KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Nick Seeuws, Maarten De Vos, Alexander Bertrand **Electrocardiogram Quality Assessment using Unsupervised Deep Learning** IEEE Transaction on Biomedical Engineering, 2021
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	N/A
Journal homepage	https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=10.
Author contact	nick[dot]seeuws[at]esat[dot]kuleuven[dot]be
Abstract	*Objective:* Noise and disturbances hinder effective interpretation of recorded ECG. To identify the clean parts of a recording, various quality indicators have been developed. Previous instances of these indicators focus on human-defined desirable properties of a clean signal. The reliance on human-specified properties places an inherent limitation on the potential power of signal quality indicators. To move away from this limitation, we propose a data-driven quality indicator. *Methods:* We use an unsupervised deep learning model, the auto-encoder, to derive the quality indicator. For different quality assessment settings we compare the performance of our quality indicator with traditional indicators. *Results:* The data-driven method performs consistently strong across tasks while performance of the traditional indicators varies strongly from task to task. *Conclusion:* This strong performance indicates the potential of data-driven quality indicators for use in ECG processing, removing the reliance on expert-specified desirable properties; *Significance:* The proposed methodology can easily be extended towards learning quality indicators in other data modalities.
IR	https://lirias.kuleuven.be/handle/123456789/xxxxxx

(article begins on next page)

Electrocardiogram Quality Assessment using Unsupervised Deep Learning

Nick Seeuws, Maarten De Vos, and Alexander Bertrand

***Abstract*— *Objective:* Noise and disturbances hinder effective interpretation of recorded ECG. To identify the clean parts of a recording, free from such disturbances, various quality indicators have been developed. Previous instances of these indicators focus on human-defined desirable properties of a clean signal. The reliance on human-specified properties places an inherent limitation on the potential power of signal quality indicators. To move away from this limitation, we propose a data-driven quality indicator. *Methods:* We use an unsupervised deep learning model, the auto-encoder, to derive the quality indicator. For different quality assessment settings we compare the performance of our quality indicator with traditional indicators. *Results:* The data-driven method performs consistently strong across tasks while performance of the traditional indicators varies strongly from task to task. *Conclusion:* This strong performance indicates the potential of data-driven quality indicators for use in ECG processing, removing the reliance on expert-specified desirable properties. *Significance:* The proposed methodology can easily be extended towards learning quality indicators in other data modalities.**

***Index Terms*— Electrocardiogram (ECG), signal quality, unsupervised learning**

I. Introduction

The electrocardiogram (ECG) is an essential tool for cardiologists. When diagnosing numerous cardiac disorders, cardiologists rely on this signal to get an objective impression of the condition of the heart. In most applications, a clean signal is a must for accurate interpretation of the ECG[1]. Capturing the ECG, however, is prone to various measurement artefacts. For example, muscle activity, electrode movement, breathing artefacts, or power line interference are common sources of such disturbances. In wearable ECG sensors for ECG monitoring in daily life, such artefacts are even more notorious, as they appear more frequently and often with a much stronger impact on the recorded ECG signal.

Different kinds of disturbances affect use cases of ECG in a different way. One can intuitively see that estimating the heart rate, for example, places less strict conditions on the quality of a signal than fine-grained analysis of ECG waveforms like, for example, the detection of atrial fibrillation. The former relies on R-peaks which, due to their high amplitude, do not easily get buried under noise. The latter requires much more detail in the signal and the slightest amount of noise can make segments unusable.

Several methods have been proposed to automatically indicate the quality level of ECG recordings. These Signal Quality Indicators (SQIs) measure the disturbance level and quantify the fit-for-purpose of the signal. In the past, they were inspired by human-defined properties of a clean signal such as skewness, kurtosis, power in certain frequency bands and many more[2]–[8]. More recently, machine learning based SQIs have been proposed[9], [10]. These SQIs measure features of the signal, often inspired by previously developed SQIs, and use machine learning to predict the quality level based on measured features. These machine learning models are trained using labels provided by humans. This reliance on human effort (both for feature design as well as for providing labels) creates a severe drawback to the machine learning SQIs. The labeling effort can be substantial for data-hungry models and requires a specific definition of quality levels to ensure consistent labeling.

Detecting noise and artefacts in data has also been a popular topic in the general field of data mining and machine learning research. In this field, it is more commonly known as outlier or anomaly detection[11]. Unsupervised anomaly detection is a specific branch of such algorithms where anomalies are detected without relying on human labels, for which auto-encoders are a popular class of models[12]–[16]. This class of models can automatically identify the important patterns in a given data set. Auto-encoders learn to reconstruct data and error measures on input data and their reconstructions can then be used to detect anomalies.

In this study, we investigate the performance of modifying such auto-encoders for unsupervised anomaly detection towards the task of unsupervised (task-agnostic) quality assessment. Unsupervised quality assessment does not rely on human input, i.e., on human-defined signal properties or human-provided labels. Additionally, unsupervised quality assessment involves a new, implicit definition of signal quality linked to

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 802895) and from the Flemish Government (AI Research Program).

N. Seeuws (nick.seeuws@esat.kuleuven.be), A. Bertrand and M. De Vos are with the Dept. of Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics (STADIUS), KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

M. De Vos is also with the Dept. of Development and Regeneration, Faculty of Medicine, KU Leuven

A. Bertrand and N. Seeuws are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

Copyright © 2021 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purpose must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org

how well a signal segment can be embedded into a lower-dimensional space. We test how well this quality definition matches with various ECG quality measures. We define two SQIs that make use of an auto-encoder trained on ECG data. For comparison, we use several classical SQIs that also do not rely on expert-provided labels. We investigate performance on two dimensions of quality assessment: detection and quantification[9]. In detection, or binary quality scoring, one aims to make a clear distinction between "good" or "bad, "clean" or "noisy." A binary decision has to be made about the usability of a given signal. Quantification is a more continuous approach to quality assessment where one tries to identify specific quality levels. This second approach can be more suited when one has to cope with varying quality needs.

The outline of this paper is as follows. In Section II we introduce the proposed machine learning model and quality indicators. We also present the experimental methodology. In Section III we show results of the experiments. In Section IV we discuss our results, underlying model assumptions and some additional remarks. With Section V we conclude the paper.

II. Methods

A. Model

1) Auto-encoder: Auto-encoders learn a data model by mapping inputs to a new representation and back to the original input space. They encode an input, $\boldsymbol{x}$, into a learned representation $\boldsymbol{z}$ with a function $f(\cdot)$, parameterized using a deep neural network, as $\boldsymbol{z} = f(\boldsymbol{x})$. In classical auto-encoders, another function $g(\cdot)$ decodes the representation back to the original input space. The full auto-encoder $r(\cdot)$ combines the encoder and decoder to compute a reconstruction $\hat{\boldsymbol{x}}$ of an input $\boldsymbol{x}$ after passing through the *latent* representation as $\hat{\boldsymbol{x}} = g(f(\boldsymbol{x})) = r(\boldsymbol{x})$. Auto-encoders are traditionally trained by improving their reconstructions over a training set.

The most common way of quantifying reconstruction errors is the mean squared error between the original data and the reconstructions. Penalizing absolute errors with the same weight in entire ECG reconstructions is, however, undesirable. One can imagine an error in reconstructing the R-peak to impact reconstruction quality much less than the same magnitude of error in the isoelectric line.

To take this issue into account, our model makes use of a decoder extension and changes the training objective, similar to the variational auto-encoder[17], by defining a distribution over reconstructions given the latent representation of a signal segment. Training the model similar to maximum likelihood estimation then gives a natural way of coping with the concern over absolute errors. The reconstruction distribution is defined as a multivariate Gaussian distribution with diagonal covariance. It is parameterized by a mean and standard deviation vector which have the same dimension as the input vector $\boldsymbol{x}$, and which are both functions of the learned representation of an input. The full model is defined as

$$\boldsymbol{\mu}(\boldsymbol{x}) = m(f(\boldsymbol{x}))$$

$$\boldsymbol{\sigma}(\boldsymbol{x}) = s(f(\boldsymbol{x}))$$

where $\boldsymbol{\mu}(\boldsymbol{x})$ and $\boldsymbol{\sigma}(\boldsymbol{x})$ define the reconstruction distribution's parameters for a specific input $\boldsymbol{x}$. In a maximum likelihood setting, $\boldsymbol{\mu}(\boldsymbol{x})$ takes the role of the reconstructed vector $\hat{\boldsymbol{x}}$, whereas $\boldsymbol{\sigma}(\boldsymbol{x})$ can be viewed as an uncertainty on the reconstructed vector $\hat{\boldsymbol{x}}$ (quantified per entry in the vector). The dependence on $\boldsymbol{x}$ for these parameters will be dropped in the remainder for legibility. The training objective, formulated as

$$\max \sum_{n=1}^{N} \log p(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}^{(n)}, \boldsymbol{\sigma}^{(n)}),$$

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{l=1}^{L} \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{(x_l - \mu_l)^2}{2\sigma_l^2}\right)$$

for a batch of N segments each of length L, measures the log likelihood of an input under the Gaussian reconstruction distribution. By defining such a distribution, we introduce a scale variable, $\boldsymbol{\sigma}$, in addition to a most likely reconstruction $\boldsymbol{\mu}$. Using this scale variable, the model can indicate where it is certain about a reconstruction or where some uncertainty exists about a precise magnitude and location, like in an R-peak. Lowering the values of $\boldsymbol{\sigma}$ can easily increase the likelihood of a signal segment, but mismatches between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ will be more severely penalized. Note that both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ depend on $\boldsymbol{x}$, and are automatically learned by the network, i.e., the model has to learn when it is safe to aim for increasing the likelihood (by reducing $\boldsymbol{\sigma}$) and when it will likely make some mistakes (and should increase $\boldsymbol{\sigma}$).

One can influence the behavior of an auto-encoder through specific constraints in the training process or the architecture. Our architecture contains such a constraint, commonly called a bottleneck. By reducing the dimensionality of the representation space compared to the input space, the auto-encoder performs data compression with the encoder. This compression forces the model to remove redundancy and focus on the most important patterns in training data to be able to accurately represent and reconstruct inputs.

2) Architecture: The architecture for our proposed network is adapted from [18], which is a fully convolutional auto-encoder for ECG signals and, as mentioned above, uses a Gaussian output inspired by [17]. The auto-encoder is trained on input segments with 1024 samples of an ECG signal sampled at 200 Hz.[1] Note that only the architecture of [18] is used as part of our approach. The training procedure of [18] involved a denoising task, whereas we aim to learn patterns directly from raw signals, both clean and noisy. Only working with the raw signals is a more general and more practical training procedure compared to [18], since we do not require a database of noise signals that are representative for the signals under test. Nevertheless, although we opted for this more general training procedure, it is in principle also possible to train our models in a more supervised fashion as in [18].

The model makes use of temporal convolutions with an Exponential Linear Unit (ELU) as non-linearity. Every kernel operates on 16 samples along the temporal axis and all features along the feature axis. A batch normalization layer [19] is

[1]If the data set is not sampled at 200 Hz, a resampling operation has to be performed if the same network dimensions are kept.

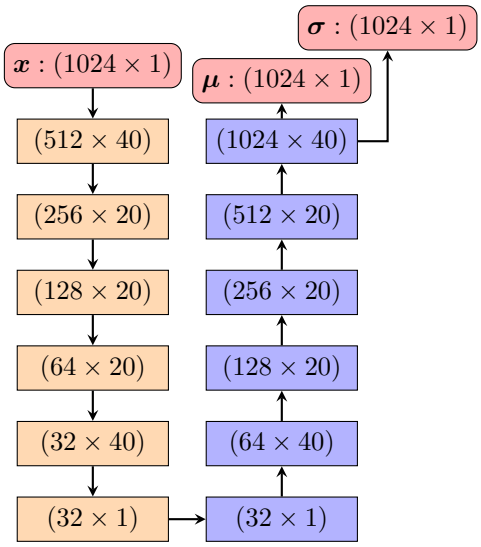


Fig. 1: Architecture of the auto-encoder with the encoder on the left and the decoder on the right. Boxes indicate intermediate data tensors of size (#Timesteps×#Features) with arrows indicating a convolution layer (details in text).

placed between the convolution operation and activation. The final layer that outputs $\boldsymbol{\mu}$ uses no non-linearity and the $\boldsymbol{\sigma}$ output makes use of a softplus non-linearity to ensure positive values. The encoder performs strided convolutions throughout to downsample the signal. The decoder uses transposed convolutions throughout for upsampling. All layers use zero-padding to ensure only the convolution strides influence the intermediate output dimensions. Figure 1 shows the complete architecture. Note that the two decoder functions $m(\cdot)$ and $s(\cdot)$ share a large part of the decoder path.

3) Training: In all experiments in this paper, the auto-encoder is trained on segments of 1024 ECG samples sampled at 200 Hz. These segments are taken from training recordings with a stride of 50 samples. A single epoch contains all such segments from the training set. During training, the loss on a separate validation set is tracked and the parameter values with the best validation loss are retained. Training is carried out for 200 epochs, which was a safe value to ensure convergence across the evaluation tasks. The models are trained using the Adam optimizer with 0.001 as learning rate. As mentioned before, a maximum likelihood-inspired training objective is used. Details on the data sets and how they are used will be described further in Subsections II-C.2 and II-C.3.

B. Quality Scoring

1) Signal Quality Indicators: Quantifying errors in the reconstruction of a new signal segment is linked to measuring signal quality. Signal quality is assumed to be good when the auto-encoder can cleanly reconstruct a new segment and poor when the auto-encoder fails to reconstruct the signal (see figure 2).

Two methods for quantifying errors are investigated in the remainder of this paper:

- The first indicator calculates the logarithm of the mean squared error (MSE) between the signal segment and the $\boldsymbol{\mu}$ vector of the output. This vector takes the role of the reconstruction in a classical auto-encoder. The indicator is calculated as follows:

$$\text{AE-logMSE}(\boldsymbol{x}) = \log \frac{1}{L} \sum_{l=1}^{L} (x_l - \mu_l(\boldsymbol{x}))^2$$

  with L the length of the signal and the subscript l indicating the l-th entry of the full vectors. The logarithm is taken to rescale the indicator values. AE-logMSE quantifies the reconstruction error; large values of AE-logMSE indicate poor signal quality and small values indicate higher signal quality.
- The second indicator makes use of the log-likelihood (LLH) values obtained from the auto-encoder output. These values are calculated at every sample in the signal and averaged over the total length of the signal. It is therefore more closely linked to the training objective than the AE-logMSE measure. The full indicator is calculated as

$$\text{AE-LLH}(\boldsymbol{x}) = \frac{1}{L} \sum_{l=1}^{L} \log p(x_l; \mu_l(\boldsymbol{x}), \sigma_l(\boldsymbol{x})).$$

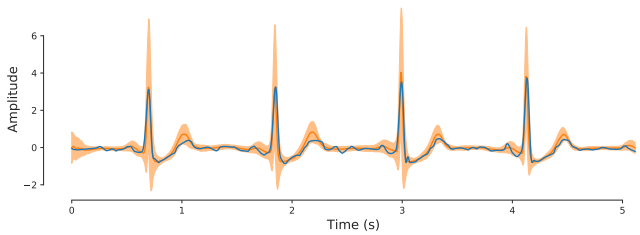
  A large value of AE-LLH requires both a good reconstruction ($\boldsymbol{\mu}$ close to $\boldsymbol{x}$) and high confidence (small $\boldsymbol{\sigma}$) of the auto-encoder in its reconstruction, indicating a segment of higher quality. A small value of AE-LLH is linked with the opposite, a bad reconstruction and/or too much confidence of the model in a relatively faulty reconstruction, indicating poorer quality.

Both AE-logMSE and AE-LLH contain the squared difference $|\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{x})|_2^2$ in their computation. The key difference between them is that AE-LLH weighs the separate elements of this difference using the additional network output $\boldsymbol{\sigma}$.

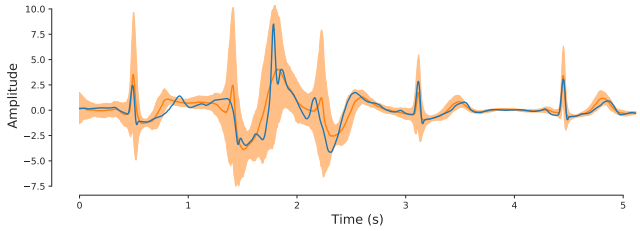
2) Time resolution: Both AE-logMSE and AE-LLH contain a contribution for every time sample of the signal, thereby allowing to quantify the quality of each and every time sample of the ECG. This can be viewed as a quality 'signal' sampled at 200 Hz. However, at this resolution the quality signal is very noisy. After smoothing it with a moving average filter with a length of 100 samples (0.5 s) a clearer quality signal arises which is illustrated in figure 3. This demonstrates that the quality indicator can be evaluated at multiple, potentially fine-grained time scales (depending on the application).

3) Edge Effect: Auto-encoders making use of convolutional layers struggle with reconstructing the edges of their inputs. These convolutional layers can work with a full context window in the center of a signal segment but cannot "see" beyond the segment's edges.

Two methods to cope with this edge effect are employed. Firstly, we ignore the first half-second and final quarter-second of a segment when computing the quality indicators (due to potential overlap in segments, the discarded parts can still be assessed in earlier/later segments). Secondly, it is noted that the auto-encoder can process larger segments for assessing quality than the ones used to train the auto-encoder. Indeed, every layer of the auto-encoder performs a filtering operation, which does not rely on a specific input length. As long as the duration of the segment being processed agrees with the



(a) Clean signal



(b) Dirty signal

Fig. 2: Examples of reconstructions with the original signal $\boldsymbol{x}$ in blue and reconstruction $\boldsymbol{\mu}(\boldsymbol{x})$ in orange with bands showing $\boldsymbol{\mu}(\boldsymbol{x}) \pm 2\boldsymbol{\sigma}(\boldsymbol{x})$. Figure 2a shows a clean signal with a good reconstruction having most probability mass tightly fit around the original. Figure 2b shows a segment of poor quality where the probability mass is clearly more spread out and the mean reconstruction does not track the signal as well as the clean signal.

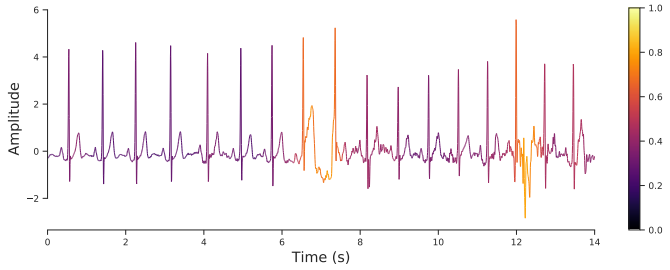


Fig. 3: Illustration of a quality signal. The input ECG is color-coded with the quality level making use of AE-logMSE without a logarithmic transformation. Darker signal parts indicate a small squared error and lighter parts indicate a larger error. The auto-encoder can report quality information at the scale of single beats.

down- and upsampling operations, the filters in the auto-encoder allow us to process segments of indeterminate length, thereby allowing to reduce the number of affected samples due to being close to an edge. The model requires the amount of time samples in a segment to be a multiple of 32 to allow the down- and upsampling to function as intended. In total, for a new recording to be processed, the largest subsegment with a number of samples that is a multiple of 32 is fed through a trained auto-encoder and the first half-second and final quarter-second are ignored for computing AE-logMSE and AE-LLH.

The specific cutoff points (the first half-second and final quarter-second) were determined empirically. On the CinC validation set (used during the training of the auto-encoder), averaged over all the ECG segments, the model showed relatively large and consistent reconstruction errors in the first half-second and final quarter-second of a segment. These are edge effects due to lack of a symmetrical surrounding context for these samples. Ignoring these parts of the signal allows the model to disregard parts of the signal we expect it to reconstruct poorly due to model structure and not due to signal quality. However, the performance impact on quality scoring of combating these edge effects was minor for our evaluations as they are only a small fraction of the total segment length.

C. Experiments

1) Benchmark indicators: To test the performance of AE-logMSE and AE-LLH, they are compared with a comprehensive suite of benchmark SQIs that are commonly used in the literature.

a) Kurtosis: Kurtosis was proposed in [2] and later used in multiple works [3]–[6]. A clean ECG is expected to not show high variance while containing large outlier values due to the R-peaks leading to a high value for the sample kurtosis. Therefore, larger values for the kurtosis of a signal are linked with higher quality.

b) Skewness: Another commonly used quality indicator based on the statistics of the signal is the normalized third-order moment (often referred to as the skewness of the distribution). A clean signal is expected to show high skewness due to the QRS complex[3], [4].

c) IOR: The in-band to out-band spectral power ratio (IOR) is a quality index based on frequency information [6], [7]. IOR assumes that the power of a clean signal is mostly contained in the 5-40 Hz band. It is calculated as

$$\mathrm{IOR} = \frac{\int_{5\ \mathrm{Hz}}^{40\ \mathrm{Hz}} P(f)\,\mathrm{d}f}{\int_{0\ \mathrm{Hz}}^{100\ \mathrm{Hz}} P(f)\,\mathrm{d}f - \int_{5\ \mathrm{Hz}}^{40\ \mathrm{Hz}} P(f)\,\mathrm{d}f}.$$

A larger value for IOR then indicates higher signal quality.

d) pSQI: The relative power in the QRS complex (pSQI)[4], [5] is the second quality index based on frequency information. It assumes that most power resides in the 5-15 Hz band. pSQI is calculated as

$$\mathrm{pSQI} = \frac{\int_{5\ \mathrm{Hz}}^{15\ \mathrm{Hz}} P(f)\,\mathrm{d}f}{\int_{5\ \mathrm{Hz}}^{40\ \mathrm{Hz}} P(f)\,\mathrm{d}f}$$

with a larger value linked to higher signal quality.

e) basSQI: A third SQI based on frequency information measures the relative power in the baseline (i.e, the frequency content below 1 Hz.)[4] as

$$\mathrm{basSQI} = \frac{\int_{1\ \mathrm{Hz}}^{40\ \mathrm{Hz}} P(f)\,\mathrm{d}f}{\int_{0\ \mathrm{Hz}}^{40\ \mathrm{Hz}} P(f)\,\mathrm{d}f}$$

with a larger value linked to higher signal quality.

f) bSQI: The first SQI that relies on beat detection (bSQI)[4], [5] compares the results of two beat detection algorithms, *wqrs*[20] and *eplimited*[21]. The former is known to be more sensitive to noise than the latter. bSQI measures the ratio of beats *wqrs* detects that match beats *eplimited* detects, with a larger value for bSQI linked to higher signal quality.

g) pcaSQI: The second SQI that relies on beat detection uses *eplimited* to detect beats in a signal segment of interest. It measures the ratio of the sum of the five leading eigenvalues of principal component analysis on time-aligned beat waveforms to the sum of all eigenvalues (pcaSQI)[4]. A larger value is linked to higher signal quality.

2) Data: We use three different ECG data sets in order to test whether the approach generalizes well to different measurement setups and subject cohorts. Furthermore, the quality labels (used here for validation purposes only) for all data sets were produced using different criteria. This will allow us to validate whether auto-encoder based SQIs generalize well to these different criteria.

a) CinC data set: The first data set is the PhysioNet Computing in Cardiology Challenge 2017 (CinC) data set[22]. It contains short single-lead ECG recordings between 30 and 60 seconds in length. The data set was constructed with the aim of developing algorithms for detection of normal rhythms, atrial fibrillations or a general class of "other rhythms". More importantly for our work, the data set also contains labels indicating that a recording is too noisy to process and properly detect rhythms. Training and testing splits are provided by the data set creators and were kept for the purpose of our work. A full distribution of the labels can be found in table I. The recordings were sampled at 200 Hz and, for our use, preprocessed using a band-pass filter with passband between 1 and 50 Hz. The authors of [22] state that the signal was originally stored with a bandwidth of 0.5 - 40 Hz, but we noted small amounts of high-frequency noise, hence the high-frequency cutoff.

b) Sleep data set: The Sleep data set, originally intended for sleep apnea research, contains over 150 hours of single-lead ECG recordings[9]. These recordings are split into one-minute segments and provided with a quality label, indicating whether the recording contains signal artefacts or not[23]. In total, 3.2% of the recordings contained artefacts (table I). The signal was sampled at 200 Hz and pre-processed using a zero phase high- and low-pass filter with cut-off frequencies at 1 Hz and 40 Hz, respectively, following the procedure of [9]. Obvious flatline recordings were removed prior to our analysis by looking at the signal power of a recording.

c) Stress data set: The Stress data set, originally part of a database of various modalities used to capture stress levels, is made up of 2879 30-second ECG segments originally sampled at 256 Hz. In [9], the authors labeled these segments as clean or noisy depending on the visibility of R-peaks. If all the R-peaks of a segment were clearly visible, a segment was deemed clean. If not, it was deemed noisy. Around a third of the segments were classified as noisy (table I). Similar pre-processing as on the Sleep data set was applied to this Stress data set. Before use, these signals were resampled to 200 Hz. Obvious flatline recordings were removed prior to our analysis by looking at the signal power of a recording.

Out of the three data sets, CinC will be our main focus as it contains atrial fibrillations and other rhythms, making it a potentially very challenging data set for our method.

3) Training and validation: Signals of all data sets are rescaled to unit variance over each individual data set. For the CinC data set, an 80/20 split is randomly made on the recordings in the training set for training and validation of an auto-encoder. This split is stratified making use of the four class labels (normal beat, atrial fibrillation, other beat, noisy). An auto-encoder is trained on this new training set and the auto-encoder weights that produce the best loss on the validation set are retained.

TABLE I: Label distribution for the different data sets

	Normal	AF	Other	Noisy
CinC - Training	5076	758	2415	279
CinC - Testing	148	47	65	40
		Clean		Noisy
Sleep		8837		295
Stress		1935		944

For the Sleep and Stress data sets, a similar random 80/20 split is made for training and validation but there is no held-out test set (see Table II). These splits are not stratified, simulating a scenario where the user does not have any labels. Various experiments are carried out to validate different aspects and use cases of SQIs, hence the two different splits.

To evaluate and compare the newly introduced quality measures with the various benchmarks, we define three different experimental settings, as shown in Table II:

a) Generalization to held-out data: This is the traditional machine learning setting with a held-out test set. It involves training on a specific measurement setup and testing on unseen signals from a similar setup. While training happens offline, the testing can in principle be done online on streaming data. While this is the main evaluation setting, for our *binary quality scoring* experiments we also consider the two other settings.

b) Data-specific model: In this setting, all signals are available at once and performance is evaluated in an *offline* setting with a large computational budget. It involves training an auto-encoder and evaluating the SQIs on the same signals. While the approach of training an auto-encoder on a set of signals and using the same signals at the testing stage might seem strange, we do want to stress the merit of the approach. The auto-encoder is an *unsupervised* model that only uses the signal data during training (without any quality labels). Only at the testing stage the quality labels are used for evaluation purposes. This setting is relevant in, e.g., a retrospective analysis, to clean up large ECG data sets or as a pre-processing step for other machine learning algorithms on such data sets.

c) Generalization to other data sets: While information obtained from the previous two settings is valuable, they do not cover all potential use cases. The performance of SQIs is also tested under a setting where, e.g., measurement setups or patient cohort change by changing the evaluation data set.

4) Binary quality scoring: In a first validation experiment, the different quality indicators are compared on their capability in predicting binary quality labels. The CinC data set contains labels indicating whether a recording is fit for interpretation or too noisy to use. Since the recorded signals were used for the detection of atrial fibrillation, the labelling process took into account the more subtle parts of the ECG morphology. A signal should be of very high quality before it is deemed fit for interpretation. Labels in the Sleep data set indicate the

TABLE II: Overview of the evaluation settings for our experiments together with the corresponding data sets

	Training	Evaluation
Binary quality scoring		
Generalization to held-out data	CinC train set	CinC test set
Data-specific model	Sleep	Sleep
	Stress	Stress
Generalization to other data sets	CinC train set	Sleep
		Stress
Correlation with quality level		
Generalization to held-out data	CinC train set	CinC test set

presence of artefacts disturbing the signals. For the Stress data set, labels tell whether all R-peaks in a recording can be identified, which is a less strict quality requirement than for, e.g., detection of atrial fibrillation.

In this experiment, we consider all three different settings (as introduced above, see also Table II):

a) Generalization to held-out data: The CinC data set, with its pre-defined training and testing split, is used in this setting. An auto-encoder is trained on the training set and SQIs are evaluated on the held-out test set, produced from the same device and a similar subject cohort.

b) Data-specific model: The Sleep and Stress data sets are used for this setting. An auto-encoder is trained using an 80/20 training-validation split for both data sets separately, and SQIs are afterwards computed for every recording in this same set. The SQIs are then evaluated on the full data set.

c) Generalization to other data sets: The Sleep and Stress data sets are used for evaluation in this setting. An auto-encoder is trained on the CinC training set and used to compute AE-logMSE and AE-LLH for the Sleep and Stress data. The SQIs are then evaluated on these two sets.

In all these settings, the area under the ROC curve (AUC) is used to score performance, and is computed using the SQI values for the signals and their respective labels. For each result, the sampling distribution is approximated using bootstrapping. From the full evaluation set of (SQI value, label) tuples bootstrap samples are created by sampling tuples with replacement. AUC is computed for many such bootstrap sets of samples to arrive at a sampling distribution.

To test significance of the predictive power of each quality indicator (in each setting), a simple classifier based on logistic regression is used, which is merely for facilitating statistical hypothesis testing. A logistic regression model is first fit for every individual SQI. On these models the likelihood ratio test is used to determine whether the SQI shows significant power in predicting the quality labels. As a second test, a logistic regression model is fit making use of all SQIs simultaneously and, using the Wald test, backwards selection is used to determine the group of indicators that jointly best predict the quality labels. Here, one can see whether other SQIs can still significantly contribute to the quality decision of individual SQIs, i.e., whether certain SQIs capture complementary information. This analysis is carried out for all the SQI values computed in our three settings (so for the full Sleep and Stress sets, and the CinC test set). The *data-specific* and *generalization to other data sets* settings are "combined" in this evaluation for Sleep and Stress data to test whether the auto-encoders trained on the respective sets (the *data-specific* setting) or on CinC data (the *generalization to other data sets* setting) capture complementary or redundant information.

Additionally, the CinC data set allows to differentiate SQI performance for different beat types. It contains labels for a *normal rhythm* class, an *atrial fibrillation* class, and a class for *other rhythms*. With the auto-encoders automatically learning the patterns in the training set, AE-logMSE and AE-LLH run the risk of mainly learning the patterns of the majority class (normal rhythms in the case of CinC) and, because of this, also run the risk of not performing well on the other beat classes. This risk is assessed by measuring the performance for every beat class separately. A class-agnostic classification threshold is chosen for AE-logMSE and AE-LLH based on their ROC curves of the test set for the binary quality scoring task. The specific threshold corresponds to the point of the ROC curve with the highest F1-score. Using this fixed threshold, sensitivity and specificity are computed for three new binary quality scoring tasks combining the CinC noisy class with either the normal rhythm, atrial fibrillation, or *other* signals. Large variations in sensitivity and specificity values for the three groups would indicate that AE-logMSE and/or AE-LLH struggle to generalize to rhythm classes that were not as well-represented in the training set as the other classes.

5) Correlation between indicators and quality level: To measure how well the SQIs correlate with signal quality, a data set was constructed in which (semi-) clean ECG signals were contaminated with different amounts of realistic noise signals. To this end, realistic ECG noise was used from the Physionet MIT-BIH Noise Stress Test Database[24]. This noise database contains examples of electrode motion artefacts, muscle artefacts and baseline wander noise. Three new data sets were constructed from the CinC test set (excluding the CinC signals that were labeled as noisy), one for each type of noise. Using the approach by [10], noise was added corresponding to four quality levels linked with four distinct SNR levels. Similar to the base CinC signals, the noise signals are also band-pass filtered. Random segments of the noise signals were taken and added to the original signals at specific SNR values for each type of noise. This results in a new signal data set with five quality labels: clean, minor noise, moderate noise, severe noise and extreme noise. Figure 4 shows an example of a signal corrupted by electrode motion noise at the various levels.

A relevant quality indicator should then change monotonically with the severity of the noise. The Kendall rank correlation coefficient based on the τ_b statistic[25] is used to measure this monotone relationship, which is explained briefly below. Do note that most SQIs decrease in value with increasing severity of noise, while AE-logMSE increases with increasing severity. For correlation tests, we are interested in absolute values of a statistic and do not focus on the sign of the correlation.

The Kendall rank correlation coefficient bases its calculation on concordance or discordance of variable pairs. For two

random variables X and Y under investigation and two joint samples (x_i, y_i), (x_j, y_j) the sample pairs are said to be concordant if the ordering is the same for both variables, either ($x_i < x_j$ and $y_i < y_j$), or ($x_i > x_j$ and $y_i > y_j$). The sample pairs are discordant if the ordering differs for the variables and tied if either $x_i = x_j$ or $y_i = y_j$. Concordance then looks for positive correlation, where discordance can capture negative correlations. In our case, the x-variable corresponds to the SQI value, and the y-variable corresponds to the noise level. The full τ_b statistic is calculated as

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_x)(n_c + n_d + n_y)}}$$

with n_c and n_d being the number of concordant and discordant pairs respectively, n_x being the number of ties in the x-variable, and n_y being the number of ties in the y-variable. Data pairs where both the x- and y-variable are tied are not counted in n_x and n_y

This is an experiment in the *generalization to held-out data* setting. AE-logMSE and AE-LLH are both calculated using an auto-encoder trained on the CinC training data, and the corrupted signals were all drawn from the CinC test set. Bootstrapping is used to estimate the τ_b sampling distributions, using a similar approach to Section II-C.4, by randomly sampling (indicator value, SNR level) tuples with replacement and computing τ_b on each bootstrap set.

III. Results

A. Binary quality scoring

Figure 5 and 6 show the AUC results for the binary quality scoring experiment in our three settings.

a) Generalization to held-out data: Figure 5 shows the AUC results for binary quality scoring on (held-out) CinC test data. Both auto-encoder based indicators performed well in predicting the quality labels. The different SQIs span a wide range of AUC values. The best score was obtained for the AE-LLH SQI (median of 0.88 AUC) with the lowest scoring SQI, pSQI, showing very weak predictive power (median of 0.53 AUC).

b) Data-specific model: Figure 6a and 6b show the data-specific results for the Sleep and Stress data respectively at the labels AE-logMSE$_{specific}$ and AE-LLH$_{specific}$, compared against the benchmarks. For the Sleep data set, AE-logMSE achieves a near-perfect median AUC of 0.98 and outperforms all SQIs. AE-LLH and bSQI share second place with a median AUC of 0.91. On the Stress data set, AE-logMSE also shows a near-perfect median AUC of 0.96. AE-LLH (median of 0.90 AUC) is slightly outperformed by bSQI (median of 0.93 AUC).

c) Generalization to other data sets: Figure 6a and 6b show results in this setting for Sleep and Stress data at the labels AE-logMSE$_{general}$ and AE-LLH$_{general}$. The benchmark results in this setting are identical to the data-specific setting, since the underlying data set doesn't change, only the auto-encoder changes (as it is now trained on a different data set). For Sleep data, AE-logMSE and AE-LLH show a substantial drop in performance compared to the data-specific case, with both indicators only outperforming kurtosis and pSQI. For Stress data, AE-logMSE and AE-LLH are more competitive, with only bSQI outperforming both.

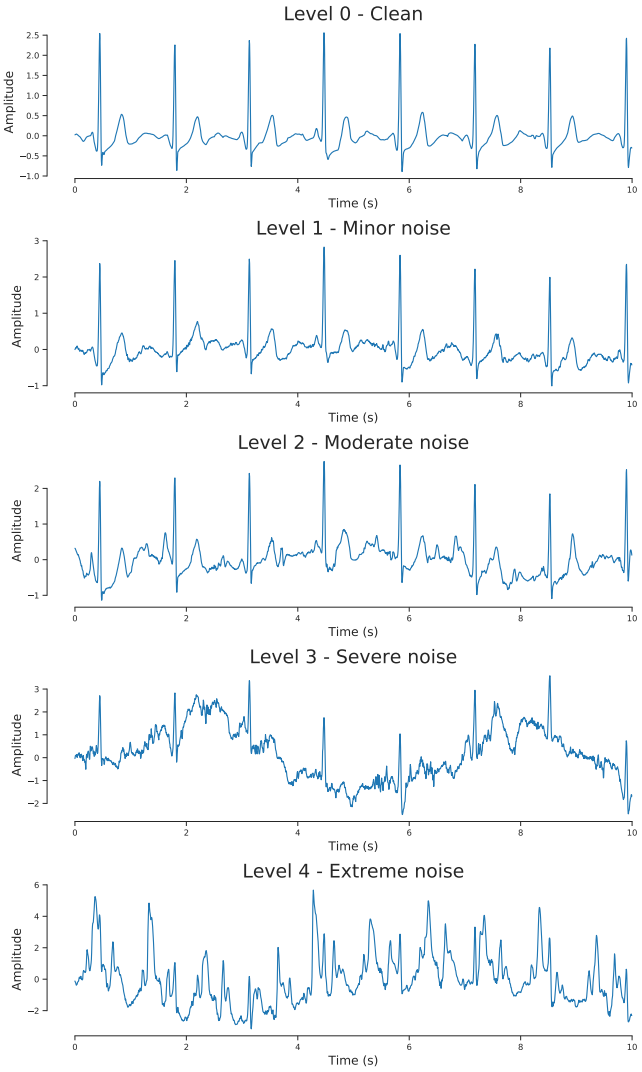


Fig. 4: Illustration of the different quality levels for electrode motion noise added to a single ECG segment.

When performing logistic regression on the individual SQIs in the *generalization to held-out data* setting (for CinC test data), each indicator besides kurtosis shows a significant effect based on the likelihood ratio test (Table III). The tests show a highly significant effect for most of the SQIs; AE-logMSE, AE-LLH, skewness, IOR, basSQI, bSQI, and pcaSQI all show $p < 0.001$. For pSQI, a p-value of 0.02 was obtained and kurtosis showed an insignificant effect in logistic regression ($p > 0.7$). Multiple logistic regression shows most SQIs can add a significant contribution to the proposed regression model in this setting. Backwards selection drops kurtosis, pSQI and bSQI from the suite of SQIs. A final combination of AE-logMSE, AE-LLH, skewness, IOR, pcaSQI, and basSQI shows a substantial drop in negative loglikelihood, from the best fit of 81.6 for AE-LLH individually to 40.2 for the group.

For logistic regression on Sleep data (combining the *data-specific* and *generalization to other data* settings), all individual SQIs had a significant effect (Table III). Even though a

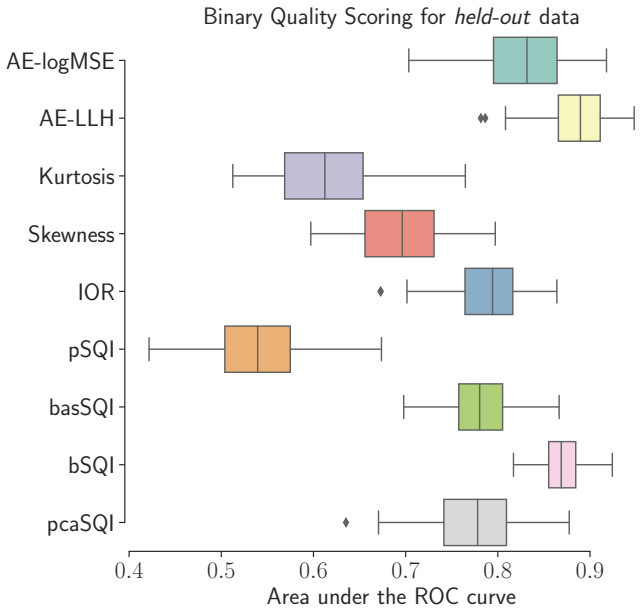


Fig. 5: Binary quality scoring results for the *held-out* CinC test set, boxplots show bootstrap estimates of the sampling distribution

significant effect was observed, large differences exist between the SQIs. Fitting logistic regression for (*data-specific*) AE-logMSE gave a negative loglikelihood of 351.1 compared to kurtosis with 921.3, barely improving upon only fitting an intercept. Backwards selection resulted in a group consisting of AE-logMSE$_{specific}$, AE-logMSE$_{general}$, kurtosis, IOR, basSQI, bSQI, and pcaSQI (so combining the AE-logMSE indicators from both settings).

Logistic regression for Stress data (combining the *data-specific* and *generalization to other data* settings) showed similar results to the *held-out* results on CinC data: kurtosis was also the sole SQI not having a significant effect (Table III). Here, backwards selection resulted in a smaller group than the other data sets with AE-logMSE, AE-LLH, basSQI, and bSQI being selected, only including the *data-specific* versions of AE-logMSE and AE-LLH. This did lead to a smaller drop in negative loglikelihood than observed with the other two data sets. The *data-specific* AE-logMSE by itself got a fit of 105.0 and the combination only dropped this to 97.5.

Table IV shows sensitivity and specificity results for a subgroup analysis of the CinC test set. Performance remains similar across the three subgroups.

B. Quality correlation

Results for the correlation between SQI values and quality levels (for held-out data) can be found in Figure 7. For electrode motion noise, AE-logMSE and AE-LLH both outperform the benchmarks. For motion artefacts, AE-logMSE and AE-LLH are outperformed by pSQI and pcaSQI. Both report a near-perfect correlation. For baseline wander, performance of AE-logMSE drops to the middle of the pack and AE-LLH again scores third-best, being outperformed by pcaSQI and pSQI. Note that AE-LLH is the only SQI that scores consistently above 0.6 for the three types of noise.

IV. Discussion

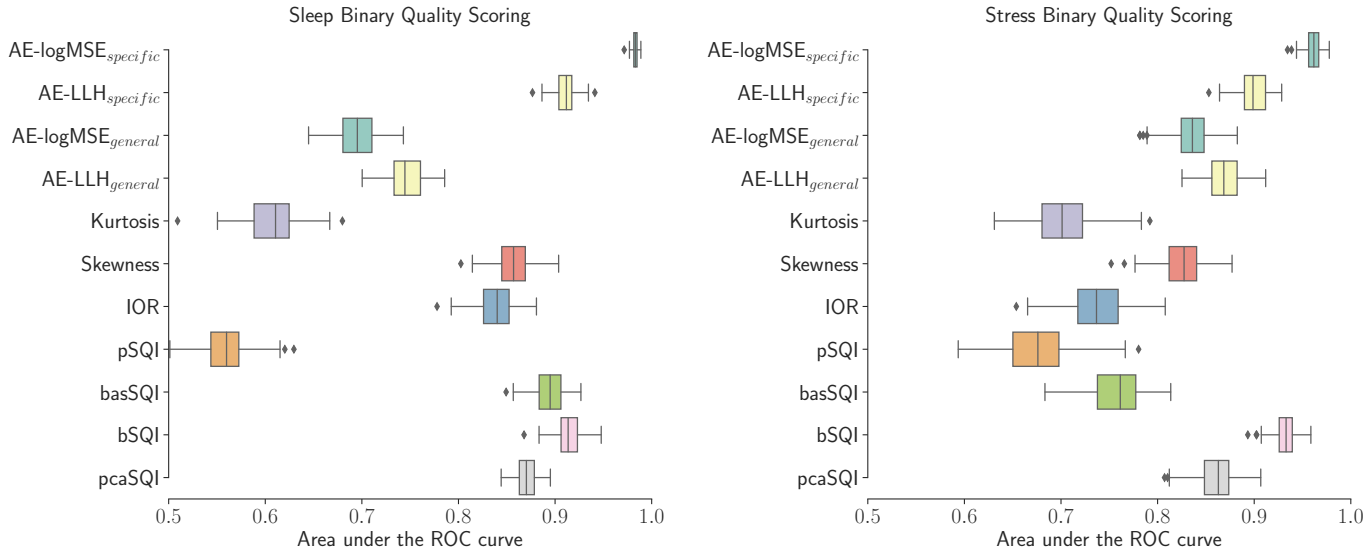
A. Discussion of experimental results

As a first experiment, the binary quality scoring capabilities of AE-logMSE and AE-LLH were evaluated. In the *generalization to held-out data* setting (Figure 5), AE-LLH outperformed the benchmarks and AE-logMSE only got outperformed by bSQI. In the *data-specific* setting (Figure 6), AE-logMSE showed very strong performance, clearly outperforming the benchmarks. AE-LLH scored on par, or slightly worse than bSQI. Results for AE-logMSE (median AUC of 0.98 and 0.96 for Sleep and Stress respectively) are nearing the performance of the supervised model of [9], where the authors reported 1.00 AUC for both of these data sets. Note that the auto-encoder achieves these results without any knowledge of the target labels. Results for the *generalization to other data sets* setting were more varied (Figure 6), with AE-logMSE and AE-LLH performing well on the Stress set, but less on the Sleep set.

To obtain a good quality indicator in the *generalization to held-out data* setting, the auto-encoder needs to transfer well to held-out data. If the model had overfit on the training set or failed to learn important ECG characteristics during training, its reconstructions would be poor. Next, AE-logMSE and AE-LLH have to accurately capture quality information. In a test set, reconstruction errors can arise either due to the novelty of the signals (being unseen in the training set) or due to quality issues in the signals. A good quality indicator has to isolate these quality issues while not being mistaken on novel signals. The strong performance in binary quality scoring on the CinC test set shows these two properties of AE-logMSE and AE-LLH.

Looking at the *data-specific* and *held-out* results, AE-logMSE outperforms AE-LLH on Sleep and Stress data, while the reverse is true on CinC data. We hypothesize that this is caused by the difference in definition of the quality labels. For CinC data, a high-quality ECG recording had to show the finer details of an ECG recording like P and T waves to allow for diagnosis of atrial fibrillation. Sleep and Stress data on the other hand, linked the quality of a recording with clearly defined R-peaks. AE-LLH allows to weigh parts of the signal depending on the confidence of the model. The auto-encoder showed high uncertainty around the R-peaks of a signal, not clearly predicting the magnitude of the peaks. The uncertainty bands shrank outside of the QRS complex, penalizing reconstruction errors at this part of the signal higher than around the R-peaks. AE-logMSE does not incorporate this local weighing and is mainly driven by reconstruction errors on the R-peaks due to the high amplitude of the peaks compared to the rest of the ECG. These properties lead to the hypothesis that AE-LLH is more suited for tasks where the finer details of the ECG matter, and AE-logMSE is more suited for a focus on R-peaks.

The CinC results indicate strong performance of AE-logMSE and AE-LLH in an online setting when training data is



(a) (b)

Fig. 6: Binary quality scoring results for the *data-specific* and *generalization to other data* settings (evaluated on the Sleep and Stress data sets), boxplots show bootstrap estimates of the sampling distribution. AE-logMSE$_{specific}$ and AE-LLH$_{specific}$ show the *data-specific* results, AE-logMSE$_{general}$ and AE-LLH$_{general}$ show results in the *generalization to other data* setting.

TABLE III: Negative loglikelihood of the logistic regression fit for SQIs on the different data sets (lower is better), also including fit of the SQI group obtained using backwards selection. Note that we combine two evaluation settings (*data-specific* and *generalization to other data*, the latter trained on CinC data, see Table II) for determining the best group fit. For the benchmark SQIs, the *data-specific* and *generalization to other data* settings are identical. Intercept indicates a model fit using only an intercept parameter, giving a baseline to perform the likelihood ratio test. Results for the likelihood ratio test:$^{*}p < 0.05$, $^{**}p < 0.01$

Evaluation setting	*Held-out data*	*Data-specific*		*Generalization to other data*	
Evaluation data	CinC	Sleep	Stress	Sleep	Stress
Intercept	130.1	926.7	293.2	926.7	293.2
AE-logMSE	96.0**	**351.1****	**105.0****	834.6**	197.1**
AE-LLH	**81.6****	601.8**	165.2**	790.3**	182.3**
Kurtosis	130.1	921.3**	291.6	921.3**	291.6
Skewness	124.7**	565.4**	201.4**	565.4**	201.4**
IOR	114.6**	737.8**	276.1**	737.8**	276.1**
pSQI	127.4*	919.5**	280.6**	919.5**	280.6**
basSQI	113.9**	461.7**	231.5**	**461.7****	231.5**
bSQI	93.7**	746.2**	159.1**	746.2**	**159.1****
pcaSQI	98.1**	879.1**	191.1**	879.1**	191.1**
Group fit	40.2**	207.3**	97.5**	207.3**	97.5**

TABLE IV: Sensitivity and specificity results for a binary quality scoring task between noisy signals and one of either normal rhythms, atrial fibrillation, or a catch-all *other* class with a common classification threshold for all tasks. The noisy signals are considered the negative class.

	Normal rhythm		Atrial fibrillation		Other rhythms	
	Sens	Spec	Sens	Spec	Sens	Spec
AE-logMSE	0.96	0.60	0.96	0.60	0.94	0.60
AE-LLH	0.93	0.70	0.93	0.70	0.88	0.70

available from the measurement setup and subject population (*held-out* setting). When changing measurement setup, results on the Sleep and Stress data set in the *generalization to other data sets* setting indicate that, while satisfactory performance can be obtained (like in the Stress case), retraining might be required for optimal performance. If a retraining step is not an option, a user might want to opt for one of the more traditional SQIs, e.g., bSQI, depending on the measurement setup in question (as it is difficult to know a priori whether the results for the Sleep or Stress data are more representative for a new use case). Note that this retraining step does not require human labels, the auto-encoder retrains in an unsupervised manner. In an offline setting with a data-specific model, our results indicate that this retraining can lead to near-perfect binary quality scoring.

Multiple logistic regression showed room for improvement on a single SQI. For every data set, the best performing individual SQI could achieve significantly better performance when joined by another indicator. Even though AE-logMSE or AE-LLH could outperform the reference SQIs by themselves, some aspects of signal quality seem to be missed by the auto-encoder but can be identified by relying on other SQIs,

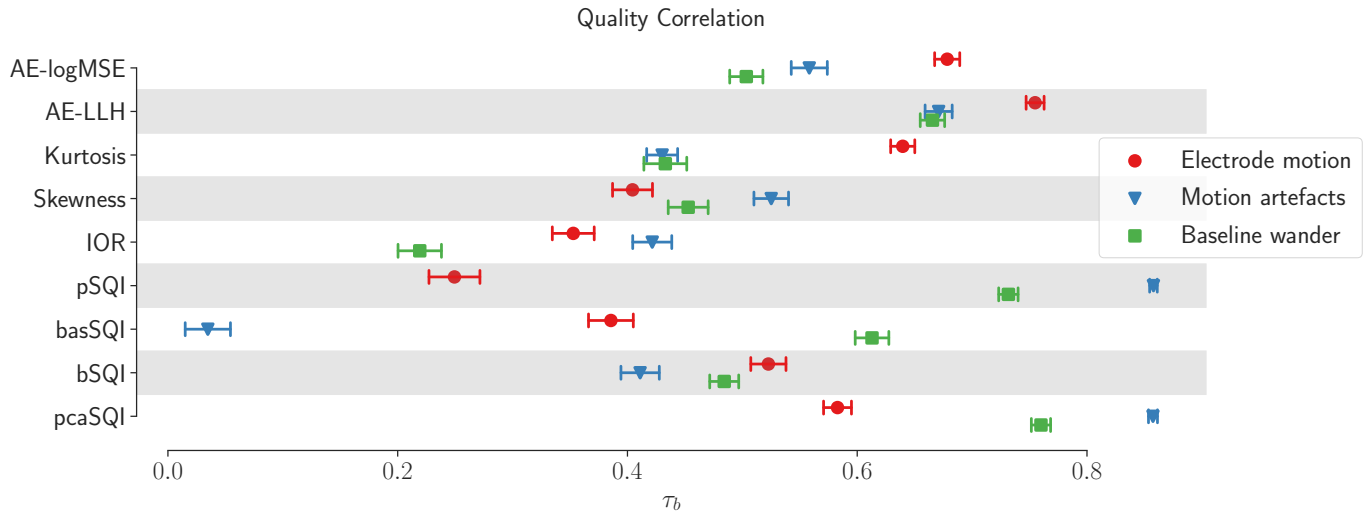


Fig. 7: Absolute values for Kendall's rank correlation (τ_b) of the quality indicators with five quality levels using electrode motion noise, motion artefacts, and baseline wander (tested on held-out data). Error bars indicate the sampling standard deviation estimated using bootstrapping.

indicating that they are complementary in nature (with the auto-encoder based indicators being the most informative). For CinC and Stress data, significant gains could even be obtained by joining AE-logMSE and AE-LLH, which are both computed from the same auto-encoder model.

When comparing results from binary scoring with the correlation test, it becomes clear that performance of the reference SQIs varies strongly from task to task. AE-logMSE, AE-LLH, and kurtosis show similarly strong correlation for the three noise types. For the binary scoring, however, kurtosis was the only SQI to show insignificant predictive power in logistic regression for two out of three data sets. On the other hand, the best performing reference SQI in the CinC binary scoring task, bSQI, shows mediocre results for the correlation test. AE-LLH outperformed it for all noise types, and AE-logMSE only scored similarly for baseline wander noise. For the other noise types, AE-logMSE clearly outperformed bSQI. The auto-encoder based quality indicators show strong performance on both tasks and on all noise types in the correlation test.

The temporal resolution of SQIs is worth considering. The data sets under investigation all contain signal segments with duration between 30 and 60 seconds. Quality assessment, however, can also be necessary at finer temporal resolution for, e.g., beat classification. In this case one is interested in resolutions of around a single second. Figure 3 shows that our auto-encoder based indicators can convey meaningful quality information at this resolution or even finer. Similar analysis indicated that the SQIs based on signal statistics (kurtosis and skewness) also already capture some quality information at such short time scales. SQIs based on frequency information are expected to struggle at finer temporal resolution due to spectral resolution issues for the lower frequencies, e.g., basSQI requiring <1 Hz frequency information. Finally, SQIs that rely on beat detectors will struggle most at these finer resolutions. pcaSQI requires more than five detected beats, and we were able to obtain reasonable results on CinC subsegments of down to five seconds. This was not possible for bSQI, where the underlying *wqrs* beat detector failed for segments of 5 seconds or less.

Our results show that the implicit definition of signal quality by using an auto-encoder (that high-quality signals can be represented on a low-dimensional manifold) is a good match for human definitions of ECG signal quality across tasks. In our experiments, AE-LLH generally outperformed AE-logMSE. We want to restate the hypothesis that AE-LLH is better suited for tasks involving fine-grained details of the signal. This already shows in the binary quality scoring results. The additive noise in the quality quantification test first distorts P and T waves before distorting the QRS complex. Quality labels in the temporal resolution test also rely on P and T waves. For these tests, AE-LLH did outperform AE-logMSE, while on the Sleep and Stress data (where quality labels relied on visibility of R peaks) the reverse was true.

To summarize, while the benchmark SQIs show varying performance on different tasks, we show that AE-logMSE and AE-LLH perform consistently well across the different tasks. However, our results adhere to the "no free lunch" principle. For binary quality scoring, bSQI is a strong contender and the more reliable SQI in a *generalization to other data* setting. To quantify the quality level of a signal, pSQI or pcaSQI might be more relevant if it is known beforehand that the signal mainly suffers from motion artefacts. However, across the board (across settings in the binary quality scoring task and across noise types in the quality quantification task) AE-logMSE and AE-LLH perform consistently, and perform well.

B. Model assumptions

Using auto-encoders to assess signal quality relies on two key assumptions at the data set level and signal level. Firstly, in order to obtain a model that is attuned to high-quality signals the majority of the training data needs to consist of examples of clean signals. In addition, these clean examples

should make up a diverse set containing all variations of the signal that could arise during use of the auto-encoder based SQIs. Secondly, the type of signal should lend itself well to unsupervised learning using auto-encoders. There should be grounds to assume a lower-dimensional manifold can be identified for the signal in question with realistic noise not lending itself well to such a lower-dimensional representation.

Both assumptions seem to hold. ECG is a very structured signal, giving confidence to the possible existence of a low-dimensional representation that auto-encoders try to identify. The typical noise in ECG recordings also shows less structure than the actual signal. This allows the auto-encoder to more easily learn desired ECG characteristics and ignore noise which is more difficult to model with a bottleneck layer. The data sets also lend themselves well to this approach, with labels indicating a majority of clean data.

There is, however, a property of the ECG that deserves closer attention. The heart is prone to various arrhythmias which deviate from the expected, normal heart rhythm. These arrhythmias, while part of the expected range of forms a clean ECG can take, can easily be deemed an anomalous pattern by the auto-encoder. To limit the risk of mistaking arrhythmias for noise, enough examples of these deviating patterns should be available in a training set for the model. This risk was the main motivation to put more focus on the CinC data set, since this data set contained not only atrial fibrillation, but other rhythms as well (labeled as "other beats"). Table IV, together with the strong performance of AE-logMSE and AE-LLH for this data set shows the risk is manageable, but has to be taken into account when applying our methodology to new data.

In the proposed approach, the model only looks at ECG signals in a single lead. CinC data consist of lead I signals, and Sleep and Stress data contain lead II signals. Our experiments show that AE-logMSE and AE-LLH can cope with single-lead signals coming from different leads, but might require retraining to improve performance further. Additionally, since the auto-encoder driving AE-logMSE and AE-LLH relies on signal structure, we expect the model to also work on multi-lead ECG where signal structure can be even more pronounced, yet this is beyond the scope of this study.

C. Additional remarks

It is noted that this study goes beyond traditional auto-encoder based anomaly detection. Purely detecting anomalous signals corresponds to the binary quality scoring setting. For this task, auto-encoders have become a popular approach in other modalities to build such a data-driven, unsupervised anomaly detector with minimal expert input. In this work, however, we showed that the auto-encoder approach can be taken further. When testing the correlation of our indicators with noise levels or with the ordinal relabeling, AE-logMSE and AE-LLH have to show a monotonic relation with these labels. This is in contrast to the anomaly detection setting, where we only look for a clear distinction between "normal" and "anomalous" classes. The strong performance of AE-logMSE and AE-LLH on these tasks show that auto-encoders are also fit for extending anomaly detection to quality assessment (at least for the particular case of ECG).

Applying auto-encoders instead of classical SQIs changes the required expertise. Detecting quality issues using the reference SQIs requires knowledge of the type of noise that might be present and what characteristics a fit-for-analysis signal should have. Our analysis shows that the reference SQIs react differently to different kinds of noise, and differ in performance depending on the specific task. Determining the best indicator requires knowledge of these differences and which differences matter for a specific use case. Auto-encoders, on the other hand, perform strongly across types of noise and tasks. They require, however, a different kind of expertise. One needs to build and train an auto-encoder that succeeds at learning a meaningful representation of the signal of interest. The model has to focus on desirable properties while ignoring various kinds of noise. Our results for binary quality scoring on Sleep and Stress data show that auto-encoders can perform well, even without retraining, for new data. It should also be noted that no part of the model architecture was changed between the different data sets, indicating that the architecture transfers well to new data. While applying our methodology for a new use case (other than ECG) might require further tuning of the auto-encoder, this architecture can nonetheless be used as a good starting point.

V. Conclusion

In this paper, we discussed the use of auto-encoders, a class of unsupervised deep learning models, in ECG signal quality assessment. Two quality indicators based on a trained auto-encoder, AE-logMSE and AE-LLH, consistently performed well on the investigated evaluation tasks compared to their benchmarks. These evaluation tasks went further than the typical anomaly detection setting for auto-encoders. Not only did AE-logMSE and AE-LLH perform well on binary quality scoring, they also showed strong performance when testing correlation with different noise levels.

In contrast to developing the benchmark indicators, no expert input is needed for our indicators. The auto-encoder automatically detects patterns in the data without additional human input. Our methodology can easily be extended to other data modalities by training an auto-encoder on these modalities, leading the way for data-driven quality assessment instead of relying on desirable data properties defined by humans.

Acknowledgment

We would like to thank the authors of [9] for sharing their data sets and for their labeling effort.

References

[1] S. Nizami *et al.*, "Implementation of artifact detection in critical care: A methodological review," *IEEE reviews in biomedical engineering*, vol. 6, pp. 127–142, 2013.

[2] T. He *et al.*, "Application of independent component analysis in removing artefacts from the electrocardiogram," *Neural Computing & Applications*, vol. 15, no. 2, pp. 105–116, 2006.

[3] G. Clifford *et al.*, "Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms," *Physiological measurement*, vol. 33, no. 9, p. 1419, 2012.

[4] J. Behar *et al.*, "Ecg signal quality during arrhythmia and its application to false alarm reduction," *IEEE transactions on biomedical engineering*, vol. 60, no. 6, pp. 1660–1666, 2013.

[5] Q. Li *et al.*, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter," *Physiological measurement*, vol. 29, no. 1, p. 15, 2007.

[6] T. H. Falk, M. Maier, *et al.*, "Ms-qi: A modulation spectrum-based ecg quality index for telehealth applications," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1613–1622, 2014.

[7] G. D. Clifford *et al.*, "Ecg statistics, noise, artifacts, and missing data," *Advanced methods and tools for ECG data analysis*, vol. 6, p. 18, 2006.

[8] L. Johannesen and L. Galeotti, "Automatic ecg quality scoring methodology: Mimicking human annotators," *Physiological measurement*, vol. 33, no. 9, p. 1479, 2012.

[9] J. Moeyersons *et al.*, "Artefact detection and quality assessment of ambulatory ecg signals," *Computer methods and programs in biomedicine*, vol. 182, p. 105 050, 2019.

[10] Q. Li *et al.*, "A machine learning approach to multi-level ecg signal quality classification," *Computer methods and programs in biomedicine*, vol. 117, no. 3, pp. 435–447, 2014.

[11] H. Wang *et al.*, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, pp. 107 964–108 000, 2019.

[12] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, vol. 2, no. 1, pp. 1–18, 2015.

[13] J. Chen *et al.*, "Outlier detection with autoencoder ensembles," in *Proceedings of the 2017 SIAM international conference on data mining*, SIAM, 2017, pp. 90–98.

[14] R. Chalapathy *et al.*, "Robust, deep and inductive anomaly detection," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2017, pp. 36–51.

[15] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 665–674.

[16] J. T. Andrews *et al.*, "Detecting anomalous data using auto-encoders," *International Journal of Machine Learning and Computing*, vol. 6, no. 1, p. 21, 2016.

[17] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114 [cs, stat]*, May 2014, arXiv: 1312.6114. [Online]. Available: `http://arxiv.org/abs/1312.6114` (visited on 11/22/2019).

[18] H.-T. Chiang *et al.*, "Noise Reduction in ECG Signals Using Fully Convolutional Denoising Autoencoders," English, *Ieee Access*, vol. 7, pp. 60 806–60 813, 2019, WOS:000468683800001, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2912036.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[20] W. Zong *et al.*, "A robust open-source algorithm to detect onset and duration of qrs complexes," in *Computers in Cardiology, 2003*, IEEE, 2003, pp. 737–740.

[21] P. Hamilton, "Open source ecg analysis," in *Computers in cardiology*, IEEE, 2002, pp. 101–104.

[22] G. D. Clifford *et al.*, "Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017," in *2017 Computing in Cardiology (CinC)*, IEEE, 2017, pp. 1–4.

[23] C. Varon *et al.*, "Robust artefact detection in long-term ecg recordings based on autocorrelation function similarity and percentile analysis," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2012, pp. 3151–3154.

[24] G. B. Moody *et al.*, "A noise stress test for arrhythmia detectors," *Computers in cardiology*, vol. 11, no. 3, pp. 381–384, 1984.

[25] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, pp. 239–251, 1945.