AN ENSEMBLE OF DEEP RECURRENT NEURAL NETWORKS FOR P-WAVE DETECTION IN ELECTROCARDIOGRAM

Abdolrahman Peimankar and Sadasivan Puthusserypady

Department of Health Technology, Technical University of Denmark, Kgs. Lyngby 2800, Denmark
{apeima,sapu}@dtu.dk

ABSTRACT

Detection of P-waves in electrocardiogram (ECG) signals is of great importance to cardiologists in order to help them diagnosing arrhythmias such as atrial fibrillation. This paper proposes an end-to-end deep learning approach for detection of P-waves in ECG signals. Four different deep Recurrent Neural Networks (RNNs), namely, the Long-Short Term Memory (LSTM) are used in an ensemble framework. Each of these networks are trained to extract the useful features from raw ECG signals and determine the absence/presence of P-waves. Outputs of these classifiers are then combined for final detection of the P-waves. The proposed algorithm was trained and validated on a database which consists of more than 111000 annotated heart beats and the results show consistently high classification accuracy and sensitivity of around 98.48% and 97.22%, respectively.

Index Terms— Deep learning, Ensemble learning, Long-Short Term Memory, Electrocardiogram, P-waves detection.

1. INTRODUCTION

Analysis of electrocardiogram (ECG) signals is considered an important step in diagnosing cardiac diseases, especially the atrial fibrillation (AFIB), which is one of the most common cardiac arrhythmias among elderly population [1]. Demographics of western countries is alarming with respect to cardiac health issues [2, 3]. Since P-wave absence in ECG is one of the clinically useful informations for the detection of AFIB, P-wave delineation is of great importance in practice.

One of the most common ways for physicians to delineate P-waves is through visual examination of the ECG recordings. However, it is not always easy and in most cases cumbersome to analyse these huge amounts of data. Therefore, it is required to develop analytic methods to automatically analyse these ECG signals, which help accelerating the process of accurate detection of P-waves. Various state-of-the-art algorithms for detecting P-waves have been introduced in the literature. Dubois et al. used a machine learning method, namely, the generalized orthogonal forward regression with Gaussian mesa function models for automatic ECG wave extraction [4]. In [5], the possibility of automatic ECG delineation using phasor transform was studied. Lin et al. proposed a Bayesian model for P- and T-waves detection which showed a higher accuracy compared to previously published algorithms but at a higher computational cost [6]. In a recent study, Gonzáles et al. investigated ECG waveform segmentation using adaptive slope Gaussian detection [7]. In [8], a template based model has been applied for detecting the P- and T-waves.

Ensemble learning methods are being used for prognostics and decision making in various applications [9, 10]. The three main parts of ensemble learning systems are [11]: (i) sampling from a dataset to make a training set, (ii) training a group of classifiers, and (iii) combining the output of classifiers. It has been shown in the literature that using ensemble learning increases the chance of selecting more accurate classifiers by avoiding selection of a single weak classifier [11, 12]

In this work, four different deep Long-Short Term Memory (LSTM) classifiers have been trained using raw ECG signals to distinguish between heart beats with P-waves (P) and without P-waves (non-P). The first two classifiers are the conventional LSTM networks [13] with different classification layers; cross entropy (LSTM-CE) and sum of squares error (LSTM-SSE). The other two are bidirectional LSTMs [14] (BiLSTM-CE and BiLSTM-SSE), which also uses CE and SSE as classification layers. Each of these classifiers is trained separately using 5-fold cross validation before the outputs are combined using the Dempster-Shafer theory (DST) [15] to enhance the classification accuracy.

2. MATERIALS AND METHODS

2.1. Preprocessing

ECG signals are filtered to remove noise and baseline wanders [16]. PhysioNet WFDB Toolbox is used to detect the R peaks and RR intervals (RRI) before segmenting the ECG signals beat by beat [17].

This work is supported by the Innovation Fund Denmark (REAFEL, IFD Project No: 6153-00009B).

2.2. Classification networks structure

Recurrent Neural Networks (RNNs) are designed to work with sequential time-series which are capable of learning dependencies in sequential information. However, it has been shown that learning long-term dependencies are very challenging [18]. LSTM networks, a special type of RNNs, are capable of addressing the problem of unstable gradient and can handle long-term dependencies [13]. As shown in Fig. 1, there are three main parts in a LSTM block: (i) forget gate (f_n), (ii) input gate (i_n), and (iii) output gate (o_n). Forget and output gates are mainly responsible to remove or add information to the memory block in the following way:

$$f_n = \varphi(b_f + \mathbf{u}_f^T \mathbf{x}_n + \mathbf{w}_f^T \mathbf{h}_{n-1}), \quad (1)$$

$$i_n = \varphi(b_i + \mathbf{u}_i^T \mathbf{x}_n + \mathbf{w}_i^T \mathbf{h}_{n-1}), \quad (2)$$

where $\mathbf{x}_n$ is the input sequence at time step n and $\mathbf{h}_{n-1}$ is the output sequence at time step $n-1$. The $\mathbf{u}_f$, $\mathbf{w}_f$, $\mathbf{u}_i$, $\mathbf{w}_i$, represent the weight matrices and b_f and b_i are bias terms. These should be learned in the training phase of the LSTM. In addition, since $0 \leq \varphi(\cdot) \leq 1$, this controls the contribution of each unit in the memory block. Therefore, the memory c_n is updated as, $c_n = f_n c_{n-1} + i_n \tilde{c}_n$, where $\tilde{c}_n = \tanh(b_c + \mathbf{u}_c^T \mathbf{x}_n + \mathbf{w}_c^T \mathbf{h}_{n-1}) \subseteq \{-1, +1\}$. Finally, the output vector h_n is computed as, $h_n = o_n \tanh(c_n)$, where $o_n = \varphi(b_o + \mathbf{w}_o^T \mathbf{x}_n + \mathbf{u}_o^T \mathbf{h}_{n-1})$, $\mathbf{u}_o$ and $\mathbf{w}_o$ are the weight matrices of the output gate, and b_o is the output bias.

BiLSTM is a variant of the LSTM which can look at a sequence of data in both directions. It consists of two hidden layers which are fed forward to the output layer [19]. In a recent study, it has been shown that choosing a suitable cost function has a major impact during the training of a deep neural network [20]. Therefore, for training the LSTM and BiLSTM networks, two different cost functions are used. These are cross-entropy (CE) and sum of square errors (SSE), which are formulated as follows:

$$\mathcal{J}_{CE}(\mathbf{X}, \mathbf{Y}) = -\frac{1}{M} \sum_{i=1}^{M} [\mathbf{y}_i \ln(a(\mathbf{x}_i)) + (1 - \mathbf{y}_i) \ln(1 - a(\mathbf{x}_i))], \quad (3)$$

$$\mathcal{J}_{SSE}(\mathbf{X}, \mathbf{Y}) = -\frac{1}{M} \sum_{i=1}^{M} (\mathbf{y}_i - a(\mathbf{x}_i))^2, \quad (4)$$

where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ are the training inputs and targets, a is the activation of the output layer, and M is the number of training inputs. There are two LSTM or BiLSTM layers in each network with 150 hidden units. The output of the last LSTM or BiLSTM layer is followed by a fully connected layer. Finally, the output of the fully connected layer with the size of 2 for the binary classification is fed into the classification layer which is a softmax function and is performed as follows:

$$P_i = softmax(\mathbf{z}_i) = \frac{e^{\mathbf{z}_i}}{\sum_i e^{\mathbf{z}_i}}, \quad i = 1, 2, \quad (5)$$

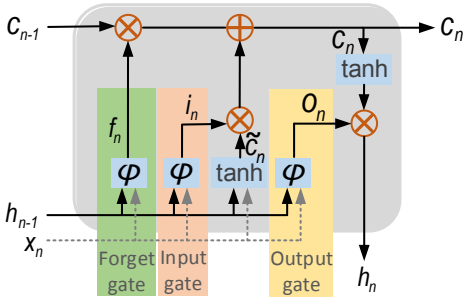


Fig. 1: Schematic diagram of a LSTM memory block.

where $\mathbf{z}_i$ is the output vector of the fully connected layer for the two classes (P and non-P) and P_i is the predicted probability for the corresponding class.

2.3. Dempster-Shafer combination rule

The Dempster-Shafer theory (DST) is a powerful technique for information fusion [15]. It is also capable of capturing the degree of certainties from different information sources [21]. In DST, there are three main functions that are used. They are: (i) mass probability function (m), (ii) belief function (Bel), and (iii) plausibility function (Pl). Of these, m is the most important in binary classification problems and should meet the following conditions [15]:

$$m(X) : 2^X \rightarrow [0, 1] \ni m(\emptyset) = 0, \text{and} \sum_{A \subseteq X} m(A) = 1, \quad (6)$$

where X is the universal set and $\emptyset$ is the empty set. In this work, $X = \{P, non\text{-}P\}$.

One advantage of DST is its capability for combining independent evidences (mass probability functions). For example, for two mass probability functions, m_1 and m_2, in order to produce more informative evidence, which is denoted as $m_1 \oplus m_2$ and is calculated as,

$$(m_1 \oplus m_2)(A) = \frac{1}{1 - L} \sum_{\substack{B \cap C = A \neq \emptyset \\ B, C \subseteq X}} m_1(B) m_2(C), \quad (7)$$

where $L = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$. In the proposed ensemble method (Section 3), the outputs of the single classifiers are actually normalized the mass functions between 0 and 1 which can be used by DST to compute a combined output.

3. ENSEMBLE P-WAVE DETECTION

The proposed algorithm is described in eight different steps as follows:

1. *Noise reduction*: The ECG signals are filtered to remove noise and baseline wanders.

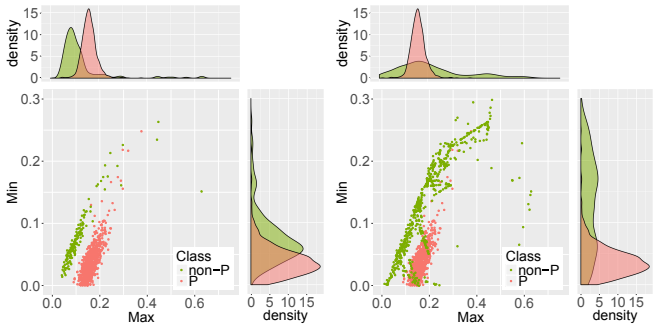
2. *R peaks detection*: As mentioned in Section 2.1, the PhysioNet WFDB Toolbox is used to read R peaks annotations from recorded ECG signals.
3. *Segmentation*: In this step, the RRIs are calculated and if the the RRI is less than 0.4s as suggested in [22], the corresponding heart beat is not considered since its P-wave and the T-wave from the previous beat may be overlapped. Otherwise, a segment of length equal to 1/3rd of the RRI, which starts from $RRI - (2/3)RRI$ and ends at $RRI - 0.07RRI$, is extracted for each ECG beat. Since the lengths of these segments may be different, they have been padded accordingly. For this purpose, each segment is randomly padded from the beginning, end, or both sides in a way that each segment has a length of 150 samples, which corresponds to 1.67s. By doing this, the training process will be more smooth for the networks and we prevent the models to have bias over a certain area of the extracted segments.
4. *Creating a synthetic dataset*: In this study, the number of extracted segments with P in the dataset are much more than non-P. Thus, the size of non-P class is not sufficient for effectively training the classification algorithms. In addition, using this imbalanced dataset increases the chance of biased classification, which in turn leads to a higher error rate on the minority class [23]. Adaptive synthetic oversampling technique (ADASYN) [23] is applied to make the dataset balanced and improve the training of the classifiers. This will also enable the classification algorithms achieving their desirable performance. For brevity, the detailed description of ADASYN is omitted here but can be found in [23].
5. *Cross validation*: The whole dataset is divided into training and validation sets using 5-fold cross validation.
6. *Train all classifiers*: The four different classifiers are trained separately using the 5-fold training sets.
7. *Evaluate all the trained classifiers*: The 5-fold validation sets are provided to the trained networks to evaluate the performance of the four different classifiers.
8. *Combine the outputs of the classifiers*: The outputs of all the classifiers, which are posterior probabilities assigned to the two classes (*P* & *non-P*), are combined using DST.

4. EXPERIMENTAL VALIDATION

In this study, we have used the QT Database (QTDB) on PhysioNet [24] to evaluate the performance of the proposed method. It consists of 105 ECG recordings of length 15 minutes each and sampled at 250 Hz. The imbalanced QTDB dataset is composed of 111041 heart beats. Table 1 shows the total number of beats for each of the two classes. It can be seen that the number of beats per class is distributed evenly after applying ADASYN method. In order to show the effect of making the dataset balanced, the distribution of the imbalanced and balanced samples of one of the signals (records #10) are plotted in Fig.2a and Fig.2b, respectively. This clearly show how synthetic data are distributed with respect to two arbitrary features, which are the maximum (*Max*) and minimum (*Min*) values of the extracted segments.

Classes	*P*	*non-P*
Number of segments (imbalanced)	97117	13924
Number of segments (balanced)	97117	98043

Table 1: Number of segments for balanced and imbalanced datasets.



(a) (b)

Fig. 2: Scatter plots of extracted segments for two arbitrary features (Min and Max): (a) Imbalanced, (b) Balanced.

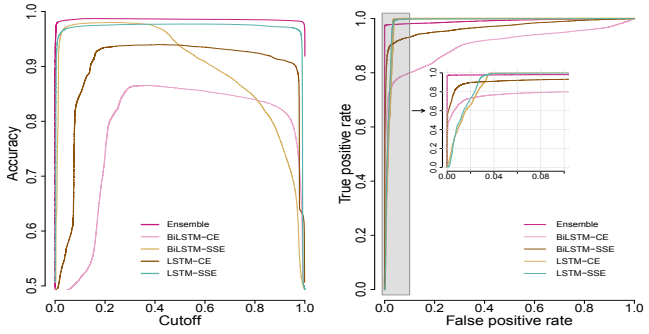
The raw extracted segments are used as inputs for training the classifiers. The 5-fold cross validation accuracy on the training and validation sets for the four single classifiers are reported in Table 2. The LSTM-SSE outperforms other single classifiers on both training and validation sets followed by BiLSTM-SSE, LSTM-CE, and BiLSTM-CE.

Algorithms	LSTM-CE	LSTM-SSE	BiLSTM-CE	BiLSTM-SSE
Training (%)	89.97	**97.58**	87.28	93.37
Validation (%)	87.96	**97.27**	85.21	91.31

Table 2: 5-fold cross validation accuracies of classifiers.

In the final step, outputs of the single classifiers are combined using the DST. The ensemble model improves the classification accuracy by more than 1% (98.49% vs 97.27%). In addition, accuracy of single classifiers along with the ensemble model at different cutoff points are shown in Fig.3a. The ensemble model has the most robust performance and its accuracies remain stable irrespective of the increase in the cutoff points. Although, LSTM-SSE as the best single classifier is relatively insusceptible to cutoff point changes compared to other algorithms, there is however a slight decrease at around cut off point 0.85. Fig.3b illustrates the comparison between true positive rate (TPR) and false positive rate (FPR) using receiver operating characteristics (ROC) curves. The zoomed area represents the FPR equals to 0.1, which means that only 10% of all incorrect classification cases are actually false positive. In other words, only 10% of incorrect classifications are the segments with absence of P-waves but classified as having P-waves. Despite the overall high accuracy, in healthcare

research, it is very important for a model to achieves a higher TPR at a lower FPR as shown in Fig. 3b.



(a) (b)

Fig. 3: Comparison of the classification performance: (a) accuracy at different cutoffs, (b) ROC curves.

The area under the curve (AUC) of the ROC curve is a numerical measure, which evaluates the capability of a classifier. The AUC, partial AUC (pAUC) at FPR equal to 0.1, and accuracy of all the classifiers at cutoff point equal to 0.9 are given in Table 3. It may be noted that the proposed ensemble model has significantly enhanced the performance of the classification task compared to the single classifiers.

Measure	Accuracy (%)	AUC	pAUC
LSTM-CE	91.26	0.97	0.090
LSTM-SSE	97.21	0.98	0.086
BiLSTM-CE	82.34	0.91	0.075
BiLSTM-SSE	76.69	0.98	0.085
Ensemble (proposed)	**98.48**	**0.99**	**0.098**

Table 3: Comparison of the accuracy, AUC, and pAUC on the validation set.

In order to further analyse the performance of the proposed algorithm, other classification measures such as $F1$-score, sensitivity (Se), and positive predictive value (PPV) (defined below) are also calculated and reported in Table 4.

$$Se = \frac{TP}{TP+FN}, \quad (8)$$

$$PPV = \frac{TP}{TP+FP}, \quad (9)$$

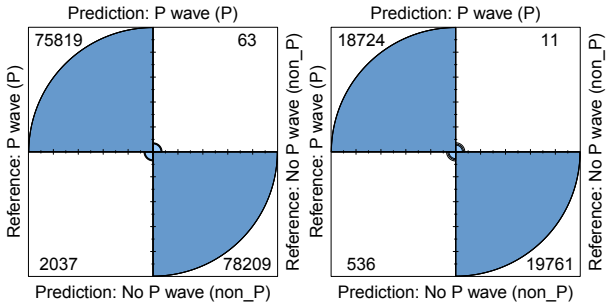
$$F = (1+\beta^2)\frac{PPV \cdot Se}{(\beta^2 \cdot PPV)+Se}. \quad (10)$$

where TP, FN, and FP are the number of true positive, false negative, and false positive cases, respectively. F-score is nothing but a weighted harmonic mean of Se and PPV, which takes both Se and PPV into account equally when $\beta = 1$ and is called the $F1$-score. The value of $F1$-score is in the range of 0 and +1, in which +1 represents the perfect classification while 0 represents the worst classification performance.

The confusion matrices for the training and validation phases of the proposed ensemble algorithm are shown in Fig. 4a and 4b, respectively. This shows the high rate of TPs and TNs compared to FPs and FNs.

Measure	F1 score	Se (%)	PPV (%)
Training	0.9863	97.38	99.92
Validation	0.9856	97.22	99.94

Table 4: Classification measures of the proposed algorithm on the training and validation sets.



(a) (b)

Fig. 4: Confusion matrix. (a) Training. (b) Validation.

It should be mentioned that there are two main advantages for the proposed ensemble deep learning model. First, unlike most of the state-of-the-art algorithms, which use feature engineering approaches, there is no need to define any features for the P-waves detection in the proposed algorithm. Second, using an ensemble of deep recurrent networks improves the performance of the single classifiers for detection of the P-waves. To the best of our knowledge, this is the first study that investigates the P-waves detection in ECG signals using an ensemble deep learning framework.

5. CONCLUSION

P-waves detection has been one of the most challenging tasks in ECG waveform delineation. In this paper, an ensemble of deep recurrent networks has been proposed to detect P-waves in ECG recordings. First, four different classifiers (LSTM-CE, LSTM-SSE, BiLSTM-CE, and BiLSTM-SSE) were trained using 5-fold cross validation on PhysioNet QTDB dataset. The laborious feature extraction step was omitted and the raw ECG segments were directly used as inputs for training the networks. The trained networks were then tested on the validation sets. Finally, the DST combination rule was used to combine the outputs of the classifiers, which were the posterior probabilities assigned to the two classes (P and non-P). The very impressive results obtained in our work (even without the feature extraction step) provide us with the opportunity to use this algorithm in-house by cardiologists to diagnose cardiac arrhythmias. This algorithm is currently being combined with our developed model to classify AFIB [25].

6. REFERENCES

[1] Y. Iwasaki, K. Nishida, T. Kato, and S. Nattel, "Atrial fibrillation pathophysiology: implications for management," *Circulation*, vol. 124, no. 20, pp. 2264–2274, 2011.

[2] S. Stewart, N. Murphy, A. Walker, A. McGuire, and J.J.V. McMurray, "Cost of an emerging epidemic: an economic analysis of atrial fibrillation in the UK," *Heart*, vol. 90, no. 3, pp. 286–292, 2004.

[3] M. Zoni-Berisso, F. Lercari, T. Carazza, and S. Domenicucci, "Epidemiology of atrial fibrillation: European perspective," *Clinical epidemiology*, vol. 6, pp. 213 – 220, 2014.

[4] R. Dubois, P. Maison-Blanche, B. Quenet, and G. Dreyfus, "Automatic ECG wave extraction in long-term recordings using Gaussian mesa function models and nonlinear probability estimators," *Computer Methods and Programs in Biomedicine*, vol. 88, no. 3, pp. 217–233, 2007.

[5] A. Martìnez, R. Alcaraz, and J.J. Rieta, "Automatic electrocardiogram delineator based on the phasor transform of single lead recordings," in *2010 Computing in Cardiology*, Sep 2010, pp. 987–990.

[6] C. Lin, C. Mailhes, and J. Tourneret, "P- and T-wave delineation in ECG signals using a bayesian approach and a partially collapsed Gibbs sampler," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2840–2849, Dec 2010.

[7] Francisco González, Raúl Alcaraz, and José J Rieta, "Electrocardiographic P-wave delineation based on adaptive slope Gaussian detection," *Computing*, vol. 44, pp. 1, 2017.

[8] A. Karimipour and M.R. Homaeinezhad, "Real-time electrocardiogram P-QRS-T detection - delineation algorithm based on quality-supported analysis of characteristic templates," *Computers in Biology and Medicine*, vol. 52, pp. 153–165, 2014.

[9] A. Peimankar and S. Puthusserypady, "Ensemble learning for detection of short episodes of atrial fibrillation," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep 2018, pp. 66–70.

[10] A. Peimankar, S.J. Weddell, T. Jalal, and A.C. Lapthorn, "Evolutionary multi-objective fault diagnosis of power transformers," *Swarm and Evolutionary Computation*, vol. 36, pp. 62–75, 2017.

[11] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and systems magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[12] A. Peimankar, S.J. Weddell, T. Jalal, and A.C. Lapthorn, "Multi-objective ensemble forecasting with an application to power transformers," *Applied Soft Computing*, vol. 68, pp. 233–248, 2018.

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

[14] M. Schuster and K.K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov 1997.

[15] G. Shafer, *A mathematical theory of evidence*, vol. 42, Princeton university press, 1976.

[16] M.J. Shensa, "The discrete wavelet transform: wedding the a trous and mallat algorithms," *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2464–2482, Oct 1992.

[17] A.L Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, and H.E. Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[18] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157 – 166, 1994.

[19] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[20] Anonymous, "Cross-entropy loss leads to poor margins," in *Submitted to International Conference on Learning Representations*, 2019, under review.

[21] L.A. Klein, *Sensor and data fusion: a tool for information assessment and decision making*, vol. 324, SPIE press, 2004.

[22] L. Maršánová, A. Němcová, R. Smíšek, T. Goldmann, M. Vítek, and L. Smital, "Automatic detection of P wave in ECG during ventricular extrasystoles," in *World Congress on Medical Physics and Biomedical Engineering 2018*. Springer, 2019, pp. 381–385.

[23] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks*, June 2008, pp. 1322–1328.

[24] P. Laguna, R.G Mark, A. Goldberg, and G.B. Moody, "A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG," in *Computers in cardiology 1997*. IEEE, 1997, pp. 673–676.

[25] 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.