Epidemiology
Rollins School of Public Health, Emory University
Atlanta, Georgia, USA
ajshah@emory.edu

Gari D. Clifford
Department of Biomedical Informatics
Emory University
Department of Biomedical Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA
gari@gatech.edu

Shamim Nemati
Department of Biomedical Informatics
Emory University
Atlanta, Georgia, USA
shamim.nemati@emory.edu

ABSTRACT

Detection of atrial fibrillation (AF), a type of cardiac arrhythmia, is difficult since many cases of AF are usually clinically silent and undiagnosed. In particular paroxysmal AF is a form of AF that occurs occasionally, and has a higher probability of being undetected. In this work, we present an attention based deep learning framework for detection of paroxysmal AF episodes from a sequence of windows. Time-frequency representation of 30 seconds recording windows, over a 10 minute data segment, are fed sequentially into a deep convolutional neural network for image-based feature extraction, which are then presented to a bidirectional recurrent neural network with an attention layer for AF detection. To demonstrate the effectiveness of the proposed framework for transient AF detection, we use a database of 24 hour Holter Electrocardiogram (ECG) recordings acquired from 2850 patients at the University of Virginia heart station. The algorithm achieves an AUC of 0.94 on the testing set, which exceeds the performance of baseline models. We also demonstrate the cross-domain generalizablity of the approach by adapting the learned model parameters from one recording modality (ECG) to another (photoplethysmogram) with improved AF detection performance. The proposed high accuracy, low false alarm algorithm for detecting paroxysmal AF has potential applications in long-term monitoring using wearable sensors.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; *Transfer learning*; • **Applied computing** → **Health informatics**;

KEYWORDS

atrial fibrillation, convolutional neural network, recurrent neural network, deep learning, transfer learning

1 INTRODUCTION

Atrial fibrillation (AF), an arrhythmia that arises from irregular atrial contraction, has a prevalence of 2% in the adult population [4, 7]. AF can often lead to dangerous complications such as ischemic stroke and heart failure [53, 58], but these complications can be treated or can be avoided by anticoagulation and control of heart rate or rhythm [19]. This highlights the need for developing accurate AF detection methods, which would allow for early treatment and diagnosis of AF.

The difficulty in the diagnosis of AF stems from its asymptomatic and paroxysmal nature. Moreover, when it is eventually diagnosed, clinicians must make multifactorial decisions regarding the optimal course of therapeutic interventions, which, among other factors, may require monitoring the burden of AF (defined as the amount of time spent in AF). Clinically, AF is diagnosed by the absence of P waves in the Electrocardiogram (ECG) with a rapid, irregular rhythm. However, clinical adjudication based on ECG alone is difficult in the presence of recording noise and distortion, which is typical in ambulatory monitoring. Over the last four decades, numerous methods have been proposed in the literature for AF detection based on beat-to-beat timing and morphological and frequency domain properties of the ECG. These range from hidden Markov modeling of the beat-to-beat timing [42], to QRST subtraction and frequency analysis of the residual waveform [54], as well as many other threshold-based, heuristic and machine learning approaches [13, 28, 34, 44, 56]. More recently, we have developed deep learning approaches applied to cardiac signals recorded in real world noisy

environments [51] in an attempt to discover novel patterns indicative of AF. Similar deep learning approaches were subsequently reported (along with many other standard approaches) in the recent PhysioNet Challenge, an international competition focused on the classification of over 10,000 publicly available ECGs [11]. This competition, organized over 9 months, led to 45 publications and 75 independent pieces of software, most of which were open sourced [12]. Notably, this competition provided the first slew of deep learning based ECG rhythm detectors, along with some simultaneous pre-prints. However, all these methods were developed within a standard machine learning paradigm, ignoring the long tracts of non-arrhythmic data that often exist between sporadic episodes of arrhythmia. This leads to wholly optimistic estimates of the performances of these algorithms when applied to long term recordings of several hours or days, and significantly underestimates the false alarm rate. The work presented here addresses the issue of false alarms through the use of an attention model applied to a deep learning framework.

Our goal during the development of the algorithm was three fold:

- To assess whether feature vectors extracted from the deep learning pipeline in combination with time series covariates, extracted using traditional methods and hand-crafted features, can achieve improved performance in comparison to other ECG based approaches.
- To test whether employing a soft attention mechanism would improve the performance of the algorithm in the presence of paroxysmal AF, by learning to put more weight on the windows with higher potential prevalence of AF, as opposed to noise and other types of rhythm irregularity.
- To empirically evaluate the potential of the model for transfer learning across different source domains.

To achieve these objectives, we designed a deep learning model based on a Bidirectional Recurrent Neural Network (BRNN) for detection of AF from a three lead ECG. The proposed model extracts spectral features from ECG signal using wavelet analysis, and a convolutional neural network (CNN) is used to process these spectral features, followed by a BRNN to analyze these features in temporal order to detect AF from 30 second windows spanned over 10 minutes of recording. In addition, we also implement a soft attention mechanism on top of the BRNN to enable the algorithm to prioritize ECG segments that are predictive of AF (given the paroxysmal nature of AF). The entire pipeline of the proposed model is illustrated in Figure 1.

2 DATA

2.1 Holter ECG dataset

The dataset consists of 24 hour Holter ECG recordings collected from 2850 patients at the University of Virginia (UVA) Heart Station from 12/2004 to 10/2010. The age of the patients varied from a few months to 100 years, with an average (± standard deviation) value of 47 ± 25 years. AF labels were obtained from an automatic classifier (Philips Holter Software), and each record was examined by a clinical adjudicator to confirm presence or absence of AF. The dataset was found to consist of various heart rhythms including AF, Normal Sinus Rhythm (NSR), and Sinus Rhythm with ectopy.

After excluding segments with a low signal quality index (*SQI*) [33], which accounted for less than 2% of all segments, the remaining recordings were divided into 364,012 10 minute segments. Each such segment was classified as *AF* if the burden of AF was greater than 5% (i.e. for more than 30 seconds), and *Other Rhythm* otherwise, however for completeness we also provide results for different burdens of AF. Although, labels obtained from automatic systems can be noisy at the level of beat-by-beat annotation, our goal was to quantify presence of AF over consecutive 10 minute segments, which is less prone to noise due to averaging of labels and removal of low *SQI* segments. Moreover, we utilize ECG-based AF detection as a starting point for cross-domain transfer learning to demonstrate the utility of using external data and data types. In particular, we use a database of pulsatile photoplethysmogram (PPG) recordings of AF from a smart watch, as described next, for this purpose.

2.2 Smart Watch PPG dataset

All subjects were adult patients (18-89 years old) who were hospitalized and were undergoing telemetry monitoring at the Emory University Hospital (EUH), Emory University Hospital Midtown (EUHM), and Grady Memorial Hospital (GMH). The study was approved by the institutional review board (IRB # 00084629) of the Emory University Hospital and Grady Memorial Hospital in Atlanta, GA. Patients were recruited at random; the rhythms were reviewed by an ECG technician, physician study coordinator, and cardiologist. The study took place from October 2015 through March 2016. Using a modern research watch-based wearable device (the Samsung Simband [48]), we recorded ambulatory PPG data from 97 subjects, 44 with *AF* and 53 with *other rhythms* for approximately 5-10 minutes. Simultaneous 128 Hz data were captured form a single channel ECG, multi-wavelength (8 channel) PPG and a tri-axial accelerometer.

3 METHODS

3.1 Model Architecture

An overview of the proposed AF detector is shown in Figure 1. The algorithm included the following stages:

- *Data preparation*: Splitting longitudinal recordings into non-overlapping 10 minute multi-channel ECG segments sampled at 200Hz.
- *Pre-processing*: Filtering to reduce and remove noise from the ECG signals, and selecting the channel with the highest *SQI*. Further split the 10 minute segment to consecutive 30 second windows.
- *Time series covariates*: Computing covariates based on beat-to-beat interval variations from the 10 minute segments.
- *Frequency analysis*: extracting time-frequency information from the processed ECG signals by applying a wavelet transform.
- *Feature extraction using CNN*: Feeding the wavelet spectrograms into a CNN for feature extraction.
- *Bidirectional Recurrent Neural Network*: Feeding features extracted from the CNN into a BRNN to capture temporal patterns.
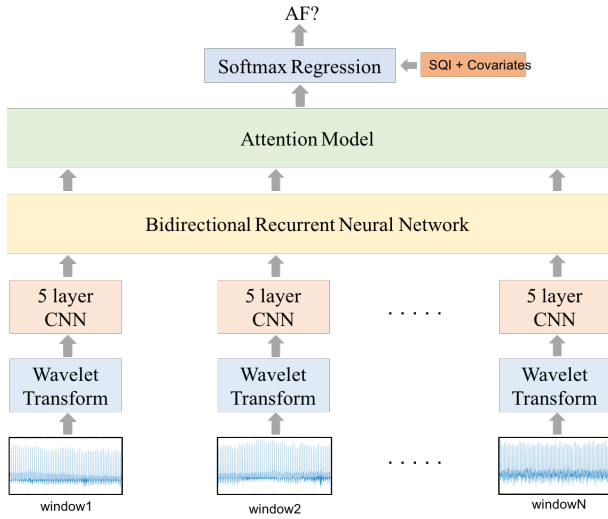
Figure 1: Schematic diagram of the AF detection algorithm. Each 10 minute ECG segment was split into 30 second windows, and a CNN was used to extract time-frequency features from the spectrograms obtained from each of the 30 second windows. The output feature vectors from the CNN were then fed sequentially into a BRNN, a soft attention mechanism prioritized the segments of ECG to attend, and then a softmax regression layer detected AF in the ECG segment by combining the deep learning features with other time series covariates.

- *Attention layer*: Passing the data from the previous step through a soft attention mechanism that assigns more weights to the windows with higher prevalence of AF.
- *Classification layer*: Feeding the weighted feature vector from the attention layer and the time series covariates into a fully connected layer. The classification layer then computes the likelihood of presence of AF in the input ECG segment.

3.2 Data preparation and pre-processing

Each ECG channel was lowpass filtered to remove frequencies above 40Hz, using a Butterworth filter of order five. Next, QRS (beat) detection was performed on each channel individually using two QRS detectors - *jQRS* [5, 25, 26] and *waveletQRS* [38] and the resultant beats were used to compute the signal quality index $bSQI$ [33]. To remove the noisy segments, we excluded all segments with $bSQI < 0.90$, which eliminated roughly 2% of the data segments. In the remaining segments, the channel with the highest $bSQI$ was then chosen for further analysis.

3.3 Time series covariates

We incorporated several metrics based on beat-to-beat interval variations in our AF detection algorithm. We calculated the standard deviation of the beat-to-beat interval time series (STD), as well as a robust version of the standard deviation (STD_r), after discarding the intervals outside the 0.02-0.98 percentile range (we assumed that the extreme intervals were due to erroneous pulse onset detection). We

also calculated the sample entropy [46] of the beat-to-beat interval time series with the embedding dimensions of $m = 1$, and 2 (denoted $SampEn_1$ and $SampEn_2$, respectively).
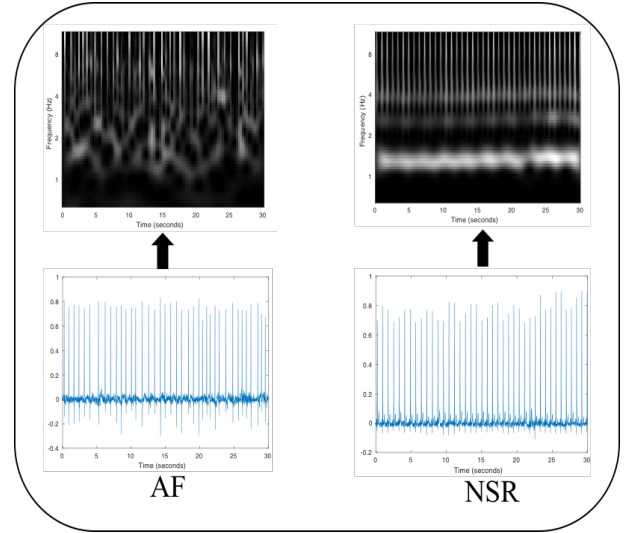
3.4 Wavelet decomposition

Figure 2: Wavelet power spectrum for 30 second ECG sample. The left panel shows an example of AF segment and the right panel represents a Normal Sinus Rhythm segment

Wavelet transform is an ideal tool for extracting time-frequency information from signals of a non-stationary nature, due to its optimal time-frequency resolution trade-offs [16]. AF is characterized by irregular heart beat rhythms [19], which can be captured through the spectral analysis of the ECG signal (See Figure 2). Shashikumar et al. [51] proposed an AF detection approach that was based on employing wavelet decomposition to extract features from the frequency domain. In our algorithm, the 10 minute segment was split into non-overlapping 30 second windows, and the wavelet transform [51] was applied to each of the windows, resulting in a wavelet spectrogram matrix of size 20 by 300. The hyper-parameters used during the wavelet decomposition were as follows: maximum scale of 20; 1/15 octaves per scale; minimum scale of 8.3; and a down-sampling factor of 20 along the time axis. The mathematics behind the wavelet analysis is well documented in Grinsted et al. [22]. In practice, recording noise and artifacts might corrupt the spectral content present in the signal which makes it difficult to distinguish AF rhythm from the other rhythms. To reduce the effect of the recording noise and other artifacts, we thresholded the spectrograms using surrogate data analysis as described in Shashikumar et al. [51].

3.5 Convolutional neural network for feature extraction

CNNs have been shown to be very effective in image-based classification tasks [27, 30] due to their translation-invariance properties, ability to capture local features, shared-weights architecture that
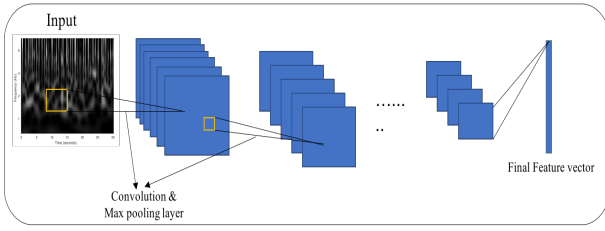
Figure 3: Schematic diagram of the deep CNN layer. The spectrogram obtained from the 30 second ECG window is fed into a 5 layer CNN and the output feature vector is then sequentially fed into BRNN

enhances generalizability [31]. Time-frequency representation of pulsatile cardiac timeseries of NSR and AF tend to exhibit distinct patterns (see Figure 2), although the exact location of the dominant frequency depends on the underlying heart rate. The translation-invariance property of the CNN can handle such spectral shifts in the patterns of interest. Our CNN architecture, comprised of two successive convolutional layers (each layer having a kernel size of 3x21) and a max pooling layer, followed by two more convolutional layers (each having a kernel size of 4x21), and one fully connected layer. Each convolutional layer in the architecture was followed by a layer of activation with Rectified Linear Unit (ReLU) nonlinearity. All the pooling layers had pooling region of size 2x2 with a stride of 2 along both the directions. The number of filters used for each of the convolutional layers was 10, with the fully connected layer having a total of 50 filters. The output of the CNN was a 50 dimensional feature vector that summarized the entire wavelet power spectrum.
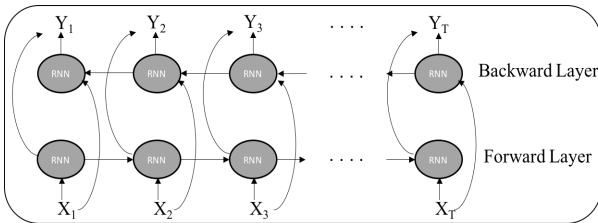
3.6 Bidirectional recurrent neural network

Figure 4: Schematic diagram of the bidrectional recurrent neural network

Recurrent neural networks (RNNs) are capable of capturing temporal patterns in the data [21, 32]. Standard RNNs are unidirectional, in the sense that input data is processed in temporal order. A short-coming of this approach is that the RNNs are restricted to the use of previous context. BRNNs [49] provide a solution by processing the data in both forward and backward directions. Figure 4 shows a BRNN architecture unfolded in time for T time steps. The BRNN consists of a forward layer and a backward layer. The forward layer h_t^f is computed by processing the input data from $t = 1, ..., T$, and the backward layer h_t^b is computed by processing the input data

from $t = T, ..., 1$ with output from both layers combined as follows:

$$h_t^f = tanh(W_{xh}^f x_t + W_{hh}^f h_{t-1} + b_h^f) \tag{1}$$

$$h_t^b = tanh(W_{xh}^b x_t + W_{hh}^b h_{t+1} + b_h^b) \tag{2}$$

$$y_t = W_{hy}^f h_t^f + W_{hy}^b h_t^b + b_y) \tag{3}$$

In our model, the size of the hidden state in the forward and the backward layer was 50, with the output at each time step being a 100 dimensional vector. The output vectors y_t obtained by processing the sequence of input data from $t = 1, ..., T$ was then fed into an attention layer.

3.7 Attention layer

Paroxysmal atrial fibrillation (PAF) occurs as intermittent periods of AF scattered with episodes of normal sinus rhythm [35]. Thus, a soft attention mechanism [3, 61, 62] was employed on top of the BRNN so that more emphasis (or attention) could be put on windows with higher prevalence of AF. The attention mechanism can be formulated as follows - Let the outputs from the BRNN $[y_1, y_2, ..., y_T]$ be combined into a matrix Y which is of size N x T, where T is the length of the input sequence and N is the size of output vector y_t. The weighted output vector h_{att} of the attention layer is as follows -

$$\alpha = softmax(w_{att}^T Y) \tag{4}$$

$$h_{att} = Y\alpha^T \tag{5}$$

That is, α is a weight vector computed from matrix Y, and the output h_{att} calculated as a weighted sum of all the output vectors from the BRNN (h_{att} was a 100 dimensional vector).

3.8 Classification layer

The attention layer was then followed by a fully connected layer, the output of which was fed into a softmax regression layer. In addition to the features obtained from the BRNN, time series covariates were fed as inputs to the softmax regression layer. The output from the softmax layer corresponded to the likelihood of presence of AF in the input ECG segment.

4 RESULTS

4.1 Experimental setup

Of the 2850 patients, 80% of them were used for developing the model (training set) and the remaining 20% of the patients were used as the hold out test set. The training set consisted of a total of 2290 patients out of which 217 patients contained at least one episode of AF, and the testing set consisted of a total of 575 patients out of which 55 patients contained at least one episode of AF (we refer to an episode of AF as a 10 minute segment with >5% of AF burden). Splitting the recordings of each patient to 10 minute segments resulted in a total of 291,240 segments (21,542 AF segments) in the training set and 72,772 segments (5,419 AF segments) in the testing set.

Details regarding the training and development of the model are as follows: The batch size was fixed at 90 patients (70% *Other Rhythm* patients, 30% *AF* patients), with data randomly sampled (with replacement) in every epoch. The model was trained end-to-end for a total of 200 epochs. RMSProp [57] was used as the

Table 1: Summary of classifier performance for different % of AF burden. The Area Under the Curve (*AUC*), Area under the precision recall curve (*AUC_pr*), Specificity (*SPC*) and Accuracy (*ACC*) are reported for both training set and testing set

	Testing set				Training set			
% AF burden	AUC	AUC_{pr}	SPC	ACC	AUC	AUC_{pr}	SPC	ACC
>5	0.94	0.84	0.95	0.94	0.96	0.93	0.96	0.96
>25	0.93	0.82	0.92	0.92	0.96	0.93	0.95	0.95
>50	0.95	0.82	0.93	0.95	0.97	0.91	0.95	0.94
>75	0.97	0.84	0.96	0.96	0.98	0.96	0.98	0.96

optimizer and L2 regularizer with regularization parameter λ = 0.01 was used. The learning rate was set at 0.001. All the hyper-parameters of the model: number of filters in each of the CNN layer, size of the hidden layer in the BRNN, learning rate, regularization parameter λ were optimized using Bayesian Optimization technique [20]. A subset of the training set was used for hyper-parameter optimization. Area under the receiver operating characteristic (*AUC*) curve, Area under the precision recall (*AUC_pr*) curve, accuracy, specificity were calculated for both the training and the test sets. The sensitivity level was fixed at 0.85. All pre-processing of the data, time series covariates computation and wavelet decomposition were performed in Matlab [39]. The rest of the pipeline was implemented in python with CNN and BRNN implemented using Tensorflow [1].

4.2 AF detection performance

First, a comparative analysis was done to evaluate the performance of the model based on % of AF burden. The % of AF burden determines the presence of AF in the 10 minute segments. The performance of the model with various thresholds on the AF burden are shown in Table 1. It can be seen that labeling segments as AF with AF burden >5% resulted in an AUC of 0.94 on the testing set. It should be noted that with increasing threshold on the AF burden, a majority of the segments containing PAF will be filtered out. The model trained on segments with >75% AF burden performed the best in comparison to the others, likely due to the absence of PAF segments (which are more difficult to detect). Since our goal was to detect both PAF and AF segments, we labeled segments with AF burden >5% as AF in all the following experiments.

Conventionally, when RNNs are used in classification tasks the hidden state vector from the last time step alone is used for classification (no pooling). An alternative would be to compute the mean of all the hidden state vectors across all time steps, which is then used for classification (Mean pooling). A third alternative would be to use attention pooling which is described in Section 3.7. The performance of the model for the three kinds of pooling is shown in Table 2 . It can be observed that using an attention layer (AUC of 0.94 on the testing set) provided improvement in performance over the other two pooling methods. An interesting observation to note is that mean pooling (AUC of 0.91 on the testing set) had a lower performance compared to performing no pooling (AUC of 0.92 on the testing set) at all.

Table 2: Summary of classifier performance for different types of pooling. The Area Under the Curve (*AUC*), Area under the precision recall curve (*AUC_pr*), Specificity (*SPC*) and Accuracy (*ACC*) are reported for both training set and testing set

	Testing set				Training set			
Model	AUC	AUC_{pr}	SPC	ACC	AUC	AUC_{pr}	SPC	ACC
Attention pooling	**0.94**	0.84	0.95	0.94	**0.96**	0.93	0.96	0.96
Mean pooling	0.91	0.82	0.93	0.92	0.95	0.89	0.92	0.92
No pooling	0.92	0.83	0.93	0.93	0.97	0.92	0.94	0.94

Finally, a comparative study was performed to assess the importance of Spectrogram based features and time series covariates towards detection of AF. Three separate models were trained and tested - a) Using only the features extracted from the spectrogram and feeding them into the BRNN, b) Using only the time series covariates and feeding to the softmax regression layer (baseline model) and c) Combining both the spectrogram based features and time series covariates. In addition to the above models, we also compared the performance of our model with the AF detection algorithm proposed by Carrara et al. [9]. As proposed in Carrara et al. [9], the Coefficient of Sample Entropy, the average Detrended Fluctuation Analysis value and the Local Dynamics score were computed for every 10 minute segments in the dataset, and were fed to a softmax regression layer for AF detection. The results for the above outlined experiments have been tabulated in Table 3. The baseline model had the lowest performance (AUC of 0.87 on the testing set) as compared to the other two models. It can be observed that by combining the spectrogram based features and the time series covariates, the model achieved an AUC of 0.94 on the testing set.

Table 3: Summary of classifier performance for the three feature groups, or models. The Area Under the Curve (*AUC*), Area under the precision recall curve (*AUC_pr*), Specificity (*SPC*) and Accuracy (*ACC*) are reported for both training set and testing set

	Testing set				Training set			
Model	AUC	AUC_{pr}	SPC	ACC	AUC	AUC_{pr}	SPC	ACC
Spectrogram	0.92	0.80	0.92	0.92	0.94	0.92	0.93	0.93
Covariates (Baseline)	0.87	0.67	0.77	0.78	0.89	0.80	0.80	0.81
Combined	**0.94**	0.84	0.95	0.94	**0.96**	0.93	0.96	0.96
Carrara et al. [9]	0.91	0.80	0.91	0.91	0.93	0.90	0.94	0.93

4.3 Visualization of the extracted features

To understand the reason for the model to mis-classify the input ECG segments, we performed clustering on the features extracted from the deep learning pipeline and the time series covariates over 10 minute segments. We used t-SNE to perform the clustering [37]. The visualization of the clusters obtained on the testing data is shown in Figure 5a. Each point in the plot corresponds to a two-dimensional representation of features extracted from a 10 minute segment. The False Positive (FP) cases (i.e. *Other Rhythm* classified as *AF*) have been highlighted in Figure 5b. The False Negative (FN) cases (i.e. *AF* classified as *Other Rhythm*) have been highlighted in

(a) Visualization of data belonging to *AF* and *Other Rhythm* class

(b) The False Positive (FP) cases highlighted

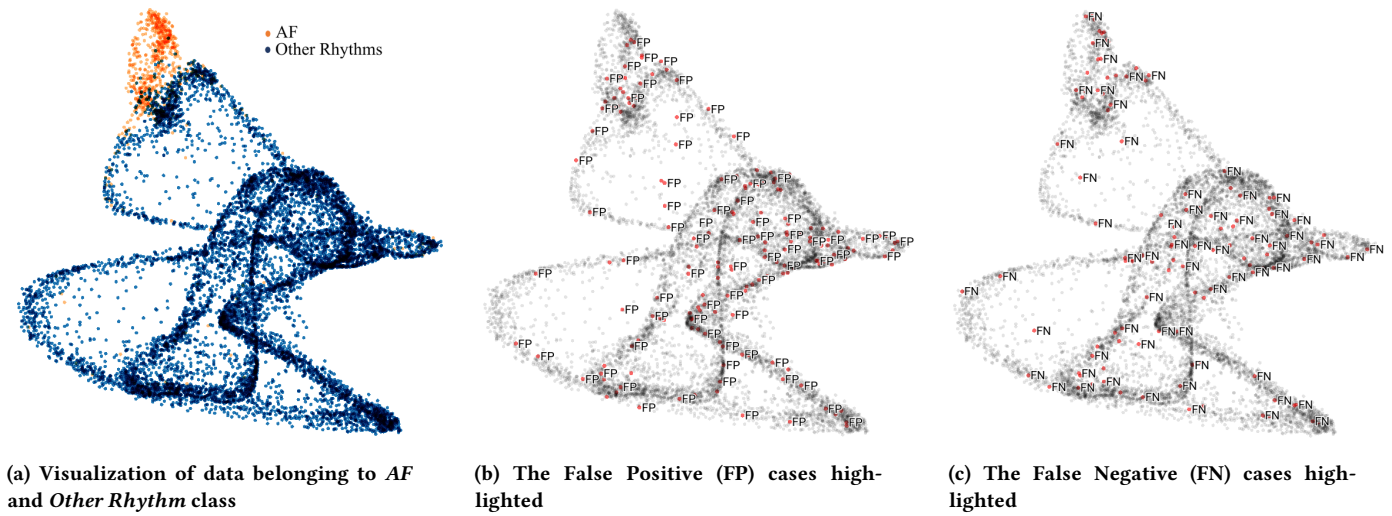**(c) The False Negative (FN) cases highlighted**

Figure 5: Visualization of the features extracted from the deep learning pipeline and the time series covariates by performing t-SNE [37] based clustering. Clustering was performed on the testing dataset

Figure 5c. Although, the two-dimensional embedding of the features only provides an approximate picture of the position of FP and FN cases, it sheds light on where the algorithm is most likely to disagree with the Holter monitor labels, which are known to be noise-prone [18]. Such points could make for potential candidates for active learning [50], wherein a meta-learning algorithm can iteratively refine the labels by querying an oracle (one or more clinicians) for ground truth labels, and re-train the model accordingly.

4.4 Knowledge transfer across different source domains

In recent times, the success of deep learning has been largely in part due to the availability of extensively labeled datasets [23, 41, 52, 59]. Deep learning models are known to perform better when large sets of data is available. However, in practical scenarios especially in health care, access to such data is limited or not practical due to the amount of effort required. Transfer learning with deep neural networks provides a powerful framework to deal with such situations [8, 23, 24, 36, 55]. In particular, we were interested in exploring how to transfer knowledge between domains, when the distribution of the input features and labels change, but the task (in our case, AF detection) remain the same. Given two datasets, one containing ECG recordings from 2850 patients and the other containing PPG recordings from only 97 patients, we aimed to transfer the knowledge learned from the ECG recordings (source domain) to PPG recordings (target domain) for detecting AF.

The target domain in our experiment was a dataset of PPG signals recorded using Samsung Simband watches (described in Section 2.2). We followed the same procedure as described in Section 3.1 for the wavelet decomposition and extraction of time series covariates from the PPG signal. We randomly split the patients as 80% for training and the remaining 20% for testing. We chose the same architecture that was used for the Holter ECG dataset, and we performed two experiments: a) We initialized the model with random weights and

trained only on the Smart watch PPG dataset b) We initialized the variables in the model with the values learned from training on the Holter ECG dataset (Pre-trained model), and further refined the values with the Smart watch PPG training dataset.

Table 4: Comparison of classifier performance depending on the type of initialization of weights in the model for the PPG dataset. The Area Under the Curve (AUC), Area under the precision recall curve (AUC_{pr}), Specificity (SPC) and Accuracy (ACC) are reported for both training set and testing set

	Testing set				Training set			
Type of initialization	AUC	AUC_{pr}	SPC	ACC	AUC	AUC_{pr}	SPC	ACC
Random	0.94	0.93	0.81	0.85	0.94	0.93	0.90	0.88
Pre-trained model	**0.97**	0.97	1.0	0.95	**0.94**	0.92	0.90	0.89

The improvement in performance obtained by using the pre-trained weights on Holter ECG dataset is shown in Table 4. We observe that when the model was **learned from scratch** (Model1), the model achieved an AUC of 0.94 on the testing set. Whereas when the model was **fine tuned from a pre-trained model** (Model2), the model achieved an AUC of 0.97 on the testing set. Comparing testing set predictions of both the models, the p-value of comparing AUC curves [14] was 0.0075, indicating a significant improvement. We achieved two important goals from these experiments - **First**, we have shown that the model that was initially trained on ECG data could be fine tuned to detect AF in PPG data. Even though the two datasets were based on completely different types of input signal, the deep learning based model was able to learn features that was generalizable across both the domains and which were predictive of AF. **Second**, we have shown that in scenarios where limited data is available, we can successfully employ a model that is trained on a much larger cohort, and fine-tune it and still obtain significant improvements in performance. In our experiment, by using a pre-trained model that was trained on a large cohort of 2850

patients, to train on the Smart watch PPG dataset which had only 97 patients, we have showed an improved performance compared to training the model from scratch.

5 DISCUSSION AND CONCLUSION

The major finding of this study is that combining spectral representation of cardiac pulsatile recordings with traditional indices of heart rhythm irregularity in a deep neural network framework results in better AF classification. This approach facilitates transferring of learned model parameters across recording modalities such as ECG and PPG, thus enabling accurate AF classification in settings with limited access to large patient cohorts for model training purposes. Moreover, our results indicate that the hierarchical architecture of a deep neural network with one or more image-based feature extraction layers, a sequential layer capable of passing temporal information, and an attention mechanism allows for accurate classification of paroxysmal AF.

A key property of deep neural networks that contribute to their ability to generalize well is learning of good representations of the data [6]. Here the notion of *good* is with respect to robustness to noise and invariance to factors that may contribute to irregularity of heart rhythms but may not be relevant to AF prediction. In the proposed method, this invariance is achieved through a combination of time-frequency representation, translation-invariance of CNN (due to the convolution operation and max pooling), and the attention mechanism that is capable of detecting important features independent of their actual position within a sequence of windows. For instance, having a higher heart rate shifts the position of the dominant frequency in the time-frequency plane representation of an ECG waveform, but the translation-invariance of the CNN makes the overall architecture robust to differences in baseline heart rates. Similarly, the time-frequency representation of a PPG waveforms exhibits dominant frequencies at the frequency of heart rate and the AF signature appears as an irregular distribution of power around this dominant frequency, and the CNN can effectively learn these irregular spectral patterns without getting confused by differences in heart rate across subjects.

Various methods have been proposed in the literature for AF detection based on studying the dynamics of beat-to-beat heart rate time series extracted from ECG [13, 15, 28, 34, 42, 44, 56] and PPG [43]. More recently, deep learning based methods have been shown to perform well in the context of AF detection. These deep learning based methods have harnessed the power of CNNs and RNNs [11, 45, 51] to capture the heart beat dynamics for detecting AF. All the previous works were designed for detection of AF over shorter segments of data (30 second to 90 second segments), and were not designed for the detection of paroxysmal AF which would require longer duration of data. In contrast, we consider ECG segments of 10 minute duration so as to capture paroxysmal AF which might occur as scattered episodes of AF in these segments. Rajpurkar et al. [45] trained a 34 layer CNN to detect arrythmia from single lead ECG time series. In their proposed algorithm, the authors intentionally selected patients exhibiting abnormal rhythms to deal with class imbalance for both training and testing data, however this does not represent the actual prevalence of the rhythms in a real world setting, and will lead to optimistic results in the performance of

the model. It is mentioned in the paper that the dataset consisted of 64,121 ECG records from 29,163 patients, with each ECG record being 30 seconds in duration, which is roughly around two 30 second windows for every patient. This method of choosing the ECG records was rather ambiguous. In the Physionet 2017 challenge, Andreotti et al. [2] reimplemented the model proposed by Rajpurkar et al. [45] and they were ranked 30th in the challenge. This showed that employing a vanilla neural network (in this case a vanilla CNN) on raw input data will not always provide the best performance in AF detection, and suggests that it would be more effective to use a combination of domain specific features and neural networks for AF detection.

Attention mechanisms in neural networks are inspired from the mechanism of visual attention found in humans. Attention mechanisms for neural networks have been well studied in the context of image recognition [17, 29, 40], and have only recently been applied to sequential prediction tasks. Some of the areas that they have been successfully applied to include neural machine translation [3], image caption generation [60], speech recognition [10], and text summarization [47]. We employ a soft attention mechanism, which has the advantage that artificial segmentation of waveforms into positive and negative classes is not a pre-requirement of learning, but rather the algorithm is able to weight the different 30 seconds data segments to arrive at a final diagnosis of AF over the entire 10 minute window, in spite of occasional noise and non-AF rhythms within this window. Moreover, the attention mechanism can provide a measure of the burden of AF, by quantifying the proportion of the 10 minute window where the attention weights are significantly positive.

A major obstacle in application of machine learning techniques to clinical problems is the lack of acceptable clinical gold standards and scarcity of granular labels. Although, the PPG dataset utilized in this study was diligently annotated by multiple clinical adjudicators, the ECG dataset labels were obtained from automatic bedside Holter software, with a coarse overview by a single adjudicator for potential mislabeled cases of AF. Therefore, the ECG-based AF detection results provided in this work should be taken with some caution [11, 63, 64]. In fact, as demonstrated in Table 1 re-defining the labels as a function of AF burden had a significant impact on the performance of the classifier (AUC of 0.94 at 5% or more versus AUC of 0.97 at 75% or more AF burden over a 10 minute segment). Although, our main goal for using the ECG dataset was to utilize the resulting model, trained on a relatively large patient cohort, in a transfer learning framework for prediction of AF in a much smaller cohort of patients with PPG recordings from a smart watch. Nevertheless, the proposed ECG-based algorithm, in concert with the visualization tool presented in Figure 5, can be used in an active learning framework to further refine the labels by consulting with an expert panel of clinicians.

The paradigm presented here sets the stage for a new generation of real-time predictive analytics with broad applications across the growing numbers of devices with heart rate sensors. The Holter ECG dataset is a large scale, diverse training set with many different sub-categories of AF (fast, slow, paroxysmal, etc). As we proved in this study, such a database enables the production of more accurate algorithms in other devices in which such large datasets may not exist. Most notably, AF detection from PPG signal acquired from a

wristband device (such as the Simband) is extremely challenging, given the low signal to noise ratio from such devices. Utilization of large datasets such as that from Holter ECG dataset with transfer learning help to enable algorithms with accuracies that may provide real-time AF detection from PPG wristband devices.

In summary, this study provides a well-motivated deep learning architecture for detection of paroxysmal AF, and demonstrates clinically acceptable AF detection accuracies across different recording modalities. To the best of our knowledge, this is the first study to consider the important problem of paroxysmal AF detection over long hours of recording using a deep learning framework with an attention mechanism. Our future work will be focused on prospective evaluation of the proposed technique in both intensive care unit patients and long-term recordings of patients with paroxysmal AF based on wearable technology.

6 ACKNOWLEDGMENTS AND FUNDING

REFERENCES

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A system for Large-scale Machine Learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 265–283.
[2] Fernando Andreotti, Oliver Carr, Marco A. F. Pimentel, Adam Mahdi, and Maarten De Vos. 2017. Comparing Feature-Based Classifiers and Convolutional Neural Networks to Detect Arrhythmia from Short Segments of ECG. *Computing in Cardiology* 44 (2017).
[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473* (2014).
[4] Jocasta Ball, Melinda J Carrington, John JV McMurray, and Simon Stewart. 2013. Atrial fibrillation: Profile and Burden of an Evolving Epidemic in the 21st Century. *International Journal of Cardiology* 167, 5 (2013), 1807–1824.
[5] Joachim Behar, Julien Oster, Qiao Li, and Gari D Clifford. 2013. ECG Signal Quality during Arrhythmia and its Application to Aalse Alarm Reduction. *IEEE Transactions on Biomedical e=Engineering* 60, 6 (2013), 1660–1666.
[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.
[7] A John Camm and others. 2012. 2012 Focused Update of the ESC Guidelines for the Management of Atrial Fibrillation. *European Heart Journal* 33, 21 (2012), 2719–2747.
[8] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulò. 2017. Autodial: Automatic Domain Alignment Layers. In *International Conference on Computer Vision*.
[9] Marta Carrara, Luca Carozzi, Travis J Moss, Marco de Pasquale, Sergio Cerutti, Manuela Ferrario, Douglas E Lake, and J Randall Moorman. 2015. Heart Rate Dynamics Distinguish Among Atrial Fibrillation, Normal Sinus Rhythm and Sinus Rhythm with Frequent Ectopy. *Physiological Measurement* 36, 9 (2015), 1873.
[10] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based Models for Speech Recognition. In *Advances in Neural Information Processing Systems*. 577–585.
[11] Gari Clifford, Chengyu Liu, Benjamin Moody, L Lehman, Ikaro Silva, Qiao Li, A Johnson, and R Mark. 2017. AF Classification from a Short Single Lead ECG Recording: The Physionet Computing in Cardiology Challenge 2017. *Computing in Cardiology* 44 (2017).
[12] Gari Clifford, Chengyu Liu, Benjamin Moody, L Lehman, Ikaro Silva, Qiao Li, A Johnson, and R Mark. 2017. Physionet Challenge 2017. *https://physionet.org/challenge/2017/* (2017).
[13] Roberta Colloca, Alistair EW Johnson, Luca Mainardi, and Gari D Clifford. 2013. A Support Vector Machine Approach for Reliable Detection of Atrial Fibrillation

[14] Nancy R Cook and Paul M Ridker. 2009. The Use and Magnitude of Reclassification Measures for Individual Predictors of Global Cardiovascular Risk. *Annals of Internal Medicine* 150, 11 (2009), 795.
[15] S Dash, KH Chon, S Lu, and EA Raeder. 2009. Automatic Real Time Detection of Atrial Fibrillation. *Annals of Biomedical Engineering* 37, 9 (2009), 1701–1709.
[16] Ingrid Daubechies. 1990. The Wavelet Transform, Time-frequency Localization and Signal Analysis. *IEEE Transactions on Information Theory* 36, 5 (1990), 961–1005.
[17] Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. 2012. Learning where to attend with Deep Architectures for Image Tracking. *Neural Computation* 24, 8 (2012), 2151–2184.
[18] Barbara J Drew, Patricia Harris, Jessica K Zègre-Hemsey, Tina Mammone, Daniel Schindler, Rebeca Salas-Boni, Yong Bai, Adelita Tinoco, Quan Ding, and Xiao Hu. 2014. Insights into the Problem of Alarm Fatigue with Physiologic Monitor Devices: a Comprehensive Observational Study of Consecutive Intensive Care Unit Patients. *PloS One* 9, 10 (2014), e110274.
[19] Valentin Fuster, Lars E Rydén, David S Cannom, Harry J Crijns, Anne B Curtis, Kenneth A Ellenbogen, Jonathan L Halperin, Jean-Yves Le Heuzey, G Neal Kay, James E Lowe, and others. 2006. ACC/AHA/ESC 2006 Guidelines for the Management of Patients with Atrial Fibrillation: full text: a Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the European Society of Cardiology Committee for Practice Guidelines (Writing Committee to Revise the 2001 guidelines for the management of patients with atrial fibrillation) developed in collaboration with the European Heart Rhythm Association and the Heart Rhythm Society. *Europace* 8, 9 (2006), 651–745.
[20] Mohammad Ghassemi, Li-wei Lehman, Jasper Snoek, and Shamim Nemati. 2014. Global Optimization Approaches for Parameter Tuning in Biomedical Signal Processing: A Focus on Multi-scale Entropy. In *Computing in Cardiology Conference, 2014*. IEEE, 993–996.
[21] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech Recognition with Deep Recurrent Neural Networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 6645–6649.
[22] Aslak Grinsted, John C Moore, and Svetlana Jevrejeva. 2004. Application of the Cross Wavelet Transform and Wavelet Coherence to Geophysical Time Series. *Nonlinear Processes in Geophysics* 11, 5/6 (2004), 561–566.
[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[24] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. 2017. Learning to Cluster in Order to Transfer Across Domains and Tasks. *arXiv preprint arXiv:1711.10125* (2017).
[25] Alistair EW Johnson, Joachim Behar, Fernando Andreotti, Gari D Clifford, and Julien Oster. 2014. R-peak Estimation using Multimodal Lead Switching. In *Computing in Cardiology Conference*. IEEE, 281–284.
[26] Alistair EW Johnson, Joachim Behar, Fernando Andreotti, Gari D Clifford, and Julien Oster. 2015. Multimodal Heart Beat Detection using Signal Quality Indices. *Physiological Measurement* 36, 8 (2015), 1665.
[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. 1097–1105.
[28] Douglas E Lake and J Randall Moorman. 2010. 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* 300, 1 (2010), H319–H325.
[29] Hugo Larochelle and Geoffrey E Hinton. 2010. Learning to Combine Foveal Glimpses with a third-order Boltzmann Machine. In *Advances in Neural Information Processing Systems*. 1243–1251.
[30] Yann LeCun, Yoshua Bengio, and others. 1995. Convolutional Networks for Images, Speech, and Time Series. *The Handbook of Brain Theory and Neural Networks* 3361, 10 (1995), 1995.
[31] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based Learning Applied to Document Recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[32] C Lee Giles, Gary M Kuhn, and Ronald J Williams. 1994. Dynamic Recurrent Neural Networks: Theory and Applications. *IEEE Transactions on Neural Networks* 5, 2 (1994), 153–156.
[33] Qiao Li, Roger G Mark, and Gari D Clifford. 2007. Robust Heart Rate Estimation From Multiple Asynchronous Noisy Sources Using Signal Quality Indices and a Kalman Filter. *Physiological Measurement* 29, 1 (2007), 15.
[34] David Thor Linker. 2009. Long-term Monitoring for Detection of Atrial Fibrillation. (Dec. 8 2009). US Patent 7,630,756.
[35] Gregory YH Lip and FL Li Saw Hee. 2001. Paroxysmal Atrial Fibrillation. *QJM: An International Journal of Medicine* 94, 12 (2001), 665–678.
[36] Mingsheng Long, Jianmin Wang, and Michael I Jordan. 2016. Deep Transfer Learning with Joint Adaptation Networks. *arXiv preprint arXiv:1605.06636* (2016).

[37] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[38] Juan Pablo Martínez, Rute Almeida, Salvador Olmos, Ana Paula Rocha, and Pablo Laguna. 2004. A Wavelet-based ECG Delineator: Evaluation on Standard Databases. *IEEE Transactions on Biomedical Engineering* 51, 4 (2004), 570–581.

[39] MATLAB. 2016. *version 9.1 (R2016b)*. The MathWorks Inc., Natick, Massachusetts.

[40] Volodymyr Mnih, Nicolas Heess, Alex Graves, and others. 2014. Recurrent Models of Visual Attention. In *Advances in Neural Information Processing Systems*. 2204–2212.

[41] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, and others. 2015. Human-level Control through Deep Reinforcement Learning. *Nature* 518, 7540 (2015), 529.

[42] George Moody. 1983. A New Method for Detecting Atrial Fibrillation using RR Intervals. *Computers in Cardiology* (1983), 227–230.

[43] Shamim Nemati, Mohammad M Ghassemi, Vaidehi Ambai, Nino Isakadze, Oleksiy Levantsevych, Amit Shah, and Gari D Clifford. 2016. Monitoring and Detecting Atrial Fibrillation Using Wearable Technology. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 3394–3397.

[44] Andrius Petrenas, Vaidotas Marozas, Leif Sornmo, and Arunas Lukosevicius. 2012. An Echo State Neural Network for QRST Cancellation during Atrial Fibrillation. *IEEE Transactions on Biomedical Engineering* 59, 10 (2012), 2950–2957.

[45] Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. 2017. Cardiologist-level Arrhythmia Detection with Convolutional Neural Networks. *arXiv preprint arXiv:1707.01836* (2017).

[46] Joshua S Richman and J Randall Moorman. 2000. Physiological Time-series Analysis Using Approximate Entropy and Sample Entropy. *American Journal of Physiology-Heart and Circulatory Physiology* 278, 6 (2000), H2039–H2049.

[47] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. *arXiv preprint arXiv:1509.00685* (2015).

[48] Samsung. 2017. Simband's Official Website. *https://www.simband.io/* (2017).

[49] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[50] Burr Settles. 2012. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6, 1 (2012), 1–114.

[51] Supreeth Prajwal Shashikumar, Amit J Shah, Qiao Li, Gari D Clifford, and Shamim Nemati. 2017. A Deep Learning Approach to Monitoring and Detecting Atrial Fibrillation Using Wearable Technology. In *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*. IEEE, 141–144.

[52] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and others. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* 529, 7587 (2016), 484–489.

[53] Simon Stewart, Carole L Hart, David J Hole, and John JV McMurray. 2002. A Population-based Study of the Long-term Risks Associated with Atrial Fibrillation: 20-year Follow-up of the Renfrew/Paisley Study. *The American Journal of Medicine* 113, 5 (2002), 359–364.

[54] Martin Stridh and L Sornmo. 2001. Spatiotemporal QRST Cancellation Techniques for Analysis of Atrial Fibrillation. *IEEE Transactions on Biomedical Engineering* 48, 1 (2001), 105–111.

[55] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of Frustratingly Easy Domain Adaptation.. In *AAAI*, Vol. 6. 8.

[56] K Tateno and L Glass. 2001. Automatic Detection of Atrial Fibrillation Using the Coefficient of Variation and Density Histograms of RR and ΔRR Intervals. *Medical and Biological Engineering and Computing* 39, 6 (2001), 664–671.

[57] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for Machine Learning* 4, 2 (2012), 26–31.

[58] Christopher X Wong, Anthony G Brooks, Darryl P Leong, Kurt C Roberts-Thomson, and Prashanthan Sanders. 2012. The Increasing Burden of Atrial Fibrillation Compared with Heart Aailure and Myocardial Infarction: a 15-Year Study of All Hospitalizations in Australia. *Archives of Internal Medicine* 172, 9 (2012), 739–741.

[59] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, and others. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144* (2016).

[60] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*. 2048–2057.

[61] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical Attention Networks for Document Classification.. In *HLT-NAACL*. 1480–1489.

[62] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based Bidirectional Long Short-term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 207–212.

[63] Tingting Zhu, Nic Dunkley, Joachim Behar, David A Clifton, and Gari D Clifford. 2015. Fusing continuous-valued medical labels using a Bayesian model. *Annals of Biomedical Engineering* 43, 12 (2015), 2892–2902.

[64] Tingting Zhu, Alistair EW Johnson, Joachim Behar, and Gari D Clifford. 2014. Crowd-sourced annotation of ECG signals using contextual information. *Annals of Biomedical Engineering* 42, 4 (2014), 871–884.