The Negative Impact of Denoising on Automated Classification of Electrocardiograms

Federica Granese*
UMMISCO
IRD, Sorbonne Université
Bondy, France
federica.granese@ird.fr

Ahmad Fall*
UMMISCO
IRD, Sorbonne Université
Bondy, France
Université Cheikh Anta Diop
Dakar, Sénégal
ahmad.fall@ird.fr

Alex Lence
UMMISCO
IRD, Sorbonne Université
Bondy, France
alex.lence@ird.fr

Joe-Elie Salem
Vanderbilt University Medical Center, Nashville, TN, USA
Clinical Investigation Center Paris-Est, INSERM, UNICO-GRECO Cardio-Oncology Program
Pitié-Salpêtrière University Hospital,Sorbonne Université, Paris, France
joe-elie.salem@aphp.fr

Jean-Daniel Zucker
UMMISCO
IRD, Sorbonne Université, Bondy, France
INSERIM, NutriOmique, AP-HP
Hôpital Pitié-Salpêtrière
Paris, France
jean-daniel.zucker@ird.fr

Edi Prifti
UMMISCO
IRD, Sorbonne Université, Bondy, France
INSERIM, NutriOmique, AP-HP
Hôpital Pitié-Salpêtrière
Paris, France
edi.prifti@ird.fr

Abstract

We present an evaluation of recent state-of-the-art electrocardiogram denoising methods and assess their impact on the performance of convolutional deep learning-based classifiers, with a focus on the risk prediction of Torsade-de-Pointes arrhythmia. Our findings indicate that the traditional approach of evaluating denoising methods independently of the application is insufficient. This is particularly the case for applications where the signals are used for phenotype prediction. We observed that when classifiers are fed denoised data instead of raw data, their performance significantly deteriorates, with a decline of up to 40 percentage points in accuracy and up to 27 percentage points in AUROC when a misclassification detection method is further applied, underscoring a notable reduction in model reliability. These findings highlight the importance of considering the downstream impact of denoising on automated classification tasks and shed light on the complexities of trustworthiness in the context of healthcare applications.

1 Introduction

An electrocardiogram (ECG) is a century-old test used to assess cardiac health. It involves placing two or more electrodes at specific locations on the chest, arms, and legs and recording the

*Equal contribution

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

heart's electrical signals through a central unit [1]. The ECG waveform (as in Figure 1) bears important information and subtle variations can be indicative of a large spectrum of diseases.

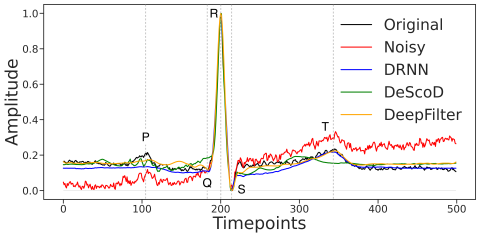


Figure 1: Example of a (heartbeat) recorded by an ECG along with waves annotation and denoised outputs.

However, recording such signals can be often susceptible to noise from various sources, including electrode loose contact, patient movement, and muscle contractions [2, 3]. This can alter the waveform resulting in challenging analyses and subsequent diagnoses. Broadly speaking, when we refer to *denoising of a signal*, we are alluding to the process of recovering the raw signal from the noisy one. Denoising ECG signals have been extensively studied in the literature and numerous methods have been proposed. These methods stemmed from different fields, including signal processing [4] as well as more recently deep learning [2, 5, 6, 7]. The evaluation of the proposed methods typically treats denoising as an independent task, while comparing the dissimilarity between the 'clean' signal and its noisy counterpart.

In this paper, we argue that assessing the denoising performance independently of the final downstream tasks is insufficient and can have important unforeseen consequences. Indeed, signals that may appear visually similar to the human eye can be fundamentally distinct when viewed from a neural network's perspective [8]. Specifically, in this work, we connect denoising with the prediction of the risk of a life-threatening type of arrhythmia known as Torsades-de-pointes (TdP) for which Prifti et al. [9] previously developed a high-performing deep convolutional neural network model. This model was trained on ECGs recorded from healthy individuals before and after the intake of Sotalol, a drug known to increase the risk of TdP.

Starting from the intuition that an effective denoiser should retain as much valuable information as possible while reducing noise in the data, we analyzed five recent state-of-the-art (SOTA) denoisers on a real-life ECG dataset `Generepol` [10]. We show that while providing good results in the denoising task, the application of these methods impact negatively the performance of the TdP risk classification model. Indeed, **the results reveal a significant decrease in the classifier's accuracy when the denoised ECG is classified as compared to the original ECG**. Additionally, **an examination of the distributions of correctly and incorrectly classified samples reveals a decrease in the model's confidence when classifying denoised data**, particularly concerning correctly classified samples. We performed the evaluation on a variety of denoisers, which include a *diffusion model* [6], a *deep recurrent neural network* (DRNN) [2], an *autoencoder* [7], and a *convolutional neural network* (CNN) [5]. Additionally, we considered a signal processing technique i.e., *wavelet transform* [4].

2 Problem formalization

The ECG denoising problem can be seen as a specific instance of the total-variation (TV) denoising problem [11] in the context of one-dimensional signals. In particular, we are given a *noisy signal* $\widetilde{\mathbf{x}} \in \mathbb{R}^d$, with $d > 1$, defined as $\widetilde{\mathbf{x}} \triangleq \mathbf{x} + \boldsymbol{\delta}$ where $\boldsymbol{\delta} \in \mathbb{R}^d$ is some unknown perturbation, and $\mathbf{x} \in \mathbb{R}^d$ is the *original signal*. The goal is to recover the underlying *denoised signal* $\widehat{\mathbf{x}} \in \mathbb{R}^d$ that best approximates the original signal $\mathbf{x}$. However, finding an accurate approximation of $\mathbf{x}$ from $\widetilde{\mathbf{x}}$ is generally ill-posed [12]. Therefore, the TV denoising problem is typically formulated as [13, 12]: $\min_{\widehat{\mathbf{x}} \in \mathbb{R}^d} F(\widehat{\mathbf{x}}, \widetilde{\mathbf{x}}) + \lambda R(\widehat{\mathbf{x}})$, where $F : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ represents the *data fidelity term*, $R : \mathbb{R}^d \to \mathbb{R}$ is the *regularization term* that narrows down the space of candidate solutions, and $\lambda > 0$ is the *regularization parameter* controlling the trade-off between permissible noise and the regularity imposed by R (e.g., the smoothness of the underlying distribution).

A *denoiser* is expected to solve the minimization by learning $\widehat{\mathbf{X}} \sim q_\psi(\widehat{\mathbf{X}}|\widetilde{\mathbf{X}})$ where ψ are the parameters to learn. In the following paper, we will investigate how SOTA denoisers behave when denoised signals are given as input for a predefined classification task. Therefore, throughout the paper, we refer to the *target classifier* as $p_\theta(\widehat{Y}|\mathbf{X})$ where $\widehat{Y}$ is the random variable representing the classifier's inference and θ are the learned parameters. Its induced hard decision is defined as $f_\theta : \mathcal{X} \to \mathcal{Y}$ s.t. $f_\theta(\mathbf{x}) \triangleq \arg\max_{y \in \mathcal{Y}} p_\theta(y|\mathbf{x})$, where $\mathcal{X} \subset \mathbb{R}^d$ is the feature space corresponding

Table 1: The 'Noised signal' column presents the results for the signal following perturbation and prior to any denoising. The results are expressed in terms of `mean±std` across all heartbeats.

	Noised signal	DeScoD [6]	DRNN [2]	DeepFilter [5]
SSD	7.62±9.45	2.35±4.39	4.00±4.25	**1.45**±2.95
MAD	0.23±0.12	0.13±0.07	0.23±0.08	**0.12**±0.05
PRDN	84.41±52.82	46.89±26.14	66.69±20.05	**37.09**±18.25
SNR	7.98±6.45	12.31±4.06	8.56±3.14	**14.10**±3.41
CosS	0.93±0.06	0.96±0.04	0.93±0.04	**0.98**±0.02

to the set of original ECGs that have not been subjected to any perturbations and $\mathcal{Y} = \{1, \dots, C\}$ represent the concept of the label space related to some task of interest, such as the detection of the risk for developing some form of arrhythmia. Formally, given $\mathbf{x} \in \mathcal{X}$ we will be interested in studying whether $f_\theta(\mathbf{x}) = f_\theta(\widehat{\mathbf{x}}) \equiv f_\theta(g_\psi(\widetilde{\mathbf{x}}))$ where $g_\psi(\widetilde{\mathbf{x}})$ is the sampling of $q_\psi(\widehat{\mathbf{x}}|\widetilde{\mathbf{x}})$.

3 Evaluation framework and results

3.1 Evaluation metrics for the standalone denoising task

We consider the distortion metrics used in the most recent literature [5, 6] to assess the capabilities of the SOTA methods for signal denoising. These metrics come from the ECG compression field [14, 15] to measure the reconstruction error between the original signal and the one obtained after compression and then decompression (encoding, decoding).

We employ the **sum of square distances** (the lower the better)

$$\text{SSD}(\mathbf{x}, \widehat{\mathbf{x}}) \triangleq \sum_{i=1}^{d} \left(\mathbf{x}(i) - \widehat{\mathbf{x}}(i)\right)^2; \tag{1}$$

the **absolute maximum distance** (the lower the better)

$$\text{MAD}(\mathbf{x}, \widehat{\mathbf{x}}) \triangleq \max_{i \in \{0,\dots,d\}} |\, \mathbf{x}(i) - \widehat{\mathbf{x}}(i) \,|; \tag{2}$$

a normalized version of the **percentage root-mean-square difference** (the lower the better)

$$\text{PRDN}(\mathbf{x}, \widehat{\mathbf{x}}) \triangleq \sqrt{\frac{\text{SSD}(\mathbf{x}, \widehat{\mathbf{x}})}{\sum_{i=1}^{d} \left(\mathbf{x}(i) - \mu\right)^2}} \cdot 100, \tag{3}$$

where μ is the mean of the clean signal, i.e., $\mu = \frac{1}{d}\sum_{i=1}^{d} \mathbf{x}(i)$. Notice that the **signal-to-noise ratio** can be easily calculated from the percentage root-mean-square difference (PRD) without normalization [15] – that corresponds to eq. (3) with $\mu = 0$ – as follows

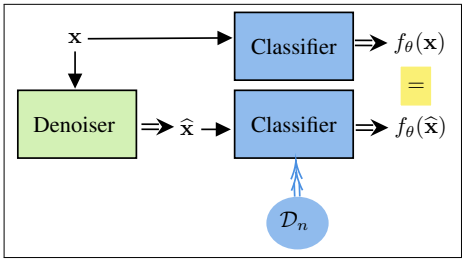
$$\text{SNR}(\mathbf{x}, \widehat{\mathbf{x}}) \triangleq 10 \cdot \log_{10}\left(\frac{\sum_{i=1}^{d} \left(\mathbf{x}(i)\right)^2}{\text{SSD}(\mathbf{x}, \widehat{\mathbf{x}})}\right) = 40 - 20 \cdot \log_{10}(\text{PRD} * 0.01). \tag{4}$$

Finally, we check the similarity between the two signals with **cosine similarity** (closer to 1 the better)

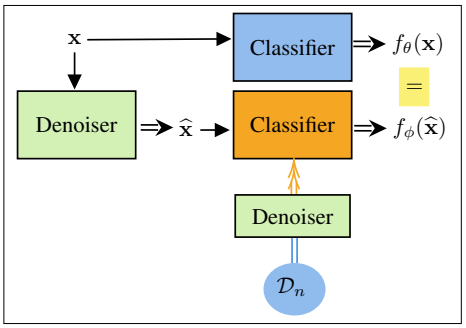
$$\text{CosS}(\mathbf{x}, \widehat{\mathbf{x}}) \triangleq \frac{\mathbf{x} \cdot \widehat{\mathbf{x}}}{||\mathbf{x}|| \cdot ||\widehat{\mathbf{x}}||} = \frac{\sum_{i=1}^{d} |\, \mathbf{x}(i) \cdot \widehat{\mathbf{x}}(i) \,|}{\sum_{i=1}^{d} |\mathbf{x}(i)| \cdot \sum_{i=1}^{d} |\widehat{\mathbf{x}}(i)|}. \tag{5}$$

3.2 Considered dataset, TdP risk prediction, and review of the related methods

`Generepol` dataset [10] and Torsades-de-Pointes (TdP) risk prediction [9]. We use the `Generepol` [10] dataset consisting of 10-seconds 8-lead ECGs sampled at 500Hz. For the experiments, we focus on lead II, commonly used to record the rhythm strip [16]. Since all the SOTA are tuned to work on *heartbeats*, i.e., on one single pulsation of the heart at a time, we segmented the original signals into chunks of 1s (500-points), centered around the R peaks. Finally, the training



(a) 'Original setting': Evaluation of the denoiser and classifier pre-trained on original ECGs.



(b) 'Denoised setting': Evaluation of the denoiser and classifier trained on denoised ECGs.

Figure 2: In Figure 2a, the classifier f_θ is trained on the original training set $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ for TdP risk prediction. On the other hand, in fig. 2b, the classifier f_ϕ, where ϕ are the learned parameters, is trained with the original training dataset but denoised using the method under consideration. In both cases, $\mathbf{x}$ and the obtained denoised version $\widehat{\mathbf{x}}$ represent a sample from the testing set.

set consisted of 30009 Sot+ and 32448 Sot- ECGs heartbeats, the validation set has 3659 Sot+ and 4188 Sot- ECG heartbeats, and the testing set comprised 7221 Sot+ and 7543 Sot- ECG heartbeats. Specifically, we use Sot- and Sot+ to refer to ECGs recorded in healthy individuals respectively before and after the intake of 80mg Sotalol, a drug known to strongly increase the risk of developing Torsade-de-Pointes events. We consider the CNN model originally developed by Prifti et al. [9] for TdP risk prediction and we retrained it to work on single heartbeats. This DenseNet model with six blocks (each having eight dense convolutional layers) was trained for 100 epochs using Adam optimizer, learning rate of 0.001, dropout rate of 0.2.

DeepFilter [5]. The model consists of six Multi-Kernel Linear And Non-Linear (MKLANL) filter modules. Each module contains two groups of four convolutional layers, where each layer is followed by a linear activation function or by a rectified linear unit (ReLU) depending on the group's type. The training loss is a combination of the sum of the squared distance and the maximum absolute distance between the clean ECG and the denoised one (cf. section 3.1). The idea is therefore to learn "smart" filters in order to discriminate between the desired ECG signal and the undesired noise.

DeScoD-ECG [6] is a novel approach that utilizes a conditional-score diffusion model. The generative model begins with Gaussian white noise and proceeds to iteratively reconstruct the signal through a fixed Markov Chain. Each step of the reconstruction involves Gaussian translation, which is conditioned on the previous step's reconstructions and the noisy ECG observations.

DRNN [2]. A DRNN was proposed here as a denoiser. Initially, the signal undergoes processing through a recurrent layer comprising Long Short-Term Memory (LSTM) [17] units. Then, the signal is further processed through a specified number of dense layers with ReLU activation. The final layer is linear and responsible for aggregating the outputs from the preceding layer by summation.

3.3 Results

Due to space constraints, the results on FCN+DAE [7] and Wavelet transform [4] are presented in Appendices A.1 and A.2.1. We refer to Appendix A.2.2 for the results conducted on an additional publicly available dataset (`PTB-XL` [18]). We relegate to Appendix A.1, a deeper exploration of TdP risk classification.

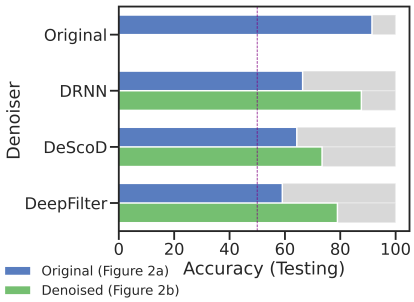


Figure 3: The vertical dashed line is at 50% accuracy.

Evaluation of the denoising task standalone. We conducted simulations of real-life noise (different levels of) *baseline wander* [4] inspired by original data obtained from the Physionet MIT-BIH NSTDB dataset [19, 20]. We used this data to assess the robustness of the evaluated methods when denoising the `Generepol` dataset. While the comprehensive discussion is relegated in Appendix A.2.1, we provide in Table 1 the summary of the results.

Effect of the denoising on TdP risk prediction. Given the TdP model trained on data without prior denoising, our objective was to assess the model's performance when the testing data is denoised using one of the SOTA denoisers. To achieve this, we begin with the original testing set from `Generepol`, which had not been subjected to any additional noise. Subsequently, we applied these denoisers to clean the signals before feeding them to the TdP risk classifier (see Figure 2a). **It is important to note that an effective denoiser should be capable of removing noise while preserving essential signal information and relevant features**. We provide in Figure 3 the accuracies of the model w.r.t. the different denoisers. The label 'Original' refers to the performance on the testing data, without any denoising applied to it. Interestingly, we observe a decrease in performance across all cases. This suggests that, despite the favorable results shown by the denoisers in Table 1, they are, in fact, removing valuable information from the signal, essential for the classification task.

We further explored the effect of denoising on the classification task by applying prior denoising to all partitions within `Generepol`. Consequently, we trained a new model for each denoiser (see Figure 2b). The updated accuracies are in Figure 3. Although the new models did not replicate the original performance, we observed a slight improvement compared to the previous setting. We finally analyzed the distribution of correctly and wrongly classified samples in Figure 4. Notably, neural networks are recognized for being overly *confident* in their predictions [21] (we call *confidence* the probability associated with the model's predicted class for a given sample). However, regardless of the denoisers used, the models tend to be less confident with the denoised data. **This can pose risks, as model confidence is commonly used for subsequent tasks related to model reliability**, such as *misclassification detection* [21, 22]. One technique used in computer vision for this task involves estimating the probability of classification error starting from the softmax outputted by the classifier [21]. Given an input sample, if this score exceeds a predefined threshold, the prediction is considered wrong, otherwise correct. The results, in Appendix A.1, show that alterations in the posterior distribution of the classifier have a detrimental effect on the method's ability to distinguish correctly and incorrectly classified samples with 40 percentage-point increase in False positive rate at 95% True positive rate (FPR, shortly) and 27-point reduction in AUROC.

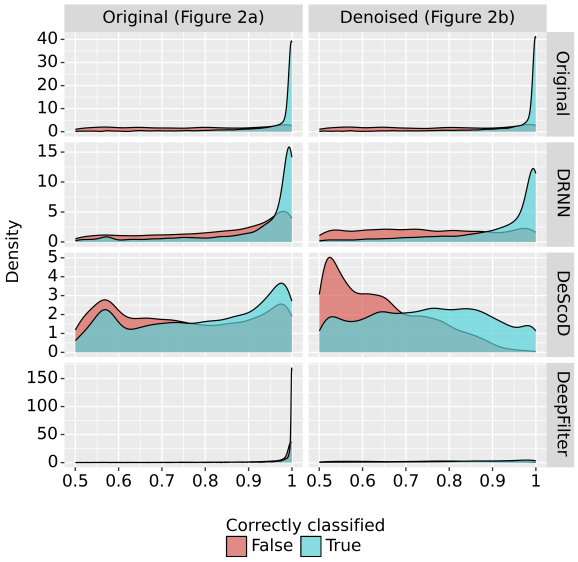


Figure 4: We split samples according to their true labels in blue and red. We examine the predicted class probability for each sample.

4 Conclusions

We examined ECG denoising methods and their effect on the automated classification of arrhythmia risk prediction. Our findings reveal that assessing denoising methods without considering downstream classification tasks yields overly optimistic results. We observed a reduction in classifier accuracy (up to 40 percentage points) when provided with denoised data compared to the original data. Further experiments have shown how alterations in the posterior distributions of the classifier on denoised data can have a detrimental impact on the misclassification detection task. These results stress the serious implications of denoising signals for the reliability of automated tasks, most notably within essential sectors such as healthcare, where even minor inaccuracies can have life-threatening outcomes.

Acknowledgments and Disclosure of Funding

This study was supported by the ANR-20-CE17-0022 DeepECG4U funding from the French National Research Agency.

References

[1] Electrocardiogram. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/electrocardiogram.

[2] Karol Antczak. Deep recurrent neural networks for ecg signal denoising. *arXiv preprint arXiv:1807.11551*, 2018.

[3] Sarang L Joshi, Rambabu A Vatti, and Rupali V Tornekar. A survey on ecg signal denoising techniques. In *2013 International Conference on Communication Systems and Network Technologies*, pages 60–64. IEEE, 2013.

[4] Rahul Kher et al. Signal processing techniques for removing noise from ecg signals. *J. Biomed. Eng. Res*, 3(101):1–9, 2019.

[5] Francisco P Romero, David C Piñol, and Carlos R Vázquez-Seisdedos. Deepfilter: An ecg baseline wander removal filter using deep learning techniques. *Biomedical Signal Processing and Control*, 70:102992, 2021.

[6] Huayu Li, Gregory Ditzler, Janet Roveda, and Ao Li. Descod-ecg: Deep score-based diffusion model for ecg baseline wander and noise removal. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[7] Hsin-Tien Chiang, Yi-Yen Hsieh, Szu-Wei Fu, Kuo-Hsuan Hung, Yu Tsao, and Shao-Yi Chien. Noise reduction in ecg signals using fully convolutional denoising autoencoders. *Ieee Access*, 7:60806–60813, 2019.

[8] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[9] Edi Prifti, Ahmad Fall, Giovanni Davogustto, Alfredo Pulini, Isabelle Denjoy, Christian Funck-Brentano, Yasmin Khan, Alexandre Durand-Salmon, Fabio Badilini, Quinn S Wells, et al. Deep learning analysis of electrocardiogram for risk prediction of drug-induced arrhythmias and diagnosis of long qt syndrome. *European Heart Journal*, 42(38):3948–3961, 2021.

[10] Joe-Elie Salem, Marine Germain, Jean-Sébastien Hulot, Pascal Voiriot, Bruno Lebourgeois, Jean Waldura, David-Alexandre Tregouet, Beny Charbit, and Christian Funck-Brentano. Genome wide analysis of sotalol-induced ikr inhibition during ventricular repolarization, "generepol study": Lack of common variants with large effect sizes. *PLOS ONE*, 12(8):1–16, 08 2017.

[11] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[12] Samuel Vaiter, Charles-Alban Deledalle, Gabriel Peyré, Charles Dossal, and Jalal Fadili. Local behavior of sparse analysis regularization: Applications to risk estimation. *Applied and Computational Harmonic Analysis*, 35(3):433–451, 2013.

[13] Mila Nikolova. Minimizers of cost-functions involving nonsmooth data-fidelity terms. application to the processing of outliers. *SIAM Journal on Numerical Analysis*, 40(3):965–994, 2002.

[14] M Sabarimalai Manikandan and Samarendra Dandapat. Ecg distortion measures and their effectiveness. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 705–710. IEEE, 2008.

[15] Andrea Němcová, Radovan Smíšek, Lucie Maršánová, Lukáš Smital, and Martin Vítek. A comparative analysis of methods for evaluation of ecg signal quality after compression. *BioMed research international*, 2018, 2018.

[16] Steve Meek and Francis Morris. Introduction. i—leads, rate, rhythm, and cardiac axis. *Bmj*, 324(7334):415–418, 2002.

[17] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[18] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

[19] George B Moody, W Muldrow, and Roger G Mark. A noise stress test for arrhythmia detectors. *Computers in cardiology*, 11(3):381–384, 1984.

[20] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[21] Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, and Pablo Piantanida. Doctor: A simple method for detecting misclassification errors. *Advances in Neural Information Processing Systems*, 34:5669–5681, 2021.

[22] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

[23] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[24] SZ Mahamoodabadi, A Ahmedian, and MD Abolhasani. Ecg feature extraction using daubechies wavelet. In *Proceedings of the fifth IASTED International Conference VISUALIZATION, IMAGING and IMAGE PROCESSING, September*, pages 7–9, 2005.

[25] Yee Guan Yap and A John Camm. Drug induced qt prolongation and torsades de pointes. *Heart*, 89(11):1363–1372, 2003.

[26] Esseim Sharma, Brian McCauley, Wasiq Sheikh, Anshul Parulkar, and Antony Chu. Torsades de pointes secondary to sotalol: Predictable but not always preventable. *Journal of the American College of Cardiology*, 73(9S1):2789–2789, 2019.

[27] Eduardo Dadalto, Marco Romanelli, Georg Pichler, and Pablo Piantanida. A data-driven measure of relative uncertainty for misclassification detection. *arXiv preprint arXiv:2306.01710*, 2023.

[28] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Openmix: Exploring outlier samples for misclassification detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12074–12083, 2023.

Table 2: The 'Noised signal' column presents the results for the signal prior to any denoising and following perturbation. The results are expressed in terms of `mean±std` across all heartbeats.

	Noised signal	FCN+DAE [7]	DeepFilter [5]	Wavelet [4]
SSD	7.62±9.45	27.78±16.89	**1.45**±2.95	7.23±9.39
MAD	0.23±0.12	0.54±0.08	**0.12**±0.05	0.21±0.12
PRDN	84.41±52.82	184.48±65.63	**37.09**±18.25	81.81±52.51
SNR	7.98±6.45	-0.08±3.55	**14.10**±3.41	8.26±6.22
CosS	0.93±0.06	0.86±0.07	**0.98**±0.02	0.93±0.07

A Appendix

A.1 Supplementary results of section 3.2

FCN+DAE [7]. The authors introduce a novel denoising algorithm utilizing a 13-layer Fully Convolutional Network (FCN) based Denoiser Autoencoder (DAE), where the decoder's objective is to reconstruct a signal based on the low-dimensional features generated by the encoder.

Wavelet transforms [4]. Wavelet transforms is a well-known technique used in signal processing for analyzing signals by cutting them up into different frequency components [23]. Because of the nature of the ECG signals, we look at the *Daubechies wavelets* family as they are similar in shape to the QRS complex, and their energy spectrum is concentrated around low frequencies [24]. The wavelet transforms approach leverages the signal's energy at different scales to effectively separate the noise from the ECG signals [6]. Nonetheless, it is worth noting, as highlighted in the literature [6, 5], that this method tends to be effective primarily in cases where the ECG signal is not severely corrupted, often struggling when confronted with high-amplitude noise.

Torsades-de-Pointes (TdP) risk prediction [9]. Torsades-de-pointes (TdP) is a life-threatening arrhythmia, which can be congenital or drug-induced, and it is associated with long QT intervals. Given the potential for sudden death associated with this condition [25], its study has garnered significant interest within the scientific community. Notably, in the study conducted using the `Generepol` dataset, the subjects were recorded ECGs both before and 1, 2, 3, and 4 hours after receiving an oral dose of 80mg of *Sotalol*, an anti-arrhythmic drug known to be associated with TdP [26].

A.2 Supplementary results of section 3.3

Publicly available code at `https://git.ummisco.fr/open/2023-denoising_impact`.

A.2.1 Evaluation of the standalone denoising task

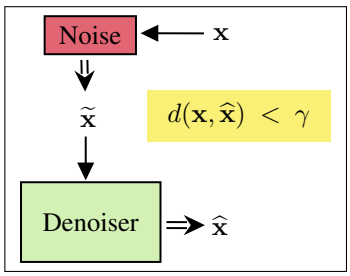


Figure 5: We denote with d any of the metric presented in section 3.1 and with $\gamma \in \mathbb{R}$ the corresponding threshold parameter.

We analyze how the methods described in section 3.2 perform when it comes to denoising on the `Generepol` dataset. In Figure 5 we show the evaluation pipeline. All the denoisers requiring a training phase have been trained as in [5]. In particular, the noise we consider throughout this paper is *basal wander* (or *basal drift*) with variable intensities, i.e., the effect where the base axis (x-axis) of a signal appears to 'wander' or move up and down rather than be straight [4].

To maintain consistency with other SOTA methods, we used the noise obtained from the Physionet MIT-BIH Noise Stress Test Database (MIT-BIH NSTDB) [19][20]. The dataset contains three types of noise including *baseline wander*, noise signals are recorded alongside clean ECGs. We superimposed the `Generepol` signals to

calibrate the amount of noise signals using the *nst* tool provided by Physionet. The amount of noise added is estimated in decibels dB. We added noise signals at 18dB and 24dB.

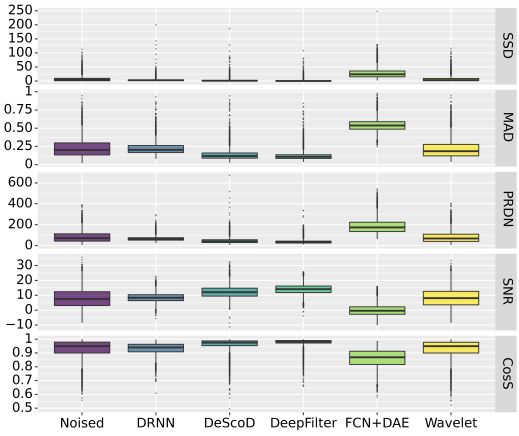
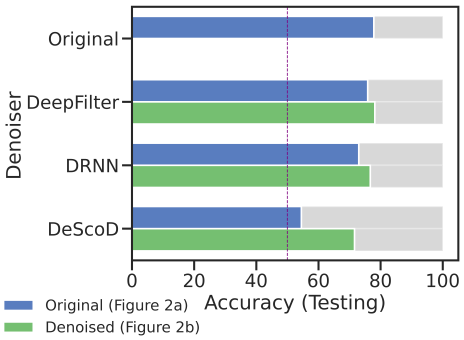


Figure 6: We remind that for `SSD` eq. (1), `MAD` eq. (2), and `PRDN` eq. (3) the lower the better; for `SNR` eq. (4) the higher the better; `CosS` eq. (5) closer to 1 the better.

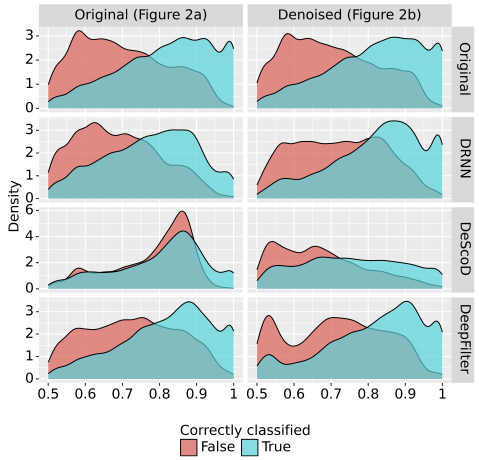
As we can see from Figure 6, even when evaluated solely for the denoising task on `Generepol`, the denoisers exhibit results consistent with those reported in their original papers. The poor performance of FCN+DAE may be attributed to a possible shift in signal reconstruction. It is worth noting that the model generates heartbeats with 512 points, even when provided with input heartbeats of only 500 points. To mitigate this discrepancy, we performed additional signal down-sampling. Finally, Wavelet transform yielded results closely resembling the original signal with noise Table 2. While this demonstrates its effectiveness when the signal has minimal noise, it offers limited denoising capabilities when confronted with higher levels of noise. Due to the unstable performances of the FCN+DAE and Wavelet, we have excluded these methods from the subsequent analysis with the TdP classifier.

A.2.2 Effect of the denoising on TdP risk prediction

We extend our analysis to a publicly available dataset `PTB-XL` [18] containing 10-second 12-lead ECGs sampled at 500Hz labeled as *normal*, *myocardial infarction*, *ST/T change*, *conduction disturbance* and *hypertrophy*. A binary classification task was created by putting all samples labeled as not normal in a single category. The final dataset consists of 96533 *normal*, class 0, and 82747 *abnormal*, class 1. Therefore we trained the same DenseNet architecture as for the TdP risk prediction in `Generepol`. We show in fig. 7a the results in terms of accuracy, and in fig. 7b the distribution of the predicted class for correctly and incorrectly classified samples.



(a) Accuracy



(b) Distributions of correctly and wrongly classified samples.

Figure 7: `PTB-XL` results. We remind that with 'Original' we indicate the setting in Figure 2a; with 'Denoised' we indicate the setting in Figure 2b.

A.2.3 Implication on misclassification detection

Table 3: Training TdP on original data.

Generepol	AUROC↑	FPR$\downarrow_{95\%}$
Original	**84.68**	**53.21**
DRNN	68.92	80.48
DeepFilter	64.81	82.03
DeScoD	58.04	92.55
PTB-XL	AUROC↑	FPR$\downarrow_{95\%}$
Original	**73.77**	**74.17**
DRNN	69.76	80.29
DeepFilter	70.26	78.66
DeScoD	55.84	82.43

Table 4: Training TdP on denoised data.

Generepol	AUROC↑	FPR$\downarrow_{95\%}$
Original	**84.68**	**53.21**
DRNN	79.91	62.23
DeepFilter	72.09	74.69
DeScoD	71.66	74.52
PTB-XL	AUROC↑	FPR$\downarrow_{95\%}$
Original	**73.77**	**74.17**
DRNN	71.67	75.32
DeepFilter	70.51	75.46
DeScoD	67.53	82.41

Misclassification detection is a hot topic in Machine Learning (ML) safety, focusing on identifying instances with potentially incorrect model predictions [21, 27, 28]. DOCTOR [21] is a simple method that aims to identify whether the prediction of a classifier should (or should not) be trusted so that, consequently, it would be possible to accept it or reject it. We conducted simulations using DOCTOR on the TdP risk classifiers in order to study the effect of denoising when the misclassification task is involved. In particular, we chose such a method to be applied to any pre-trained model, and it did not require prior information about the underlying dataset. The detector was defined as

$$D_\alpha(\mathbf{x}, \gamma) = \begin{cases} 1, & \text{if Gini}(\mathbf{x}) \geq \gamma' \cdot (1 - \text{Gini}(\mathbf{x})) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $\text{Gini}(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{y \in \mathcal{Y}} p_\theta(y|\mathbf{x})(1 - p_\theta(y|\mathbf{x}))$ is the probability of incorrectly classifying the feature $\mathbf{x}$ if it was randomly labeled according to the model distribution. Therefore, the higher $\text{Gini}(\mathbf{x})$, the higher the probability of $f_\theta(\mathbf{x})$ is of being wrong. The performances of the detector are evaluated in terms of False positive rate at 95% of True positive rate (FPR$\downarrow_{95\%}$), the lower the better, and AUROC (AUROC↑), the higher the better. We refer to [21] for a complete discussion.

In Table 4 and **??** we provide the summary of the results. Interestingly, even when the TdP models are trained on denoised data DOCTOR is not able anymore to distinguish the correctly classified samples from the incorrect ones due to the alteration in the posterior distribution of the models.