A two-staged classifier to reduce false positives: On device detection of atrial fibrillation using phase-based distribution of poincaré plots and deep learning

A Two-Staged Classifier to Reduce False Positives: On Device Detection of Atrial Fibrillation using Phase-Based Distribution of Poincaré Plots and Deep Learning

Peter Doggart (1,2)

peter.doggart@pulseai.io

Dr Alan Kennedy (1,2)

alan.kennedy@pulseai.io

Dr Raymond Bond (2)

rb.bond@ulster.ac.uk

Dr Dewar Finlay (2)

d.finlay@ulster.ac.uk

Stephen W. Smith, MD (3,4)

smith253@umn.edu

(1) PulseAI, 58 Howard Street, Belfast, Northern Ireland, BT1 6PL.

(2) Ulster University, Shore Road, Northern Ireland, BT37 OQB.

(3) Department of Emergency Medicine, Hennepin County Medical Center, Minneapolis, NM, USA.

(4) University of Minnesota, Department of Emergency Medicine, USA.

Introduction

Atrial Fibrillation (AF) is the most common chronic arrhythmia associated with an adverse prognosis. It is associated with heart failure, stroke and excess mortality [1]. AF affects around 33.5 million individuals worldwide, including around 3 million in the USA [2]. The ratio of asymptomatic ("silent" AF) to symptomatic AF has been estimated to be approximately 12:1 [3]. It is therefore critical to capture asymptomatic episodes in high risk patient groups who might otherwise be undiagnosed. Mobile Cardiac Outpatient Telemetry (MCOT) is often used in these scenarios, as it is suited to capture both symptomatic and asymptomatic AF episodes. MCOT is used both in pre-ablation monitoring, to document AF burden and initiating triggers, and in post-ablation to document freedom from AF episodes. [3]

MCOT devices rely primarily on the algorithmic detection of AF events, which are then stored and transmitted to a clinician for review. Prior studies of these algorithms have demonstrated frequent under diagnosis [4] as well as over diagnosis of AF [5]. Under diagnosis is much more common in MCOT devices, as devices typically opt for low sensitivity to maintain acceptable levels of false positives, and therefore PPV. Low PPV creates additional workload for over-reading clinicians, who would have to over-read many normal recordings to find true arrhythmia events. This poor algorithmic classification performance comes from two sources. Firstly, it is very difficult to select an optimal threshold for traditional algorithms, as they are often developed using small data sets. Secondly, human engineered features often fail to capture the underlying complexity of the human physiology for classification [6].

The use of Artificial Intelligence (AI), more specifically Deep Convolutional Neural Networks (DCNNs) for ECG classification, is well documented in the literature. DCNNs have previously been applied to increase the PPV of Implantable Loop Recorders (ILRs) by acting as a filter to remove false positive results prior to clinical review [5]. ILRs collect, analyse and transmit

data in much the same way as MCOT devices. The DCNN was, however, significantly too large and computationally demanding to be embedded within the ILR itself.

The purpose of this study is to build and analyse the performance of a miniaturized DCNN that could be embedded onto a MCOT device to reduce the number of transmitted false positives whilst maintaining a high sensitivity for AF detection.

Materials and Methods

Datasets

We created two non-overlapping training datasets from the proprietary PulseAI worldwide ECG database. This database contains labelled ECGs from over 1 million patients from 7 countries. The ECGs were labelled as part of standard clinical care, with a cardiologist or emergency medicine physician over-reading the automated ECG machine interpretation. The first dataset consisted of all records of 10 seconds in length and where Lead II was present in the recording. We selected four classes (1) atrial fibrillation and atrial flutter, (2) sinus arrhythmia, (3) sinus rhythm with ectopy and (4) all other rhythms. Sinus arrhythmia and sinus rhythm with ectopy are commonly misclassified as AF, and therefore we extracted examples of these classes to increase the classification performance of the DCNN on these classes. 491,727 ECGs were extracted from the database for this dataset. The second dataset was created using 20-second recordings where Lead II was present, and simply comprised of two class labels; (1) atrial fibrillation and flutter (2) all other labels. This dataset contained 58,450 recordings. The distribution of class labels for both of these datasets is shown in Table 1.

	Dataset One		**Dataset Two**	
Class Label	**Count**	**Prevalence (%)**	**Count**	**Prevalence (%)**
Atrial Fibrillation / Flutter	33,591	6.83	3,585	6.13
Sinus Arrhythmia	23,109	4.70	-	-
Sinus Rhythm with Ectopy	26,907	5.47	-	-
All Other Rhythms	408,120	83.00	54,865	93.87

Table 1 - Distribution of class labels in dataset one and two.

Three Physionet datasets [7] were used to test the performance of our classifier: MIT-BIH Arrhythmia Database (MITDB) [8], MIT-BIH Atrial Fibrillation Database (AFDB) [9] and, Long-Term Atrial Fibrillation Database (LTAFDB) [10].

Methods

Even in high risk populations, the pre-test probability of an AF episode is relatively low. We therefore theorize that the majority of classification windows for any classifier will be 'normal'. These periods of sinus rhythm can be easily classified by existing low complexity embedded algorithms. However, other irregular rhythms, such as frequent, complex atrial or ventricular ectopy are more difficult to classify [11]. We therefore propose a two-stage classifier for AF detection. Firstly, a low complexity, rule-based, continuous classifier based on QRS detection and RR interval analysis to identify periods of irregularity that could be AF. If a period is flagged as irregular, then further analysis is undertaken by the on-device DCNN.

Statistical analysis

All 95% confidence intervals (CIs) were computed using Wilson score intervals. Differences of positive predictive values and F1 scores between the one-stage and two-stage classifier (unpaired data) were tested with 2-sided proportion Z tests. A value of $p < 0.05$ was considered to be statistically significant.

Rule-based classifier

We implemented the classifier detailed by Luo et al. [12] which utilizes a simple but effective analysis of first-order Poincaré Plots. It computes a polar coordinate transfer of the Poincaré Plot to create a histogram-like phase distribution. Two features, the distribution width *Dw* and the average distribution height *Dh* are extracted and used to classify each episode as AF or not AF. We did not make any modifications to this algorithm and applied the published threshold values for classification of each 20-second window of ECG data. The choice of RR interval classifier is not important as long as the classifier is capable of high detection

sensitivity. Although not reported here, we achieved very similar results when using coefficient of sample entropy [11].

Deep convolutional neural network

We trained a small residual network (resnet) with 3 residual blocks with up to 64 convolutional filters in each layer. The structure of the residual blocks is shown in Figure 1. The output of the residual blocks was fed through a global average pooling layer, then classified using a dense layer with softmax activation. A resnet was selected to decrease the depth of the model and also to decrease training time. The model classifies each ECG at a sampling frequency of 256 Hz and contains 128,612 parameters quantized to int8 requiring just 158 KB of flash storage. This network was trained on dataset one and no hyperparameter tuning was performed. To train and assess the model performance, we held-out 10% of the dataset for unseen test data and then split the remaining data into 80% training and 20% validation with blind oversampling of the smaller classes for training.

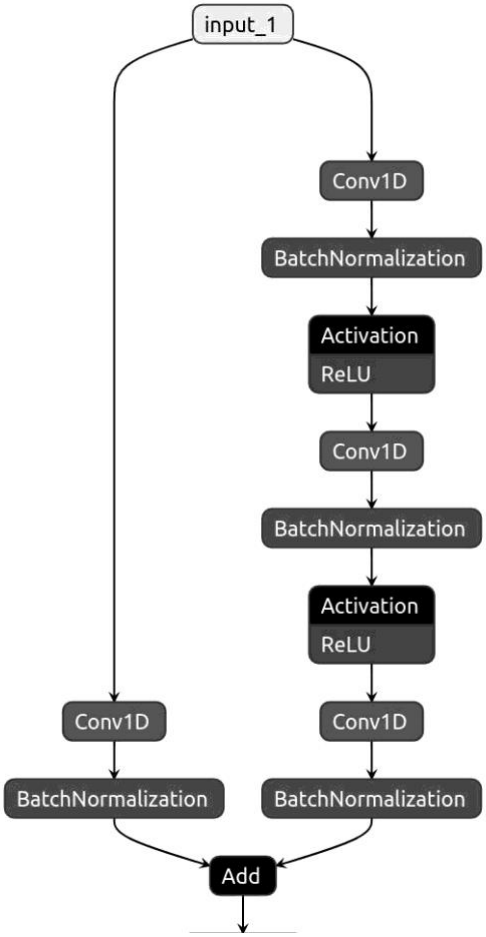

Figure 1 - Residual block structure in the DCNN.

Setting a detection threshold for this DCNN requires an additional step. The held-out test data for training the DCNN has a 6% prevalence of AF. As PPV is the primary performance metric for this study, the impact of prevalence is important to consider. In this case, the DCNN only processes ECGs after they have been labelled as AF by the rule-based classifier. As a result, the DCNN processes a dataset that has a much larger prevalence of AF, as the majority of the normal ECGs have already been discarded. To set our detection threshold for the DCNN, we computed the rule-based labels for all ECGs in dataset two, and then provided only the AF labelled examples to the trained DCNN. We then used the output scores to determine the threshold using the Precision-Recall (PR) curve, with a target sensitivity of 95%.

Results

Once training had completed, the DCNN had an area under the precision recall curve (AUPRC) of 0.853 (95% CI: 0.850 to 0.856) on the held-out dataset one ECGs. On dataset two, the prevalence of AF is increased and so too does the classifier performance, with an AUPRC of 0.918 (95% CI: 0.916 to 0.920). We used the PR curve for dataset two in Figure 2 to select a threshold of 0.15 for the output of the DCNN, giving a 95% sensitivity and 78% PPV. The precision recall curves are only used to help set the detection thresholds for the DCNN, and ultimately do not reflect the true performance of our dual classifier model.

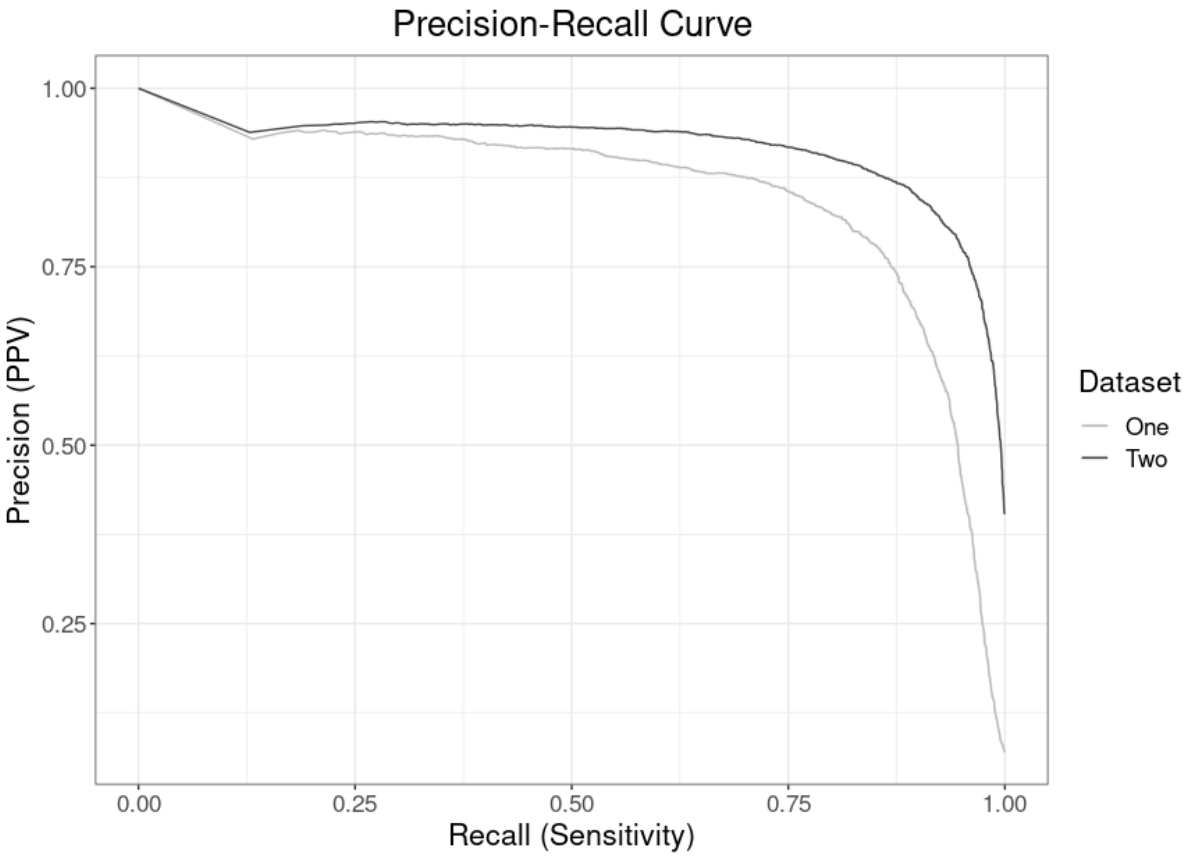

Figure 2 - The precision-recall curves for the trained DCNN on the held out test set of dataset one and the RR-interval labelled AF data in dataset two.

For a fair comparison with other classifiers, the performance of our dual classifier on the standardized datasets is reported in line with AAMI/ANSI EC57:2012 [13]. We present

episode and duration sensitivities and positive predictive values. The classification was completed on entire records without training periods, the minimum AF episode length was 30 seconds and atrial flutter in the reference annotations was excluded. The results are shown in Table 2 and Figure 3.

	Gross				Average			
Dataset	**Episode SE (%)**	**Episode PPV (%)**	**Duration SE (%)**	**Duration PPV (%)**	**Episode SE (%)**	**Episode PPV (%)**	**Duration SE (%)**	**Duration PPV (%)**
RR Classifier								
MITDB*	100	20	98	60	100	28	98	27
AFDB	97	69	97	95	98	71	96	79
LTAFDB	94	76	95	95	97	75	93	84
RR Classifier + Neural Network (change)								
MITDB*	98 (↓2)	59 (↑39)	92 (↓6)	83 (↑23)	98 (↓2)	60 (↑32)	93 (↓5)	57 (↑30)
AFDB	92 (↓5)	92 (↑23)	87 (↓10)	98 (↑3)	95 (↓3)	85 (↑14)	86 (↓10)	90 (↑11)
LTAFDB	89 (↓5)	87 (↑11)	93 (↓2)	97 (↑2)	93 (↓4)	87 (↑12)	88 (↓5)	91 (↑7)

Table 2 - Performance of the RR classifier and RR classifier with subsequent neural network analysis. Episode SE - Episode Sensitivity, Episode PPV - Episode Positive Predictive Value, Duration SE - Duration Sensitivity, Duration PPV - Duration Positive Predictive Value. (* paced records excluded).

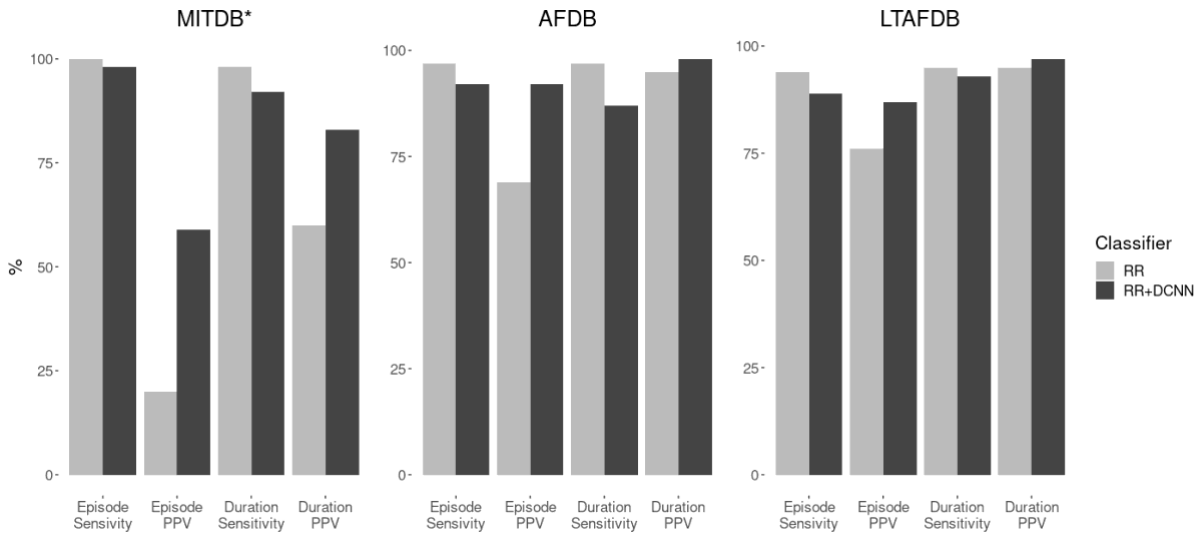

Figure 3 - The gross performance of the RR classifier and the RR classifier with subsequent DCNN analysis. (* paced records excluded)

The PPV of the one-stage classifier detected AF episodes was 20.5% (95% CI: 13.2% to 30.4%) for MITDB, 69.3% (95% CI: 65.7% to 72.6%) for AFDB and 76.4% (95% CI: 75.3% to 77.4%) for LTAFDB. After incorporation of the neural network in the two-stage classifier, the PPV of detected AF episodes increased significantly to 58.7% (95% CI: 44.3% to 71.7%, $p < 0.0001$) for MITDB, 91.6% (95% CI: 89.6% to 93.3%, $p < 0.0001$) for AFDB and 86.7% (95% CI: 85.9% to 87.5%, $p < 0.0001$) for LTAFDB.

With the 2-stage classifier there is a reduction of 2.4% (95% CI: 0.4% to 12.3%), 5.0% (95% CI: 2.8% to 8.7%) and 5.5% (95% CI: 4.6% to 6.4%) in true positive detection for MITDB, AFDB and LTAFDB respectively. However, the harmonic mean of sensitivity and positive predictive value (F1 score) is significantly increased across all datasets. For MITDB the F1 score increased from 0.33 (95% CI: 0.25 to 0.43) to 0.74 (95% CI: 0.62 to 0.83, $p < 0.0001$), for AFDB the F1 score increased from 0.81 (95% CI: 0.77 to 0.84) to 0.93 (95% CI: 0.90 to 0.96, $p < 0.0001$) and for LTAFDB the F1 score increased from 0.84 (95% CI: 0.83 to 0.85) to 0.88 (95% CI: 0.87 to 0.89, $p < 0.0001$). This shows that the two-stage classifier is

improving the classification performance and not simply trading reduced sensitivity for increased positive predictive value.

The number of analysis windows where the RR interval classifier was triggered and therefore a neural network analysis was performed was also recorded. This was used to calculate a percentage of time the neural network was active and is shown in Table 2.

	MITDB*	**AFDB**	**LTAFDB**
Total Analysis Windows	4,320	42,167	352,908
Neural Network Analysis Windows	787	17,249	183,479
Neural Network Active Time	18%	41%	52%

Table 2 - Analysis of windows that resulted in DCNN analysis. (* paced records excluded).

Discussion

The principal findings of this study are that: 1) Low complexity RR-interval classifiers can offer very high sensitivity for AF detection but the PPV of detected AF is low; 2) A small DCNN embedded on the device can successfully increase PPV by removing false positive events that are mislabelled by the low complexity classifier, with only a small loss of true-positive AF episodes; 3) Even in patients with large amounts of RR interval irregularity (caused by AF or other irregular rhythms), there is a significant computational saving by not analysing all ECG data with the DCNN.

In this study, we applied a simple Poincaré Plot based RR interval classifier using published threshold values. It is clear from our results that such classifiers highly generalizable and can offer high detection sensitivities (>93% across all datasets) for AF classification. As predicted, the classifier performs poorly when the PPV is measured. On AFDB and LTAFDB, approximately 1 in every 4 reported episodes is a false positive. The PPV on MITDB is particularly low due to the number of other irregular rhythms in the dataset, as well as a very small number of true AF episodes. This performance is to be expected, as not all the diagnostic criteria for AF are visible in the RR series, such as missing p-waves.

When our DCNN was applied to automatically review the positive cases from the RR interval classifier, PPV increased substantially across all datasets. The two-stage classifier increased the episode PPV by 39%, 23% and 11% on MITDB, AFDB and LTAFDB respectively. At the same time, the high episode sensitivity of the RR classifier was reduced by 5% or less. This increase in PPV is due to a decrease in the number of false positives, and therefore would reduce the workload for over-reading physicians. Figure 4 shows an example of a false positive detection by the RR classifier that is subsequently removed by the DCNN. It is important to note that the DCNN can never increase the detection sensitivity of the system. Therefore, a high sensitivity RR interval classifier is essential for optimal operation.

Figure 4 - A 20-second rhythm strip extracted from MITDB record 209 showing a short run of atrial tachycardia, followed by sinus tachycardia with premature atrial complexes. The RR classifier labels this as a false positive for AF. However, the DCNN correctly identifies it is not AF.

The two-step nature of this classifier allows for large computational savings compared with running the DCNN on every 20-second window. Analysis of the classification model shows that the DCNN was triggered on 18%, 41% and 52% of windows on MITDB, AFDB and LTAFDB respectively. It is clear that the number of windows analysed by the DCNN is dependent on the prevalence of irregular rhythms in the dataset, with LTAFDB having large periods of AF alongside other rhythms with large RR interval variability, such as atrial and ventricular bigeminy. Meanwhile, AFDB consists of atrial fibrillation, atrial flutter and normal sinus rhythm only. Despite large amounts of RR variability in these data sets, the low complexity classifier is able to confidently classify at least 48% of windows are not AF. This means that the large computations of the DCNN are not required, potentially reducing power usage and therefore increasing device battery life.

Atrial fibrillation is not the only arrhythmia commonly detected by MCOT devices. In future work, we plan to investigate applying this new concept of two-stage classifiers to a wider selection of arrhythmias.

Conclusion

This study has shown that DCNNs for ECG arrhythmia detection can be miniaturized to the extent that they could be deployed on MCOT devices, whilst maintaining a high level of performance. In fact, such networks can be paired with low power and low complexity RR interval classifiers for AF detection. This removes the need for the DCNN to analyse every ECG case, potentially increasing battery longevity. The two-stage classifier shows a large increase in PPV when compared to only the low complexity RR classifier, due to a decrease in the number of false positive events. This reduces the review burden for physicians and has been achieved with only a modest decrease in sensitivity.

References

[1] Benjamin, E.J., Wolf, P.A., D'Agostino, R.B., Silbershatz, H., Kannel, W.B. and Levy, D., 1998. Impact of atrial fibrillation on the risk of death: the Framingham Heart Study. *Circulation*, *98*(10), pp.946-952.

[2] Chugh, S.S., Havmoeller, R., Narayanan, K., Singh, D., Rienstra, M., Benjamin, E.J., Gillum, R.F., Kim, Y.H., McAnulty Jr, J.H., Zheng, Z.J. and Forouzanfar, M.H., 2014. Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study. *Circulation*, *129*(8), pp.837-847.

[3] Zweibel S., Trelfa M., The Use of Mobile Cardiac Telemetry to Improve Diagnostic Accuracy and Enable More Efficient Patient Care, *US Cardiology 2012;9(1)*, pp.43-66

[4] Willcox, M.E., Compton, S.J. and Bardy, G.H., 2021. Continuous ECG monitoring versus mobile telemetry: A comparison of arrhythmia diagnostics in human-versus algorithmic-dependent systems. Heart Rhythm O2, 2(6), pp.543-559.

[5] Mittal, S., Oliveros, S., Li, J., Barroyer, T., Henry, C. and Gardella, C., 2021. AI filter improves positive predictive value of atrial fibrillation detection by an implantable loop recorder. Clinical Electrophysiology, 7(8), pp.965-975.

[6] Somani, S., Russak, A.J., Richter, F., Zhao, S., Vaid, A., Chaudhry, F., De Freitas, J.K., Naik, N., Miotto, R., Nadkarni, G.N. and Narula, J., 2021. Deep learning and the electrocardiogram: review of the current state-of-the-art. *EP Europace*, *23*(8), pp.1179-1191.

[7] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P.C., Mark, R., Mietus, J.E., Moody, G.B., Peng, C.K. and Stanley, H.E., 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. E215–e220.

[8] Moody, G.B. and Mark, R.G., 2001. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, *20*(3), pp.45-50.

[9] Moody, G., 1983. A new method for detecting atrial fibrillation using RR intervals. *Computers in Cardiology*, pp.227-230.

[10] Petrutiu, S., Sahakian, A.V. and Swiryn, S., 2007. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. *Europace*, *9*(7), pp.466-470.

[11] Lake, D.E. and Moorman, J.R., 2011. 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), pp.H319-H325.

[12] Luo, C., Li, Q., Rao, H., Huang, X., Jiang, H. and Rao, N., 2021. An improved Poincaré plot-based method to detect atrial fibrillation from short single-lead ECG. *Biomedical Signal Processing and Control*, *64*, p.102264.

[13] AAMI/ANSI EC57:2012, Testing and reporting performance results of cardiac rhythm and st-segment measurement algorithms, American National Standard (2012).