SN Computer Science (2024) 5:699
https://doi.org/10.1007/s42979-024-03017-7

ORIGINAL RESEARCH

Cleaning ECG with Deep Learning: A Denoiser Tested in Industrial Settings

Mariana Dias[1] · Phillip Probst[1] · Luís Silva[1] · Hugo Gamboa[1]

Received: 30 October 2023 / Accepted: 29 May 2024

Abstract
As the popularity of wearables continues to scale, a substantial portion of the population has now access to (self-)monitorization of cardiovascular activity. In particular, the use of ECG wearables is growing in the realm of occupational health assessment, but one common issue that is encountered is the presence of noise which hinders the reliability of the acquired data. In this work, we propose an ECG denoiser based on bidirectional Gated Recurrent Units (biGRU). This model was trained on noisy ECG samples that were created by adding noise from the MIT-BIH Noise Stress Test database to ECG samples from the PTB-XL database. The model was initially trained and tested on data corrupted with the three most common sources of noise: electrode motion artifacts, muscle activation and baseline wander. After training, the model was able to fully reconstruct previously unseen signals, achieving Root-Mean-Square Error values between 0.041 and 0.023. For further testing the model's robustness, we performed a data collection in an industrial work setting and employed our model to clean the noisy data, acquired from 43 workers using wearable sensors. The trained network proved to be very effective in removing real ECG noise, outperforming the available open-source solutions, while having a much smaller complexity compared to state-of-the-art Deep Learning approaches.

Keywords GRU · Denoiser · ECG · Industry

Introduction

The electrocardiogram (ECG) is a widely used acquisition procedure to visualize and assess cardiovascular activity. The ECG waveform can be separated into distinct sections that can be attributed to physiological processes occurring during the cardiac cycle: (1) the P wave generated by atrial depolarization, (2) the QRS complex arising from ventricular depolarization, and (3) the subsequent T wave that forms due to ventricular repolarization [1].

It is well-known that ECG is used for cardiovascular diseases detection, but as the shape and temporal distribution of the ECG waveform changes according to the current physiological state, there is more information that can be extracted from ECG monitoring, namely changes in heart rate variability (which reflects one's neurophysiological state) and other cardiac events [2]. The development of wearables and their increasing popularity has made ECG assessments more accessible. They enable users and researchers to take ECG monitoring out of the clinical and laboratory environment to assess cardiovascular health at virtually any environment and moment in time, resulting in a higher fidelity long-term assessment. These more accessible wearable data acquisition systems have implications on health care quality as these continuous monitoring schemes allow for early detection, prediction and possibly prevention of developing CVDs [3].

The acquisition of ECG signals with wearables is relatively straightforward; however, potential noise sources that can distort the signal to such an extent that a correct analysis and interpretation is made difficult must be considered.

✉ Mariana Dias
mag.dias@campus.fct.unl.pt

Phillip Probst
p.probst@campus.fct.unl.pt

Luís Silva
lmd.silva@fct.unl.pt

Hugo Gamboa
hgamboa@fct.unl.pt

[1] Department of Physics, LIBPhys (Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics), NOVA School of Science and Technology, NOVA University Lisbon, Largo da Torre, 2829-516 Caparica, Portugal

Published online: 03 July 2024

The most common sources of noise in ECG signals stem from muscle activation (MA) in regions surrounding the ECG electrode placement, electrode movement (EM), and baseline wander (BW) [4]. Artifacts that overlap with the frequency spectrum of the ECG signal (~ [1 Hz, 45 Hz]) are especially challenging because these cannot be removed with linear filters (low-, band-, nor high-pass filters) [1]. Such noise sources need to be especially considered in factory settings where workers execute high amount of fast repetitive movements and electromagnetic interference from surrounding machines is common. Thus, to acquire reliable ECG signals, effective noise removal techniques have to be developed and implemented.

Noise removal techniques can generally be separated into two categories: traditional and Machine/Deep Learning based approaches. Traditional noise removal techniques encompass empirical mode decomposition, wavelets, sparsity-based models, Bayesian filters, non local means denoisers, and hybrid models [5, 6]. In comparison to traditional approaches, deep neural networks have shown enhanced performance in ECG denoising, having the advantage of being adaptable to different noise amounts and sources, if trained with sufficient and relevant data. The most common approach in this field is utilizing denoising autoencoders (DAEs) [7–10]. More recently generative adversarial networks have been employed, producing state-of-the-art results [11]. Nevertheless, current deep learning approaches result in relatively large networks that aim at producing results with very low reconstruction error, meaning that the focus lies more on performance than on efficiency. In real-time processing pipelines networks that are able to denoise the ECG signal reliably and that are at the same time efficient (i.e., have low latency) are needed. Furthermore, performance of networks on real-world data is currently lacking.

Industry 4.0 has ushered in novel opportunities to collect data in real-world contexts. However, these opportunities also present challenges concerning signal quality, once the environment is not under the researcher's control. ECG is of significant relevance in industries, especially on assembly lines: it provides direct insights into the cardiovascular response to repetitive tasks, which are often linked to cardiovascular disorders and biomechanical risks. These tasks are often performed in awkward postures, lifting of potentially heavy components, and under high levels of noise and stress, which impacts overall occupational and cardiovascular health [12, 13]. These wearable setups allow for continuous data acquisition during work tasks without hindering the workers in their task execution. Monitoring occupational health through ECG to assess cardiovascular health should have positive outcomes on both an individual and organisational level, as the onset of work-related CVDs has direct implications on productivity through higher absenteeism and lower performance [12, 14]. The data quality is fundamental, since ergonomic and clinical interpretations will be based on these metrics. For this reason, the present work includes data from a project that was centered around the theme of Digital Transformation in Industry with a particular emphasis on the concept of Operator 4.0., referred to as OPERATOR. The primary objective was to create digital tools for monitoring an operator's cognitive, motor, and physiological responses through the use of wearable technology.

In this work, we present one component that could be integrated into these frameworks, namely, a comparatively small bidirectional Gated Recurrent Unit (biGRU) model for removing MA, EM and BW noise from ECG data. The model was trained on the PTB-XL database [15] using additive noise samples from the MIT-BIT Noise Stress Test database [16]. Additionally, we performed a data collection in a real world scenario, namely in an assembly line in automobile industry, where tasks involve numerous motor strategies and the signal could be subject to several noisy factors. This way, we were able to robustly evaluate the networks performance and efficiency using data collected in a manufacturing plant. Thus, the main contributions of our work can be summarized as follows: (1) using small GRU-based networks for ECG denoising tasks, (2) utilizing a heterogeneous database (PTB-XL), containing records from more than 18 thousand subjects encompassing data from various CVDs and healthy patients, (3) evaluating the model on real-world noisy data, and (4) providing the denoising model as an open-source tool for public usage. The developed code repository (including the weights of the trained network) was made open-source and it is available at https://github.com/marianaagdias/ECG_Denoiser.

Related Work

Deep Learning (DL) methods are becoming more popular in recent years as they have been successfully applied to ECG denoising generating state-of-the-art results. With regards to traditional demonising techniques, extensive reviews on these can be found in [5, 17]. Most ECG denoising approaches that employ DL focus on autoencoders (DAE), for which several model architectures have been proposed, including feed forward neural networks (NN), convolutional neural networks (CNN), and long short-term memory (LSTM) cells.

Rodrigues and Couto [7] presented a feed forward NN with three hidden layers with 1000 neurons for the reduction of electrode motion artifacts. The network was initialized using a stack of Restricted Boltzmann Machines and was then subsequently fine tuned. With their model, they showed that a variety of QRS detectors perform better after using it to denoise ECG signals. In [8], two fully-connected DAEs, one with a single hidden layer (SHL) and one with

multiple hidden layers (MHL) were trained on ECG signals to which Gaussian white noise was added. For both networks the number of neurons and the activation function for the layers was investigated. They showed that the MHL-DAE, consisting of three hidden layers with a 50-L/2-50 neuron architecture, where $L = 80$ represents the length of the input signal, together with a *tanh* activation function performed the best.

Arsene, Hankins, and Yin [9] proposed both a CNN and an LSTM based model for ECG denoising. The CNN model consists of six blocks that contain a convolutional layer with 36 filters of kernel size 19x1 and a stride of [1 1] followed by batch normalization layer, a rectified linear unit layer (both of size 36), and average pooling layer with a stride of four and a pooling size of two. The six blocks are followed by a fully-connected regression output layer that reconstructs the original signal. The LSTM model is composed of two blocks that include a LSTM cell with 140 neurons followed by fully-connected rectified linear unit layer, also with 140 neurons. The same regression layer as in the CNN model is used to generate the output. Both models use an input layer of 30,000 neurons which is equal to the number of samples in the ECG input signal. To train and test their models they used two synthetic datasets and a real ECG dataset. In the synthetic datasets they added random and drift noise, while for the real ECG dataset a focus was put on motion artifact noise. The two models were compared with a traditional wavelet approach, showing that the CNN model outperformed the LSTM and the wavelet approach, when comparing the root mean square error between the reconstructed and original signal. Antczack [10] developed an architecture consisting of an LSTM cell with 140 neurons followed by two ReLU layers of 64 neurons each that was trained on ECG sequences with a length of 600 samples. Their network was pre-trained with synthetic data to find an optimized architecture and subsequently fine-tuned with real data. Both datasets were corrupted by white noise. They compared their network to a wavelet approach and a band-pass filter and their model outperformed these approaches. Pre-training with synthetic data helped the model to converge faster when being fine-tuned.

A DAE that combines CNN and LSTM was proposed by Dasan and Panneerselvam [18]. The encoder of their model consists of eight convolutional layers of with max-pooling layers in between. The final component of the encoder consists of a LSTM cell followed by a fully-connected layer. The decoder mirrors the convolutional layers of the encoder with the difference that no batch normalization is applied and max-pooling is replaced by 1D up-sampling. They evaluated their model by comparing it to other convolutional as well stacked DAEs when applying white gaussian noise to ECG signals, showing that their approach outperformed the comparative networks.

Recently, an approach using a generative adversarial network (GAN) was proposed in [11]. The generator of the GAN receives signals corrupted with noise and is set to recreate the original noise-free signal. It consists of three fully-connected layers; the input and output layers are of size 310 which corresponds to the number of samples in the signals fed to the network. The generator of the network is trained with different loss functions that are composite combinations of the default generator loss, a distance function, a maximum local difference, and a mean square error. The discriminator receives either the output of the generator or a real noise-free ECG signal. It is set to distinguish if the provided input is either a real noise-free signal or a denoised signal.

While the presented approaches are all valid in their own right, they have certain limitations. With the exception of [8, 10], all presented papers used either synthetic data and/or the MIT-BIH arrhythmia database [19], which consists of data from only 47 subjects. There is the possibility that these networks are biased towards arrythmia signals and their denoising results would not translate to other pathologies. Furthermore, in all works noise was applied to the entire sample, which is not necessarily representative of real-world scenarios (e.g., motion or muscle artifacts do generally appear for short amounts of time, unless the subject is moving constantly). Finally, while [8, 10] used the PTB-XL database they only applied white noise to their ECG samples. This shows that there is a lack of models that are trained with (1) more diverse datasets (i.e., containing more pathologies and data from healthy subjects), (2) realistic noise characteristics and (3) that are tested and evaluated on data collected in out-of-laboratory environments.

Methods

Gated Recurrent Units

In this work we present models based on GRU-cells, which are a way to implement Recurrent Neural Networks (RNNs). RNNs are neural networks that were designed for sequence analysis. Generally speaking, sequences are made of data points that are sequentially dependent, meaning that a data point at a specific time step is affected by previous points and might influence subsequent points [20]. Physiological signals such as the ECG are considered to be sequential data or in other words, time series. To model the sequential dependencies between time steps, RNNs include feedback connections among hidden units, which allow a "memory" of previous inputs to persist in the network's internal state. However, when trying to model long-term dependencies with basic RNN models, the gradients calculated for updating the network's weights during the back-propagation step,

tend to vanish [21]. One way to overcome the this vanishing gradient problem is to utilized GRU-cells [22]. The main architectural component of the GRU cell are two non-linear gates, the reset and update gates, that regulate the flow of information into and out of the cell, making it possible to maintain the state over time. Equations 1–4 are the mathematical formulas that describe the information flow through the GRU cell, which is represented in Fig. 1.

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (update\ gate) \tag{1}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (reset\ gate) \tag{2}$$

$$\tilde{h}_t = tanh(W \cdot [h_{t-1} * r_t, x_t] + b_h) \quad (candidate\ activation) \tag{3}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (output) \tag{4}$$

Part 1: Model Development

Databases

For training our model and performing an initial evaluation, we used two databases: the PTB-XL database [15] and the MIT-BIH Noise Stress Test database [16], both publicly available in PhysioBank [23].

The PTB-XL database contains 21,837 10-s clinical ECG records from 18,885 patients. The subjects' age ranges between 0 and 95 years and are evenly distributed in terms of gender. The 12-lead ECG signals were originally acquired at 400 Hz and two versions of the data are provided in PhysioBank: an upsampled version at 500 Hz and a downsampled version at 100 Hz. The database encompasses pathologies associated with conduction disturbance, myocardial infarction, hypertrophy, and ST/T changes. Furthermore, it includes 9528 instances (43.6%) of normal ECG records.

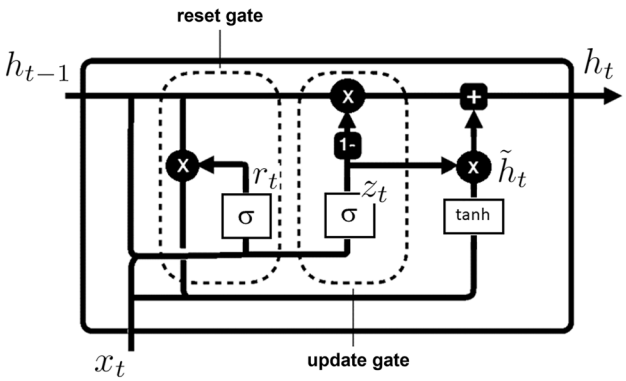

Fig. 1 Illustrative representation of a GRU cell

According to the authors, the data is of high quality, with only a few instances exhibiting static noise (14.9%), baseline drift (7.4%), burst noise (2.8%), or electrode-related issues (0.1%).

The MIT-BIH Noise Stress Test database contains three half-hour recordings of BW, MA, and EM noise. Two channels of noise were collected at 250 Hz, by placing the electrodes in such a way that the ECG was not visible.

Data Pre-processing

The following pre-processing steps were applied to the ECG signals: (i) the 500 Hz version of the PTB-XL database was down-sampled to 360 Hz (same as the MIT-BIH arrhythmia database); (ii) a second-order bandpass butterworth filter was applied to all ECG signals, with a high-pass cutoff frequency of 1 Hz and a low-pass cutoff frequency of 45Hz to eliminate high-frequency noise [1]; and (iii) a *Min–Max* normalization was performed (normalization based on the minimum and maximum values of each record). The noise data was also resampled to 360 Hz.

Training, Validation, and Test Sets

The ECG data from PTB-XL was divided in training, validation, and test sets, with a proportion of 70%, 15%, and 15%, respectively. For each signal, the three leads with the lowest number of peaks were selected and, for the training set, the 3 leads were used as separate signals; for the validation and test sets, 1 of the 3 leads was randomly selected. The rationale of this procedure is: as all leads are records of the same heart activity, a higher number of peaks in one lead is likely due to the presence of noise. This way, we had 45,777 (15,259 × 3 leads) ECG samples for training, 3270 for validation, and 3270 for testing. Regarding the noise data, both provided channels were used, totaling 60 min of BW, MA, and EM noise. 80% (48 min) was used for training, 10% (6 min) for validation, and the remaining 10% for testing.

The Model'S Input: Noisy Samples

For training our model, we added to each 10-s clean ECG signal a random portion of noise, with a length of 2–6 s. The EM and MA noises were added to a random part of the clean ECG signal while the BW noise was always added to the entire 10 s, as illustrated in Fig. 2. Additionally, to some records, mixed samples of different types of noise were added. The set of all possible combinations was: (MA, EM, BW, BW + MA, BW + EM, MA + EM, BW + MA + EM). There was a probability of 1/3 that the added noise was MA, of 1/3 that the added noise was EM and of 1/3 that the added noise was either BW or any combination. The reason for this is the fact that BW noise is the easiest to remove and having combinations of

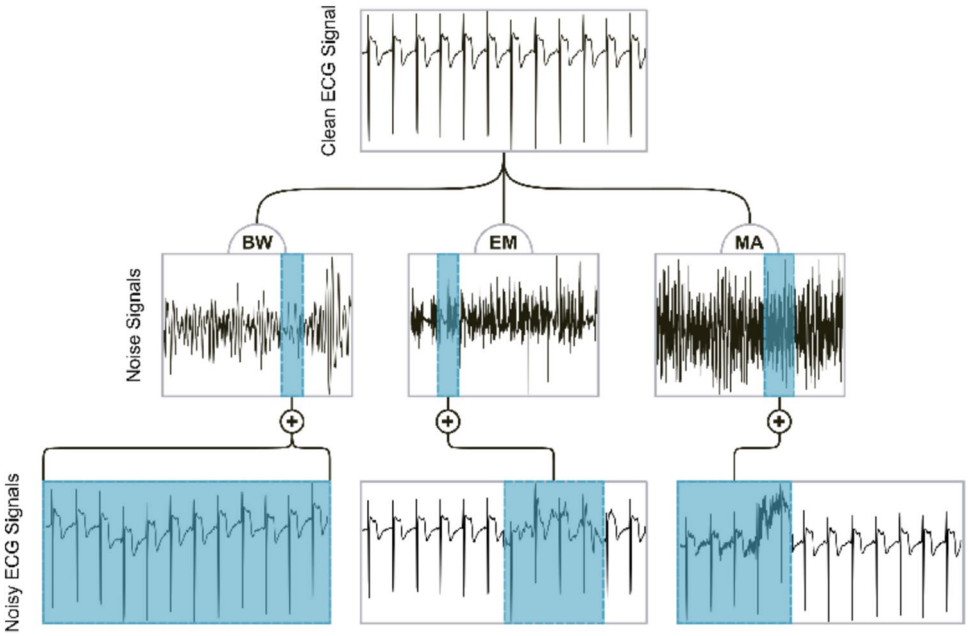

Fig. 2 Schematic representation of the creation of noisy samples from adding MA, EM and BW noise to the clean ECG samples from PTB-XL

different noise simultaneously is less likely in the real world scenarios. In the training set, since we used 3 leads from the same signal separately, a different type of noise was added to each lead. In the training and validation sets, in order to include different proportions of noise, the noise was multiplied by a randomly chosen factor between [0.7, 1.5] before being added to the ECG signal. These boundaries were chosen to ensure that the model was exposed to neither too low nor unrealistically high amounts of noise. In the test set, the added noise had known SNR_{in} values of either 0, 5, 7, or 10 dB.

Architecture Definition and Training

During the development of the GRU model, various architectures were tested in the validation set. The optimized hyperparameters are presented in Table 1 and the two principal versions of the architecture are illustrated in Fig. 3: the unidirectional and the bidirectional GRU. The model was fed with ECG signals following pre-processing and noise addition steps. Prior to inputting the data into the model, a Min–Max normalization process was applied, resulting in sequences of 3600 samples each (representing 10 s recorded at 360 Hz) with values ranging from 0 to 1. The desired output consisted of the clean signals after pre-processing. While the model's number of layers, neurons, and the dropout rate (applied to the output of GRU layers, except for the last one) were subject to optimization, certain parameters of the model were preestablished. In the bidirectional approach, only the first or first two layers (in the cases where there were three) were configured to be bidirectional. The loss function employed was the Root Mean Square Error, measuring the discrepancy between the predicted and desired outputs. All models were trained using the *Adam* optimizer, for 130 epochs with a batch size of 256 and an initial learning rate of 0.005. For each approach, the model parameters from the epoch with the lowest validation loss during training were saved. Finally, validation losses were compared and the model achieving the lowest value was tested on the test set.

We conducted model training using a Nvidia GeForce GTX 1080 Ti GPU and the code was implemented in the deep learning framework Pytorch [24].

Metrics for Evaluation

To evaluate model performance on the PTB-XL database, the main metrics found in the literature were used: the Root-Mean-Square Error (RMSE), the improvement in the Signal-to-Noise Ratio (SNR_{imp}), and the Percentage-Root-Mean-Square Difference (PRD) [5]. Equations 5–9 describe these metrics, where y is the original clean ECG signal, $\tilde{y}$ is the noisy signal, and $\hat{y}$ is the denoised signal. It is important to note that all these metrics imply the existance of a clean ground truth signal to compare the denoised output with, which means that they are not useful for real use cases. The equations that define these metrics are the following:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}[y(n) - \hat{y}(n)]^2} \tag{5}$$

$$SNR_{imp} = SNR_{out} - SNR_{in} \tag{6}$$

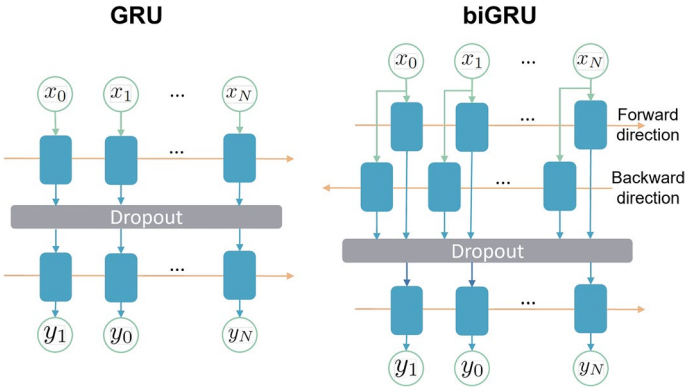

Fig. 3 Overview of the proposed GRU based framework for ECG denoising

Table 1 Hyperparameters of the model and corresponding optimized values

Hyperparameter	Tested values
Number of layers	{2, 3}
Number of neurons	{32, 64, 128, 256}
Dropout rate	{0, 0.3}
Bidirectional	{True, False}

$$SNR_{in} = 10 \times log_{10}\left(\frac{\sum_{n=0}^{N-1}[y(n)]^2}{\sum_{n=0}^{N-1}[\tilde{y}(n) - y(n)]^2}\right) \quad (7)$$

$$SNR_{out} = 10 \times log_{10}\left(\frac{\sum_{n=0}^{N-1}[y(n)]^2}{\sum_{n=0}^{N-1}[\hat{y}(n) - y(n)]^2}\right) \quad (8)$$

$$PRD = \sqrt{\frac{\sum_{n=0}^{N-1}[\hat{y}(n) - y(n)]^2}{\sum_{n=0}^{N-1}[y(n)]^2}} \times 100\% \quad (9)$$

Part 2: Field Application-Testing the Model on Industrial Setting Data

The main goal of the present study is to build a model that is effective in the task of removing noise from ECG data acquired with wearable sensors in real-world scenarios, namely during work in a factory, where noise sources are various and extremely difficult to avoid. As such, After developing our model, we performed a test on data acquired in an industrial setting, with the aim of testing the robustness of the model. In "Data Acquisition Protocol" section, the data acquisition protocol is described; in "Pipeline for ECG Denoising" section the applied method for testing the model with these data is defined; and in "Model Evaluation on Industrial Setting Data" section we justify the parameters used for evaluating its efficacy in the absence of clean ground truth data.

Data Acquisition Protocol

A total of 46 subjects working in an automotive assembly line participated in the study. Data from 3 subjects were lost due to equipment problems during the acquisition, resulting in a final sample size of 43 participants (Table 2 contains the demographics information of the participants). Each session had a duration of 20 min to 1 h where the workers performed their regular work tasks (it was not possible to standardize the time duration of the acquisitions because the regular assembly line functioning could not be compromised). The data collection was performed using the acquisition software *OpenSignals* (Plux, Lisbon, Portugal). A single-lead electrocardiography sensor with triode configuration (Plux, Lisbon, Portugal) was placed on the operator's chest, as represented in Fig. 4, and the sensor was acquiring with a sampling rate of 350 Hz. The electrode placement was chosen to minimize the muscle activation noise, once the workers are performing intense upper body work and there was the need of reducing the risk of obtaining irreversibly noisy data. In addition to the ECG, other sensors were used to acquire electromyography, respiration and movement data, nonetheless, this data was not used for the present study. The protocol was clearly explained to each participant, and all subjects gave their informed consent for inclusion before they participated

Table 2 Participants' demographics data

Sex	4 F 39 M
Age	38.0 ± 7.7 yrs
Height	176.7 ± 6.9 cm
Mass	77.9 ± 11.1 kg

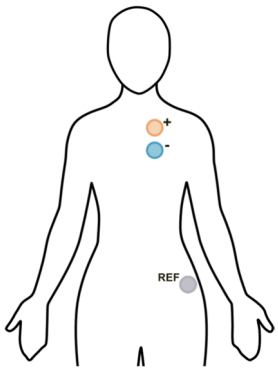

Fig. 4 An illustrative representation of the ECG electrode placement

in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of University of Porto.

The acquired raw data were stored and visually inspected. From the visual inspection, it was possible to see that, although noisy as predicted, the data was in general of satisfying quality considering the conditions under which the acquisitions were performed.

Pipeline for ECG Denoising

The main pipeline that was applied to the ECG data for testing our model can be described by the following sequence of steps: (i) randomly selecting 10-min from each subject's data, (ii) preparing the signal, which includes resampling to 360 Hz, and *MinMax* normalization; (iii) running the model giving the prepared signal as input; (iv) post-processing—the *post-alignment* step—which is the application of a 3rd order highpass butterworth filter with a cutoff frequency of 0.5 Hz. The last step was added to our pipeline due to a bias that the model was adding to the output when long signals were given as input, causing a misalignment.

For time optimization purposes, we evaluated the impact of segmenting the ECG signal prior to its use as input to the model. The segmentation process, depicted in Fig. 5, can be outlined as follows: (i) dividing the signal into 31-s segments with a 1-s overlap with the preceding one (excluding the initial 30-s segment), (ii) applying denoising to each segment using the biGRU denoiser, and (iii) reassembling the segments. To reconstruct the signal, the initial 0.5 s of each output segment are cut, and the remaining overlapping 0.5 s undergo mean overlap computation, as indicated in green in Fig. 5.

Model Evaluation on Industrial Setting Data

Testing on real-world data is a necessary step to understand if a model has the generalization power to properly execute the task it was designed for or if it only learned the features of the data it was trained with (i.e., whether it is overfitted or has good generalization power). Nonetheless, this evaluation is not straightforward, since there is no ground truth to compare our results with. The literature on ECG quality assessment can be divided in 3 main approaches [25]: the first uses

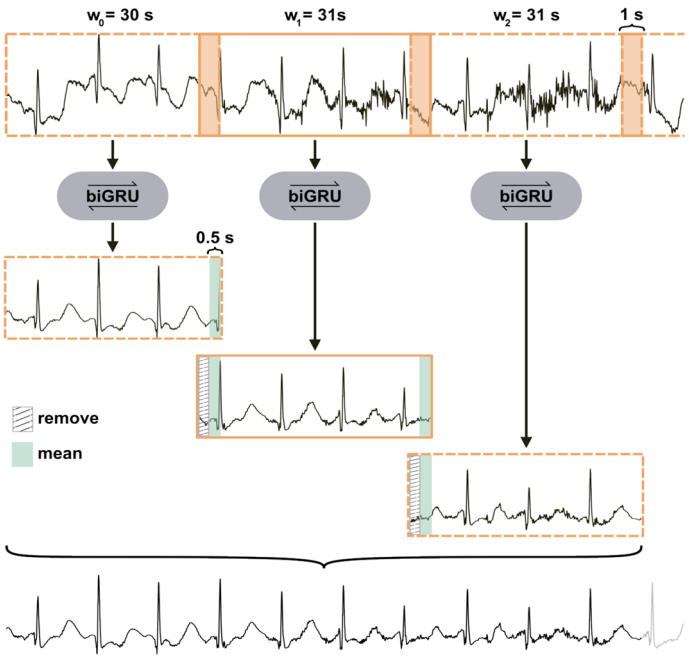

Fig. 5 Illustration of the segmentation process applied to long ECG signals. Within the reconstruction phase of the process, "remove" and "mean' correspond to two key actions performed on output signal segments: "remove" indicates the portions of the signal that are cut, while "mean" designates the segments overlapped to compute the final reconstructed signal, depicted at the bottom of the figure

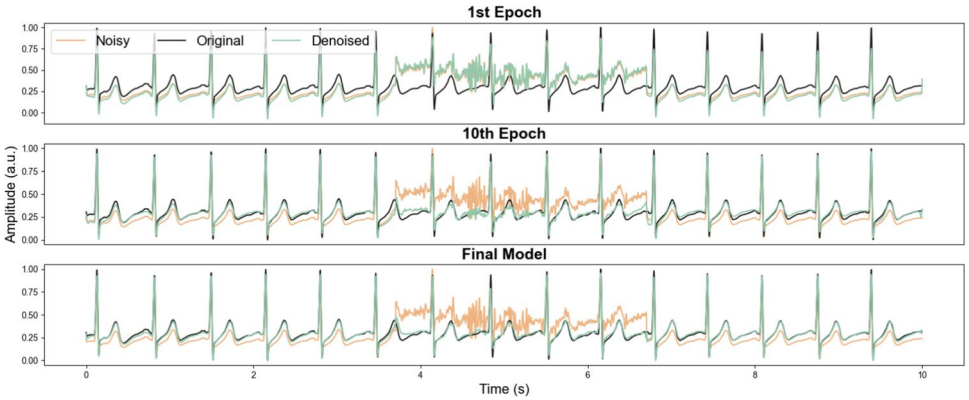

Fig. 6 Denoiser performance after the first, the tenth and the 58th (best) epochs

statistical metrics to assess if the signal is of good quality or not, based on the expected shape of the histogram of a clean ECG signal; the second is based on distance metrics from template signals; the third comprises machine learning models that perform classification distinguishing noisy from clean signals. The first two approaches fall short when we are dealing with data that are not acquired in a clinical setting, because the shape of the ECG signal and, consequently, of the histogram depend on the placement of the electrodes while acquiring the ECG. This way, clean ECG signals can have high distance values from ECG templates or have different histogram shapes not meaning that it is due to the presence of noise. The most recent studies on ECG quality assessment methods focus on distinguishing noisy from not noisy segments of signal and not on quality quantification [26] [27]. This can be useful for automatic deletion of noisy data (for example, for clinical purposes) or for selecting the parts of the data which should be denoised. For the purpose of this paper, nonetheless, it is not applicable as we are dealing with data that is almost continuously subject to noise and we are trying to improve its quality in order to allow the study of cardiovascular response in workers using wearables. Moreover, it's crucial for these algorithms to be made publicly available, enabling wider testing and application, and this is frequently not the scenario [26].

Previous results obtained within the scope of the OPERATOR project have shown a significant correlation between heart rate variability (HRV) and fatigue state of subjects performing repetitive work in a laboratory setting [28, 29]. These results suggest that from HRV we can obtain meaningful information from a subject's physiological state during work. Nonetheless, the various sources of noise that were present in an industrial workstation contaminate the ECG data and, hence, lead to higher difficulties while performing R-peak detection for the computation of HRV metrics. As such, in order to evaluate the performance of the biGRU model in this new dataset, the chosen approach was to evaluate the improvement in R-peak detection after cleaning the noise with our model, similarly to the strategy implemented by Rodrigues and Couto in [7].

In summary, for evaluating the model, we used a random 10-min segment from the data of each participant and computed the number of missing peaks (MP) and wrongly detected peaks (WP) before and after performing the noise removal. To do that, we did the following steps: (i) R-peak detection, for which we employed a widely used Python algorithm from a well-know toolbox named *Neurokit* [30] (we used the default method, where the QRS complexes are detected based on the steepness of the absolute gradient of the ECG signal); (ii) compute the distance between each two consecutive peaks; (iii) compute the upper and lower limits for outlier detection, based on the 25% (Q1) and 75% (Q3) quartiles (and the Inter-Quartile Range (IQR)), as represented in Eqs. 10–12; (iv) compute the derivative of the distances vector and check if it is higher than a minimum value of 50 (experimentally determined). The last criterion was added to make it robust to large (yet physiologically acceptable) heart rate fluctuations, avoiding erroneous outlier detections.

$$IQR = Q3 - Q1 \tag{10}$$

$$Upper\ limit = Q1 + 1.25 \times IQR \tag{11}$$

$$Lower\ limit = Q3 - 1.25 \times IQR \tag{12}$$

Results

Part 1: Model Development

From the models described in "Architecture Definition and Training" section, the one that achieved the lowest RMSE on the validation set was the bidirectional GRU with 2 layers of 64

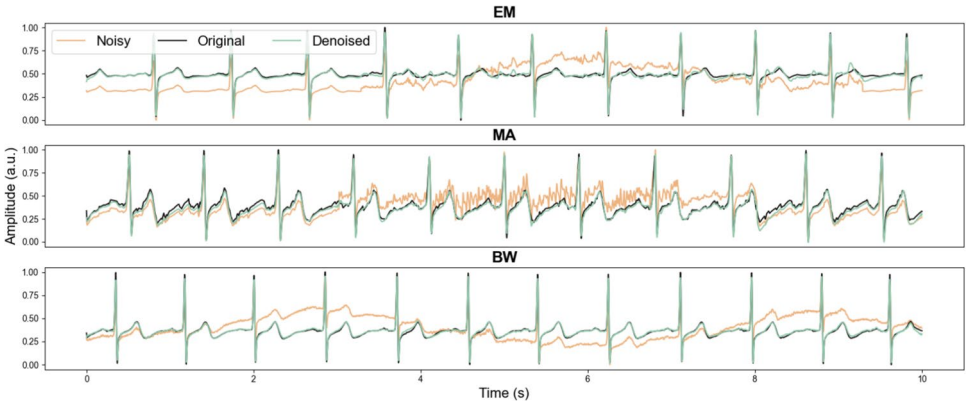

Fig. 7 Denoiser performance on samples from the PTB-XL test set: electrode motion (EM), muscle artifact (MA), and baseline wander (BW). Right: samples corrupted with different combinations of noise

hidden units with 0 dropout rate. This model has 26,121 parameters and achieved the best performance on the validation set after training for 58 epochs. Figure 6 illustrates the evolution of the model performance throughout the training phase: after the first epoch, the model outputs a timeseries that is almost entirely the same as the input; after 10 epochs it gets closer to the clean signal and after training for 58 epochs the noise is essentially removed from the input signal. Figures 7 and 8 show some examples of the performance of the final model on different signals from the test set, corrupted with each type of noise and combination of different noises, respectively.

Table 3 demonstrates the mean results over all test signals according to the SNR_{in}. When higher amounts of noise are added to the input (i.e., lower SNR_{in}) the SNR_{imp} is higher. However, PRD and RMSE increase, indicating that the output has a higher denoising error (Fig. 8).

Regarding the model's performance on different types of noise, described in Table 4, the results show that the model is most and least effective in cleaning BW and EM, respectively.

In Fig. 9 we compare our model's performance with two traditional denoising methods: the *Neurokit* filter (a 0.5 Hz high-pass butterworth filter of 5th order, followed by powerline filtering) and the Pan-Tompkins algorithm. These denoisers are open source programs which are publicly available in the Neurokit Python Toolbox [30]. The RMSE is 0.029, 0.393, and 0.405 when using the proposed GRU model, the Neurokit and the Pan-Tompkins denoisers, respectively, which shows that our model clearly outperforms these traditional denoisers.

Part 2: Field Application

As explained in "Pipeline for ECG Denoising" section, the results from the data acquired in the industrial work setting were obtained by segmenting the input signals. This was done to decrease the runtime of the model pipeline. Figure 10 reveals the generated output when performing versus not performing segmentation around an example of a segmentation limit, showing that the results are similar, having little influence in the signal morphology around the segments limits. In the subsequent figures of the present section, all the denoised signals correspond to outputs obtained using the segmentation step.

The peak detection process, before and after denoising the signal, is represented in Fig. 11. This image illustrates the presence of outliers in the BPM signal, which was computed from the distance between each two consecutive peaks divided by the sampling frequency, 360 Hz, multiplied by 60 s/min. The peak detection performance is extremely improved after applying the proposed algorithm.

Figure 12 exhibits an example of the output of the model in comparison with the noisy input signal. In Table 5, we present the mean results across all 43 subjects, including the counts of missing and wrongly detected peaks (MP and WP, respectively) before and after denoising, along with the corresponding relative improvements (%Imp) achieved after signal cleaning using the proposed algorithm. These results distinguish between outcomes obtained before and after the implementation of the segmentation step. On average, more than 70% of the missing peaks are correctly detected after denoising the signal. Concerning WP, more than 50% of the peaks that were erroneously detected are corrected. These results demonstrate that the efficacy of the model does not change considerably due to the segmentation step. Nevertheless, performing signal segmentation significantly reduced processing time by about 50% for 10-min sequences.

Discussion

Regarding the model's performance on different types of noise, the results from Table 4 show that the model is most and least effective in cleaning BW and EM, respectively,

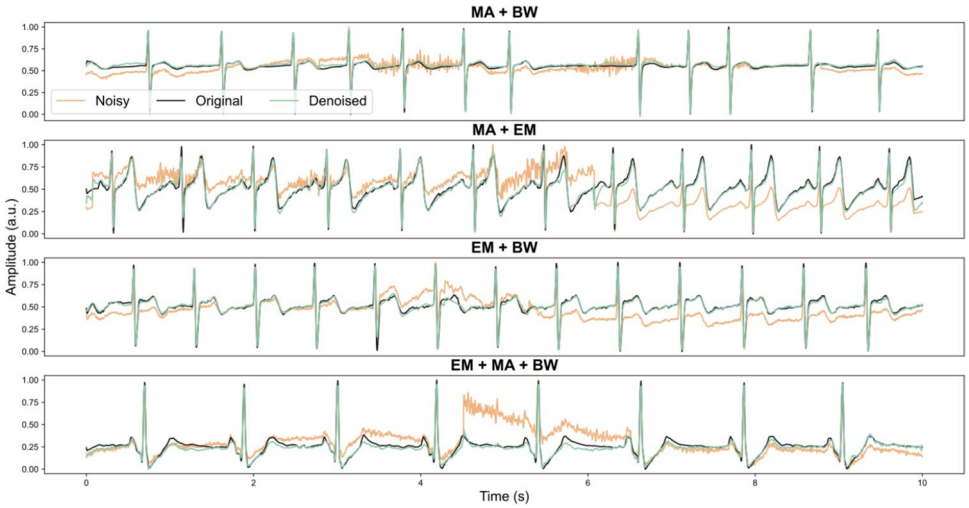

Fig. 8 Denoiser performance on samples from the PTB-XL test set: samples corrupted with different combinations of noise

Table 3 Mean results obtained (SNR_{imp}, PRD, and RMSE) for each SNR_{in}

SNR_{in} (dB)	SNR_{imp} (dB)	PRD (%)	RMSE
0	21.7	8.9	0.041
5	19.5	6.5	0.029
7	18.6	5.7	0.025
10	16.5	5.4	0.023

Table 4 Results obtained (SNR_{imp}, PRD, and RMSE) per noise type

Noise Type	SNR_{imp} (dB)	PRD (%)	RMSE
MA	19.2	6.4	0.029
EM	18.2	7.0	0.032
BW	20.6	5.5	0.024
MA + BW	19.4	6.5	0.027
EM + BW	19.5	6.1	0.027
MA + EM	19.2	6.4	0.029
MA + EM + BW	18.9	6.6	0.029

which is in accordance with results found in the literature. Unexpectedly, the model's performance does not necessarily decline when more than one type of noise corrupts the data. For example, when denoising MA+BW noise, the RMSE is lower than when removing just MA noise and higher than when removing just BW noise. This can be interpreted as if the denoiser removes each type of noise independently and, for the same amount of SNR_{in}, when multiple noise sources are present, the error is around the mean of the obtained error for each type of noise alone. Overall, the model is able to efficiently remove a great part of the noise from the noisy ECG signals, achieving SNR_{imp} values of up to 30.6, 28.5, 26.5, and 24.2 dB on records with SNR_{in} of 0, 5, 7 and 10 dB, respectively.

Regarding employment of the biGRU model to the data acquired in a real industrial setting, the results were extremely satisfying. The are two main reasons that make these results impressive. The first is the fact that the new data in which the model was tested differed greatly from the data with which it was trained: different environment (clinical

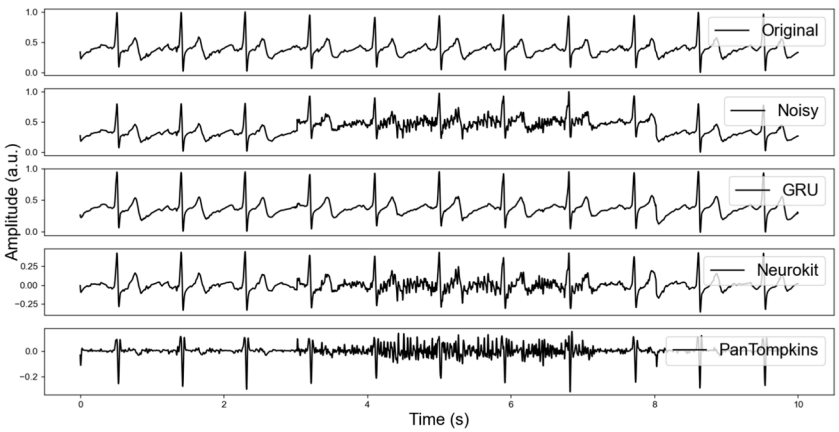

Fig. 9 Comparative results of the proposed denoiser (GRU) and two denoisers publicly available through the *Neurokit* Python toolbox [30]: *Neurokit* and Pan-Tompkins

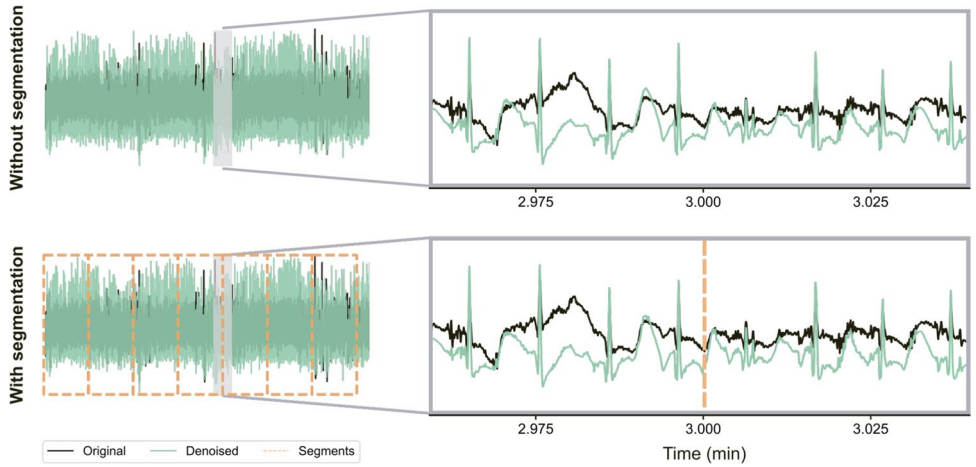

Fig. 10 Comparison of the output signal around a segment limit when the segmentation of the signal was not performed (top) versus when it was (bottom)

setting versus factory), different protocol (steady versus in movement), and different electrode placement in the body (which alters the morphology of the signal). The second is that the workers were performing a variety of upper body work tasks (and some of them were very intense) which inherently results in a great amount of muscle activation noise summing to other noise sources from the workplace context (previously unseen by the model). The fact that the model was able to properly remove the main noise components of the signal (as it is visible in Fig. 11, for example) shows that it has a good generalization power and that it is not over-fitted neither for the PTB-XL ECG nor for the MIT-BIH NST noise sources.

Given that we are the first to use the PTB-XL database together with MIT-BIH Noise Stress database as basis for developing a denoiser to clean MA, EM and BW noise, there

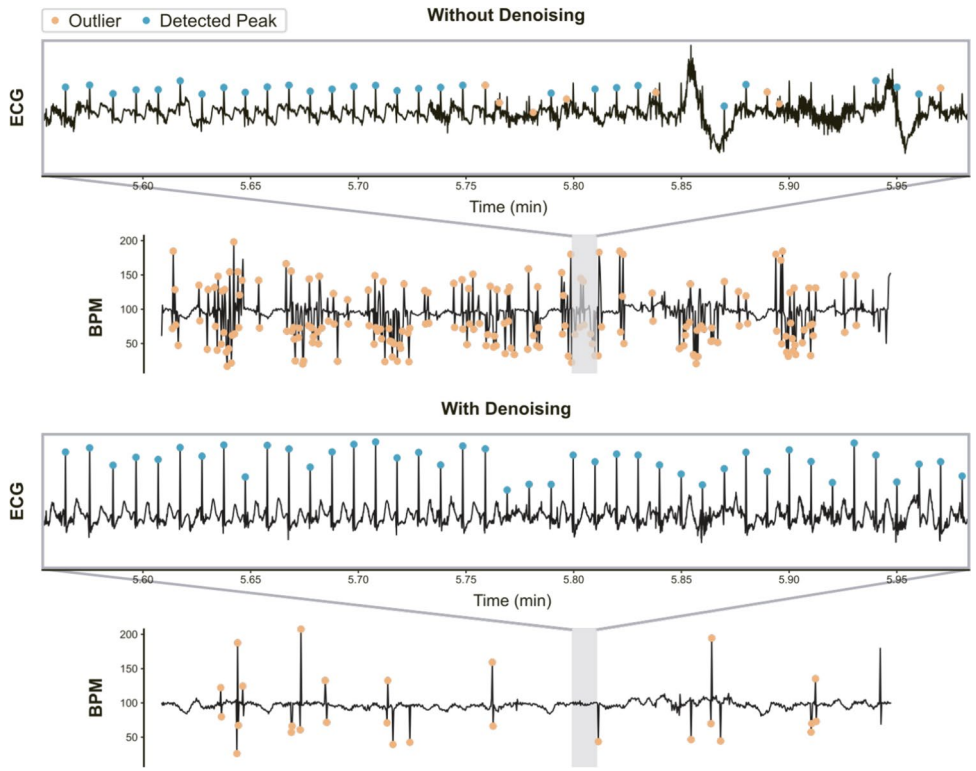

Fig. 11 Presence of outliers in the beats per minute (BPM) plot, from the peak detection process in the ECG signal, before and after signal denoising

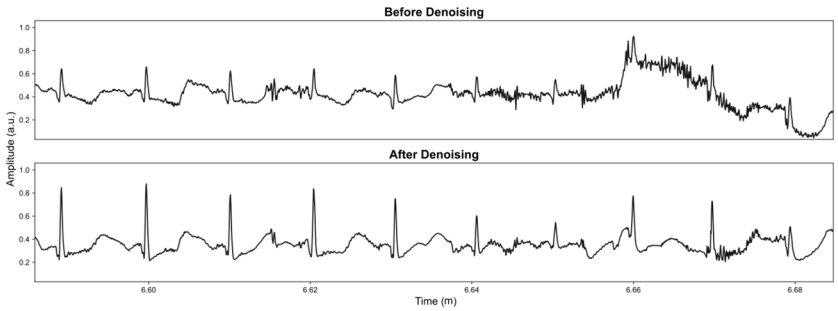

Fig. 12 Example of the result obtained when applying the proposed model to a noisy input signal (shown in the upper plot) from the factory setting collection

Table 5 Results obtained in R-peak detection before and after applying the proposed model

	Before		After (no segmentation)		After (with segmentation)	
	MP	WP	MP	WP	MP	WP
Mean ± std	60 ± 50.7	36.2 ± 28.5	17.0 ± 18.0	16.2 ± 17.2	16.7 ± 16.0	16.5 ± 17.5
% Imp	–	–	71.7	55.2	72.2	54.4

Two different approaches (with and without segmentation) are compared with the results prior to cleaning the signals acquired in an industrial setting

MP and WP denote Missing Peaks and Wrongly detected Peaks, respectively. % Imp denotes the relative improvement of the corresponding metric. Std stands for standard deviation

are no benchmarks available from the literature to which our results can be directly compared. For this reason, we opted for: (i) comparing our models to traditional denoising approaches and (ii) performing a qualitative comparison with state-of-the-art models. From Fig. 9, it is observable that our model's performance significantly surpasses widely used publicly available tools for ECG denoising. Regarding the state-of-the-art approaches reported in the literature and described in "Related Work" section, we undertook a qualitative comparison, summarized in Table 6.

As already mentioned in the "Related Work" section, in most studies the models were trained using the MIT-BIH Arrhythmia database. Apart from this one the other databases that were used are: the MGH Waveform database, with data from 250 patients [32], the QT database, which included 105 recordings [33], the PTB Diagnostic, from 290 subjects [34], and the MIT-BIH Apnea, consisting of 18 records [35]. Compared to the size and inherent diversity of PTB-XL, the mentioned databases lack representative power. Regarding the noise addition step that is common to all approaches, limiting the training and testing of the model to the presence of gaussian white noise (as in [8, 9, 18]) is very restrictive. In [7], the authors perform tests on one record from the MGH database, which contains real noisy signals (ie, there was not the need of artificially adding noise to it); however, the only paper that also performed tests on a non-clinical environment was [31]. Even so, the tests performed in the present study go a step further when the data collection is performed in a non-controlled environment, where the researchers had no interference on the tasks that were being performed.

In terms of computational complexity, our model has a parameter count of 2.6×10^4. Among the related works, from our estimates, the DAE from [8] has an even lower parameter count; however, its training and testing were limited to Gaussian noise, significantly limiting its adaptability to other prevalent sources of noise. Our proposed model outperforms the remaining by at least a tenfold reduction in the number of parameters when compared to other studies. This advantage is critical, particularly for reducing testing time, especially on lightweight devices, where computational resources are constrained. It's worth noting that most of the existing models are not publicly accessible, rendering testing and utilization impossible. The FFNN model [7] and DeepRTSNet [31] are publicly accessible. While the former provides the code but not the trained model, the latter requires a paid subscription for access.

Conclusions and Future Work

In this work, a simple GRU-based DL architecture is proposed to execute the task of ECG denoising. For the first time, a DL denoiser was trained and tested on the PTB-XL

Table 6 Overview of key attributes of main related work

Model	Parameters	ECG database	Noise added	Real noisy ECG tests	Open-source
FFNN [7]	3,003,000[a]	MIT-BIH Arrhythmia; MGH; QT[b]	EM[c]	Yes	Yes
DAE [8]	12,220[a]	PTB diagnostic	GWN	No	No
CNN; LSTM [9]	8,764,758; 21,266,320[a]	Synthetic data; MIT-BIH arrhythmia	Synthetic random noise; EM[c]	No	No
LSTM [10]	9,915,352[a]	Synthetic data; PTB diagnostic	GWN	No	No
CDAE-LSTM [18]	10,919,980	MIT-BIH arrhythmia	GWN	No	No
GAN [11]	281,060	MIT-BIH arrhythmia	EM, MA, BW[c]	No	No
DeepRTSNet [31]	346,034	MIT-BIH Apnea	EM, MA, BW[c]	Yes	Yes
biGRU	26,121	PTB-XL	EM, MA, BW[c]	Yes	Yes

Acronyms: FFNN: Feed Forward Neural Network; DAE: Denoising Auto-Encoder; CNN: Convolutional Neural Network; LSTM: Long Short-Term Memory; CDAE: Convolutional Denoising Auto-Encoder; GAN: Generative Adversarial Network; DeepRTSNET: Deep Robust Two-Stage Network; biGRU: bidirectional Gated Recurrent Units; GWN: Gaussian White Noise; EM: Electrode Motion; MA: Muscle Activation; BW: Baseline Wander

[a]Estimated based on the provided information (minimum estimate)

[b]MGH: record mgh124; QT: record sele0106

[c]Noise data from MIT-BIH Noise Stress Database

database for the task of removing MA, EM and BW noise from ECG signals. The developed approach cleans noisy ECG signals and outputs signals with good quality, while simultaneously preserving the trace of abnormal signals of patients. Furthermore, the model's robustness for application in real-world scenarios was tested on a large data collection that was performed in an industrial setting, during work. The results obtained from the new dataset emphasize the model's efficacy and its generalization power, further reinforcing the notion of integrating the biGRU denoiser into a comprehensive framework for monitoring the occupational health of workers through wearable data.

As part of our future work, we plan to enhance the denoising pipeline by including an ECG quality assessment tool. This tool will be responsible for identifying and selecting segments of the signal requiring cleaning, ensuring that the denoiser is applied only where necessary. Additionally, we will investigate the influence of segment size on the segmentation step, aiming to further improve the model's efficacy and efficiency.

Limitations

The primary limitation of this study pertains to the R-peak detection tool utilized, as it is not infallible. Nonetheless, we contend that the impact of this limitation is mitigated. Also, it's worth noting that the clean signals used for model training are not entirely devoid of noise (although the database is of very good quality); in the context of our study, the metrics employed in the initial phase compute the divergence between the provided output and the original signal. Consequently, an output that surpasses the original signal in terms of cleanliness might erroneously receive a lower performance score. Another limitation pertains to the absence of a ground-truth reference for evaluating the denoiser's performance on real-world noisy data. Nevertheless, it's worth noting that this challenge extends to the evaluation of most machine learning models, as real-world scenarios often lack definitive ground-truth references. Addressing this issue is imperative to effectively assess the model's generalization capabilities.

Funding Open access funding provided by FCT|FCCN (b-on). This work was supported by Project OPERATOR (NORTE01-0247-FEDER-045910), cofinanced by the European Regional Development Fund through the North Portugal Regional Operational Program and Lisbon Regional Operational Program and by the Portuguese Foundation for Science and Technology, under the MIT Portugal Program. M. Dias and P. Probst were supported by the doctoral Grants SFRH/BD/151375/2021 and RT/BD/152843/2021, respectively, financed by the Portuguese Foundation for Science and Technology (FCT), and with funds from State Budget, under the MIT Portugal Program.

Data Availability The developed code repository was made open-source and it is available at https://github.com/marianaagdias/ECG_Denoiser.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

Consent to Participate All participants gave their informed consent for inclusion before they participated in the study.

Ethics Approval The study was conducted in accordance with the Declaration of Helsinki, and OPERATOR projected procedures were reviewed by the Ethics Committee of University of Porto, as partner of the project, under the Approval Number reference 2020/09-6.

References

1. Li HZ, Boulanger P. A survey of heart anomaly detection using ambulatory electrocardiogram (ECG). Sensors (Basel). 2020. https://doi.org/10.3390/S20051461.
2. Khan MG. Rapid ECG interpretation. New York: Springer; 2008.
3. Cosoli G, Spinsante S, Scardulla F, D'Acquisto L, Scalise L. Wireless ECG and cardiac monitoring systems: state of the art, available commercial devices and useful electronic components. Measurement. 2021;177:109243.
4. Joshi SL, Vatti RA, Tornekar RV. A survey on ECG signal denoising techniques. In: Proceedings - 2013 international conference on communication systems and network technologies, CSNT 2013. 2013; p. 60–4. https://doi.org/10.1109/CSNT.2013.22.
5. Chatterjee S, Thakur RS, Yadav RN, Gupta L, Raghuvanshi DK. Review of noise removal techniques in ECG signals. IET Signal Process. 2020;14:569–90. https://doi.org/10.1049/IET-SPR.2020.0104.
6. Tracey BH, Miller EL. Nonlocal means denoising of ECG signals. IEEE Trans Biomed Eng. 2012;59(9):2383–6.
7. Rodrigues R, Couto P. A neural network approach to ECG denoising; 2012. https://doi.org/10.48550/arXiv.1212.5217.
8. Marwan B, Samann F, Schaanze T. Denoising of ECG with single and multiple hidden layer autoencoders. Curr Direct Biomed Eng. 2022;8:652–5. https://doi.org/10.1515/cdbme-2022-1166.
9. Arsene CTC, Hankins R, Yin H. Deep learning models for denoising ECG signals. In: European signal processing conference 2019-September; 2019. https://doi.org/10.23919/EUSIPCO.2019.8902833.
10. Antczak K. Deep recurrent neural networks for ECG signal denoising; 2018. https://doi.org/10.48550/arxiv.1807.11551.
11. Wang J, Li R, Li R, Li K, Zeng H, Xie G, Liu L. Adversarial denoising of electrocardiogram. Neurocomputing. 2019;349:212–24. https://doi.org/10.1016/J.NEUCOM.2019.03.083.
12. Spook SM, Koolhaas W, Bültmann U, Brouwer S. Implementing sensor technology applications for workplace health promotion: a needs assessment among workers with physically demanding work. BMC Public Health. 2019;19(1):1–9.
13. Dias M, Silva L, Folgado D, Nunes ML, Cepeda C, Cheetham M, Gamboa H. Cardiovascular load assessment in the workplace: a systematic review. Int J Ind Ergonom. 2023;96:103476. https://doi.org/10.1016/j.ergon.2023.103476.
14. Adams JM. The value of worker well-being. Public Health Rep. 2019;134(6):583–6.
15. Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, Schaeffter T. Ptb-xl, a large publicly available electrocardiography dataset. Sci Data. 2020;2020(17):1–15. https://doi.org/10.1038/s41597-020-0495-6.
16. Moody GB, Muldrow W, Mark RG. A noise stress test for arrhythmia detectors. Comput Cardiol. 1984;11(3):381–4.
17. Tripathi PM, Kumar A, Komaragiri R, Kumar M. A review on computational methods for denoising and detecting ECG signals to detect cardiovascular diseases. Arch Comput Methods Eng. 2021. https://doi.org/10.1007/s11831-021-09642-2.
18. Dasan E, Panneerselvam I. A novel dimensionality reduction approach for ECG signal via convolutional denoising autoencoder with LSTM. Biomed Signal Process Control. 2021;63:102225. https://doi.org/10.1016/J.BSPC.2020.102225.
19. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag. 2001;20(3):45–50. https://doi.org/10.1109/51.932724.
20. Schmidt RM. Recurrent neural networks (RNNS): a gentle introduction and overview. Preprint; 2019. arXiv:1912.05911.
21. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncertain Fuzziness Knowl Based Syst. 1998;6(02):107–16.
22. Cho K, Merriënboer BV, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: EMNLP 2014–2014 conference on empirical methods in natural language processing, proceedings of the conference; 2014. p. 1724–34. https://doi.org/10.48550/arxiv.1406.1078.
23. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):215–20.
24. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inform Process Syst. 2019;32:1.
25. Karmakar C, Rahman S, Natgunanathan I, Yearwood J, Palaniswami M. Robustness of electrocardiogram signal quality indices; 2022. https://doi.org/10.1098/rsif.2022.0012.
26. Bijl K, Elgendi M, Menon C. Automatic ECG quality assessment techniques: a systematic review. Diagnostics. 2022;12(11):2578.
27. Yuan S, He Z, Zhao J, Yang Z, Yuan Z. Long-term electrocardiogram signal quality assessment pipeline based on a frequency-adaptive mean absolute deviation curve. Appl Intell. 2023;53(17):20418–40.
28. Carvalho D, Silva L, Carvalho M, Dias M, Costa N, Folgado D, Lua M, Gamboa H, Edelman E. Heart rate variability during repetitive work in the presence of fatigue. In: Ergonomics in design. AHFE international; 2023. 14th international conference on applied human factors and ergonomics (AHFE 2023).
29. Carvalho D, Silva L, Carvalho M, Dias M, Costa N, Folgado D, Nunes ML, Gamboa H, Andza K, Edelman E. Cardiovascular reactivity (CVR) during repetitive work in the presence of fatigue. Int Hum Syst Integr (IHSI 2023) Integr People Intell Syst. 2023. https://doi.org/10.54941/ahfe1002833.
30. Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SHA. NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav Res Methods. 2021;53(4):1689–96. https://doi.org/10.3758/s13428-020-01516-y.
31. Aghaomidi P, Mohammadisarab A, Mazloum J, Akbarzadeh MA, Orooji M, Mokari N, Yanikomeroglu H. Deeprtsnet: deep robust two-stage networks for ECG denoising in practical use case. IEEE

Access. 2022;10:128232–49. https://doi.org/10.1109/ACCESS.2022.3225899.
32. Welch J, Ford P, Teplick R, Rubsamen R, et al. The Massachusetts general hospital-Marquette foundation hemodynamic and electrocardiographic database-comprehensive collection of critical care waveforms. Clin Monit. 1991;7(1):96–7.
33. Laguna P, Mark RG, Goldberg A, Moody GB. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. In: Computers in cardiology. IEEE; 1997, p 673–6.
34. Bousseljot R, Kreiseler D, Schnabel A. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet; 1995.
35. Ichimaru Y, Moody G. Development of the polysomnographic database on CD-ROM. Psychiatry Clin Neurosci. 1999;53(2):175–7.