Investigating the Robustness of Deep Learning to Electrocardiogram Noise

Jenny Venton[1], Philip J Aston[2,1]

[1] Department of Data Science, National Physical Laboratory, Teddington, UK
[2] Department of Mathematics, University of Surrey, Guildford, UK

Abstract

Deep learning models for electrocardiogram (ECG) classification can be affected by the presence of physiological noise on the ECG, as shown in previous work.

In this study, we explore the impact of different physiological noise types, and differing signal-to-noise ratios (SNRs) of noise on classification performance.

We find that classification performance is impacted differently by different noise types. In addition, the best classification performance comes from using a network trained on clean ECGs to classify clean ECGs.

In conclusion, this study has revealed several questions regarding inclusion or exclusion of noise on the ECG for training and classification by deep learning models.

1. Introduction

Deep learning models are an increasingly successful and popular way to detect abnormalities on an electrocardiogram (ECG) that are indicative of a heart condition. There are vast numbers of ECG classification deep learning models described in the literature [1], and studies that look to improve the trustworthiness, reproducibility, generalisability and robustness of these models are becoming more common. In a machine learning context, robustness describes how well model performance is preserved when perturbed input data is passed to the model to be classified. For deep learning models that classify ECGs, this perturbation can be in the form of equipment noise, adversarial noise, or physiological noise, all of which are known to cause misclassification errors [2].

While studies that develop deep learning models for classifying ECGs sometimes consider the generalisability of the model, that is how well it performs on another dataset, robustness as defined in the previous paragraph is rarely addressed. Note that sometimes the term 'robustness' is used to mean 'generalisability' in the ECG deep learning model literature [3, 4]. As deep learning models are developed, trained, and released into real world situations, it is imperative that the end user (clinicians or otherwise) can rely on the output the model is giving. One study that gave an overview of the deep learning pipeline for ECG analysis noted the importance of removing noise and baseline wander prior to model training and testing, but did not comment on noise in the context of robustness [5]. Addressing the robustness of ECG deep learning models is one aspect of improving their reliability, as in a real world situation outside of a controlled research environment, ECG data collected from humans is often subject to varying levels of physiological and equipment noise.

When noisy or perturbed input data is passed to a trained ECG model, this can affect performance of the model. Existing studies have looked at ECG perturbations due to adversarial attack [6] and physiological noise [7]. Here, we explore the robustness of trained models to physiological noise further, looking at different levels of noise, and whether different types of physiological ECG noise (ambulatory, non-ambulatory) have a different impact.

2. Methods

Data used were 2,678 12-lead ECG signals taken from the first source of data made available for the PhysioNet/Computing in Cardiology Challenge 2020 [8]. All signals were 8 to 138 seconds long and were recorded at 500 Hz. This dataset is hereafter referred to as the raw dataset. For full details of the data used, see [7].

All ECGs in the raw dataset were filtered to remove as much existing physiological noise as possible, creating the clean dataset. Filtering was carried out using the ECGdeli toolbox for Matlab [9]. This filtering included: baseline wander removal, low pass filter (150 Hz), high pass filter (0.05 Hz), notch filter (49 Hz to 51 Hz) and isoline correction. Two types of noise were added to the clean dataset:

1. **Recorded Noise**. Prerecorded physiological ambulatory ECG noise, taken from the MIT-BIH noise stress test database [10].
2. **Filtered Noise**. ECG noise filtered from the current ECG dataset.

See Figure 1 for details of how the Recorded Noise and Filtered Noise datasets were generated. Both noise types were applied to the clean dataset, after rescaling to an SNR selected randomly from one of three ranges. The raw

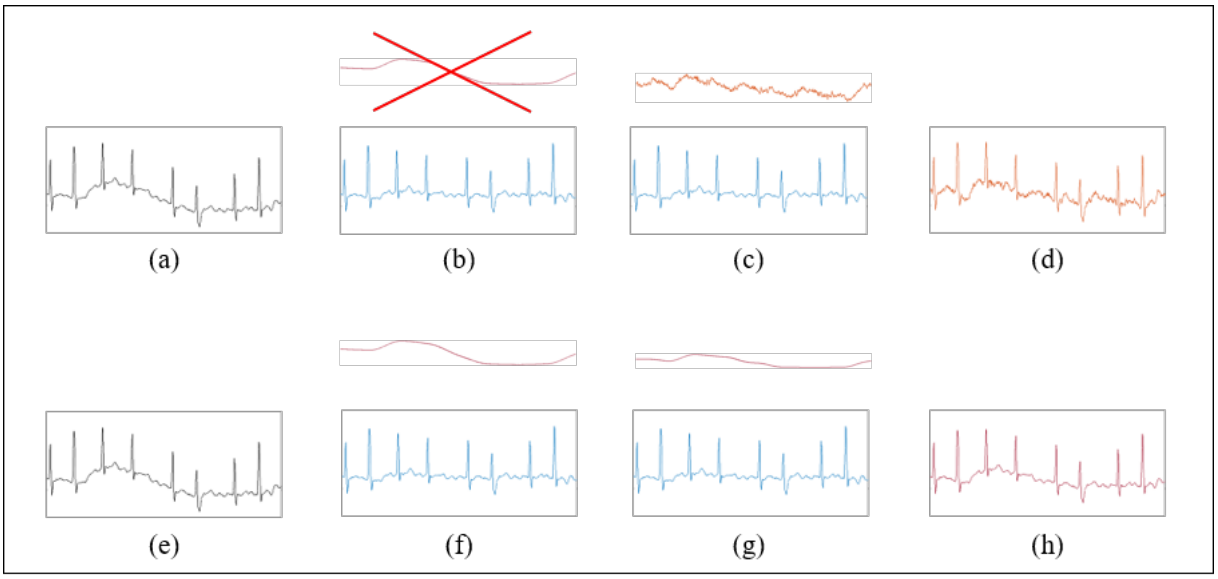

Figure 1. How the Recorded Noise (top) and Filtered Noise (bottom) datasets were generated. For the Recorded Noise dataset, (a) obtain raw data, (b) filter to obtain clean signal and noise, discard noise, (c) select segment of recorded noise and scale to specified SNR, (d) add recorded noise to the clean signal. For the Filtered Noise dataset, (e) obtain raw data, (f) filter to obtain clean signal and noise, (g) scale noise to specified SNR, (h) add scaled noise to the clean signal.

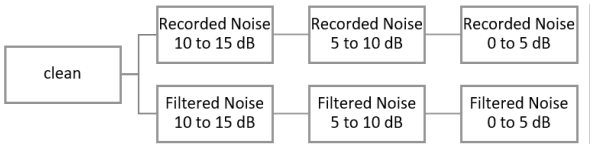

Figure 2. Details of the seven datasets generated, from cleanest to noisiest. All signal-to-noise ratios (SNRs) given in dB.

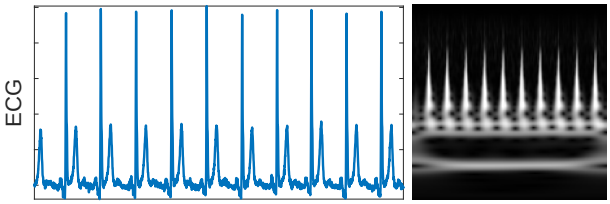

Figure 3. Example ECG signal and corresponding scalogram image.

dataset had a mean SNR of 18.5 dB and standard deviation of 6.1 dB, and SNR ranges were chosen to provide progressively noisier signals. From cleanest to noisiest these were: 10 to 15 dB, 5 to 10 dB and 0 to 5 dB, creating six noise datasets in total. See Figure 2 for details of all seven different datasets used.

Prior to classification, the continuous wavelet transform of each ECG in all seven datasets was generated, and the resulting coefficients plotted to obtain a scalogram image transform, see Figure 3. For each dataset, the scalogram images were split into training and validation images, and unseen test images. A ResNet-50 convolutional neural network (CNN) architecture pretrained using ImageNet images was adapted to classify the scalogram images using transfer learning. Classification performance of the networks was measured using the macro F1 score.

Seven CNNs were trained with 5-fold cross validation using each of the seven datasets. The trained networks were then used to classify the other datasets to gain understanding of how different levels of noise in the training data and input (test) data affect classification performance.

3. Results

The results for all seven networks were looked at, but only results for the following three networks are presented as these were representative: clean, 0 to 5 dB Recorded Noise and 0 to 5 dB Filtered Noise. When using the network trained on the noisiest Recorded Noise dataset to classify the various levels of Recorded Noise datasets, the performance was fairly consistent across the input datasets, with even a slight increase in performance in some cases, in comparison with a network trained on clean data used to classify the same datasets (Figure 4). Notably, the F1 score of the clean network dropped off considerably from classifying the cleanest to noisiest test dataset ($\sim$0.3).

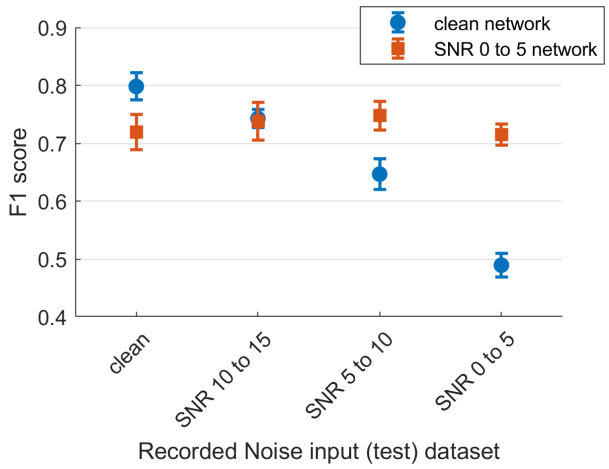

Figure 4. F1 scores for network trained using the 0 to 5 dB SNR Recorded Noise training data, and the network trained using the clean data, when classifying the Recorded Noise test datasets.

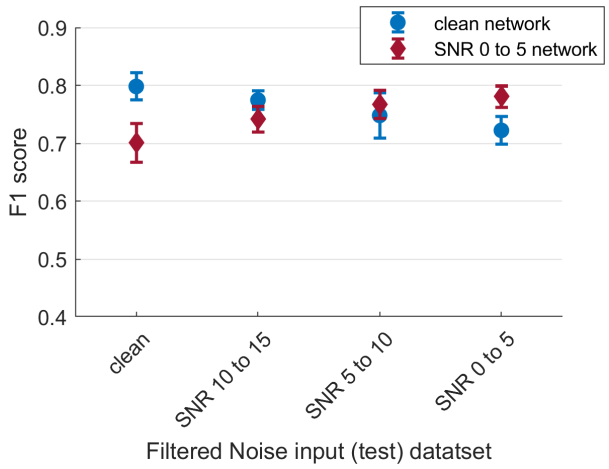

Figure 5. F1 scores for the network trained using the 0 to 5 dB SNR Filtered Noise training data, and the network trained using the clean training data, when classifying the Filtered Noise test datasets.

The Recorded Noise was recorded ambulatory noise. To see whether different noise types produced similar results, we created the Filtered Noise datasets. The results using the network trained on the noisiest Filtered Noise dataset and the network trained on clean data can be seen in Figure 5. In this case, the drop off for the clean network from classifying the cleanest to the noisiest data was much less than for the previous case ($\sim$0.1), but there was a definite decrease in performance of the network trained on the noisiest Filtered Noise dataset when classifying clean data.

4. Discussion

In this study, robustness of deep learning networks trained on different noisy datasets was investigated. This was carried out by evaluating the impact of noisy training data on network robustness and by evaluating the robustness of trained networks to noisy input data. It was found that the networks trained on the noisiest Recorded Noise and Filtered Noise datasets were most robust to the noisiest data, which is not surprising as this was the data they were trained on, but suffered performance decrease on clean input data. In the remaining text, Recorded Noise network and Filtered Noise network refer to the noisiest version of both.

The Recorded Noise network had a fairly consistent performance across all Recorded Noise input datasets, however performance of the clean network on the Recorded Noise input datasets dropped significantly with increasing noise levels, reaching an F1 score of $\sim$0.5 for the noisiest dataset. For the Filtered Noise input datasets, performance of the Filtered Noise and clean networks were both in the range 0.7 to 0.8 F1 score. The performance of the clean network dropped off as the noise level increased, but to a much lesser extent than with the Recorded Noise. The big difference in clean network performance between the Filtered Noise and Recorded Noise datasets was at least in part due to the fact that the Recorded Noise input data was recorded separately to the dataset and added, whereas the clean dataset may have had traces of the Filtered Noise remaining despite the filtering. This meant that there may have been residual Filtered Noise in the clean training data.

Overall, the best F1 score performance was obtained using a network trained on clean data to classify clean input data. This implies that when training networks for deployment in the real world, care must be take to ensure that all training data and input data has been cleaned. At present, it is unclear whether this result would still hold when using training data from multiple sources, with various methods of preprocessing inevitably present in each. For input (test) data, applying a standardised cleaning method to all future input data may pose a risk of obscuring useful features on an ECG, for example ST depression can be altered by filtering [11]. The concept of post-collection adversarial adjustments affecting classification has been addressed [6], but the possibility of post-collection standard filtering affecting classification should also be borne in mind.

As deep learning models for ECG classification progress, provable guarantees of performance are vital for safety in real world settings. In terms of robustness, we wanted to understand how performance of the networks was affected by noise on the input data, and also how this robustness could be influenced by the inclusion or exclusion of noise in the training data. We found that networks trained on very noisy data were robust to very noisy input data, as would be expected. When using these networks

with clean data performance held up for the Recorded Noise network but decreased for the Filtered Noise network. While there are no apparent similar studies that used different noise levels that we can compare the present results to, one study used varying SNR levels of the same Recorded Noise we used to evaluate the impact this has on heart rate variability (HRV) performance [12]. They found that performance deteriorated from SNR values around 5 dB and below.

Clean network performance was markedly different on the Recorded Noise and Filtered Noise input datasets. As already mentioned, this may have been in part due to some residual noise being left on the clean signal. Another contributing factor is that two different noise types were used: Recorded Noise was ambulatory ECG noise and Filtered Noise was non-ambulatory hospital recordings. Interestingly, when using the noisiest Filtered Noise network to classify the Recorded Noise datasets and vice versa, the performance pattern was similar to that of the clean network (not shown). This suggests that noise type, as well as noise level, is important for robustness. Further work to understand how components of the different noise types (such as distribution or frequency spectra) may affect the classification performance is underway.

Here we have examined how including or excluding different amounts of noise in the training data can impact robustness of the trained network to noise on the input data. Alternative methods to improve robustness to physiological noise could be to train a network to identify and disregard the noise, or methods to improve robustness more generally include encouraging networks to prefer coarser features of an ECG, mimicking how a clinician would assess an ECG [6]. Networks developed for the 2017 CinC challenge identified noisy signals [13], excluding these from classification as a healthy or pathology class. The dataset in the current study contained signals that were mostly noise, and in a real world setting it would be preferable for these signals to be excluded from classification.

In conclusion, the inclusion of noisy ECG data to train an ECG network provides robustness to noisy input data, however this may come at the cost of classification performance on clean input data. Furthermore, we have shown that to provide robustness to a specific noise type on input data, it is important to include similar noise types in training data.

Acknowledgments

This project 18HLT07 MedalCare has received funding from the EMPIR programme co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation programme.

References

[1] Hong S, Zhou Y, Shang J, Xiao C, Sun J. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. Comput Biol Med 2020;122.

[2] Luo S, Johnston P. A review of electrocardiogram filtering. J Electrocardiol 2010;43(6):486–496.

[3] Alfaras M, Soriano MC, Ortín S. A Fast Machine Learning Model for ECG-Based Heartbeat Classification and Arrhythmia Detection. Front Phys 2019;7(July):1–11.

[4] Inan OT, Giovangrandi L, Kovacs GT. Robust neural-network-based classification of premature ventricular contractions using wavelet transform and timing interval features. IEEE T Bio Med Eng 2006;53(12):2507–2515.

[5] Somani S, Russak AJ, Richter F, Zhao S, Vaid A, Chaudhry F, De Freitas JK, Naik N, Miotto R, Nadkarni GN, Narula J, Argulian E, Glicksberg BS. Deep learning and the electrocardiogram: review of the current state-of-the-art. Europace 2021;1179–1191.

[6] Han X, Hu Y, Foschini L, Chinitz L, Jankelson L, Ranganath R. Deep learning models for electrocardiograms are susceptible to adversarial attack. Nat Med 2020;26(3):360–363. ISSN 1546170X.

[7] Venton J, Harris PM, Sundar A, Smith NAS, Aston PJ. Robustness of convolutional neural networks to physiological ECG noise. Philos T R Soc A 2021;(Advanced Computation in Cardiovascular Physiology: New Challenges and Opportunities).

[8] Perez Alday EA, Gu A, Shah A, Liu C, Sharma A, Seyedi S, Bahrami Rad A, Reyna M, Clifford G. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol Meas 2021;41.

[9] Pilia N, Nagel C, Lenis G, Becker S, Dössel O, Loewe A. ECGdeli - An Open Source ECG Delineation Toolbox for MATLAB, 2020. URL `https://github.com/KIT-IBT/ECGdeli`.

[10] Moody GB, Muldrow WK, Mark RG. A noise stress test for arrhythmia detectors. In Computers in Cardiology, volume 11. 1984; 381–384.

[11] Lenis G, Pilia N, Loewe A, Schulze WH, Dössel O. Comparison of Baseline Wander Removal Techniques considering the Preservation of ST Changes in the Ischemic ECG: A Simulation Study. Comput Math Method M 2017;2017.

[12] Cavalieri RN, Filho PB. Determination of Maximum Noise Level in an ECG Channel Under SURE Wavelet Filtering for HRV Extraction. Rev Mex Ing Biomed 2020;41(2):66–72.

[13] Clifford GD, Liu C, Moody B, Lehman LH, Silva I, Li Q, Johnson AE, Mark RG. AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017. In Comput. Cardiol. 2017, volume 44. 2017; .

Address for correspondence:

Jenny Venton
Dept. of Data Science, National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK
jenny.venton@npl.co.uk