Device agnostic AI-based analysis of ambulatory ECG recordings

Device Agnostic AI-based analysis of ambulatory ECG recordings.

Alan Kennedy[1], Peter Doggart[1], Stephen W. Smith[2], Dewar Finlay[3], Daniel Guldenring[4], Raymond Bond[3], Christopher McCausland[3], James McLaughlin[3]

[1] PulseAI Ltd, Belfast, United Kingdom,
[2] Hennepin County Medical Center, Minnesota, United States,
[3] Ulster University, Belfast, United Kingdom,
[4] HS Kempten, Germany,

Correspondence:

alan.kennedy@pulseai.io
+447714640144

Abstract

Deep Convolutional Neural Networks (DCNNs) have been shown to provide improved performance over traditional heuristic algorithms for the detection of arrhythmias from ambulatory ECG recordings. However, these DCNNs have primarily been trained and tested on device-specific databases with standardized electrode positions and uniform sampling frequencies. This work explores the possibility of training a DCNN for Atrial Fibrillation (AF) detection on a database of single-lead ECG rhythm strips extracted from resting 12-lead ECGs. We then test the performance of the DCNN on recordings from ambulatory ECG devices with different recording leads and sampling frequencies.

We developed an extensive proprietary resting 12-lead ECG dataset of 549,211 patients. This dataset was randomly split into a training set of 494,289 patients and a testing set of the remaining 54,922 patients. We trained a 34-layer convolutional DCNN to detect AF and other arrhythmias on this dataset. The DCNN was then validated on two Physionet databases commonly used to benchmark automated ECG algorithms (1) MIT-BIH Arrhythmia Database and (2) MIT-BIH Atrial Fibrillation Database. Validation was performed following the EC57 guidelines, with performance assessed by gross episode and duration sensitivity and positive predictive value (PPV). Finally, validation was also performed on a selection of rhythm strips from an ambulatory ECG patch that a committee of board-certified cardiologists annotated.

On MIT-BIH, The DCNN achieved a sensitivity of 100% and 84% PPV in detecting episodes of AF. and 100% sensitivity and 94% PPV in quantifying AF episode duration. On AFDB, The DCNN achieved a sensitivity of 94% and PPV of 98% in detecting episodes of AF, and 98% sensitivity and 100% PPV in quantifying AF episode duration. On the patch database, the DCNN demonstrated performance that was closely comparable to that of a cardiologist.

The results indicate that DCNN models can learn features that generalize between resting 12-lead and ambulatory ECG recordings, allowing DCNNs to be device agnostic for detecting arrhythmias from single-lead ECG recordings and enabling a range of clinical applications.

1. Introduction

Automated ECG interpretation dates back to the 1960s [1]. Correct automated ECG interpretations have been shown to increase diagnostic accuracy and reduce the time spent interpreting ECGs by clinical staff [2], particularly in the case of long-term ambulatory ECG recordings [3]. However, automated ECG interpretations are still frequently incorrect, with incorrect interpretation having been shown to influence patient management negatively [4]. Detecting Atrial Fibrillation (AF) has gained significant interest in automated ECG interpretation due to the association between AF and the increased risk of ischemic stroke.

Recently there has been a rapid increase in consumer and clinical arrhythmia monitoring devices. Devices such as wearable patches, implantable monitors, and smartwatches provide automated AF detection. Unfortunately, traditional AF detection algorithms still demonstrate significant false positive and false negative rates [5]. These traditional algorithms depend largely on lead-specific features to detect, for example, P-waves and other relevant ECG characteristics [6]. Feature selection in this manner may not always be transferable to other ambulatory ECG devices, which record a different ECG lead, thus limiting the application of the algorithms to specific devices. Within Artificial intelligence, DCNNs have emerged and been applied to various clinical decision support use cases to help physicians make more accurate and faster decisions [7]. DCNNs are a class of algorithms capable of learning directly from large datasets without hand-crafted feature selection. DCNNs have improved the computerized interpretation of ECG recordings from resting 12-lead ECG [8] and ambulatory ECG devices [9].

However, the performance of DCNNs on ambulatory ECG recordings, taken from devices they were not explicitly trained on, is unknown. In this study, we sought to evaluate the performance of a DCNN (PulseAI, Belfast, United Kingdom), developed on rhythm strips taken from resting 12-lead ECG recordings, at detecting AF from single-lead ambulatory ECGs recorded using traditional holter monitors and a wearable ECG patch.

2. Methods

2.1. Deep Neural Network Development

The DCNN was trained using a database of 549,211 resting 12-lead ECGs from a private anonymized dataset, which physicians had previously annotated.

The DCNN takes ECG voltage values as an input time series and produces a sequence of classification results. The DCNN architecture is based on residual blocks and is similar to the architecture described by Hannun et al. [9]. The DCNN takes as an input the raw ECG data (sampled at 256 Hz, or 256 samples per second) in microvolts and outputs one prediction of the ECG rhythm every second. The DCNN has thirty-four layers, consisting of sixteen residual blocks with two convolutional layers per block. Every other residual block performs downsampling via max pooling. To help with regularization, we applied batch normalization, rectified linear activation and dropout. The final fully connected softmax layer produces a probability of each ECG rhythm which is then thresholded using a cutoff of >0.5 to determine the presence or absence of AF. The DCNN was trained de novo with random initialization of the weights described by He et al. [10]. We used the Adam optimizer and a mini-batch size of thirty-two. We initialized the learning rate (0.001) and reduced it by a factor of ten when the testing set loss stopped improving for two consecutive epochs. During DCNN training, the weights are altered iteratively to reduce differences between the DCNN's output and the reference targets. This study trained the DCNN on a randomly selected single-lead from the 12-lead signal for each training mini-batch to maximize the DCNN's' exposure and generalisability to different waveform morphologies and amplitudes. This process was repeated iteratively for all ECGs in the training set until the model had fully converged and the model with the lowest loss on the test set was chosen.

3. Validation Databases

The DCNN's performance on standard holter monitors was validated using the MIT-BIH Arrhythmia Database (MIT-BIH) [11] and the MIT-BIH Atrial Fibrillation Database (AFDB) [12]. MIT-BIH consisted altogether of 24 hours of ECG from 47 patients, and AFDB consisted of 234 hours of ambulatory ECG from 23 patients. The DCNN's performance on a patch-based monitor was evaluated on the publicly available

database taken from [10], which consisted of 328 30-second ECG recordings from a wearable patch monitor. Six board-certified cardiologists annotated each recording against a committee consensus of independent cardiologists, which was used as the reference annotation. The mean inter-annotator agreement on the patch validation set was 72.8%. All ambulatory ECGs; were resampled to 256Hz using linear interpolation. In the training database, 39668 patients had AF or Atrial Flutter. The AF labels for model training contained both AF and atrial flutter. 2903 Atrial Tachycardias (Ectopic Atrial Rhythms) were not included in the AF class.

Table 1. The prevalence of AF patients in each of the validation databases.

Database	**MIT-BIH Arrhythmia**	**MIT-BIH AFDB**	**Patch Database**
Patients	47	23	328
% AF Patients	17	100	18
Recording Duration	30 minutes	10 hours	30 seconds

4. Statistical Analysis

In this work, we have used two approaches to the measurement of performance on two different forms of databases: (1) continuous Holter ECG (30mins-10 hours) recordings and (2) short-term (30 seconds) rhythm strips. For Holter recordings, we assessed episode sensitivity, duration sensitivity and positive predictive value in line with the EC57 standard [13]. However, it is not possible to use episode and duration statistics on 30-second rhythm strips due to the short duration. Therefore, we decided to assess performance based on the more standard measurement of sensitivity, PPV and PR-AUC, similar to [14].

4.1. Holter Recordings

The holter recordings used in this study ranged from 30 min to 10 hours in duration. The DCNN performance was validated on those recordings following the EC57 guidelines [13]. EC57 is the FDA-recognized standard and provides instructions for determining AF detection sensitivity and positive predictive value (PPV) in the context of wearable monitors. Statistics were calculated with a minimum AF duration of 30 seconds.

$$Episode\ Sensitivity\ =\ \frac{True\ Positive\ Episodes}{True\ Positive\ Episodes/False\ Negative\ Episodes}$$

$$Episode\ PPV\ =\ \frac{True\ Positive\ Episodes}{True\ Positive\ Episodes/\ False\ Positive\ Episodes}$$

$$Duration\ Sensitivity\ =\ \frac{Duration\ of\ Overlap}{Duration\ of\ Reference\ -\ Annotated\ AF}$$

$$Duration\ PPV\ =\ \frac{Duration\ of\ Overlap}{Duration\ of\ Neural\ Network\ -\ Annotated\ AF}$$

5. Patch Recordings

The duration of the ECG patch recordings (30 seconds) did not allow for the calculation of meaningful EC57 statistics from the patch database. Therefore, to determine the performance of the DCNN on this dataset, we assessed the DCNN's performance on each ECG rhythm strip in terms of sensitivity, PPV, and F1 score for the detection of AF in comparison to the reference annotation. The AF class contained both AF and atrial flutter.

6. Results

6.1. Physionet Databases

On the MIT-BIH dataset, the DCNN achieved 100% sensitivity and 84% PPV in detecting AF episodes and 100% sensitivity and 94% PPV in quantifying the duration of the AF episodes. Similar sensitivity was discovered on the MIT-BIH AFDB database with a sensitivity of 94% and PPV of 98% for AF episode detection and 98% sensitivity and 100% PPV in quantifying episode duration.

6.2. Patch Database

On the patch database, the DCNN achieved an area under the PR curve of 0.85 and had a performance comparable to that of an individual cardiologist compared to committee consensus (reference annotation).

7. Discussion

Accurate automated ECG interpretation for AF is important, as approximately 25–30% of ischemic strokes are associated with AF [15]. However, appropriate intervention with anticoagulation in AF patients is proven to prevent stroke. In the United States, 450,000 hospitalizations yearly are due to AF, which presents a significant cost and resource burden to healthcare systems. Almost 50 million people worldwide are affected by AF, and the incidence since 1990 has been increasing. Ambulatory ECG monitoring is the primary method for detecting AF episodes in the population. It is also more effective at detecting AF in patients who have suffered a cryptogenic stroke than conventional follow-up [16].

Data collection and annotation at scale can be costly and time-consuming for ECG manufacturers, particularly at the scale required for training DCNNs. Using more readily available 12-lead ECG data to train DCNNs may allow for automated ECG analysis methods which rely on deep learning to be developed much more rapidly, providing overall cost savings and reducing time to market.

In this study, we observed that a DCNN trained on rhythm strips from 12-lead resting ECG recordings could detect AF from single-lead ambulatory ECG recordings with high levels of episode sensitivity and PPV as well as duration sensitivity and PPV. Previous studies have demonstrated the performance of DCNNs in detecting arrhythmias [9, 17], with some showing improved performance through pretraining of the neural network weights and fine-tuning on smaller databases [18, 19]. However, the current study is the first to evaluate the performance of a DCNN to the EC57 standard on a range of ambulatory ECG devices for which the training data did not come from ambulatory ECG devices but instead resting 12-lead ECG monitors. No fine-tuning of the network was performed on ambulatory ECG sources was required. Evaluation of the DCNN on the Physionet databases showed comparable or improved performance with other industry-leading algorithms [20, 21], and evaluation of the DCNN on data from ECG patches demonstrated comparable performance for AF detection between the average cardiologist (70%, F1-score) and the DCNN (73%, F1-score).

8. Conclusion

A DCNN trained on single-lead ECGs extracted from resting 12-lead ECGs can be used to detect AF from a range of ambulatory monitoring devices. Results indicate that the developed DCNN can generalize between resting 12-lead and long-term ambulatory ECGs. These findings suggest DCNNs can be device agnostic for detecting AF from single-lead ECG recordings, enabling a range of clinical and consumer-focused applications.

References

1. Macfarlane PW, Van Oosterom A, Pahlm O, Kligfield P, Janse M, Camm J, editors. Comprehensive electrocardiology. Springer Science & Business Media; 2010 Nov 5.
2. Smulyan H. The computerized ECG: friend and foe. The American journal of medicine. 2019 Feb 1;132(2):153-60.
3. Fiorina L, Marijon E, Maupain C, Coquard C, Larnier L, Rischard J, Bourmaud A, Salerno F, Horvilleur J, Lacotte J, Ait Said M. 222 AI-based strategy enables faster Holter ECG analysis with equivalent clinical accuracy compared to a classical strategy. EP Europace. 2020 Jun 1;22(Supplement_1):euaa 162-374.
4. Smith SW, Rapin J, Li J, Fleureau Y, Fennell W, Walsh BM, Rosier A, Fiorina L, Gardella C. A deep neural network for 12-lead electrocardiogram interpretation outperforms a conventional algorithm, and its physician overread in the diagnosis of atrial fibrillation. IJC Heart & Vasculature. 2019 Dec 1;25:100423.
5. Schläpfer J, Wellens HJ. Computer-interpreted electrocardiograms: benefits and limitations. Journal of the American College of Cardiology. 2017 Aug 29;70(9):1183-92.
6. Laguna P, Jané R, Caminal P. Automatic detection of wave boundaries in multilead ECG signals: Validation with the CSE database. Computers and biomedical research. 1994 Feb 1;27(1):45-60.
7. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nature Reviews Cardiology. 2021 Jul;18(7):465-78.
8. Smith SW, Walsh B, Grauer K, Wang K, Rapin J, Li J, Fennell W, Taboulet P. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. Journal of electrocardiology. 2019 Jan 1;52:88-95.
9. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine. 2019 Jan;25(1):65-9.

10. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 1026-1034).
11. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Engineering in Medicine and Biology Magazine. 2001 May;20(3):45-50.
12. Moody G. A new method for detecting atrial fibrillation using RR intervals. Computers in Cardiology. 1983:227-30.
13. ANSI/AAMI (American National Standards Institute/Association for the Advancement of Medical Instrumentation) EC57. Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms, AAMI (2012)
14. Mittal S, Oliveros S, Li J, Barroyer T, Henry C, Gardella C. AI filter improves positive predictive value of atrial fibrillation detection by an implantable loop recorder. Clinical Electrophysiology. 2021 Aug 1;7(8):965-75.
15. Marini C, De Santis F, Sacco S, Russo T, Olivieri L, Totaro R, Carolei A. Contribution of atrial fibrillation to incidence and outcome of ischemic stroke: results from a population-based study. Stroke. 2005 Jun 1;36(6):1115-9.
16. Sanna T, Diener HC, Passman RS, Di Lazzaro V, Bernstein RA, Morillo CA, Rymer MM, Thijs V, Rogers T, Beckers F, Lindborg K. Cryptogenic stroke and underlying atrial fibrillation. New England Journal of Medicine. 2014 Jun 26;370(26):2478-86.
17. Teplitzky BA, McRoberts M, Ghanbari H. Deep learning for comprehensive ECG annotation. Heart Rhythm. 2020 May 1;17(5):881-8.
18. Weimann, Kuba, and Tim OF Conrad. "Transfer learning for ECG classification." Scientific reports 11.1 (2021): 1-12.
19. Salem, Milad, Shayan Taheri, and Jiann-Shiun Yuan. "ECG arrhythmia classification using transfer learning from 2-dimensional deep CNN features." 2018 IEEE biomedical circuits and systems conference (BioCAS). IEEE, 2018.
20. Babaeizadeh S, Gregg RE, Helfenbein ED, Lindauer JM, Zhou SH. Improvements in atrial fibrillation detection for real-time monitoring. Journal of electrocardiology. 2009 Nov 1;42(6):522-6.
21. Larburu N, Lopetegi T, Romero I. Comparative study of algorithms for atrial fibrillation detection. In 2011 Computing in Cardiology 2011 Sep 18 (pp. 265-268). IEEE.

Figure Captions

Figure 1. The sequence-to-sequence neural network architecture.

Figure 2 Performance Statistics of the network in detecting episodes of AF and quantifying the duration of those episodes. MIT-BIH database has a higher number of episode false positives detection, this may be down to the number of very short AF episodes (<30 seconds) that the network may have detected.

Figure 3. The precision-recall curve of the network in detecting AF on patch-based ECG recordings compared to six board-certified cardiologists. The model achieved an area under the PR curve of 0.85 and when using a detection threshold of >0.5 achieved a comparable F1 score (0.73, F1-score) to the average F1 by the cardiologists (0.70, F1-Score).