

# Benchmarking the Impact of Noise on Deep Learning-Based Classification of Atrial Fibrillation in 12-Lead ECG

Theresa BENDER<sup>a,b,1</sup>, Philip GEMKE<sup>a</sup>, Ennio IDROBO-AVILA<sup>a</sup>, Henning DATHE<sup>a</sup>,  
Dagmar KREFTING<sup>a,b</sup> and Nicolai SPICHER<sup>a,b</sup>

<sup>a</sup>*Department of Medical Informatics, University Medical Center Göttingen,  
Göttingen, Germany*

<sup>b</sup>*DZHK (German Centre for Cardiovascular Research), partner site Göttingen,  
Göttingen, Germany*

**Abstract.** Electrocardiography analysis is widely used in various clinical applications and Deep Learning models for classification tasks are currently in the focus of research. Due to their data-driven character, they bear the potential to handle signal noise efficiently, but its influence on the accuracy of these methods is still unclear. Therefore, we benchmark the influence of four types of noise on the accuracy of a Deep Learning-based method for atrial fibrillation detection in 12-lead electrocardiograms. We use a subset of a publicly available dataset (PTB-XL) and use the metadata provided by human experts regarding noise for assigning a signal quality to each electrocardiogram. Furthermore, we compute a quantitative signal-to-noise ratio for each electrocardiogram. We analyze the accuracy of the Deep Learning model with respect to both metrics and observe that the method can robustly identify atrial fibrillation, even in cases signals are labelled by human experts as being noisy on multiple leads. False positive and false negative rates are slightly worse for data being labelled as noisy. Interestingly, data annotated as showing baseline drift noise results in an accuracy very similar to data without. We conclude that the issue of processing noisy electrocardiography data can be addressed successfully by Deep Learning methods that might not need preprocessing as many conventional methods do.

**Keywords.** Deep Learning, Electrocardiogram, Atrial Fibrillation, Noise

## 1. Introduction

Electrocardiograms (ECGs) are recordings of the electrical activity of the heart and are frequently used in emergency and in-patient care. However, different types of noise, either stemming from the patient's behaviour (e.g. motion) or the devices (e.g. power line interference), can be introduced during measurement. The presence of noise leads to a twofold problem: It impedes detection of anomalies leading to false findings and alarms [1] and, if the signal-to-noise ratio (SNR) reaches a certain level, detecting diagnostically relevant features becomes impossible [2].

---

<sup>1</sup> Corresponding Author: Theresa Bender, University Medical Center Göttingen, Robert-Koch-Str. 40, 37075 Göttingen, Germany, E-mail: [theresa.bender@med.uni-goettingen.de](mailto:theresa.bender@med.uni-goettingen.de).

One class of features with high clinical importance are the so-called “fiducial points”, i.e. the center, on- and offsets of ECG waves such as the QRS complex and the P-/T-wave. They are used for segmenting heartbeats into meaningful intervals [3] and by doing so allow for arrhythmia detection. Atrial fibrillation (AF) is the most prevalent arrhythmia which is characterized by uncoordinated electrical impulses in the atrium and might lead to severe cardiovascular issues, such as stroke or heart failure. Analyzing the interval in a heartbeat where a P-wave is expected is crucial for AF classification as its absence indicates a lack of sinoatrial node activity and is thereby a sign for AF [4]. However, so-called fibrillatory waves might occur, mimicking P-waves, impeding the assessment of sinoatrial node activity.

Many state-of-the-art algorithms for ECG classification are based on extracting semantic features derived from human expert knowledge, such as fiducial points. However, as these algorithms tend to wrong results in case of noise [5], various denoising strategies [6] have been proposed. In contrast, algorithms from the field of deep learning (DL) were explored for ECG classification tasks recently [7,8]. Instead of semantic features, they are based on agnostic features derived from fully-automatic correlation analysis between input ECGs and output classes in an end-to-end fashion. These models are based on the underlying premise that training and test datasets are stemming from the same distribution, which is often their pitfall in case of dataset shifts (variant devices, users, noise). Although initial studies indicate a better robustness to high SNRs [9,10], it remains unclear to which extent it affects these models.

Thereby, in this work we benchmark the accuracy of a state-of-the-art pre-trained DL model for 12-lead ECG classification regarding its susceptibility to different types of noise. We use the publicly available PTB-XL dataset which contains annotations for several categories of noise made by human technical experts and compare the model’s accuracy w.r.t. type of noise.

## 2. Methods

We analyze a subset of the PTB-XL dataset containing 12-lead ECGs of 10 second length with a sampling rate of 500 Hz that were acquired between 1989 and 1996 [11]. The subset contains all 1,514 ECGs annotated as showing AF (label in PTB-XL: *AFIB*) and we add the first 2,000 normal ECGs (*NORM*) as healthy controls. For each signal, we use a qualitative and a quantitative method to estimate SNR.

### 2.1. SNR Based on Annotations ( $SNR_a$ )

For each ECG we determine the number of noisy leads using the columns *baseline\_drift*, *static\_noise*, *burst\_noise* and *electrodes\_problems* provided in the PTB-XL metadata. In the majority of cases, they contain the name of a single lead (e.g. “aVL”), multiple leads (“I,aVR”) or ranges (e.g. “I-III”). Using a custom script, we convert this information to numeric values ranging from 0 to 12 for each type of noise. The labels “alles” (all) and “noisy recording” are converted to 12. We remove ECGs associated with other labels as they are of a more qualitative nature (e.g. “leicht” (light)). In this way, for each signal a qualitative, unit-less, linear SNR measure is computed, ranging from 0 (no noise reported) to  $12 \cdot 4 = 48$  (all leads are affected by all types of noise). As shown in Tbl. 1, we use this information to split the dataset in ECGs without (“w/o”) a noise label and ECGs with (“w/”) a noise label.

**Table 1.** Properties of subset extracted from PTB-XL (left) and results of DL-based AF classification (right). ECGs are grouped according to annotations: In case there is one or more noise label in the metadata, an ECG is assigned to “w/”, else to “w/o”. FP and FN denote False Positive and False Negative, respectively.

| Noise Label | AF    | Healthy controls | Noise Label | DL: FP | DL: FN |
|-------------|-------|------------------|-------------|--------|--------|
| w/o         | 1,097 | 1,581            | w/o         | 0.04 % | 3.96 % |
| w/          | 417   | 419              | w/          | 0.24 % | 7.06 % |

It has to be underlined that a value of zero does not have to mean that there is no noise, it just reflects that there is a potential for a noise-free ECG. The authors of PTB-XL also indicated that missing annotations in case of artifacts or false annotations in case of noise-free signals might occur. However, they concluded that the metadata bears the potential for ECG quality assessment [12].

## 2.2. Measured SNR ( $SNR_m$ )

Due to the limitations of the manual annotations and as they are only available for 22 % of the PTB-XL database [12], we additionally use a quantitative SNR measure for each signal. We compute the Fourier Transform of the signals as well as the ratio of energies in two frequency bands as proposed in [13]. Based on the expected heart rates during AF, we define the “signal” frequency band ranging from 40 to 150 beats-per-minute (0.66 to 2.5 Hz) and define the “noise” frequency band as  $< 40$  and  $> 150$  beats-per-minute. By scaling with  $10 \log 10$ , we arrive at an SNR expressed in logarithmic decibel scale (dB).

## 2.3. DL Classification

ECG data is classified with a pre-trained model by Ribeiro et al. [7]. The model is a residual network and was trained on more than two million ECGs that were acquired within a Brazilian telehealth network. It outputs independent probabilities for six abnormalities, but we limit our analysis to AF. We use a threshold defined by the authors<sup>2</sup>.

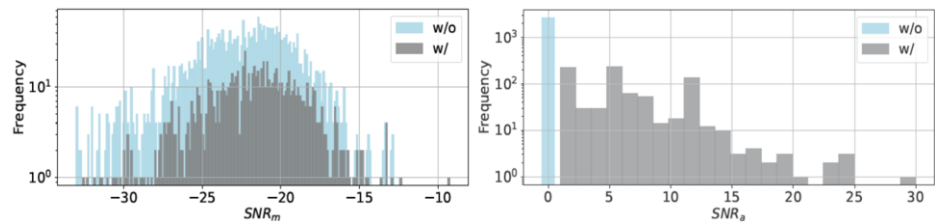
## 2.4. Data Analysis

We analyze the subset regarding differences between ECGs with and without noise labels for i) their distribution of  $SNR_m$  and  $SNR_a$  as well as ii) the accuracy of DL classification of each noise category. For ii) we compared the noisy recordings ( $SNR_a > 0$ ) with randomly drawn signals from equally sized control groups ( $SNR_a = 0$ ).

# 3. Results

Fig. 1 shows the distribution of  $SNR_a$  and  $SNR_m$  values on the left and right side. The majority of ECGs with noise labels has less than 15 with the maximum being 29. This shows that even in the duration of 10 seconds, different data quality issues per lead may occur.  $SNR_m$  values are occurring in the range of  $[-33.03, -7.78]$  dB with no clear difference between ECGs with and without noise labels.

<sup>2</sup> [https://github.com/antonior92/automatic-ecg-diagnosis/blob/master/generate\\_figures\\_and\\_tables.py](https://github.com/antonior92/automatic-ecg-diagnosis/blob/master/generate_figures_and_tables.py), commit 89f929d, line 121



**Figure 1.** Distribution of values of both SNR metrics with (grey) and without (blue) noise labels.

**Table 2.** DL accuracy w.r.t. the four types of noise. The variable  $n$  represents the number of signals with the given label (w/). For comparison to signals without a label (w/o),  $n$  ECGs are randomly drawn 100 times and accuracy is given as mean  $\pm$  standard deviation.

| Type<br>Label | Baseline Drift<br>( $n = 305$ ) | Static Noise<br>( $n = 478$ ) | Burst Noise<br>( $n = 156$ ) | Electrode<br>Problems ( $n = 6$ ) |
|---------------|---------------------------------|-------------------------------|------------------------------|-----------------------------------|
| w/o           | 96.8 % $\pm$ 0.9 %              | 96.8 % $\pm$ 0.7 %            | 96.9 % $\pm$ 1.3 %           | 96.3% $\pm$ 8.0 %                 |
| w/            | 97.7 %                          | 94.6 %                        | 94.9 %                       | 100.0 %                           |

Tbl. 1 (right) shows FP and FN rates of AF classification w.r.t. the existence of noise labels. FP is worsened by 0.2 % and FN by 3.1 % in case ECGs are annotated with noise labels. Tbl. 2 shows the DL accuracy for each type of noise compared to the same number of ECGs but randomly drawn 100 times from data without noise labels. ECGs with baseline drift or electrode problems are classified more accurately in comparison to random ECG signals without noise annotations, whereas ECGs with annotated burst and static noise reveal worse performance.

4. Discussion

In general, the DL model robustly classifies AF, even in case ECGs are labelled by human experts as having multiple leads influenced by noise. Interestingly, in presence of baseline drift or electrode problems, accuracy is not deteriorated, but within one standard deviation compared to signals without noise labels. As a limitation, it has to be underlined that annotations are non-complete [12] and the subset contains only six signals annotated with electrode problems.

As the DL model can be assumed as a “black box”, we can only speculate about the reasons for this behaviour. It could be explained by partial misinterpretation of baseline drift or static noise as P-waves. As we could show in previous work [14], the DL model was trained such that P-waves and R-peaks have a high relevance, similar to human perception, while numerous other features influence its decision. This multi-factor decision process could be robust to different kinds of noise, but this requires its presence during training. A shift between training and test datasets is always an issue for DL models. To mitigate this effect is has been suggested to intentionally include noise during training [9]. The model used in this work was trained on two million non-public ECGs.

However, since the distribution of  $SNR_m$  looks visually similar with or without noise labels,  $SNR_a$  might not be optimal for quality assessment on its own. A “no noise” label, explicitly identifying ECGs without data quality issues, and more labels in general would be a valuable addition for future experiments.

## 5. Conclusion

Results show that the DL model is able to detect AF in 12-lead ECGs with high accuracy, even in the presence of data quality issues according to human experts. We conclude that end-to-end DL models based on agnostic features can address the difficulty of processing noisy ECGs. In contrast to conventional methods based on semantic features, they might not require preprocessing methods for achieving high accuracy. However, more experiments with larger and more diverse datasets should be the subject of future work.

## References

- [1] Festag S, Spreckelsen C. Semantic Anomaly Detection in Medical Time Series. *Stud Health Technol Inform* 2021; 278:118–25, doi: 10.3233/SHTI210059.
- [2] Apandi ZFM, Ikeura R, Hayakawa S, Tsutsumi S. An Analysis of the Effects of Noisy Electrocardiogram Signal on Heartbeat Detection Performance. *Bioengineering* 2020; 7(2), doi: 10.3390/bioengineering7020053.
- [3] Spicher N, Kukuk M. Delineation of Electrocardiograms Using Multiscale Parameter Estimation. *IEEE J Biomed Health Inform* 2020; 24(8):2216–29, doi: 10.1109/JBHI.2019.2963786.
- [4] Kreimer F, Aweimer A, Pflaumbaum A, Mücke A, Gotzmann M. Impact of P-wave indices in prediction of atrial fibrillation-Insight from loop recorder analysis. *Ann Noninv Electrocard* 2021; 26(5):e12854, doi: 10.1111/anec.12854.
- [5] Kumar P, Sharma VK. Detection and classification of ECG noises using decomposition on mixed codebook for quality analysis. *Health Technol Lett* 2020; 7(1):18–24, doi: 10.1049/htl.2019.0096.
- [6] Mir HY, Singh O. ECG denoising and feature extraction techniques – a review. *Journal of Medical Engineering & Technology* 2021; 45(8):672–84, doi: 10.1080/03091902.2021.1955032.
- [7] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* 2020; 11(1):1760, doi: 10.1038/s41467-020-15432-4.
- [8] Attia ZI, Harmon DM, Behr ER, Friedman PA. Application of artificial intelligence to the electrocardiogram. *Eur Heart J* 2021; 42(46):4717–30, doi: 10.1093/eurheartj/ehab649.
- [9] Venton J, Harris PM, Sundar A, Smith NAS, Aston PJ. Robustness of convolutional neural networks to physiological electrocardiogram noise. *Philos Trans A Math Phys Eng Sci* 2021; 379(2212):20200262, doi: 10.1098/rsta.2020.0262.
- [10] Sraith M, Jabrane Y, Hajjam El Hassani A. A Robustness Evaluation of Machine Learning Algorithms for ECG Myocardial Infarction Detection. *JCM* 2022; 11(17):4935, doi: 10.3390/jcm11174935.
- [11] Wagner P, Strodthoff N, Boussejot R-D, Kreiseler D, Lunze FI, Samek W et al. PTB-XL, a large publicly available electrocardiography dataset. *Sci Data* 2020; 7(1):154, doi: 10.1038/s41597-020-0495-6.
- [12] Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL; 2020 Apr 28, doi: 10.48550/arXiv.2004.13701.
- [13] Haan G de, Jeanne V. Robust pulse rate from chrominance-based rPPG. *IEEE Trans Biomed Eng* 2013; 60(10):2878–86, doi: 10.1109/TBME.2013.2266196.
- [14] Bender T, Beinecke JM, Krefting D, Müller C, Dathe H, Seidler T et al. Analysis of a Deep Learning Model for 12-Lead ECG Classification Reveals Learned Features Similar to Diagnostic Criteria; 2022, doi: 10.48550/arXiv.2211.01738.