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. 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 level of disturbance 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 skew, 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 characteristics present in a given data set, in which the auto-encoders learn to reconstruct data and error measures on input data and 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 quality assessment. Unsupervised quality assessment allows to move away from a reliance on human input, from human-defined signal properties or human-provided labels. 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."

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

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

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. This encoding takes the form of a function $f(\cdot)$ parameterized using a deep neural network to calculate a new representation $\boldsymbol{z} = f(\boldsymbol{x})$ of the input data. In classical auto-encoders, another function $g(\cdot)$ decodes a point in the representation space 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 new 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 measuring the quality of reconstructions 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 an extension on the decoder and a different 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 by placing certain constraints on the model. These constraints can come from 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 characteristics in training data to be able to accurately represent and reconstruct inputs.

2) Architecture: The architecture for our proposed network is adapted from [18] and, as mentioned above, the Gaussian output inspired by [17]. The auto-encoder is trained on input segments containing 1024 samples of an ECG signal sampled at 200 Hz.[1]

The model makes use of temporal convolutions with an Exponential Linear Unit (ELU) as activation function. Every kernel operates on 16 samples along the temporal axis and all features along the feature axis. A batchnormalization layer [19] is placed between the convolution operation and activation. The final layer that outputs $\boldsymbol{\mu}$ uses a linear activation function and the $\boldsymbol{\sigma}$ output makes use of a softplus activation to ensure positive values. The encoder performs strided convolutions to downsample the signal. The decoder uses transposed convolutions 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: 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.

[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.

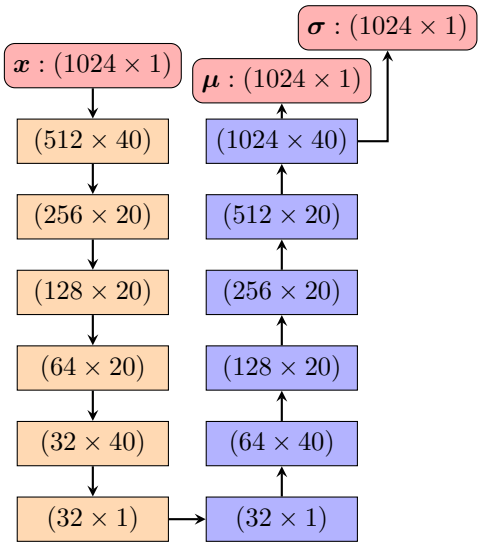
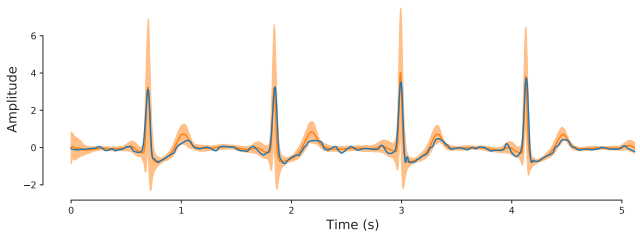
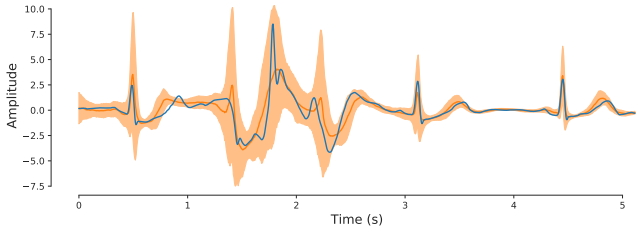


Fig. 1: Architecture of the auto-encoder with the encoder on the left and the decoder on the right. Boxes indicate intermediate data tensors of size (#Timesteps×#Features) with arrows indicating a convolution layer (details in text).



(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.

Training is carried out for 200 epochs. During training, the loss on a separate validation set is tracked and the parameter values with the best validation loss are retained after training. 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.

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 properly 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.

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 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.

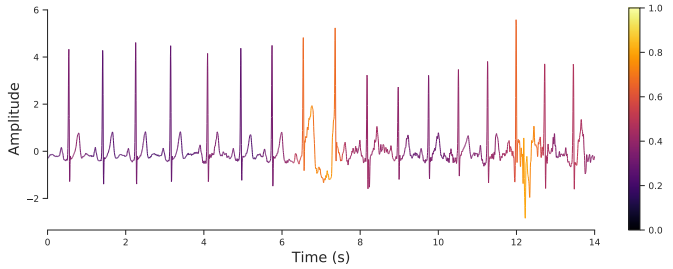


Fig. 3: Illustration of a quality signal. The input ECG is color-coded with the quality level making use of the squared error within the formulation of AE-logMSE (without a logarithmic transformation). Darker parts of the signal indicate a small squared error and lighter parts indicate a larger error.

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). These boundaries were empirically determined. Averaged over the validation sets, the first half second and final quarter second of a segment showed substantially larger reconstruction errors. 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 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 later ignored for computing AE-logMSE and AE-LLH.

C. Experiments

To test the performance of AE-logMSE and AE-LLH, they are compared with benchmark SQIs that are commonly used in the literature.

1) Benchmark indicators: Four reference SQIs are used to benchmark the auto-encoder based indicators.

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) Skew: Another commonly used quality indicator based on the statistics of the signal is the normalized third-order moment (often referred to as the skew of the distribution). A clean signal is expected to show high skew 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 quality of the signal.

d) pSQI: The relative power in the QRS complex (pSQI) [4] 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 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. 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 these different criteria.

a) CinC data set: The first data set is the PhysioNet Computing in Cardiology Challenge 2017 (CinC) data set [20]. This data set contains short single-lead ECG recordings between 30 and 60 seconds in length. The data set was constructed with the aim of developing algorithms that could detect normal rhythms, atrial fibrillations or a general class of "other rhythms". More importantly for the task of this paper, 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.

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 [21]. 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. 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

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

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.

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. This split is stratified making use of the four class labels (normal beat, atrial fibrillation, other beat, noisy). For the Sleep and Stress data sets, a similar random 80/20 split is made for training and validation. These splits are not stratified, simulating a scenario where the user does not have any labels.

3) 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. This means the signal should be of very high quality before it is deemed fit for interpretation. Labels in the Sleep data set indicate the 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.

The CinC data set is used to test the generalization character of auto-encoder based SQIs. An auto-encoder is trained on the CinC training set, with a training-validation split to monitor validation performance during training. This model is then used to assess the quality of the CinC test set, testing how well the auto-encoder can cope with new signals. For these signals, ROC curves are constructed using the indicator values for the recordings and their respective labels.

For the Sleep and the Stress data set tests are carried out under the assumption that all data that requires quality scoring is available from the start, which can then also be used for training the model. In this case, an auto-encoder is trained, with a training-validation split, on the full data set and SQIs are afterwards computed for every recording in the set. This is in contrast with the test on CinC data, where an estimate is obtained for performance on new data with the same measurement setup. For all SQIs, the AUC for predicting the binary quality labels is computed. As an additional test, the auto-encoder that was trained on CinC data is also used to calculate values for AE-logMSE and AE-LLH on both the Sleep and Stress data sets. This allows to test how well a trained model transfers to new acquisitions setups.

To test the significance of the predictive power of each quality indicator, a simple classifier based on logistic regression is used. A logistic regression model is first fit for every individual indicator. On these models the likelihood ratio test is used to determine whether the indicator shows significant power in predicting the quality labels. As a second test, a logistic regression model is fit making use of all indicators 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 the CinC, Sleep, and Stress data.

4) Correlation between indicators and quality level: To measure how well the indicators correlate with signal quality, a data set was constructed in which (semi-) clean ECG signals were contaminated with different amounts of representative noise signals. To this end, realistic ECG noise was used from the Physionet MIT-BIH Noise Stress Test Database [22]. This 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. 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 data set of ECG signals 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. Therefore, the quality indicators are evaluated based on their correlation with these quality levels. The linear relationship between the indicators and the quality levels is measured using the Pearson correlation coefficient, ρ. While a linear relationship aids interpretation, the existence of a monotone relationship between the indicator values and quality levels should also be investigated. This relationship indicates the predictive power of a SQI, even though a clear linear relation can be lacking. The Kendall rank correlation coefficient based on the τ_b statistic [23] 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 value of the quality indicator, and the y-variable corresponds

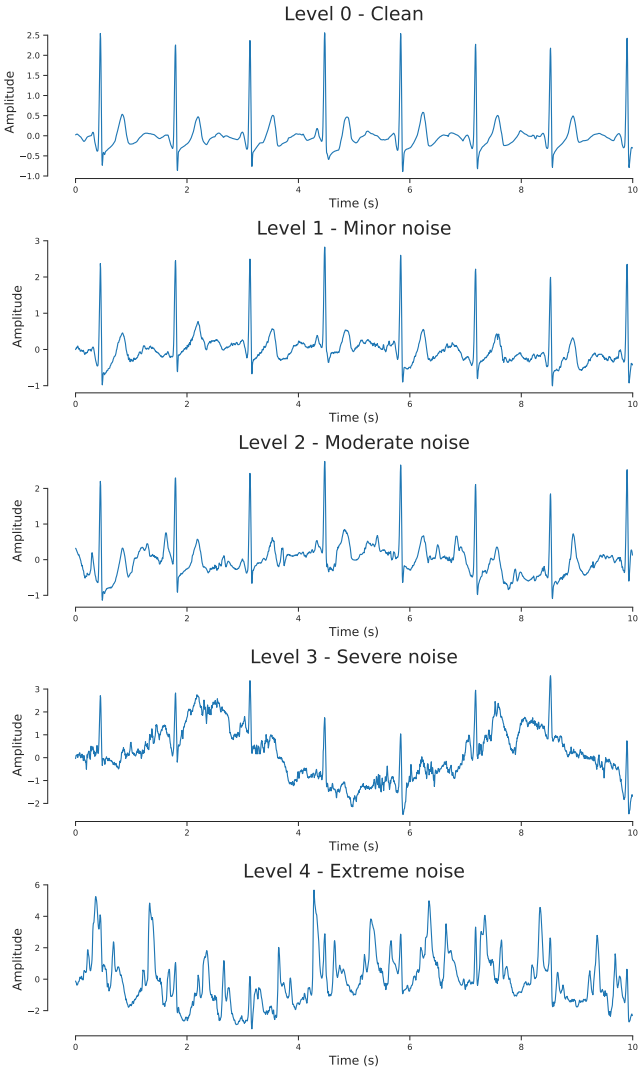


Fig. 4: Illustration of the different quality levels for electrode motion noise added to a single ECG segment.

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

AE-logMSE and AE-LLH in this test are both calculated using the auto-encoder trained for binary quality scoring on CinC data.

5) Temporal resolution of the quality indicators: Depending on the use case, a quality assessment of the signal is required at a varying time scale. When processing an entire recording like in the CinC data set, quality information is required for a full recording. Other types of beat analysis, however, can require quality assessment per beat. Different indicators can be better suited for different temporal resolutions of quality assessment.

To test the temporal resolution of the different quality indicators, signal segments of varying duration are relabeled from the CinC test set. For signal lengths of 1, 5, 10, and 20 seconds, 200 segments were selected at random from the "normal beat" recordings of the CinC data set and 100 random segments from the "bad quality" segments. These segments were each assigned one label out of three classes: good, moderate, and bad quality. Good quality segments were segments where the morphology of a full heart beat could be discerned and where P and T waves could easily be located. Moderate quality segments showed clear R-peaks but the rest of the ECG morphology was disturbed. Bad quality segments showed too much noise to accurately locate R-peaks.

For each segment duration, Kendall's tau was calculated between the quality indicator values and the ordinal scale of good, moderate, and bad quality. For the 5, 10, and 20 second segments, the auto-encoder processed the full segments to calculate AE-logMSE and AE-LLH. For the 1 second segments the auto-encoder made use of a 5 second signal segment containing the 1 second segment under investigation. The benchmark SQIs are calculated on the full 1, 5, 10, or 20 second segment.

III. Results

A. Binary quality scoring

Both auto-encoder based indicators performed strongest in predicting the quality labels of the CinC test set (Fig 5). The legend of Figure 5 shows a wide range of area under the curve (AUC) values for the different indicators. The best score was obtained for the AE-LLH indicator (0.876 AUC) with the lowest scoring indicator, pSQI, showing very weak predictive power (0.539 AUC).

Table II shows the AUC results for binary quality scoring on the Sleep and Stress data sets. For the Sleep data set, using the auto-encoder trained on CinC data results in skew and IOR showing better performance than AE-logMSE and AE-LLH. Training the auto-encoder on the Sleep data set itself, however, improves the performance of AE-logMSE and AE-LLH substantially, with both outperforming the reference indicators. A near perfect AUC of 0.98 is obtained for AE-logMSE.

On the Stress data set, AE-logMSE and AE-LLH both outperform the reference quality indicators when making use of the auto-encoder trained on CinC data. Training a new auto-encoder further improves the AUC with AE-logMSE achieving 0.96 AUC.

When performing logistic regression on the individual indicators for the CinC test set, each indicator besides kurtosis shows a significant effect based on the likelihood ratio test (table III). The tests shows a highly significant effect for AE-logMSE, AE-LLH, skew, and IOR with $p < 0.001$ for all. For pSQI, a p-value of 0.02 was obtained and kurtosis showed an insignificant effect in logistic regression ($p > 0.8$). Multiple logistic regression on CinC data shows most SQIs can add a significant contribution to the proposed model. Backwards selection only drops skew from the selection. A combination of AE-logMSE, AE-LLH, kurtosis, IOR, and pSQI shows a

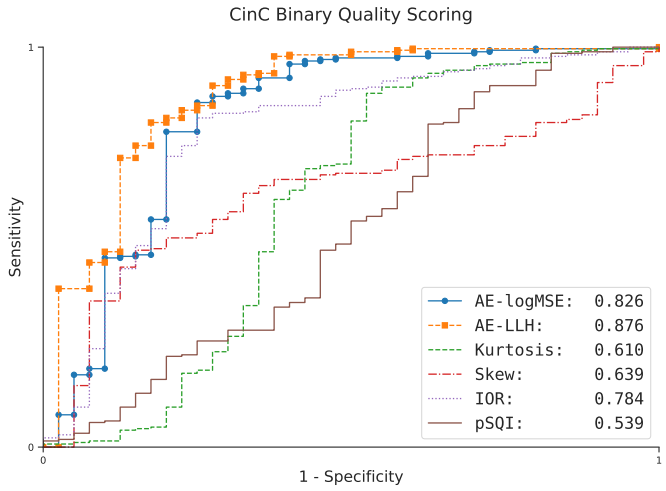


Fig. 5: Binary quality scoring ROC curves for the CinC test set. AUC values are shown in the legend.

TABLE II: Binary quality scoring performance on the Sleep and SWEET data sets measured using AUC. AE-logMSE$_a$ and AE-LLH$_a$ indicate the use of an auto-encoder trained on the CinC data set, AE-logMSE$_b$ and AE-LLH$_b$ are calculated using an auto-encoder trained on the data set under investigation.

	Sleep	Stress
AE-logMSE$_a$	0.69	0.84
AE-LLH$_a$	0.74	0.87
AE-logMSE$_b$	**0.98**	**0.96**
AE-LLH$_b$	0.91	0.90
Kurtosis	0.60	0.70
Skew	0.87	0.83
IOR	0.84	0.74
pSQI	0.56	0.67

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. Intercept indicates a model fit using only an intercept parameter, giving a baseline to perform the likelihood ratio test. AE-logMSE$_a$ and AE-LLH$_a$ are computed using an auto-encoder trained on CinC data; AE-logMSE$_b$ and AE-LLH$_b$ make use of an auto-encoder trained on either Sleep or Stress data (depending on the data set under investigation). Results for the likelihood ratio test:*p < 0.05, **p < 0.01

	CinC	Sleep	Stress
Intercept	117.8	926.7	293.2
AE-logMSE$_a$	86.2**	834.6**	197.1**
AE-LLH$_a$	**74.6****	790.3**	182.3**
AE-logMSE$_b$	-	**351.1****	**105.0****
AE-LLH$_b$	-	601.8**	165.2**
Kurtosis	117.8	921.3**	291.6
Skew	112.0**	565.4**	201.4**
IOR	103.2**	737.8**	276.1**
pSQI	115.1*	919.5**	280.6**
Group fit	52.4**	231.1**	102.4**

TABLE IV: Absolute values for Pearson correlation (ρ) and Kendall's rank correlation (τ_b) of the indicators with five quality levels constructed using electrode motion noise (EM), motion artefacts (MA), and baseline wander (BW). *p < 0.05, **p < 0.01

	EM		MA		BW	
	ρ	τ_b	ρ	τ_b	ρ	τ_b
AE-logMSE	0.77**	0.65**	0.65**	0.54**	0.62**	0.51**
AE-LLH	**0.85****	**0.73****	0.76**	0.63**	**0.68****	0.56**
Kurtosis	0.33**	0.71**	0.26**	0.57**	0.26**	**0.59****
Skew	0.16**	0.23**	0.12**	0.20**	0.16**	0.24**
IOR	0.57**	0.48**	0.06*	0.03	0.19**	0.13**
pSQI	0.26**	0.20**	**0.78****	**0.67****	0.07**	0.06**

substantial drop in negative loglikelihood, from the best fit of 74.6 for AE-LLH individually to 52.4 for the group. The Wald test indicates that all these SQIs still contribute significantly.

For Sleep data, all individual indicators had a significant effect (table III). Even though a significant effect was observed, large differences exist between performance of the SQIs. Fitting a model for AE-logMSE$_b$ 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$_a$, AE-LLH$_a$, AE-logMSE$_b$, kurtosis and pSQI.

Logistic regression for Stress data showed a similarity with the CinC data: kurtosis was again the sole SQI not having a significant effect (table III). Here, backwards selection resulted in a smaller group than the other data sets with the combination of AE-logMSE$_b$ and pSQI being selected. This did, however, lead to a smaller drop in negative loglikelihood than observed with the other two data sets. AE-logMSE$_b$ by itself got a fit of 105.0 and the combination only dropped this to 102.4.

B. Quality correlation

Results for the correlation between indicator values and quality levels can be found in table IV. The highest Pearson correlation between quality levels and indicator values was achieved by AE-LLH for electrode motion noise and motion artefacts with AE-logMSE a close second for these kinds of noise. For the motion artefact noise, pSQI showed the best Pearson correlation with quality levels, followed by AE-LLH and AE-logMSE.

For Kendall's rank correlation, AE-logMSE and AE-LLH, and kurtosis consistently reported high values across different noise types, whereas the results varied more for the other indicators. Electrode motion noise was best tracked by AE-LLH, followed by kurtosis and AE-logMSE in third place. For motion artefacts, the best results were obtained using pSQI with AE-LLH in second place. Finally, kurtosis best predicted baseline wander quality levels, followed by AE-LLH and AE-logMSE. Skew scored consistently poorly, while the performance of IOR and pSQI varied with the different types of noise. IOR showed a moderately high rank correlation of 0.48 for electrode motion noise, while for motion artefacts no significant rank correlation was found. Similar swings in performance were seen for pSQI, showing the best performance for motion artefacts while scoring very poorly for electrode motion and baseline wander.

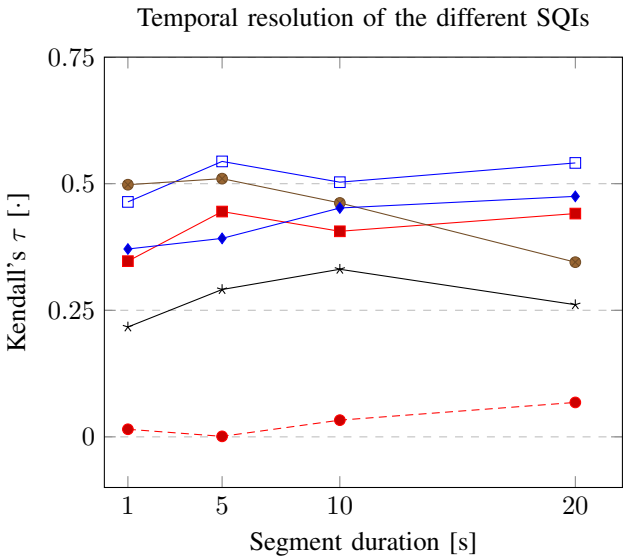


Fig. 6: Performance of the different indicators at changing segment durations (AE-logMSE —■—, AE-LLH —□—, Kurtosis —●—, Skew —⋆—, IOR —♦—, pSQI - -●- -)

C. Temporal resolution

Figure 6 shows the indicators' performance for different signal segment durations. AE-logMSE, AE-LLH and IOR all show improving performance with increasing signal duration. Kurtosis' performance drops with increasing signal duration and skew achieves a maximal performance at a signal duration of 10 seconds. No significant correlation between the ordinal labels and pSQI was observed for any duration.

IV. Discussion

A. Experiment results

As a first experiment, the binary quality scoring capabilities on a test set were measured. The auto-encoder was trained on the CinC training set and afterwards used to predict a binary quality label for the CinC test set. Here, AE-logMSE and AE-LLH both outperformed the benchmark indicators.

For AE-logMSE and AE-LLH, this was a double test. To obtain a good quality indicator on this test set, the auto-encoder itself had to transfer well to the unseen test set. If this model had overfit on the training set or not learned 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.

Retraining the auto-encoder on Sleep and Stress data showed very strong performance for AE-logMSE and AE-LLH on binary quality scoring, clearly outperforming the benchmarks. Of note, however, was the outperformance of AE-LLH by AE-logMSE. The outset of the binary quality scoring task is similar to the CinC case, but there AE-LLH outperformed AE-logMSE. We hypothesize that this is the case due to the difference in definition of the labels. For the CinC data set, a high-quality ECG recording had to show the finer details of an ECG recording like the P and T waves to allow for diagnosis of atrial fibrillation. The Sleep and Stress data sets 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 allow for 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.

Multiple logistic regression showed room for improvement on a single SQI. For every data set, the best performing individual SQI could show significantly better performance when joined by another indicator. Even though AE-logMSE and AE-LLH could outperform the reference indicators by themselves, some aspects of signal quality seem to be missed by the auto-encoder but can be identified by relying on other quality indicators indicating that they are complementary in nature (with the auto-encoder based indicators being the most informative). For CinC and Sleep data, even significant gains could be obtained by joining AE-logMSE and AE-LLH, which are both computed from the same auto-encoder model.

When comparing results from the binary scoring with the correlation test, it becomes clear that performance of the reference SQIs can vary strongly from task to task. AE-logMSE, AE-LLH, and kurtosis show similarly strong rank correlation for the three noise types. For the binary scoring, however, kurtosis was the only indicator to show insignificant predictive power in logistic regression in two out of the three data sets. On the other hand, the best performing reference indicator in the CinC binary scoring task, IOR, shows poor results for the correlation test. Results for the electrode noise were decent, but poor for motion artefact noise and baseline wander with the former case not showing a significant correlation. The auto-encoder based quality indicators, however, show strong performance on both tasks and on all noise types in the correlation test.

Performance varied on the temporal resolution test for most indicators. AE-LLH maintained a consistent score in contrast to the two best benchmarks: IOR and kurtosis. Kurtosis performed well on shorter segments but its performance dropped severely when moving to longer segments. IOR showed the opposite behavior, performing better when the segments got longer but struggling on shorter segments.

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 of these assumptions seem to hold. ECG is a very structured signal, giving confidence to the possible existence of a low-dimensional representation that auto-encoders can attempt to identify. The noise one expects to be present 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 the 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"). 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.

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 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 relationship 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 quality indicators changes the required expertise. Detecting quality issues using the reference indicators requires knowledge of the type of noise that might be present and what characteristics a signal that is fit for analysis should have. Our analysis showed that the reference indicators 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, J. R. Green, and C. McGregor, “Implementation of artifact detection in critical care: A methodological review,” *IEEE reviews in biomedical engineering*, vol. 6, pp. 127–142, 2013.

[2] T. He, G. Clifford, and L. Tarassenko, “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, J. Behar, Q. Li, and I. Rezek, “Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms,” *Physiological measurement*, vol. 33, no. 9, p. 1419, 2012.

[4] J. Behar, J. Oster, Q. Li, and G. D. Clifford, “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, R. G. Mark, and G. D. Clifford, "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, F. Azuaje, and P. Mcsharry, "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, E. Smets, J. Morales, A. Villa, W. De Raedt, D. Testelmans, B. Buyse, C. Van Hoof, R. Willems, S. Van Huffel *et al.*, "Artefact detection and quality assessment of ambulatory ecg signals," *Computer methods and programs in biomedicine*, vol. 182, p. 105050, 2019.

[10] Q. Li, C. Rajagopalan, and G. D. Clifford, "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, M. J. Bah, and M. Hammad, "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, S. Sathe, C. Aggarwal, and D. Turaga, "Outlier detection with autoencoder ensembles," in *Proceedings of the 2017 SIAM international conference on data mining*. SIAM, 2017, pp. 90–98.

[14] R. Chalapathy, A. K. Menon, and S. Chawla, "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, E. J. Morton, and L. D. Griffin, "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

[18] H.-T. Chiang, Y.-Y. Hsieh, S.-W. Fu, K.-H. Hung, Y. Tsao, and S.-Y. Chien, "Noise Reduction in ECG Signals Using Fully Convolutional Denoising Autoencoders," *Ieee Access*, vol. 7, pp. 60 806–60 813, 2019, wOS:000468683800001.

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

[20] G. D. Clifford, C. Liu, B. Moody, H. L. Li-wei, I. Silva, Q. Li, A. Johnson, and R. G. Mark, "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.

[21] C. Varon, D. Testelmans, B. Buyse, J. A. Suykens, and S. Van Huffel, "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.

[22] G. B. Moody, W. K. Muldrow, and R. G. Mark, "A noise stress test for arrhythmia detectors," *Computers in cardiology*, vol. 11, no. 3, pp. 381–384, 1984.

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