**Deep Learning to Automatically Interpret Images of the Electrocardiogram: Do We Need the Raw Samples?**

Rob Brisk[a,b,*], MBBCh, Raymond Bond[a], PhD, Elizabeth Banks[b], MBBCh, Alicja Piadlo[b], MBBCh, Dewar Finlay[c], PhD, James McLaughlin[c], PhD, David McEneaney[b], MD

[a] School of Computer Science, Ulster University, Jordanstown, Northern Ireland

[b] Dept of Cardiology, Craigavon Area Hospital, Craigavon, Northern Ireland

[c] Nanotechnology and Integrated Bioengineering Centre, Ulster University, Jordanstown, Northern Ireland

* Correspondence to:

robbrisk@hotmail.com

Dept of Cardiology,
Craigavon Area Hospital,
68 Lurgan Rd,
Portadown,
Craigavon
BT63 5QQ

+447850103054

## Introduction

Rule-based, computerised electrocardiogram (ECG) interpretation has been employed as an important diagnostic aid for over half a century.[1] Despite this, there is significant room for improvement in such systems, particularly with regards to arrhythmia detection and classification.[2-4] Over the last five years, a type of machine learning algorithm known as a deep neural network (DNN) has facilitated significant advances in the field of algorithmic data processing.[5] Within the last two years, these advances have been translated into the field of ECG signal processing and a number of so-called "deep learning" (DL)-based ECG classification algorithms have produced promising results. [6-9] It is perhaps too early to predict the extent to which DNNs will transform the practice of automated ECG analysis, but they have undoubtedly been highly disruptive in other domains such as speech recognition, computer vision and autonomous driving. [10-12] We may, as researchers from Stanford claim in their seminal work on this subject as published earlier this year, be on the cusp of truly "cardiologist-level" ECG read-outs. [6]

To date, the vast majority of research into DL-based ECG interpretation has focussed upon raw signals recorded directly from the ECG hardware. Yet, there is an enormous body of historical ECG data worldwide that exists only in paper form, or as scanned images thereof.[13] These ECGs are often associated with medical records containing years of rich clinical information: echocardiograms, angiographic findings, cardiac biomarkers, morbidity and mortality endpoints, and so on. It has long been acknowledged that such data could provide a rich source of insights to inform the science of ECG interpretation. Furthermore, the printed ECG is the universal format. Accurate, computerised analysis thereof would overcome the difficulties arising from proprietary formats and algorithms, long cited by researchers in the field as a substantial hindrance.[14]

There have, of course, been significant efforts towards converting ECG images to digital signals. These are summarized by Waits and Soliman (2017) excellent review in this journal.[15] However, regarding the current state of image-based ECG analysis, they conclude that "*certain limitations have been identified and overcome while others remain elusive*". A significant issue, noted both in the aforementioned review and by other authors, is a relatively decreased signal to noise ratio (SNR) compared with direct-from-hardware data.[15,16] Modern, sophisticated digitization methods have certainly made progress in this area, but validation of such techniques has been undertaken almost exclusively on 12-lead ECGs recorded in a controlled environment.[17] There has been little or no work exploring the digitization of ambulatory ECGs, where computerised analysis is already particularly challenging due to poorer SNRs caused by additional noise and movement artefact.[18] Furthermore, most studies have sought to validate digitization methods using metrics based on ECG intervals, amplitudes and areas, but few have examined the impact of raw signals vs image-derived signals on final diagnosis.

There is good reason to suppose that DL techniques may substantially increase the robustness of the image-based ECG interpretation pipeline and improve diagnostic quality: it has been established that DNNs, by virtue of certain regularization techniques such as "dropout" and data augmentation, can be particularly adept at handling low SNRs.[19,20] To test this hypothesis, we attempt to use DL to achieve accurate ECG interpretation of a particularly challenging dataset, consisting of images of ambulatory ECGs produced at half resolution.

## Methods

### Data acquisition

The 2017 Physionet AF Challenge (PAFC) was identified as an appropriate benchmark for our study, as the training data and results from several approaches (both rule-based and DL-based) were publicly available. The goal of the challenge was to classify each of 8528 single-lead ECG recordings into one of four rhythm categories: sinus rhythm, atrial fibrillation, other or noisy (see https://physionet.org/challenge/2017/ for competition rules and profile of training data).[21]

## Plotting ECGs to image files

To generate an image database for this study, all ECG signals were plotted as RGB image files using a standard Python library (MatPlotLib). Original signals were recorded at 300Hz on AliveCor devices, thus a 300 pixels / second resolution would have been required to maintain full resolution. In fact, a target resolution of 150 pixels / second and 75 pixels / mV was chosen, as this corresponds to an ECG printed at 25mm/s and 10mm/mV then scanned using a low-resolution, 150DPI scanner. Modern digital scanners are usually much higher resolution than this, but 150DPI scanners may still be found in developing health systems and it was felt to be an appropriate test of robustness of the computerised analysis pipeline. Figure 2 shows an example ECG image generated by this process.

## Digitization of image-based ECG signals

A number of approaches to digitising paper ECG signals for subsequent automated analysis have been explored over previous decades.[15] In order to better accommodate the characteristics of our ambulatory ECG dataset, we developed our own digitization method based upon established techniques. We hypothesised that the DNN used to interpret the signals generated by our digitization method would be more robust to noise than most rule-based approaches. We therefore omitted some noise-filtering techniques used by other authors (e.g. median filtering and interpolation, which Ravichandran et al (2013) applied to deal with the “salt-and-pepper” noise caused by thresholding).[16]

In summary, our approach consisted of scaling, thresholding, binarization and column-wise pixel searching. A thorough discussion of each of these techniques is provided by Waits and Soliman, therefore none are discussed in detail here.[15]

## DL model

Current state-of-the-art arrhythmia detection from ambulatory signals has been achieved using a 34-layer convolutional neural network (CNN) with residual connections between layers, developed by researchers at Stanford University.[6] We therefore selected this model architecture for our study.

In order to streamline the training process for the model, we were able to obtain pre-trained weights published by researchers at Oxford University, who had trained a model with the aforementioned architecture on the raw signals from the Physionet AF Challenge.[22] Their model was not among the highest competition scorers, but we expected to thoroughly retrain our model and this was simply a step to avoid randomly initialising the entire DNN, which would have substantially increased the computational and time requirements of this study. After some experimentation, we modified the model architecture slightly for handling image-derived data, with two fully connected layers each containing 512 nodes interposed between the final convolutional layer and the fully connected output layer (which contained four nodes, as this was a four-class problem). The weights of the additional fully connected layers of the model were randomly initialised.

## Training and analysis

Model performance was evaluated on the entire dataset prior to any training. This was necessary to ensure the pre-trained weights obtained from the Oxford team did not cause the model to over fit the data.

The model was then trained and evaluated using a five-fold cross validation (5FCV) process with 80% of the data used for training and 20% for validation during each 5FCV cycle. During training, the weights of the latter six layers of the network (two fully-connected layers and four convolutional layers) were progressively unfrozen. Each time a new layer was unfrozen, the model was trained until five epochs had passed without improvement in the validation accuracy.

5FCV was chosen because six of the top 10 scoring teams in the AF Challenge published results from 5FCV on the training set, so we were able to make a direct comparison with their models. It should be noted that the 5FCV results were published within papers written by each individual team; the results from the collective scoreboard were based on a hidden test set to which we did not have access. We therefore did not include any of the official competition results in our analysis.

As in the competition itself, the single performance metric used to undertake a like-for-like comparison between models was the combined F1 score, which is the harmonic mean of the F1 score for each of the four categories (see equation 1).

$$F1 = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

*Equation 1 – the F1 score*

## Results

The model was evaluated on the full image-based dataset upon initialisation with pre-trained weights. The results were in keeping with random chance, with a combined F1 score of approximately 0.5.

Following training, the mean combined F1 score and 95% confidence interval across the five cycles of this process was 0.78 (+/- 0.02). Readers can find the source code and reproduce the experiment from https://github.com/docbrisky/af-challenge. Figure 1 gives a visual report of the F1 score obtained for each of the four categories, plus error bars reflecting the 95% confidence interval across the 5FCV process.

Official scores from the 2017 AF Challenge were based on a hidden test set, to which we did not have access. However, six of the top 10 competitors published 5FCV scores obtained on the training set, which is the same data used to train and validate our model. The mean combined F1 score of those six teams was 0.83. (See https://physionet.org/challenge/2017/papers/ for a full list of publications.)

The model produced by the Oxford University team whose weights were used for initialisation of the convolutional layers of our model obtained a combined F1 score of 0.72 at 5FCV.

## Discussion

The results produced by this study suggest that DNN-based arrhythmia detection from ambulatory ECG images can be undertaken without substantial loss of accuracy compared with raw signal analysis. This is despite the fact that (i) ambulatory ECG data generally contains more noise and movement artefact than recordings in a controlled environment,[23] (ii) the ECG signals in this study were plotted into particularly low resolution images to simulate outdated hardware and (iii) several noise-filtering techniques were omitted from the digitization approach. We therefore posit that this represents a state-of-the-art result in terms of image-based ECG analysis.

A recent paper in the Lancet provides an apt context for the relevance of this finding. By undertaking a retrospective analysis of over 600,000 ECGs from nearly 200,000 patients, Attia et al (2019) used a DNN to predict incipient AF among patients currently in “normal” sinus rhythm with approximately 80% sensitivity and specificity.[24] In this case, the researchers were investigating a high-incidence endpoint (the development of AF) and were able to obtain sufficient digital ECG signals without needing to digitise historic ECG images. However, the obvious question arising from this study is whether patients deemed to be “at risk of future AF” based on an ECG in NSR have a correspondingly increased lifetime risk of stroke, and whether they should therefore be prescribed oral anticoagulation. Pending a prospective study to answer this question, which may take many decades, it is likely to be beneficial to apply Attia et al’s algorithm to historic ECGs that are already associated with a lifetime of follow-up data. Such ECGs will inevitably be images rather than digital signals, in which case the findings of our study would suggest that (i) signals generated by digitizing ECG images can be used to obtain reliable results from a DL model and (ii) weights obtained by training a DNN on raw signal data can be expected to transfer well to the task of analysing image-derived ECG data.

There are, however, important limitations to our study. Firstly, the ECG images were plotted directly from signal data, rather than being printed and scanned. They therefore contained minimal visual artefact and were unrotated (although CNNs are known to be translation invariant). It was the authors’ opinion that any additional artefact within printed and scanned ECGs compared with the direct-to-image ECGs would be easily overcome with established image processing techniques, and therefore that the printing and scanning of 8528 ECGs was unnecessary to produce meaningful results from this study. (Please see figure 2 for an example ECG image used in this study.) Nevertheless, to confirm that the results obtained herein will transfer to printed and scanned ECGs, further work in this area should be undertaken.

Secondly, the pretrained weights used to initialise the convolutional layers of the network had, presumably, been exposed to all of the ECG examples in the Physionet Challenge, albeit in raw signal form. Though three fully-connected layers were appended to the network and randomly initialised, and the performance of the newly-formed network was then confirmed to be approximately equal to a random-chance classifier, there is nonetheless a risk that the early convolutional layers of our network have overfit the data. This may explain why the results obtained from this experiment were substantially better than those obtained by the model whose weights were used for initialisation, though we propose that the improvement is down to a greater level of data augmentation and the two additional, fully-connected layers. The only way to evaluate this would be to re-train the network from randomly initialised weights, though any drop in performance of the randomly initialised model could also be ascribed to the stochastic nature of the training process.

Nonetheless, it is the authors' belief that the advent of DL-based ECG interpretation, and particularly its increased robustness to noise and resolution loss, should catalyse a renewed interest in high-quality, automated interpretation of image-based ECGs.

## References

[1] Pipberger HV, Freis ED, Taback L, Mason HL. Preparation of electrocardiographic data for analysis by digital electronic computer. Circulation. 1960;21:413-8.

[2] Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms: Benefits and Limitations. J Am Coll Cardiol. 2017;70(9):1183-1192.

[3] Shah AP, Rubin SA. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. J. Electrocardiol. 2007;40:385–390.

[4] Poon K, Okin P M, Kligfield P. Diagnostic performance of a computer-based ECG rhythm algorithm. J. Electrocardiol. 2005;38:235–238.

[5] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–444.

[6] Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65-69.

[7] Acharya UR, et al. A deep convolutional neural network model to classify heartbeats. Comput. Biol. Med. 2017;89:389–396.

[8] Shadi G, Mostafa A, Nasimalsadat M, Kamran K, Ali G. Atrial fibrillation detection using feature based algorithm and deep conventional neural network; Proceedings of the Conference on Computing in Cardiology (CinC); Rennes, France. 24–27 September 2017; Piscataway, NJ, USA: IEEE; 2017.

[9] Acharya UR, et al. Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. Inf. Sci. 2017;405:81–90.

[10] Amodei D, et al. Deep Speech 2: end-to-end Speech recognition in English and Mandarin. In Proc. 33rd International Conference on Machine Learning, 2016;173–182.

[11] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification; Proceedings of the International Conference on Computer Vision; Las Condes, Chile. 11-18 December 2015. IEEE; 2015:1026–1034.

[12] Liu S, Tang J, Zhang Z, Gaudiot JL. Computer Architectures for Autonomous Driving. IEEE Computer. 2017;50(8):18-25.

[13] Holkeri A, Eranti A, Kenttä TV, et al. Experiences in digitizing and digitally measuring a paper-based ECG archive. J Electrocardiol. 2018;51(1):74-81.

[14] Kligfield P. Overview of the ISCE ECG "genome project". J Electrocardiol. 2003;36 Suppl:163-5.

[15] Waits GS, Soliman EZ. Digitizing paper electrocardiograms: Status and challenges. J Electrocardiol. 2017;50(1):123-130.

[16] Ravichandran L, Harless C, Shah AJ, Wick CA, Mcclellan JH, Tridandapani S. Novel Tool for Complete Digitization of Paper Electrocardiography Data. IEEE J Transl Eng Health Med. 2013;1

[17] Holkeri A, Eranti A, Kenttä TV, et al. Experiences in digitizing and digitally measuring a paper-based ECG archive. J Electrocardiol. 2018;51(1):74-81.

[18] Elgendi M, Mohamed A, Ward R. Efficient ECG Compression and QRS Detection for E-Health Applications. Sci Rep. 2017;7(1):459.

[19] Srivastava, Nitish, et al. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 2014;15(1):1929-1958.

[20] Borodinov N, Neumayer S, Kalinin SV, Ovchinnikova OS, Vasudevan RK, Jesse S. Deep neural networks for understanding noisy data applied to physical property extraction in scanning probe microscopy. NPJ Computational Materials. 2019;5(25)

[21] Clifford GD, Liu C, Moody B, et al. AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge 2017. Comput Cardiol. 2017;44

[22] Andreotti F, Carr O, Pimentel MAF, Mahdi A, De Vos M. Comparing Feature-Based Classifiers and Convolutional Neural Networks to Detect Arrhythmia from Short Segments of ECG. IEEE Proceedings of the Conference on Computing in Cardiology (CinC); Rennes, France. 2017

[23] Chae DH, Alem YF, Durrani S, Kennedy RA. Performance study of compressive sampling for ECG signal compression in noisy and varying sparsity acquisition. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; Vancouver, BC. 2013;1306-1309

[24] Attia ZI, Noseworthy PA, Lopez-jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet. 2019;