# ECG quality assessment using neural network trained on synthetic data

Antti Vasankari

MSc thesis
Dec 2024

Reviewers:
Prof. Matti Kaisti
Prof. Ion Petre

DEPARTMENT OF MATHEMATICS AND STATISTICS

UNIVERSITY OF TURKU
Department of Mathematics and Statistics

VASANKARI, ANTTI: ECG quality assessment using neural network trained with synthetic data
MSc Thesis, 40 pages
Mathematics, Data Analytics
Dec 2024

---

An electrocardiogram (ECG), is a common, safe, and simple way to examine the electrical activity of the heart as a signal over time, from which heart rate can be measured and heart diseases can be detected. ECG is susceptible to interference, so it is important that the quality of the signal is sufficient for the feature being examined. The quality of the ECG should reflect its usability, and the quality assessment should be consistent. Manual signal quality assessment is laborious and depends on the annotators expertise. Training machine learning models for this task requires large amounts of quality-labeled sensitive health data. Additionally, the generalizability of the model is decreased due to the inability to distinguish the noise from the pure signal.

This master's thesis presents a method, in which a quality classifier was added to a modular synthetic ECG generator. The labeled synthetic data was then used for training a simple convolutional neural network. Additionally, methods to improve the representativeness of the quality-labeled synthetic data was examined. The validation and test data consisted of manually labeled real-world single-lead ECG measurements. The performances of the models, trained on synthetic data, were compared with a model of the same architecture that was trained on the validation data.

The results support the utility of the synthetic ECG signals for training a quality assessor. Augmenting the training data with synthetic data improved the classification performance of the model trained on validation data. Furthermore, the flexibility of the synthetic ECG generator enables the customization of quality criteria and domain randomization, which, based on the results, appear to be beneficial methods.

Keywords: ECG, electrocardiogram, synthetic data, signal quality assessment, SQA, machine learning, ML, health technology.

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Antti Vasankari: ECG quality assessment using neural network trained on synthetic data
Pro gradu -tutkielma, 40 s.
Matematiikka, Data-analytiikka
Joulukuu 2024

---

Elektrokardiogrammi (EKG), eli sydänfilmi on yleinen, turvallinen ja yksinkertainen tapa tutkia sydämen sähköistä toimintaa ajan signaalina, josta voi mitata sykettä ja havaita sydänsairauksia. EKG on altis häiriöille, ja täten tulkinnassa on tärkeää, että signaalin laatu on tarkasteltavan ominaisuuden kannalta riittävän hyvä.

EKG:n laadun tulisi heijastaa sen käyttökelpoisuutta, ja laadun luokittelun tulisi olla johdonmukaista. Manuaalisesti signaalien laadun luokittelu on työlästä sekä riippuvaista luokittelijan osaamisesta. Koneoppimismallien kouluttaminen tehtävään vaatii suuria määriä laatuluokiteltua sensitiivistä terveysdataa, ja lisäksi näiden mallien koulutuksessa esiintyy aineiston aiheuttamaa vinoumaa, koska laatua heikentävä kohina ei ole eristettävissä kohisevasta signaalista.

Tässä pro gradu -tutkielmassa esitetään menelmä, jossa modulaariseen synteettisen EKG:n generaattoriin lisättiin generoidun signaalin laatuluokittelija. Tätä laatuluokiteltua synteettistä dataa hyödynnettiin yksinkertaisen konvoluutioneuroverkon koulutuksessa. Lisäksi tarkasteltiin synteettisen datan laatuluokkien edustavuuden parantamista. Validointi- ja testidatana käytettiin manuaalisesti luokiteltuja todellisia yksikytkentäisiä EKG-mittauksia. Koulutettujen mallien luokittelukykyä vertailtiin saman arkkitehtuurin malliin, joka koulutettiin validointidatalla.

Tulokset tukevat synteettisen EKG:n käyttökelpoisuutta laatuluokittelijan koulutusaineistona. Synteettisen aineiston käyttö paransi mallin luokittelukykyä. Lisäksi synteettisen EKG:n generaattorin joustavuus mahdollistaa laatukriteerin muokattavuuden ja koulutusaineiston rikastamisen, jotka tulosten perusteella vaikuttavat hyödyllisiltä menetelmiltä.

Avainsanat: EKG, elektrokardiogrammi, synteettinen data, signaalin laadun arviointi, koneoppiminen, terveysteknologia.

# Contents

# 1 Introduction

*Electrocardiogram* (*ECG*) is a valuable non-invasive method to monitor the cardiac function by observing the electrical activity of the heart. This is used from detecting cardio vascular diseases to personal fitness purposes. Wearable ECG is also growing in medical use [1]. This further compromises the quality of the recording already susceptible to interference. In the flourishing of automated systems, to be useful for diagnostic purposes and reliably measure heart rate (HR), filtering adequate data is a necessary preprocessing step to ensure reliable performance. Additionally, it ought to be noted that the criteria for quality are not unequivocal, but should consider preserving the information relevant to the intended use of the signals.

Manually, assessing the quality of the data can be laborious, ineffective and subjective. An automatic method for signal quality assessment must be considered. This can be conducted using automatic feature extraction methods for ECG signals, whereby the extracted features are used to determine the signal quality. A highly simplified example of this is determining the quality by the value of the measured power of the signal. Clearly, methods of this kind rely on the assumption of the quality being sufficiently represented by these extracted features, i.e. the variation of the signals considered to be pure should not have an effect on these features.

Alternatively, a signal quality assessor can be conducted using *machine learning* (*ML*) methods, more specifically *supervised learning*. This introduces a problem of the sufficiency of consistently quality assessed ECG signals. Training a machine learning model for signal quality assessment requires a large quantity of quality-labeled data. The question of quality is also use-case-specific. Therefore, it would be beneficial to implement a quality criterion specific automated signal quality assessor for a synthetic ECG generator.

The labeled signals, features extracted from them or both can be used to train a machine learning model. In this thesis, a *neural network* (*NN*) is employed as the classifier, with its operation based on the idea that the network inherently performs the feature extraction. Machine learning methods utilizing neural networks is often referred to as *deep learning* (*DL*). The usage of neural networks enable the identification of quality-relevant features that are not considered in feature extraction methods.

In this thesis, the following research questions will be addressed.

- Can the parametric ECG generator be used for generating quality classified signals, aligning with determined quality criterion?
- How does an ML model trained with synthetic quality-labeled ECGs perform assessing real-world ECGs compared to a model trained with real-world ECGs?
- Is there any additional value of using synthetic signals?

In Section 2 the characteristics of ECG, the process of measuring ECG and its operating principle are described. The subject of signal quality assessment is introduced and the methods, quality criteria and quality labels are discussed. Moreover, the quality criterion used in this thesis is presented and justified. Finally, the use of synthetic data generator is motivated.

In Section 3 the framework for generating ECG signal is described and the implemented automatic quality annotator is presented.

In Section 4 the validation and test datasets and the assessment process are described.

In Section 5 the basics of convolutional neural networks are presented and the classifier used for quality assessment in this thesis is described.

In Section 6 the specifics of the model training, and the differences between the models to be compared, are explained. Moreover, the evaluation metrics are described.

In Section 7 the results of the tests and the common mistakes in classification are presented.

Finally in Section 8 the results and their reasons are discussed. Possible oversights and biases are noted and proposals for improvement regarding future studies are also considered.

This thesis builds upon the synthetic biosignal framework introduced by Karhinoja et al. in 2024 [2]. I worked for Assistant Professor Matti Kaisti (Department of Computing, Faculty of Technology, University of Turku) on this study as a research assistant, refining and testing the model code and formulating the mathematical representation of the signal generator framework. Furthermore, some of the earlier results of the research done in this thesis were presented in the study as a potential application of the framework.

# 2 Signal quality assessment

The objective of *signal quality assessment* (SQA) is to improve the reliability of physiological measurements obtained via wearable sensors, for it is not always possible to extract valid measurements from a corrupted signal solely by using noise suppressing techniques [3]. ECG recordings are often corrupted by noise and artefacts that hinders the recognition of the morphological properties of ECG causing false alarms leading to alarm fatigue and reduced attentiveness among healthcare personnel [4]. Moreover, with a reliable SQA method, unusable data can be filtered out before (automated) diagnostic processes. A trained deterministic model is also a consistent way to annotate data.

To evaluate the SQA performance, a consistent method of manually annotating the signal quality must be considered. For a measured noisy signal it is practically impossible to isolate the effect of noise or an artefact from the pure signal, therefore an expert annotation is used for labeling the real-world ECG data.

In Section 2.1, ECG is briefly explained only to a necessary extend for providing a sufficient understanding of the morphological properties that are essential for quality criteria. This is done with the guidance of the excellent book "The only EKG book you'll ever need" by Malcom S. Thaler [5]. In Section 2.3 the concept of SQA in relation to ECG is discussed and in Section 2.4 the use of modular synthetic signal generator for generating training data for machine learning models is motivated.

## 2.1 About ECG

Cardiac cells are negatively charged, with respect to their outsides, in their resting state. This internal negativity is lost when there is a positively charged adjacent cardiac cell. This process is called *depolarization* that is caused by a spontaneous depolarization of the pacemaker cells of the *sinus node* located in the right atrium. After the depolarization, cardiac cells restore their resting state in a process called *repolarization*. These processes are depicted in Figure 1. These electrical events can be detected by electrodes placed on the surface of the body and the recording of these events is called an electrocardiogram.

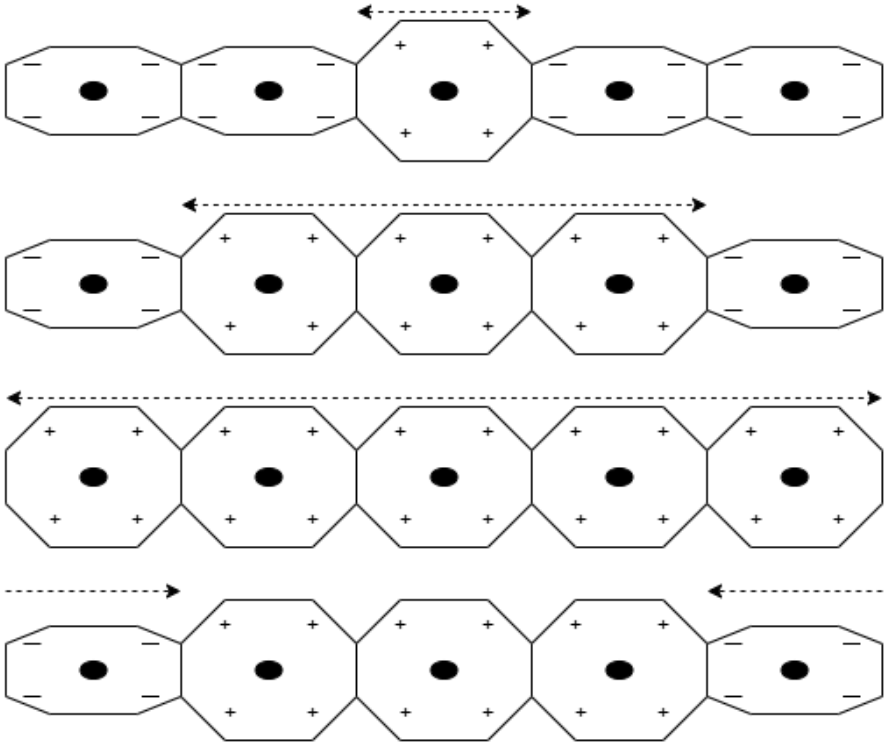

Figure 1: The octagons representing the cardiac cells, the depolarization and repolarization spread to adjacent cardiac cells in a "first in, last out" manner.

The ECG consists of three parts called *P-wave*, *QRS-complex* and *T-wave*. The spontaneous depolarization of the sinus node spreads through the atrial myocardium from right atrium to the left one, that causes the atrial contraction and is detected as the P-wave. After the P-wave the ECG is flat due to an electrical gate that in a healthy heart prevents the depolarization spreading from atria to ventricles, thus allowing the atria to finish contracting before the ventricles. The charge is conducted by the *atrioventicular (AV) node*, that branches to both ventricles' myocardia, slows down the conduction and after a fraction of a second leads to ventricular myocardial depolarization that causes the ventricular contraction. This event is detected as the QRS-complex, where the first negative deflection is the *Q-wave*, the large positive deflection is the *R-wave* and the negative deflection after R-wave is called the *S-wave*. After the ventricular contraction and a brief refractory period, the myocardial cells are ready to repolarize, usually beginning from the last depolarized cells, and this event is detected as the T-wave. These events of the ECG are depicted in Figure 2.

In this thesis the considered ECG is single-lead. The sign of the wave is determined by the direction of the depolarization or repolarization in relation to the electrode that is positioned somewhere over the left ventricle. Thus, the atrial and ventricular depolarizations are propagated towards the electrode. The negativity of the Q-wave is caused by the left-to-right depolarization of the interventricular septum. The S-wave is negative, because the depolarization of the ventricles passes by the electrode. The repolarization causes a positive wave, when receding from the electrode, thus the T-wave is positive.

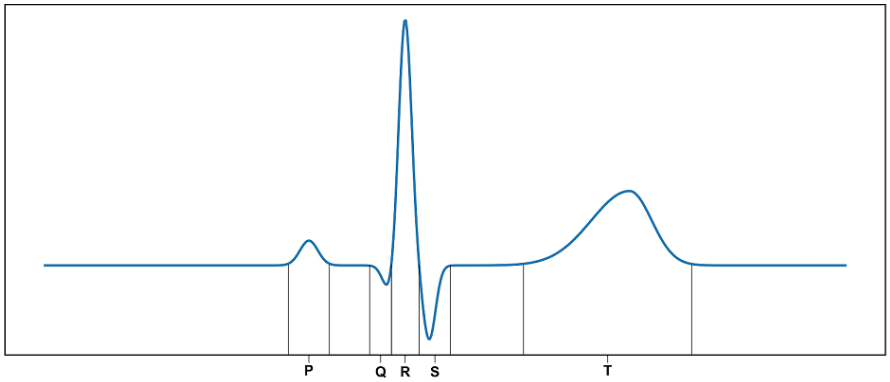


Figure 2: ECG over one cardiac cycle.

## 2.2 Previous studies

In 2022, van der Bijl et al. conducted a literature review on machine learning methods assessing ECG quality that included 19 articles from January 2012 to January 2022 [6]. All of these methods utilized real-world records or augmented measurements as training data. None of these studies published the code they used.

In the 2018 review of signal processing techniques and ECG SQA by Satija et al. review 47 SQA methods that they grouped into the five following categories [7].

- Fiducial features and heuristic rules based SQA methods.
- Fiducial features and machine learning based SQA methods.
- Nonfiducial features and heuristic rules based SQA methods.
- Nonfiducial features and machine learning based SQA methods.
- Filtering-based SQA methods.

Here the fiducial features are specified as morphological features of the ECG waveform and heartbeat interval features, and the non-fiducial features are specified as time-domain, frequency-domain, time-frequency-domain, statistical, and information theoretic features.

Most of these studies used the pubilcly available PhysioNet/Computing in Cardiology (CinC) challenge 2011 dataset [8], where a group of annotators with varying levels of expertise assessed signals into 5 categories:

- A: (0.95): excellent,
- B (0.85): good,
- C (0.75): adequate,
- D (0.60): poor,
- F (0): unacceptable.

The semantics of these categories were not elaborated on. According to the average of these grades, each record was then assigned to one of the following three groups.

- Group 1 (acceptable): The average grade is greater or equal to 0.70, and at most one grade is F.
- Group 2 (indeterminate): The average grade is greater or equal to 0.70, but two or more grades were F.
- Group 3 (unacceptable): The average grade is less than 0.70.

These records were assigned into the groups with an approximate distribution of 70-1-30 respectively.

The second most prevalent data used in these studies was from the MIT-BIH Arrhytmia Database [9]. These are long signals of which two cardiologists annotated the quality changes from clean to noisy and vice versa.

The dichotomous approach, acceptable and unacceptable, as quality groups is most widely adopted, although there are studies where quality categories are divided into three or five groups [7]. Depending on the intended use of the quality classifier, more categories or a continuous quality values would be more useful, considering the variety of information that can be extracted from ECG [10].

The review [6] divides the automatic ECG quality assessors into two categories: *feature-based* (*FB*) and *non-feature-based* (*NFB*). The FB-methods are based on *signal quality indices* (*SQI*), that are features representing different aspects of quality extracted from the signal. These features often included checks for background noise, beat consistency, amplitude range and QRS detection. These FB-methods usually employ a stepwise quality check, where each index is incrementally considered, if previous ones are considered acceptable. Some of the open-source SQI tools are found to be inconsistent with each other and with experts [11]. 14 out of the 19 reviewed classifiers employed a feature based classifier, from which two classifiers utilized both FB and NFB classifiers.

Feature extraction and selection require a deep understanding of signal processing and analysis and metrics to evaluate the importance of these features regarding the quality. FB methods also require the feature extraction to be performed on every input signal, which might be computationally too expensive for real-time quality assessment.

In this thesis, as will be presented in Section 5.2, the single-lead ECG quality classifier is developed as an NFB model. Out of the five reviewed exclusively NFB methods, only two used one lead ECGs. These were both conducted by Álvaro Huerta et al. In the first study, five well-known deep convolutional neural network models: AlexNet, VGG16, GoogLeNet, ResNet18 and InceptionV3, were compared [12], and in the second study, the training of AlexNet model with real-world records and augmented records were compared. Both of these studies utilized two-category expert annotated 5-second signals from the PhysioNet/CinC Challenge 2017 [13] database.

## 2.3 Assessment criterion

The quality of the signal is determined, not solely by the noise, but the relation of the noise to the signal; what properties of the signal are obscured by the noise. Furthermore, the quality measure is dependent on the information of the signal seen as valuable. For instance, if ECG is used to obtain heart rate, only R-peak detection is relevant. These measures are practically impossible to ascertain unequivocally, for the noise cannot be separated from the measured signal. Therefore, a criterion must defined, by which the real-world signals are assessed. The quality labels should correspond to the different use-cases the signal provides and be as distinct as possible. Naturally, visual inspection leads to a great amount of ambiguity. By these guidelines the assessment criterion for the quality of an ECG signal was determined as follows.

- Level 1 (good signal): P-waves, QRS-complex and T-waves are recognized by visual inspection.
- Level 2 (moderate noise): P-waves or T-waves are not recognizable, but the signal can be identified as an ECG and R-peaks are completely recognizable, as there should be no spikes with higher amplitude than QRS-complex, thus the heart rate can be calculated from the signal.
- Level 3 (severe noise): The signal can hardly be identified as an ECG and QRS-complexes cannot be recognized with certainty.

By this criterion, baseline wander can be disregarded. This also affects the choice of a signal filter for it is advantageous to disregard frequencies that do not contribute to these waves. Furthermore, these quality categories are not necessarily dependent on the amount of noise, but the morphological properties of the signal. For instance, the occurrence of an atrial fibrillation may affect the recognition of the P-wave, thus a pure signal would be annotated as level 2.

## 2.4 Rationale for utilizing synthetic data

In respect of quality assessment using real-world records, the classification of the train data is conducted by expert annotation. Due to the ambiguity of the labels, it is preferable to have multiple expert annotators. This is time consuming, requires discrepancy revision and is biased according to the annotators.

Regarding sensitive medical data, privacy must be considered. ML models trained with private dataset are vulnerable to model inversion attacks [14]. This can be mitigated, for instance, via implementation of differential privacy [15] at the cost of model performance [16]. The use of public data relies on the availability and quality of suitable data, and testing the performance of an ML model using data from one source overestimates its generalizability, giving an overly optimistic view of its accuracy outside of the data source [17]. Moreover, the usage of public ECG data may introduce prevalence bias due to diagnostic labels and without augmentation, the data may be imbalanced.

The synthetic ECG generators can be divided into three categories: parametric models, physiological models and deep learning models. The parametric modular biosignal generator used in this thesis, which is presented later in Section 3, provides tools for generating unlimited amount of quality-labeled data with no privacy concern. Physiological models generate synthetic ECG signals via simulation of the heart. These are good at considering the physiological variations, but are limited considering the randomization of the signals. The physiological models are also complex and computationally heavy [2]. The deep learning models, such as generative adversarial networks, are trained with real-world data. Hence, the challenges of sufficient and representative data and privacy concerns are also true for these models [18].

The advantages of the modular parametric synthetic biosignal model are numerous.

- Unlimited amount of arbitrarily distributed data.
- No privacy concern.
- Quality can be determined automatically and consistently.
- Allows *domain randomization*.

The pure signal can be configured to preserve selected morphological properties or to include anomalies. The quality can be determined by the noisy ECG signal, solely by the added noise or by the relation of the noise to the signal. This is unique for synthetic data due to the modular structure, even though with a robust database of ECG records assessed as pure, this noise module can also be used to augment a dataset of this kind with the same modular benefits. Moreover, the quality can depend on the parameters that are used to generate the noise or properties extracted from the noise signal, such as the total power of the noise.

Therefore, provided a good implementation of the desired quality criterion, the quality assessment of the synthetic signals can be automated such that large amounts of assessed signals can be generated in an instant. In addition, the quality criterion is consistent and easily replicable, for this does not depend on visual assessment.

The correct identification of the waveforms P, QRS and T is challenging because of the variability of the signals between different individuals, and also between different recordings due to the placement or type of the electrode [3]. Therefore, an automatic SQA model will require a comprehensive training data that the synthetic data model can provide. In addition, the synthetic data model enables domain randomization [19], that should increase the ability of the SQA model to generalize well on unseen signals, and is seen as beneficial method for training neural networks that use ECG data [20]. The principle of domain randomization is elaborated on in Section 3.8.

# 3 Synthetic data

Karhinoja et al. presented a framework for generating synthetic biosignals that is publicly available on GitHub [2][21]. The modular structure of this biosignal generator allows the ECG signal and the noise to be generated separately. This enables the quality labeling to be dependent solely on the noise. The noisy signal is obtained by simply adding the generated noise signal to the generated pure ECG signal.

This section describes the essential parts of the framework and its modules relevant to this thesis. Moreover, the implementation of the quality classifier and its operating principles are presented. The block diagram of this synthetic signal generator integrated with the automatic quality classifier is presented in Figure 3.

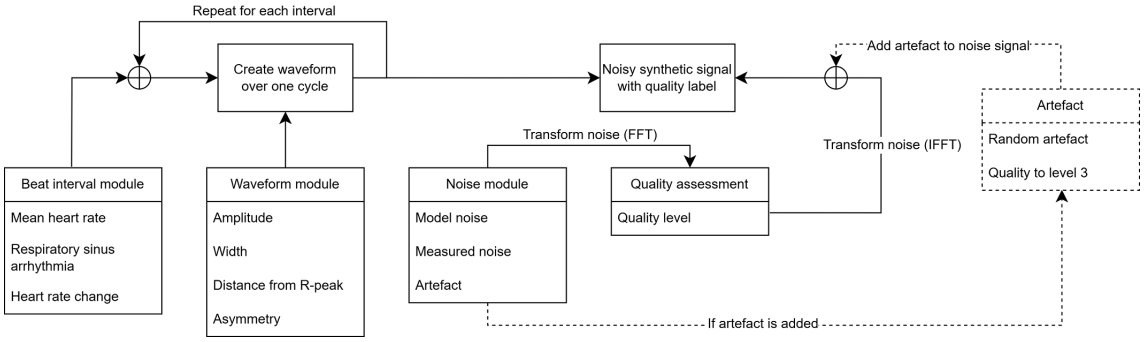


Figure 3: Block diagram of the synthetic quality-labeled signal generator.

## 3.1 Beat intervals

The purpose of the beat interval module is to generate a sequence of the lengths of cardiac cycles. The fluctuation of the generated sequence should mimic the real-world heart rate variability. In this thesis only short quality-labeled signals are considered. Therefore, the representativeness of long-term variations are not important, but rather the generation of sequences with varying beat interval lengths and short-term heart rate changes.

**Definition 1.** The length of the $i$th cardiac cycle is

$$\theta_i := \mu + \beta \sin(2\pi f_b t_{i-1}) + \gamma_i, \text{ where} \tag{1}$$
$$t_i = \sum_{j=1}^{i} \theta_j, \forall i \geq 2 \text{ and } t_0 = 0,$$

in which $\mu$ is the average length of the beat interval, $\beta$ is breathing coefficient, $f_b$ is the respiratory rate, $t_{i-1}$ is the start time of beat interval $\theta_i$, i.e. the length of the beat intervals fluctuates in the range of $[-\beta, \beta]$ due to the rate of breathing. The parameter $\gamma_i$ represents the long-term correlated changes of the beat intervals.

Gradual heart rate change can be generated by using a sigmoid function

$$S(x) = \frac{1}{1 + \exp(-x)}.$$

Denoting the new average beat interval length as $\mu'$, the model (1) can be modified as

$$\theta_i = \mu + (\mu' - \mu)S\left(\frac{x-d}{\tau}\right) + \epsilon_i,$$

where $d$ is the middle point of the transition, $\tau$ controls the rate of transition and $\epsilon_i = \beta \sin(2\pi f_b t_{i-1}) + \gamma_i$.

In this thesis the parameter $\gamma_i$ can be ignored due to the short length of the considered signals. The parameter ranges are presented in Table 1.

## 3.2 Waveform

Intuitively, the derivative of an ECG over one cardiac cycle is drawn around a circle. This circle slides on $x$-axis and rotates with angular velocity of $\frac{2\pi}{\theta_i}$ drawing a signal. The sequence of beat intervals $(\theta_k)_{k=0}^n$ represents the time of each revolution. This signal is then integrated resulting an ECG where the amplitude, width and timing of each wave is dependent on the length of its cycle. This will be presented more precisely in the following.

The *irregular sawtooth function of time* $x : \mathbb{R}_{\geq 0} \to [0, 1]$ is defined as

$$x(t) = \begin{cases} \frac{t}{\theta_1}, & \text{when } t \in [0, \theta_1] \\ \frac{1}{\theta_k}\left(t - T_{k-1}\right), & \text{when } t \in (T_{k-1},\, T_k]\,,\ \forall k \geq 2 \end{cases},$$

where $T_k = \sum_{i=1}^{k} \theta_i$. Now a *phase function*

$$\phi : [0, 1] \to [-\pi, \pi],\ \phi(x) = 2\pi x - \pi \tag{2}$$

maps each cardiac cycle as a full revolution, and $(\phi \circ x)(t)$ is a function of time.

Assume that the events $P$, $Q$, $R$, $S$ and $T$ are similar to the Gaussian function

$$a \exp\left(-\frac{(x-\mu)^2}{2w^2}\right),$$

where $a$ is the maximum, hence the amplitude, $w$ affects the width and $\mu$ is the location of the peak on $x$. The partial derivative of a modified Gaussian function of phase is denoted as

$$g(\phi) := \frac{\partial}{\partial x} a \exp\left(-\frac{m\phi^2}{2w^2}\right) = a \exp\left(-\frac{m\phi^2}{2w^2}\right)\frac{\partial}{\partial x} - \frac{m}{2w^2}\phi^2$$

where the parameter

$$m = \begin{cases} m_1 > 0, \text{ when } \phi \leq 0 \\ m_2 > 0, \text{ when } \phi > 0 \end{cases}$$

was added to introduce asymmetry for the waves. Furthermore,

$$\frac{\partial}{\partial x} - \frac{m}{2w^2}\phi^2 = -2\frac{m}{2w^2}\phi\phi' = -\frac{2\pi m}{w^2}\phi, \text{ because } \phi'(x) = 2\pi.$$

Thus,

$$g(\phi) = a \exp\left(-\frac{m\phi^2}{2w^2}\right)(-\frac{2\pi m}{w^2}\phi) = -\frac{2\pi ma\phi}{w^2}\exp\left(-\frac{m\phi^2}{2w^2}\right).$$

Each wave is generated using the function $g$. The position of each wave is determined by first expanding the phase function's (2) domain $[0, 1]$ to the set of real numbers as

$$\phi : \mathbb{R} \to [-\pi, \pi),\ \phi(x) = 2\pi(x - \lfloor x \rfloor) - \pi, \tag{3}$$

such that the phase of each event can be shifted with a parameter $d \in [0, 1]$ as $\phi(x + d)$. Now the synthetic ECG signal can be constructed via functions $g$ with wave $j \in \{P,\ Q,\ R,\ S,\ T\}$ specific parameters.

**Definition 2.** The *synthetic pure ECG signal ECG* is the sum of wave functions integrated over time:

$$\begin{aligned} ECG &:= \int ecg\ dt, \text{ where} \\ ecg &:= \sum_{j \in \{P,\ Q,\ R,\ S,\ T\}} g(\phi; j) \quad \text{and} \\ g(\phi; j) &= -\frac{2\pi m_j a_j \phi(x + d_j)}{w_j^2} \exp\left(-\frac{m_j \phi(x + d_j)^2}{2w_j^2}\right). \end{aligned}$$

The parameter ranges are presented in Table 1.

## 3.3 Noise

The noise is generated both by using a mathematical model and real-world measurements. This mathematical model combines and randomizes $1/f$*-noise*, also referred to as *pink noise*, and *white noise*. The $1/f$-noise amplifies low frequencies, introducing baseline wander to the signal, whereas the white noise amplifies all frequencies equally. These determine the frequency components of the noise. Additionally, the framework allows to add noise of an arbitrary frequency, but to simplify the preprocessing of the signals and the quality assessment method, this addition of point frequency noise is disregarded.

The real-world noise is obtained from noisy measurements either by selecting a random segment from the noisy measurement or generating randomized long-term noise from the frequency components of the noisy measurements, similarly to model noise.

In the following, the concepts essential to the operation of the noise module are presented.

**Definition 3.** The *power spectral density (PSD)* of a signal is the distribution of power over the frequency components of the signal.

Due to the scope of this thesis, the *Fourier transform* is considered solely as a signal processing method.

**Definition 4.** Fourier transform is a transformation of a signal to its frequency domain:

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x) \exp(-i\omega x)\, dx,$$

and its inverse is a transformation of the frequency components to a signal of time:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) \exp(i\omega x)\, d\omega. \tag{4}$$

For the discrete signals, this transformation is applied as *fast Fourier tranform* (FFT) and its inverse from frequency domain to time domain as *inverse fast Fourier transform* (IFFT).

### 3.3.1 Model

The *synthetic noise* is generated by randomizing a PSD of the weighted sum of 1/f noise and white noise:

$$\frac{1}{f^{\alpha}} + \sigma^2, \quad \alpha, \sigma \geq 0. \tag{5}$$

The $1/f$-noise is standardized by dividing by its average $\bar{P}$ and multiplied by a parameter $c$.

**Definition 5.** The noise model is a combination of weighted $1/f$-noise and white noise represented as a parametric function of frequency:

$$PSD(f;\, c, \alpha, \sigma^2) = \frac{c}{f^{\alpha}}/\bar{P} + \sigma^2.$$

Now, for the same frequency interval $f \in (a, b]$, in which $a = 0$ and $b = \frac{f_s}{2}$, i.e. the *Nyquist frequency*, the $1/f$ and white noise are comparable by the constant $c$. The higher values of $\alpha$ emphasizes the lower frequencies, whereas the parameters $c$ and $\sigma$ determine the overall effect of the $1/f$-noise and white noise respectively.

In discrete frequency domain, i.e. $(f_k)_{k=1}^{n} = (f_1, ..., f_n),\ 0 < f_j < f_{j+1}\ \forall j$, the randomization is conducted by multiplying the value of each frequency $PSD_j := PSD(f_j)$ by a complex Gaussian variable $z_j = a_j + b_j i$, where $a_j, b_j \sim N(0, 1)$ and $i^2 = -1$. The randomized square root of the function $PSD$ is denoted as

$$R_j := \frac{\sqrt{PSD_j z_j}}{2}.$$

Considering *Euler's formula*

$$\exp(ix) = \cos(x) + i \sin(x),$$

on the inverse Fourier transform (4),

$$\begin{aligned} \exp(i\omega x) &= \cos(\omega x) + i \sin(\omega x), \text{ and} \\ \sin(x) &= -\sin(-x). \end{aligned}$$

Thus, to obtain a real valued signal via IFFT, the sequence of randomized PSD values has to be symmetric with respect to the origin for its imaginary part. Therefore,

$$R_j := R_j^* \quad \forall j,$$

where $R^*$ is the complex conjugate of $R$ and $R_0 := 0,\ f_0 = 0$. Now

$$IFFT((R_k)_{k=-n}^{n} = (R'_k)_{k=1}^{n}, \quad \text{and} \quad R'_j \in \mathbb{R}\ \forall j,$$

ergo the transformed signal is a real valued signal of length $\frac{n}{f_s}$.

This time-domain noise is then multiplied by an amplitude coefficient $a$ and added to the signal.

### 3.3.2 Real-world measurements

In addition to the noise model, noise can also be generated from real-world signals. In the noise module, there are PSDs of noisy *photoplethysmogram* ($PPG$) records. These can be randomized and transformed into time domain exactly as the model noise. There is also an option to distort the generated signals by adding artefacts loaded from MIT-BIH Noise Stress Test database [22]. These include baseline wander and muscle artefacts. The time, duration and amplitude are determined by noise module parameters.

## 3.4 Model parameters

In comparison to the presented framework [21], the ranges for uniformly distributed waveform parameters affecting the ECG morphology were relaxed. In addition, the randomization of the model parameters through a truncated normal distribution with different values of variances and means was explored. This experiment yielded no observable changes regarding the quality representativeness of the generated signals, that is discussed in Section 2.

The set ranges are presented in Table 1 and the descriptions of the parameters are presented below.

$d$: Location of the peak relative to phase.

$a$: Amplitude of each wave.

$w$: Width of each wave.

$m$: Asymmetry parameter.

$\mu$: Mean beat interval.

$\beta$: Breathing coefficient.

$f_b$: Breathing frequency.

$\alpha$: Exponent of $\frac{1}{f}$ noise.

$c$: Scalar of $\frac{1}{f}$ noise.

$\sigma^2$: Scalar of white noise.

$\mu'$: Mean beat interval after the change in HR.

$l$: Location of the change relative to the number of beats.

$\tau$: Length of the transition.

$p$: The probability of HR change.

| **ECG waveform** | $d$ | $a$ | $w$ | $m$ |
|---|---|---|---|---|
| P | $[-0.35, -0.1]$ | $[0.1, 0.4]$ | $[0.05, 0.6]$ | $[0.8, 3]$ |
| Q | $[-0.05, -0.03]$ | $[-0.7, -0.05]$ | $[0.05, 0.1]$ | $[0.8, 3]$ |
| R | 0 | $[0.8, 1.4]$ | $[0.07, 0.1]$ | $[0.8, 3]$ |
| S | $[0.02, 0.1]$ | $[-0.4, -0.05]$ | $[0.05, 0.1]$ | $[0.8, 3]$ |
| T | $[0.15, 0.35]$ | $[0.2, 0.5]$ | $[0.22, 0.6]$ | $[1, 3]$ |
| **Beat interval** | $\mu$ | $\beta$ | $f_b$ | |
| | $[0.4, 1.6]$ | 0.1 | 0.28 | |
| **Noise** | $\alpha$ | $c$ | $\sigma^2$ | |
| | $[0, 5]$ | $[0, 4]$ | $[0, 3]$ | |
| **HR step** | $\mu'$ | $l$ | $\tau$ | $p$ |
| | $[0.4, 1.6]$ | $[0, 1]$ | $[1, 10]$ | 0.5 |

Table 1: The parameter ranges of the synthetic ECG generator.

## 3.5 Weaknesses of the model

The distances of the waves from R-peak are linearly dependent on the length of the RR interval that is not necessarily the case in real-world data.

The noise in non-deterministic, thus the implemented quality check has to be after randomization of the noise. Therefore, there is no way of generating a specific amount of signals of each category other than repeating the stochastic process until reaching the desired amount of signals for each label. Of course, the parameter values of the noise module can be adjusted, such that the qualities of the generated signals are uniformly distributed.

## 3.6 Quality assessment

The main problem in this thesis is the formulation of the heuristic quality criterion, presented in Section 2.3, into a systematic protocol that can be implemented into

the synthetic ECG model. The first approach was to generate pure signals, and inspect the PSDs of these signals. The extreme cases considering different frequency elements are from the extreme heart rates. These pure signals and their PSDs are displayed in Figure 7. This figure also demonstrates the effect of the chosen filter discussed in Section 3.9.

Low frequency components are seen as baseline wander. This does not distort the signal in a way that would prevent the recognition of waveforms. Hence, these low frequency noise can be discarded. Another consideration concerns high frequencies. By the inspection of frequency components in pure signals, these can also be filtered out, and therefore ignored in terms of quality assessment. This interval of considered frequencies is called a *passband*. In this passband, the components of the pure ECG signals in the low to medium frequencies seems to be more significant. Additionally, the level 2 quality is determined only by the recognition of the QRS-complex. Thus, the following assumptions were made:

- Extreme low and high frequency noise can be filtered out.
- The quality is more sensitive to lower to medium frequencies in the passband.
- Low frequency and high frequency noise in the passband can be tolerated without degradation in quality.
- Different frequency components have a greater impact on the degradation of quality between levels 1 and 2 than between levels 2 and 3.

These assumptions were formalized by determining the quality based on the values of the frequency components of the noise exceeding set boundaries. More precisely, two functions of the form

$$g(f) = \frac{c_1}{(x - f_1)^\alpha} + c_2 + \frac{c_3}{1 + e^{x - f_2}} \tag{6}$$

were set as the boundaries for noise to be allowed for signals of level 1 or 2. Clearly frequencies lower than $f_1$ are not considered and frequencies greater than $f_2$ are tolerated more. Parameters $c_i$ affect the toleration of the corresponding component. These parameters are adjusted for both qualities of level 1 and 2 separately and these boundary functions are visualized in Figure 4.

**Definition 6.** By the noise PSD and the boundaries (6) the quality is set as

$$q = \begin{cases} 1, \text{ when } \int h_1^+ \leq M_1 \\ 2, \text{ when } \int h_1^+ > M_1 \text{ and } \int h_2^+ \leq M_2 \\ 3, \text{ when } \int h_2^+ > M_2 \end{cases}, \text{ where} \tag{7}$$

$$h_i = |R| - a_i g_i, \; h_1(x) < h_2(x) \; \forall x > f_1,$$

$R$ is the randomized PSD of the noise, $a_i > 0$ is a coefficient for the boundary function and $h_i^+$ is the positive part of the function $h_i$.

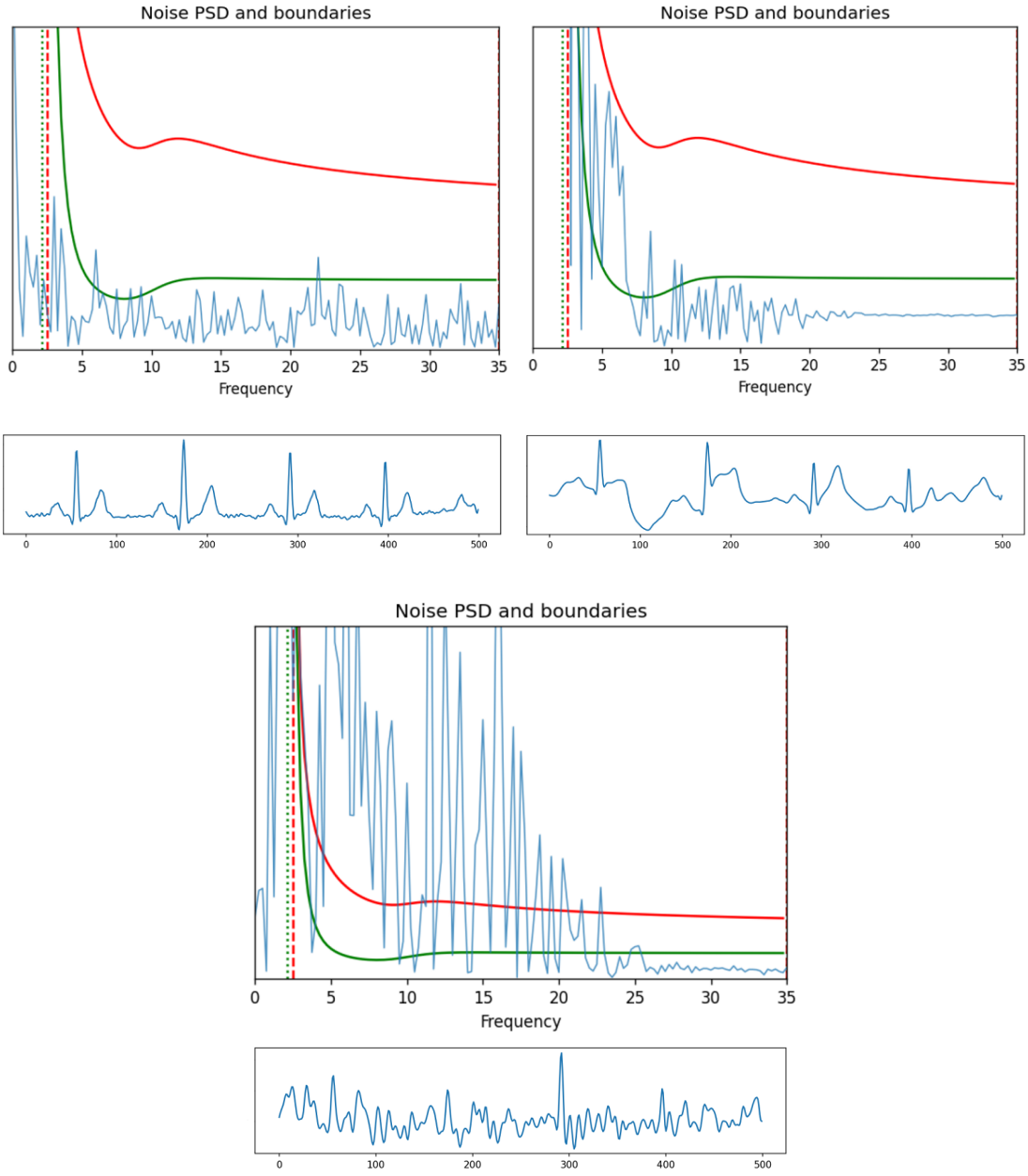


Figure 4: The noise boundaries (6), where the green plot is $g_1$ and the red plot is $g_2$. The blue plot is the PSD of the noise. Below each PSD plot, there is the corresponding noise added to the same synthetic signal. These signals are assessed as level 1, 2 and 3 respectively.

Thus, the quality is determined by the amount of noise exceeding the set limits. Again, by visual inspection, these parameters were adjusted such that the labels of the generated signals align with the criterion. The parameters are reported in Table 2. Furthermore, the parameters $a_i$ and $M_i$ were adjusted during training which is elaborated on in Section 6.3. If an artefact is added, the quality is set as level 3. Examples of these quality-labeled synthetic signals are displayed in Figure 5.

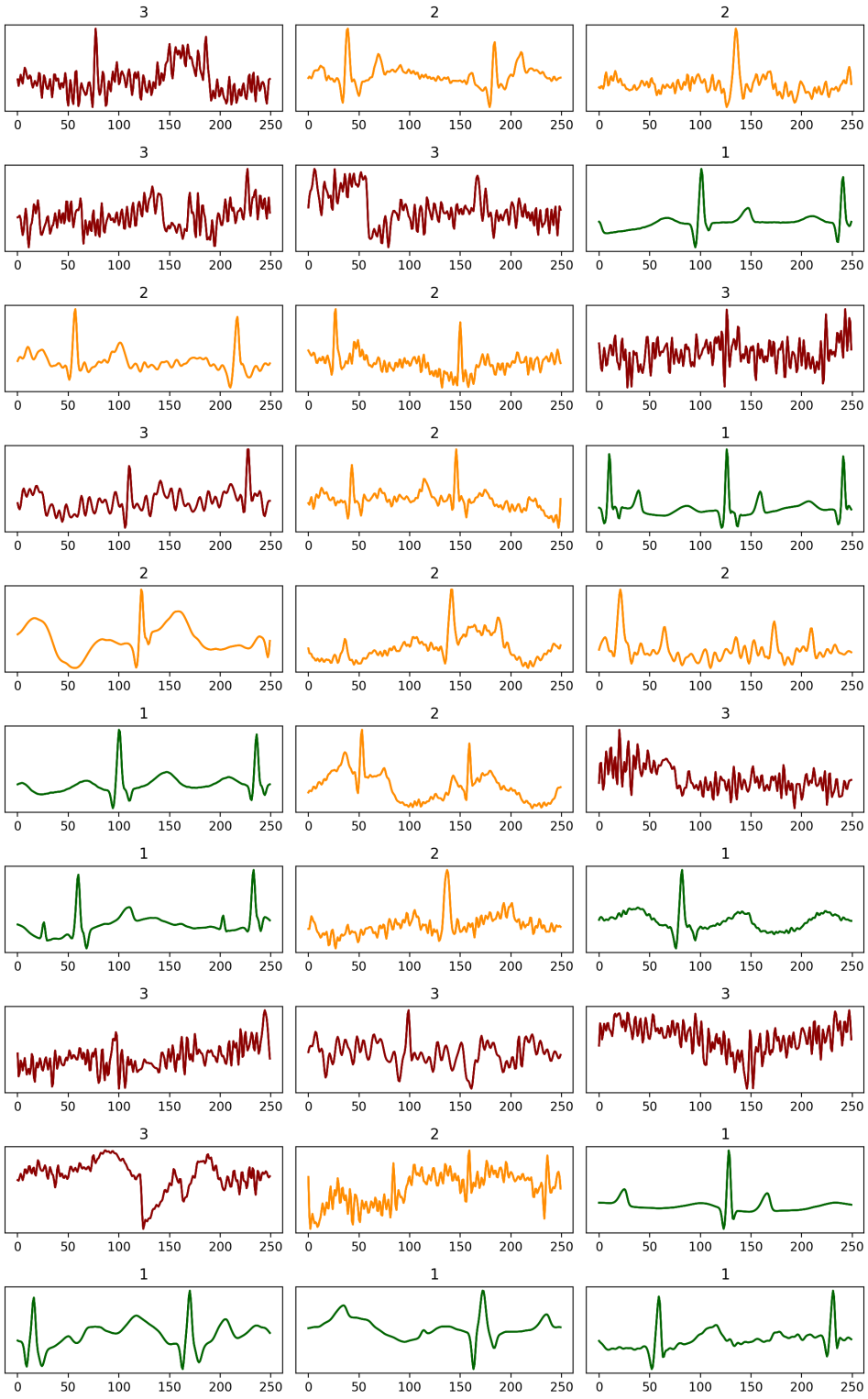


Figure 5: Automatically labeled synthetic signals. The colours green, yellow and red correspond to qualities of levels 1, 2 and 3 respectively as these plots are titled. By inspecting the generated signals, it is apparent that the automatic quality labeling does not fully correspond to the presented quality criterion. Nonetheless, the qualities are generally correct.

The descriptions of the synthetic ECG quality classifier parameters are presented below.

$\alpha$: Determines the slope for shifted $\frac{1}{f}$ noise tolerance.

$c_1$: Tolerance for shifted $\frac{1}{f}$ noise.

$c_2$: Tolerance for white noise.

$c_3$: Increase in white noise tolerance for frequencies higher than $f_2$.

$f_1$: Frequency from which noise is considered.

$f_2$: Frequency from which white noise tolerance is increased by $c_3$.

| **Boundaries** | $\alpha$ | $c_1$ | $c_2$ | $c_3$ | $f_1$ | $f_2$ |
|---|---|---|---|---|---|---|
| Level 1/Level 2 | 2 | 8 | 0.7 | 0.7 | 2.1 | 10 |
| Level 2/Level 3 | 0.7 | 8 | 1.1 | 1.1 | 2.5 | 10 |
| **Adjustable quality parameters** | | $a_1$ | $a_2$ | $M_1$ | $M_2$ | |
| | | 0.3 | 0.35 | 5 | 10 | |

Table 2: The parameters of the synthetic ECG quality classifier.

## 3.7 Alternative automatic QA methods

Two other methods of automatic quality assessment of the synthetic signals were explored. In the first method, the signal quality was assessed by *signal to noise ratio* (*SNR*):

$$\text{SNR} = \frac{P_{signal}}{P_{noise}}.$$

This does not consider the weights of different frequency components on the morphological properties of ECG, hence on the quality. Visually evaluating the generated signals, the quality labels were inconsistent.

In the other, generated synthetic signals were visually labeled and the average value for each of the following frequency bins $\{[0, 2), [2, 5), [5, 10), [10, 20), [20, 35), [35, \infty)\}$ of the PSD of the generated noise were saved as features. On these five features a principal component analysis (PCA) was performed, from which the first two components accounted for 0.98 of the total variance. Using these PCs the clusters were separated by partitioning the two-dimensional space into three parts, yet there was too much overlap for this to be a sufficiently consistent method of classifying the quality.

## 3.8 Domain randomization

Domain randomization is shown to be a useful method in image classification tasks. Using this technique, the simulated or augmented data is randomized to include non-realistic examples that maintain the key features in regards to the classification.

The implementation of domain randomization to neural network training essentially forces the learning of essential features. Or, on the other hand, reduces the learning of irrelevant but domain specific features [19][23].

The synthetic data model introduces domain randomization by simply enabling control over the signal and noise generation parameters [2]. The parameter ranges, shown in Table 1, are loosened from generating only realistic physiological ECG, but the quality criterion is still maintained. Furthermore, given the defined heuristic criterion, the two noisy quality categories can be augmented with non-realistic ECG signals that satisfy the quality constraints. Synthetic signals of quality $q \in \{1, 2\}$ by definition (6), and with P-wave or T-wave amplitude of 0 can be assessed as level 2, because clearly even with no added noise, the P-wave or T-wave or both are not recognizable. Likewise, to level 3 category, pure noise, flatline, PPG or any other non-ECG signals can be added. These augmentations are demonstrated in Figure 6.

Henceforth, in this thesis, the term 'domain randomization' refers to this extreme case of training data enrichment.

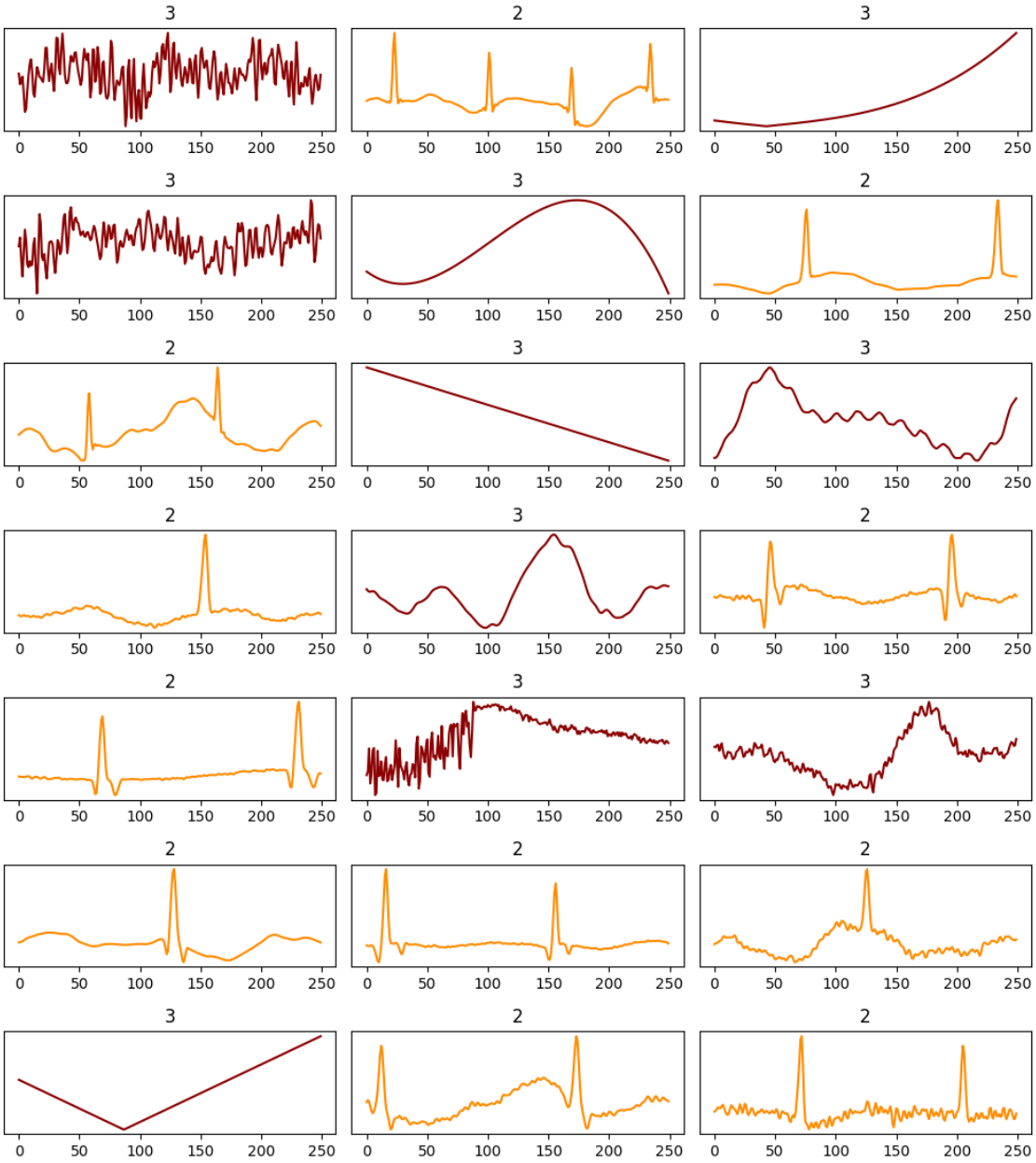


Figure 6: Non-realistic synthetic ECG signals with quality labels. The colours yellow and red correspond to qualities of levels 2 and 3 respectively as these plots are titled.

## 3.9 Butterworth filter

The usage of a signal processing filter was considered for it is disadvantageous to assess the ECG signal by properties that can be filtered without the loss of essential information. The *Butterworth filter* has maximally flat stopband and passband, therefore its usage doesn't distort the signal, when the passband is chosen to filter out frequencies outside the range of pure ECG signal's frequency components [24]. Moreover, the filtering narrows down the components affecting the synthetic quality criterion.

The filter ought to be the same for all signals, thus the range of the passband is justified by inspecting the frequency domain of pure synthetic signals of extreme heart rate. The range of the average length of the beat intervals was restricted to $[0.4,\ 1.6]$. The fluctuation parameters; the long-term correlations and breathing coefficients, were set to 0, leading to two ECG signals with consistent beat intervals: 0.4 and 1.6. These two signals were transformed to frequency domain by FFT. The effect of filtering was inspected and the filter range was set to $[0.4, 35]$. The effect of filtering is demonstrated in Figure 7.

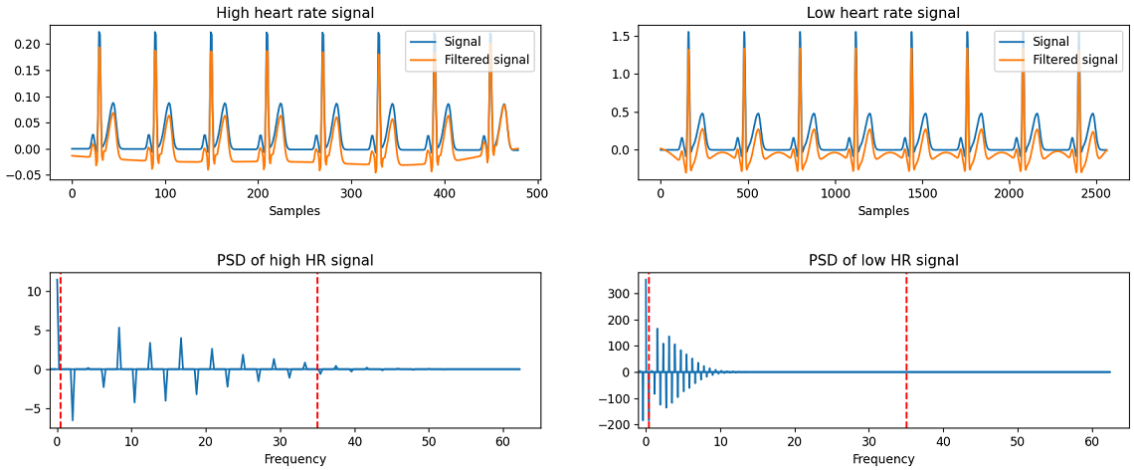


Figure 7: The distortion caused by the use of Butterworth filter on the synthetic signals of extreme heart rates and these signals transformed in to frequency domain by FFT.

# 4 Validation and test data (real-world measurements)

In the measurements the quality of the signal seems to be correlated with the heart rate, which is reasonable, because generally movement generates noise and increases heart rate. This is an example of a feature that an SQA model could learn while using real-world data, but with synthetic data the noise is not, at least necessarily, correlated with the length of the beat interval.

The measurements for validation data were recorded by eight members of the University staff in various states, for instance lying down, sitting, walking, running, cycling etc. The distribution between subjects and activities of this data is presented in Table 3. These signals are single-lead ECG signals that were recorded via Maxim Integrated MAX30003, California. The positioning of the electrode was imprecise to emulate non-professional use and obtain a broad spectrum of data variability. These signals were then resampled to 125Hz and filtered with Butterworth filter with passband of [0.4, 35], which is elaborated on in Section 3.9.

| sub | $n$ | act1 | act2 | act3 | act4 | act5 | act6 | act7 | act8 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 40 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 2 | 38 | 5 | 5 | 5 | 5 | 4 | 5 | 4 | 5 |
| 3 | 48 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 4 | 40 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 5 | 80 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 6 | 40 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 7 | 78 | 10 | 10 | 10 | 10 | 10 | 10 | 9 | 9 |
| 8 | 48 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| Total $n$ | 412 | 52 | 52 | 52 | 52 | 51 | 52 | 50 | 51 |

Table 3: Description of the validation dataset.

The glossary of the actions in Table 3 is presented below.

- act1: Lying on bed.
- act2: Sitting.
- act3: Standing.
- act4: Walking.
- act5: Walking up the stairs.
- act6: Walking down the stairs.
- act7: Jogging.
- act8: Cycling.

The test dataset was obtained from measurements of 29 university staff members (including the 8 subjects in the validation dataset) by continuously measuring while performing and changing between various activities. These signals were longer and much more inconsistent in terms of quality that those in the validation dataset. From the measured signals the first two seconds (250 samples) were ignored, for there is an artefact in the beginning of almost every record.

The 412 measured signals were cut to snippets of the length of 1000 samples (8 seconds), for the quality of the signal should be entirely of one quality category. These snippets were then assessed by visual inspection to six categories:

- Level 0 (mixed signal): The signal quality is not consistent by the assessment criterion.
- Level 1 (good signal): P-waves, QRS-complex and T-waves were recognized by visual inspection.
- Level 1.5: The quality is in between of levels 1 and 2, thus these are left out of the validation data set for not to make adjustments based on the results of ambiguous data.
- Level 2 (moderate noise): P-waves or T-waves were not recognizable, but the signal can be identified as an ECG and R-peaks are completely recognizable, as there should be no spikes with higher amplitude than QRS-complex, thus the heart rate can be calculated from the signal.
- Level 2.5: The quality is between of levels 2 and 3, thus not included to the validation dataset.
- Level 3 (severe noise): The signal can hardly be identified as an ECG and QRS-complexes cannot be recognized with certainty.

This led to 1560 signals of the length 1000 samples, that were labeled as level 1, 2 or 3. The distribution between the quality categories is shown in Table 4.

| Level 1 | Level 2 | Level 3 |
|---|---|---|
| 755 | 652 | 153 |

Table 4: The distribution of the validation dataset.

The test dataset was obtained likewise. From 5259 8 seconds (1000 samples) long snippets were acquired. By selecting randomly, these snippets were assessed such as the validation data, until there was at least a hundred of each label. The distribution of the test set is shown in Table 5.

| Level 1 | Level 2 | Level 3 |
|---|---|---|
| 123 | 287 | 105 |

Table 5: The distribution of the test dataset.

The assessment was performed by one person with only limited amount of expertise in ECG signals. Furthermore, considering the scarcity of subjects, the validation and test datasets are most likely biased. Manually labeled records from validation dataset are shown in Figure 8.

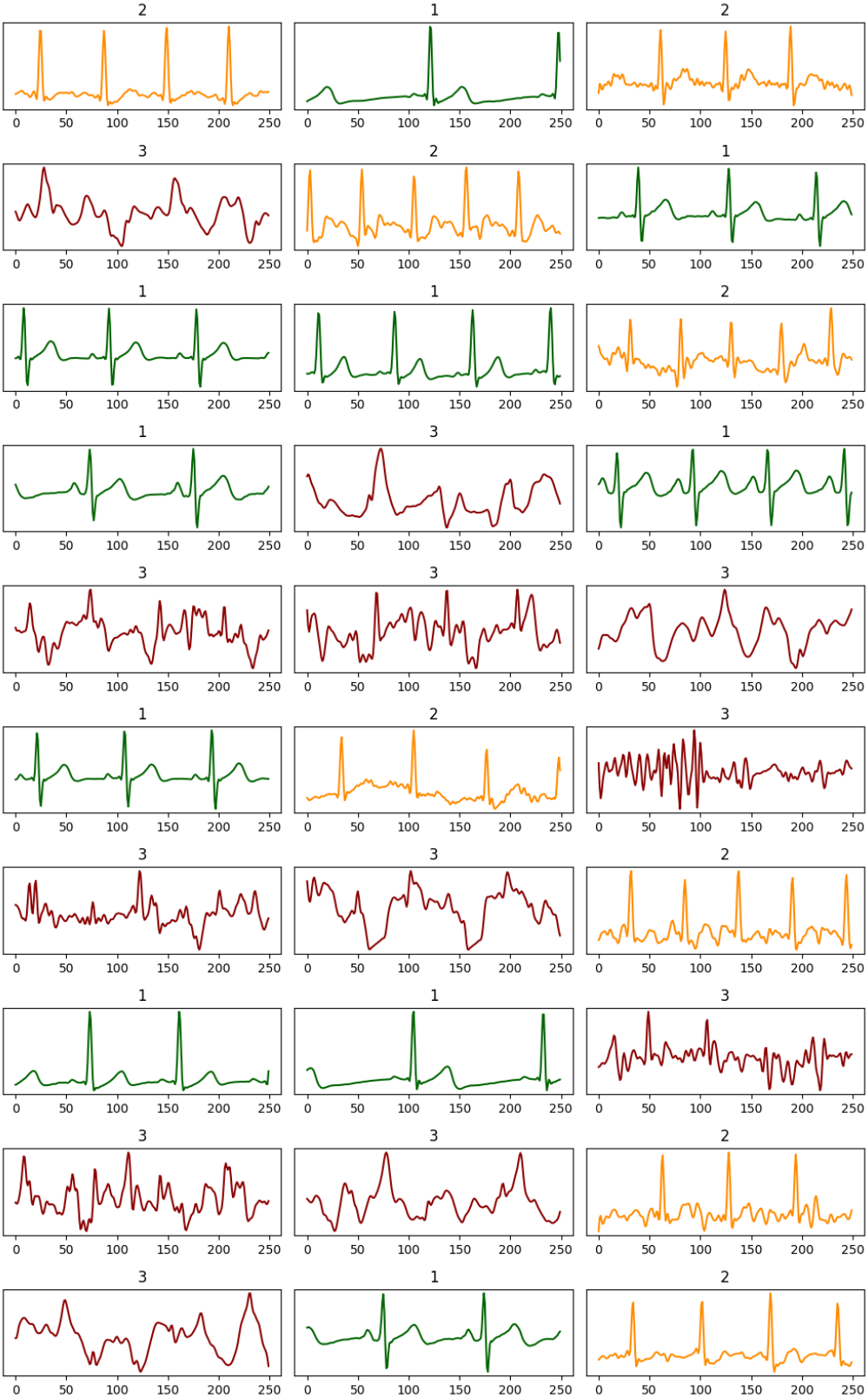


Figure 8: Visually annotated real-world ECG records. These signals were selected as an example, as they unequivocally represent their quality classes.

# 5 Classification method

The aim of this thesis being the exploration of the utility of the synthetic data model in ECG quality assessment, the model architecture was designed to remain simple and computationally efficient. In addition, the lack of available code for existing ECG SQA models [6], discouraged the usage of an existing architecture. In this section the concept of convolutional neural network is introduced based on the book *Deep learning* by Ian Goodfellow, Yoshua Bengio and Aaron Courville [25]. Additionally, the architecture of the quality classifier trained in this thesis is presented and justified.

## 5.1 Convolutional neural network (CNN)

Convolutional neural networks are neural networks that employ convolution to detect features from the input data. The convolution can be thought of as a window or a filter smaller than the amount of input data, moving across the input, which applies a transformation. Applying these filters in neural networks, the transformations are optimized such that the filters detect significant features of the inputs considering the objective. In comparison to tabular data, the CNNs are used for input data of a grid-like structure. This is due to the fact that the filters consider the proximity and order of the input, such as a segment of a signal. CNNs are widely successful in applications such as image or time series recognition and classification.

In this section the concepts of *convolution*, *pooling* and a *convolutional layer* are introduced, and moreover motivated. This section is based on Chapters 5–9 from the book "Deep Learning" [25]. The main focus is on Chapter 9 "Convolutional Networks" which is followed with the exception of focusing on one-dimensional signals of time as inputs rahter than images.

**Definition 7.** Let $x : \mathbb{R} \to \mathbb{R}$ and $w : \mathbb{R} \to \mathbb{R}$ be continuous functions of time. The convolution of the *input* $x$ and the *kernel* $w$ is

$$s(t) = \int x(a)w(t-a)da,$$

where the convolution is usually denoted as $s(t) = (x * w)(t)$ and the output is referred to as the *feature map*. In discrete time, convolution is defined as

$$s(t) = (x * w)(t) = \sum_a x(a)w(t-a).$$

Utilizing convolution in machine learning allows smaller structures, the size of the kernel, of the input to be examined. In this manner, the neural network can learn meaningful features, regarding the objective, using only subsets of the input reducing the memory and computational requirements.

Convolution is a linear operation, thus to enable a neural network model to learn non-linear structures, the output of a convolution should transformed. This is done with activation functions.

**Definition 8.** *Rectified linear unit* (*ReLU*) is a well known activation function

$$\mathrm{ReLU}(x) = \max\{0, x\}, \tag{8}$$

that maps $x$ to its positive part.

In the training of a feed forward neural network, the input $\boldsymbol{X}$ propagates through the network producing an output $\hat{\boldsymbol{y}}$. The output $\hat{\boldsymbol{y}}$ can be seen as a prediction based on the input $\boldsymbol{X}$. This is called *forward propagation*. The difference between the prediction $\hat{\boldsymbol{y}}$ and the true values $\boldsymbol{y}$ are measured with a *loss function*.

**Definition 9.** The loss function

$$l : Y \times Y \to \mathbb{R}_+,$$

assigns a positive real value $l(\hat{y}_i, y_i)$ for prediction $\hat{y}_i$.

One common loss metric for categorical models that is also used in this thesis is *cross-entropy loss*.

**Definition 10.** The *mean cross-entropy loss* is

$$L(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \sum_{n=1}^{N} l_n, \text{ where}$$
$$l_n = -\sum_i y_n \log(\hat{y}_i)$$

is the cross-entropy loss and $\hat{y_n}$ is the predicted probability of the class $n$ given by the *softmax function*.

**Definition 11.**

$$\mathrm{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}. \tag{9}$$

Clearly, minimizing the training loss is analogous to the model learning from the training data. This minimization can be conducted by *back-propagation*, that is a method for computing the gradients, that are the derivatives of the loss with respect to the parameters (weights and biases) of the neurons. These parameters are optimized via *gradient descent*.

The activation function ReLU (8), which maps negative outputs to zero, may cause *dying ReLU problem*, in which neurons become inactive for any input [26]. This can be prevented by using activation functions such as *leaky ReLU* or *exponential linear unit* (*ELU*).

**Definition 12.** Exponential linear unit is an activation function mapping negative inputs as follows:

$$\mathrm{ELU}(x) = \begin{cases} x, & \text{if } x > 0 \\ \exp(x) - 1, & \text{if } x \leq 0 \end{cases}. \tag{10}$$

ELUs are shown to lead to better generalization performance than ReLUs or LReLUs on networks deeper than 5 layers [27], but also proved to be effective choice for the SQA model discussed in Section 5.2.

The output of the convolution and activation function is often further processed.

**Definition 13.** *Pooling* gives a summary statistic of the nearby inputs. Considering only one dimension, the window of the nearby inputs is

$$W_k = \{x[i] \mid ks \leq i \leq ks + |k|\},$$

where $|k|$ is the size of the kernel, i.e. the number of nearby input indices considered, and $s$ is the *stride* of the window. Pooling is the output of a transformation $T$ for each input window:

$$p(k; s, |k|) = T_{i \in W_k}(x(i)).$$

For instance, *max pooling* returns the maximum value within a rectangular neighborhood. This can be used to summarize important characteristics of the input while ignoring the variance of small local translations. In addition, this reduces the input size, improving the computational and memory efficiency.

In this thesis convolutional layer refers to the complex layer terminology used in the book [25] that is presented in Figure 9. Using these convolutional layers in a neural network creates a CNN.

**Definition 14.** Convolutional layer consists of three stages:

- Convolution stage: Affine transform
- Activation: Non-linearity
- Pooling: Invariance to small translations

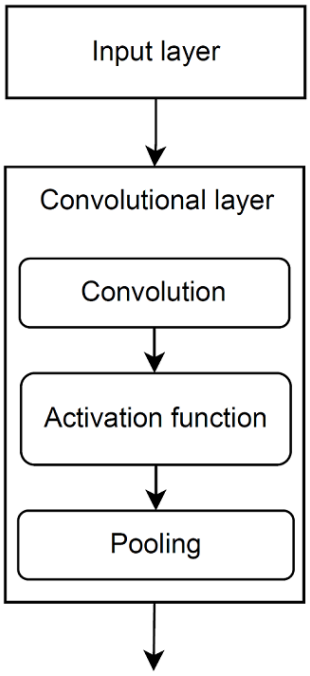


Figure 9: The components of a convolutional layer.

**Definition 15.** A convolutional neural network is a feedforward neural network, where at least one hidden layer is a convolutional layer.

Furthermore, the following regularization methods might be beneficial additions to the CNN architecture.

**Definition 16.** *Dropout layer* zeroes elements from input with set probability $p > 0$. The choice of these elements are independent and Bernoulli distributed.

Implementing dropout is an efficient way to reduce overfitting, and thus improve generalization [28].

During model training, the training error or train loss steadily decreases, because the model is adapting to the distribution of the training data. Using the validation dataset to calculate validation loss, without back-propagation, it is often observed, that there is a point at which validation loss starts increasing. This is due to overfitting to the training dataset.

**Definition 17.** *Early stopping* is a strategy, where the model training, i.e. updating the weights, is interrupted or terminated based on the increase of validation loss or the divergence of train and validation loss.

There are multiple ways to implement early stopping.

## 5.2 SQA classifier architecture

The SQA is conducted via CNN. To detect noise in specific sections of a long signal, the classifier should be trained with either short categorized snippets or a long signal with quality index for each sample. The latter is more demanding in terms of work load and expertise regarding both the assessment of real-world data and training the neural network. Also the score for quality would be less straight forward. Hence, it was decided to utilize short signals, more specifically with a duration of two seconds or 250 samples. In this manner, the classifier can be employed to assess longer signals by convolution.

The classifier is a simple three-layer convolutional neural network. The model architecture is presented in Figure 10. The input layer receives a 250 sample ECG signals that undergo max pooling with a kernel size 4, following a convolutional layer with a convolution kernel size 7, producing 32 output channels, ELU activation function and max pooling with a kernel size 2. The hidden layers follow the same structure, of which the first hidden layers employs convolution with a kernel size 13, producing 64 output channels, followed by ELU activation function and max pooling with a kernel size 19. Likewise, the second hidden layer employs convolution with a kernel size 13, producing 128 output channels leading to an ELU activation function. This output is then flattened, and a dropout layer is employed with a probability of 0.2. An affine linear transformation is then applied, producing a value for each quality class, which are transformed into weight values in the range of $[0, 1]$ using the softmax function (9).

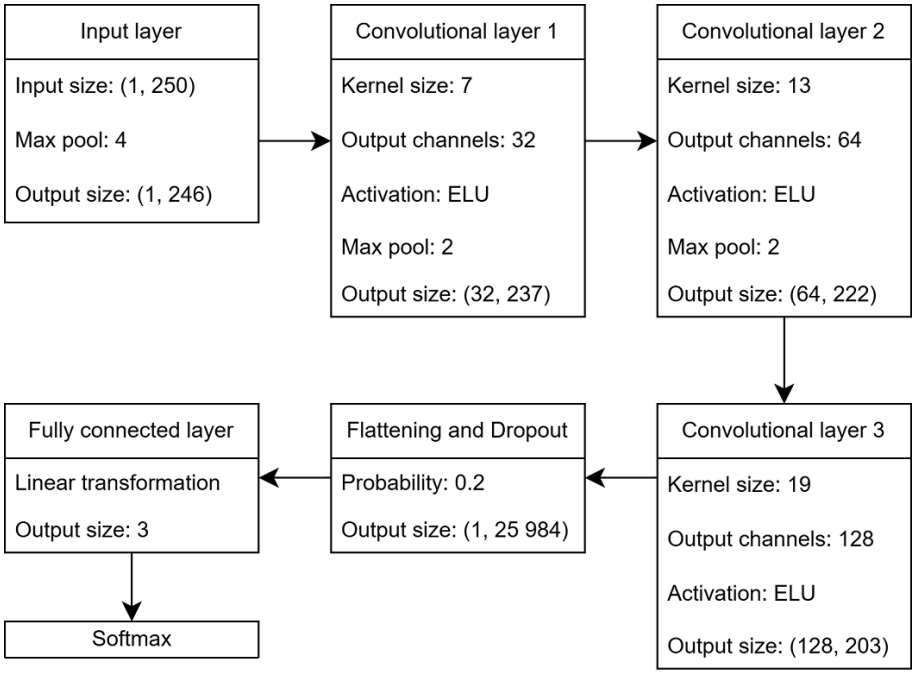


Figure 10: Model architecture.

The choice of this architecture was justified via random search varying in the number of layers, the size of the kernels, strides, activation functions etc, conducted with the training and evaluating with the validation data that was divided into train and test sets. The same signals were used multiple times by drawing random two-second snippets from the 8-second validation signals to increase training data and to prevent the learning of the phase of the signals. Due to the scarce subjects of the validation dataset, this method of choosing model architecture is most likely biased.

# 6 Experiments

In this section the different experiments and training specific parameters and protocols are presented. In addition the evaluation metrics are discussed. The training was conducted with Python version 3.12.4 using PyTorch version 2.4.1.

## 6.1 Training parameters

The gradient decent was performed using *Adam* (from adaptive moment estimation), introduced in 2015, that is an effective optimization algorithm. Adam optimizer was chosen as it typically requires minimal hyperparameter tuning [29].

The training loop consisted of generating 1500 synthetic signals, training the model with *learning rate* of $\frac{1}{10000}$ and using *minibatch* with batch size of 64. The *number of epochs*, i.e. the number of iterations over the whole train set of 1500 signals, was set to 100, but with early stopping the usual number of iterations were less than 5. Once the early stopping conditions were met, the model performance was evaluated using validation dataset and the evaluation metrics, introduced later in Section 6.2. Following this, a new round of iteration began by generating 1500 new signals, and so forth.

The early stopping was implemented as a counter of over how many consecutive epochs the validation loss was greater than train loss. In addition, the counter was decremented by one every time the validation loss was decreased from previous epoch. The *patience*, i.e. the threshold of the counter was configured to three for models trained on synthetic signals, and to five for the bench mark model trained with validation data. The small patience value for models using synthetic data was to prevent excessive growth of the validation loss, which was notably more typical for a model trained on synthetic data than for the benchmark model. The behavior of training and validation losses is illustrated in Figures 13 and 14.

The trend of accuracy on validation data during training is displayed in Figure 11. The training was stopped after 200 iterations and the model with best accuracy on validation data was saved.

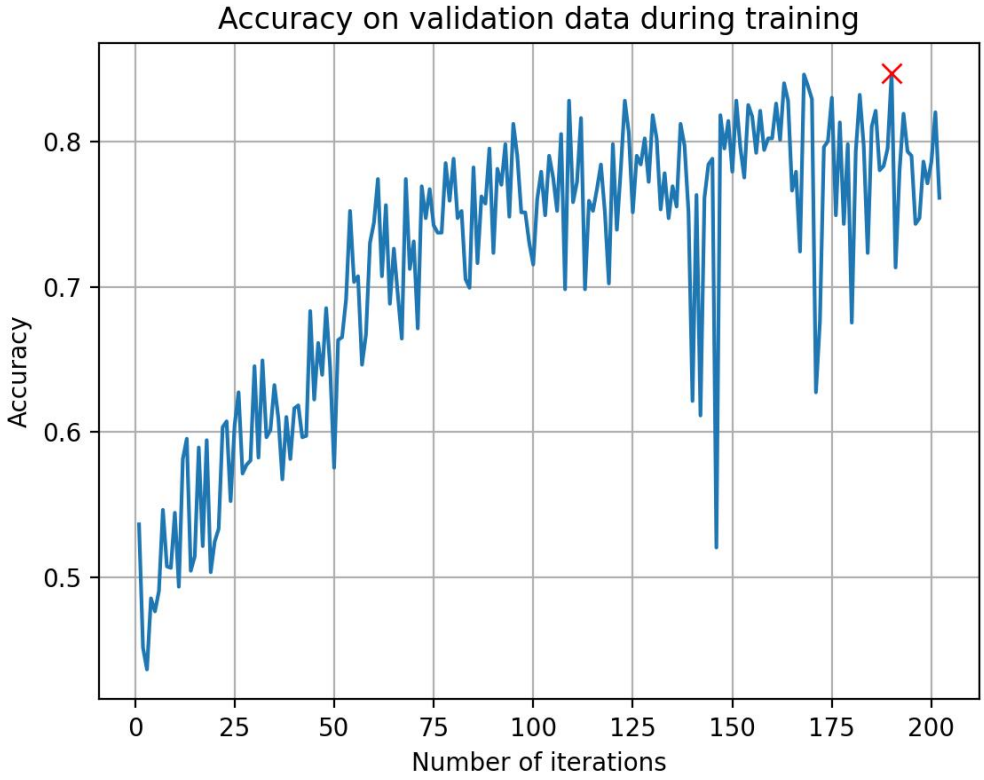


Figure 11: The accuracy on validation data during the training of SQA model on only synthetic data. The red X symbol marks the accuracy on validation data of the chosen model.

Furthermore, another model was trained in the same manner as the first one, but the training data was augmented by domain randomization as demonstrated in Section 3.8. This was done by adding 400 synthetic non-realistic ECG signals of quality levels 2 and 3, 200 of each. Also the training was stopped after 100 iterations. The trend of accuracy on validation data during training is displayed in Figure 12.

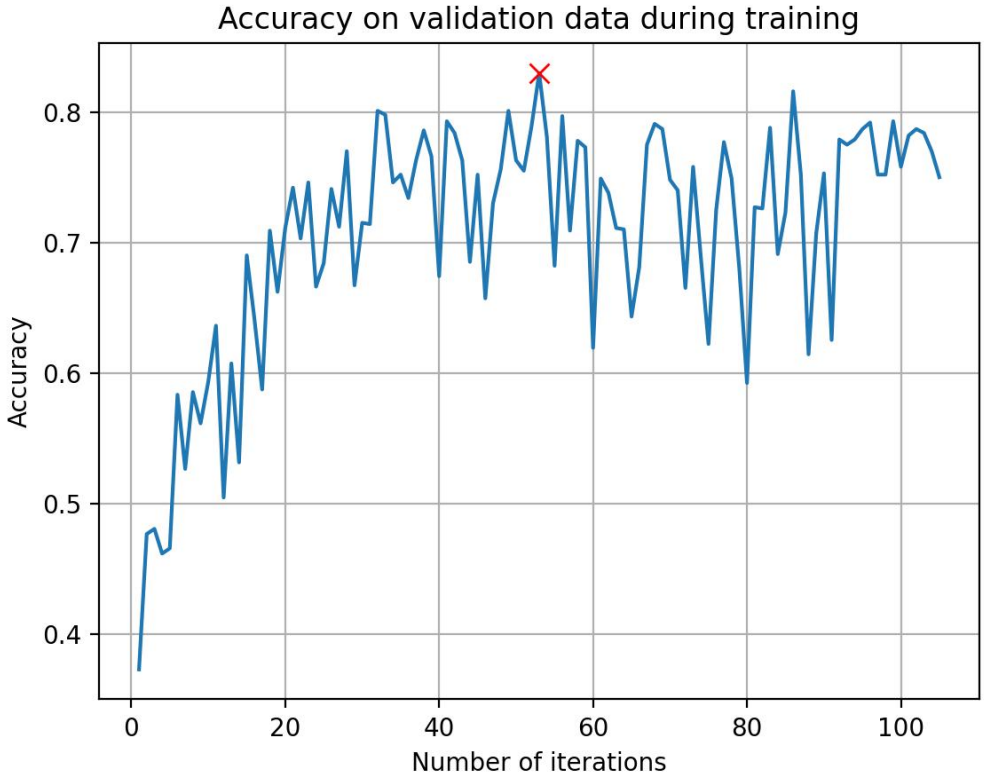


Figure 12: The accuracy on validation data during the training of SQA model on domain randomized synthetic data. The red X symbol marks the accuracy on validation data of the chosen model.

## 6.2 Evaluation metrics

To measure model performance, accuracy and balanced accuracy, that is accuracy weighted to consider class prevalence, are used. Clearly, in three-class classification, accuracy score of 0.33 corresponds to coin toss.

Moreover, *AUCs* for each quality class and *macro averaged AUCs* are used. These are obtained as follows.

**Definition 18.** *True positive rate* (*TPR*) is the probability of true labeling and *false positive rate* (*FPR*) is the probability of falsely labeling as positive:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} \quad \text{and} \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}},$$

where TP is the number of true positives, FP is the number of false positives, P is he number of all positives and N is the number of all negatives.

**Definition 19.** *Receiver operating characteristic* (*ROC*) curve is the relation of TPR and FPR given any threshold, for which the probability value for given label is considered to be positive.

**Definition 20.** *Area under the receiver operating characteristic* (*AUROC*), abbreviated as AUC in this paper, is the integral of the ROC curve. An AUC score of 1 corresponds to perfect classification and a score of 0.5 corresponds to a coin toss.

Macro averaged AUC is the average of the AUCs for each class. AUC is a reasonable metric to evaluate on each class compared to accuracy, because model predicting all inputs as one label is perfectly accurate for that label, whereas AUC accounts for FPR.

## 6.3 Feedback loop

The training cycle was perfected by using the evaluation metrics. According to the TPR and FPR of the level 1 and 3 signals, the parameters $M_i$ and $a_i$ of the quality (7), where $M_i$ is the maximum power of boundary exceeding noise and $a_i$ is the coefficient for the boundary function (6), were adjusted in the following manner:

$$a_{1,i+1} = \begin{cases} 0.98a_{1,i}, & \text{if } \mathrm{FPR}_1 > 0.1 \\ 1.03a_{1,i}, & \text{if } \mathrm{TPR}_1 < 0.5 \\ a_{1,i}, & \text{otherwise.} \end{cases}$$

$$a_{1,i+1} = \begin{cases} 0.98a_{1,i}, & \text{if } \mathrm{FPR}_1 > 0.1 \\ 1.03a_{1,i}, & \text{if } \mathrm{TPR}_1 < 0.5 \\ a_{1,i}, & \text{otherwise.} \end{cases}$$

$$M_{1,i+1} = \begin{cases} 1.05M_{1,i}, & \text{if } \mathrm{TPR}_1 < 0.5 \\ M_{1,i}, & \text{otherwise.} \end{cases}$$

$$a_{2,i+1} = \begin{cases} 0.9a_{2,i}, & \text{if } \mathrm{TPR}_2 < 0.5 \\ a_{1,i}, & \text{otherwise.} \end{cases}$$

$$M_{2,i+1} = \begin{cases} 1.1M_{2,i}, & \text{if } \mathrm{FPR}_2 > 0.3 \\ M_{2,i}, & \text{otherwise.} \end{cases}$$

This aimed to enforce the recognition of each quality class and specifically to reduce the FPR for level 1 signals.

From training with feedback, two models were saved: the model with the highest validation accuracy and the model with the lowest FPR for level 1 signals and the validation accuracy greater than 0.8. These will be referred to as the *feedback model* and the *high specificity model*, respectively.

## 6.4 Benchmark and data augmentation

A benchmark model was trained exclusively on real-world data, both to justify the choice of model architecture and hyperparameters, and to assess the representativeness of the synthetic data. The validation dataset was divided into train and validation sets with 80/20 split. From these 8-second-long signals five 2-second-long snippets were drawn randomizing the phase of the signals. Therefore, the training set consisted of 6235 signals, and the validation set consisted of 1565 signals, from

1247 and 313 unique recordings respectively. The relative prevalence displayed in Table 4 was maintained. More precisely, the validation dataset consisted of 151 unique recordings of level 1 signals, 131 of level 2 and 31 of level 3. The corresponding values for the training set were 604, 521 and 122.

The benchmark model training was almost identical to the synthetic model, with the exception of not generating more training data of course, and also the patience of early stopping was set to five. This is due to the fact that the validation loss fluctuated more moderately on real-world training data than on only synthetic data. The losses are displayed in Figure 13. The model with the highest accuracy on validation data was saved.

An augmented model was trained with the exact same real-world signals in train and validation datasets as the benchmark model. The training data was augmented with 6235 synthetic signals, thus the amount of training data was doubled. Again, the patience of early stopping was set to five and the best model according to validation accuracy was chosen. The losses of the training are displayed in Figures 13 and 14.

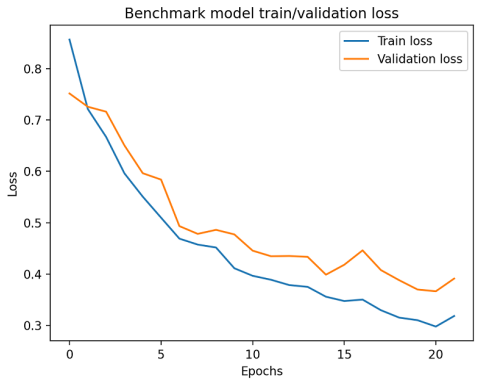


Figure 13: The train and validation loss during training of the benchmark model.

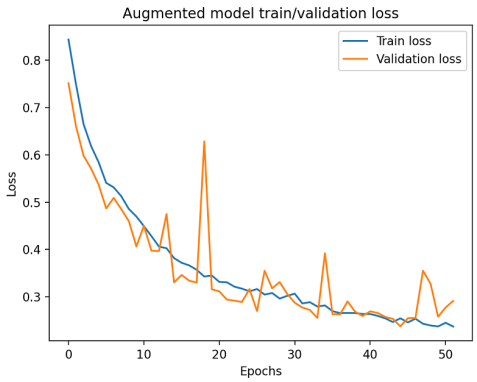


Figure 14: The train and validation loss during training of the augmented model.

# 7 Results

Five SQA classifiers were developed:

- Synthetic: Trained exclusively on synthetic data.
- Domain randomized: Trained exclusively on synthetic data that was augmented with signals presented in Section 3.8.
- Feedback: Trained exclusively on synthetic data with validation feedback.
- High specificity: Trained in the same way as feedback model, emphasizing the specificity of level 1 quality.
- Benchmark: Trained exclusively on validation data.
- Augmented: Trained on both validation data and synthetic data.

These classifiers were tested on dataset formed by selecting 10 random two-second segments from every 8-second signal in the dataset. On this test data, accuracies, balanced accuracies and macro averaged AUCs and AUCs for every quality class were calculated. These scores are presented as bar charts in Figure 15.

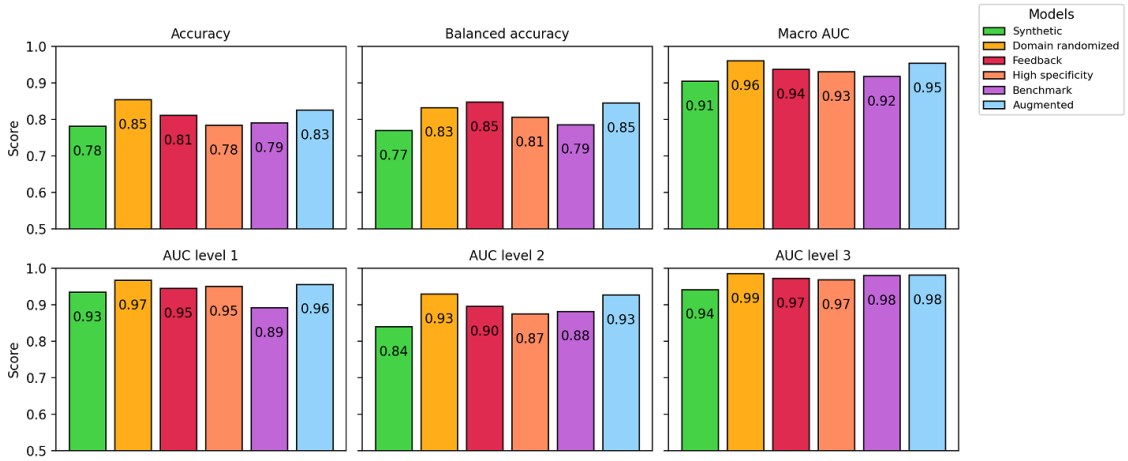


Figure 15: Accuracy, balanced accuracy and AUC scores on test dataset for all models. The order of the bars is synthetic, domain randomized, feedback, high specificity, benchmark and augmented, corresponding to colours green, yellow, red, orange, purple and blue respectively.

The AUC score of the level 2 signals is the lowest for all classifiers. This is reasonable, for the level 2 signals are clearly the most ambiguous in relation to their quality due to being in between qualities of level 1 and 3. The level 3 signals were mostly correctly recognized with the exception of the synthetic model that performed slightly worse than other models. The worst performing classifier in relation to the level 1 signals is the benchmark model. In comparison to the benchmark classifier, the data augmentation significantly improved the model performance.

In the test set, the prevalence of level 2 signals is substantially greater than that of the other two classes as shown in Table 5. Thus, the domain randomized and augmented classifiers that performed the best on AUC of level 2 signals also performed the best on accuracy score.

All of the adjustments implemented for models trained on synthetic data improved the model performances compared to the synthetic classifier. The high specificity classifier is disregarded for there is no increase in AUC score of level 1 signals compared to the feedback classifier. The feedback loop and domain randomization resulted in scores as good as the augmented classifier.

Some of the mislabeled test set signals of the best performing models trained on synthetic signals, domain randomized and feedback, are displayed in Figures 16 and 17. This illustrates some of the real-world signals, in which the synthetic data shows deficiencies, or the model architecture fails to capture these quality affecting features. Moreover, one might have differing views on the "true qualities" of some of these these signals. All of the misclassified signals are adjacent to their visually assessed quality category, i.e. for these two classifiers there were no signals of level 1 quality to be assessed as level 3 or vice versa.

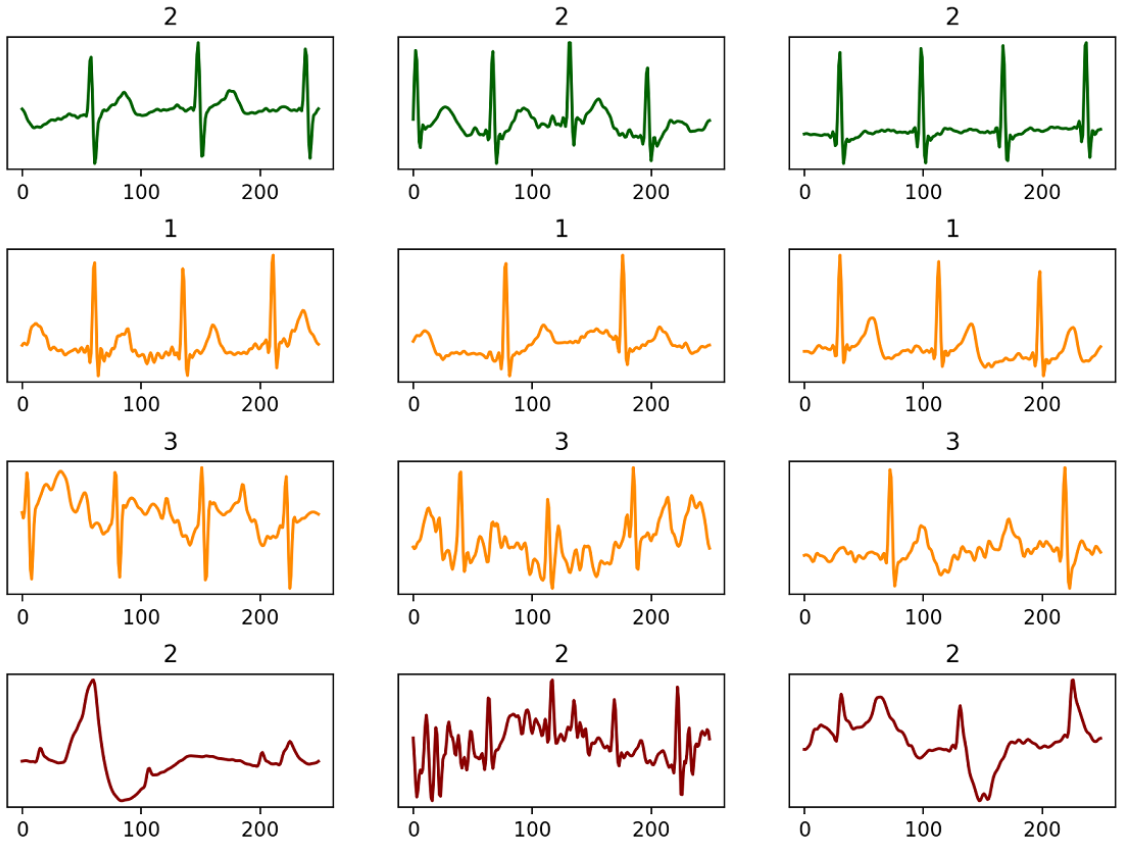


Figure 16: Signals misclassified by the domain randomized model. The title of the plot is the predicted quality class and the colors green, yellow and red correspond to true qualities of level 1, level 2 and level 3 respectively.

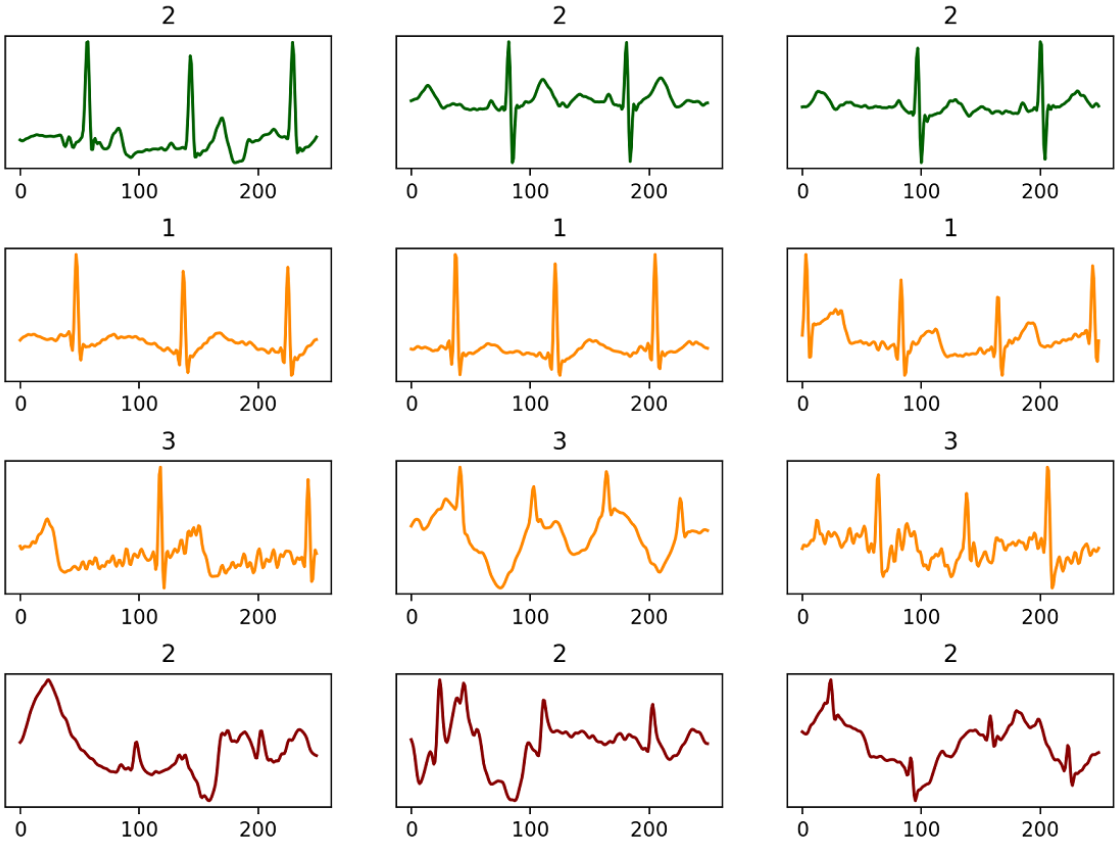


Figure 17: Signals misclassified by the feedback model. The title of the plot is the predicted quality class and the colors green, yellow and red correspond to true qualities of level 1, level 2 and level 3 respectively.

# 8 Discussion

The poor performance of the benchmark model of level 1 signals is surprising, because that is the most prevalent quality class in the validation dataset. This might be caused by an unfortunate data split or a suboptimal implementation of the early stopping. The performance drop on level 2 signals for all classifiers is significant, but in most use-cases for this quality criterion, the qualities would likely to be grouped as either of the two following ways.

- Level 1 and not level 1, i.e. can every waveform be detected?
- Level 3 and not level 3, i.e. can heart rate be detected?

Therefore, the specificity on the signals of level 2 quality is not as important as for the two other quality classes.

The increase in scores from synthetic model to feedback model suggests that the feedback loop leaked essential features from the validation set explaining the quality. The implementation of domain randomization presented in this thesis appears to enhance the representativeness and generalizability of the training data, as the results were similar to those of the augmented model. This also reinforces the idea that in generating synthetic data, it is not essential to create data that aligns with the test set. Instead, the generated data should be consistent with the specific property under consideration, where flexible, parametric models that generate synthetic data excel. The control over the train data, including during training, allows for specific models to be adapted to the intended use.

The mislabeled signals, some of which are represented in Figures 16 and 17, suggest that smoother level 3 signals are incorrectly assessed as level 2 by all classifiers trained on synthetic data. This could probably be improved by adding a wider range of non-ECG signals labeled as level 3 to the training set. In addition, a high amplitude of the negative S-waves in level 1 or 2 signals contributed to the model being more prone to classify these into a lower quality class, suggesting that the maximum absolute value of the range for amplitude parameter of the S-wave should be increased. With the exception of the domain randomized classifier, the level 2 signals assessed as level 1 are mainly signals that are quite smooth between QRS-complexes, i.e. the P-waves or T-waves are not recognizable, but the parts between QRS-complexes do not appear particularly noisy. Presumably, this is caused by the P-waves and T-waves being obscured by the added noise to be qualified as level 2 signals in the synthetic training data, whereas the domain randomized classifier performs significantly better in these instances. Additionally, this review of mislabeled signals highlighted some of the inconsistencies in the test dataset, further emphasizing the importance of reliable SQA methods.

It is not convincing to state that the automatic quality classifier for ECG conducted in this thesis is objectively better than the other ones mentioned in Section 3.7. The SNR method could be improved by weighing the frequencies, for that would be very similar to the excess noise over boundary functions. Similarly, the PCA quality method could be improved by higher frequency resolution. Moreover, with a representative validation dataset, a quality classifier that considers the essen-

tial frequency components with a reasonably concise parameter space, can be fitted by an automatic method, such as the feedback loop presented in this thesis.

In future studies, it ought to be beneficial to utilize expert annotated, more robust data from different sources by multiple annotators to reduce bias and ensure the model utility. Also, the validation dataset and test dataset should be completely from different sources diminishing the increase in scores caused by over fitting. These methods should also be validated by multiple runs to ensure that the stochastic nature of the ECG generator is still predictable. Feature based methods, or state of the art SQA neural networks could be employed to increase performance. In addition, more specified quality categories with specific utilities, multiple SQIs and continuous quality values should be researched.

Furthermore, for diagnostic purposes, the revision of the quality heuristic is necessary, such that certain anomalies, distortions in ECG caused by CVDs, are included in the signal generator. In this way, the SQA can be used as a filter, when important features considering the disease detection are not filtered out.

The biosignal framework also provides PPG generation, hence it is reasonable to suggest that SQA models could also be developed for PPG quality assessment using the same framework.

The utility and applicability of the trained classifiers is demonstrated in Figure 18, where long, quality-wise inconsistent signals are assessed by sliding over the signal and averaging the quality label, giving a continuous quality value for every sample.

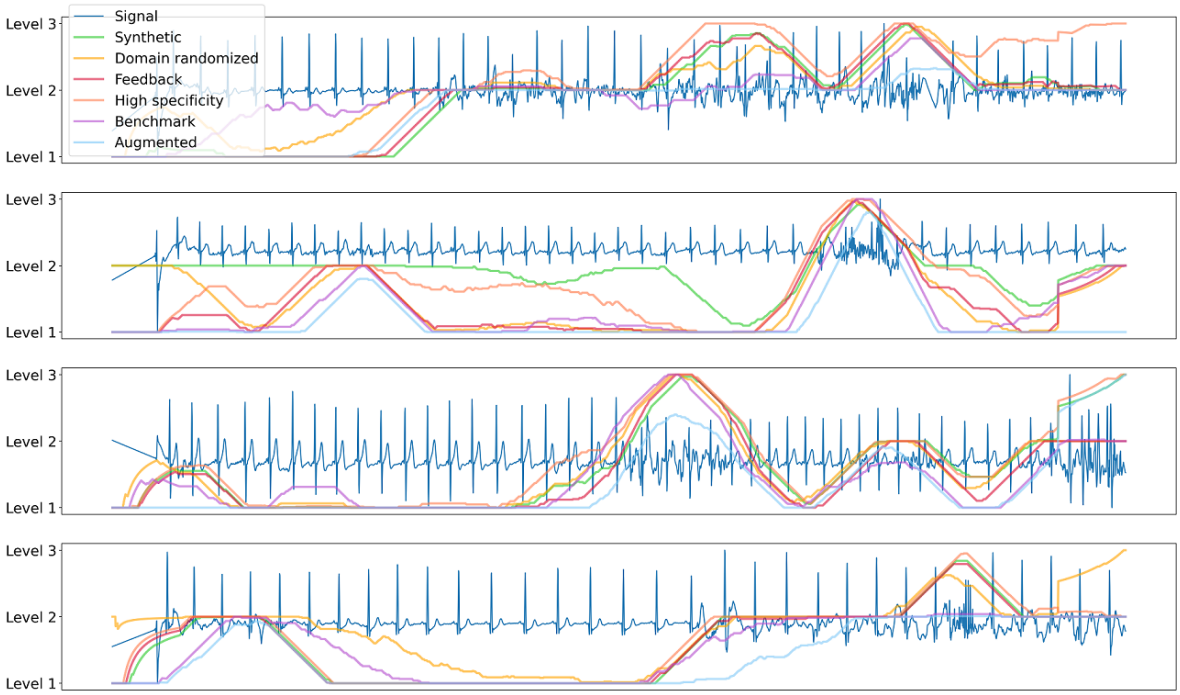


Figure 18: Demonstration of the quality assessment for long signals using the trained models.

The modular parametric signal generator [2] offers a wide range of possibilities to adjust the criterion for the quality classes of the generated signals. This seam between the quality heuristic and its precise implementation is not closed, since the heuristic lies on the perception of the quality, but the framework has provided a parametric mapping for this abstraction.

Exploits, such as signal generator parameter tuning due to feedback according to validation scores, can significantly increase the model performance when only trained on synthetic signals, leaking valuable information on the desired properties without overfitting. Moreover, the utility of synthetic ECGs in data augmentation and domain randomization appear convincing.

# References


[1] Lukasz Piwek et al. “The rise of consumer health wearables: promises and barriers”. In: *PLoS medicine* 13.2 (2016), e1001953.

[2] Katri Karhinoja et al. “Flexible framework for generating synthetic electrocardiograms and photoplethysmograms”. In: *arXiv preprint arXiv:2408.16291* (2024).

[3] Christina Orphanidou. “Signal Cuality Cssessment in Chysiological Conitoring: State of the Srt and Practical Considerations”. In: (2017).

[4] Nicolò Gambarotta et al. “A review of methods for the signal quality assessment to improve reliability of heart rate and blood pressures derived parameters”. In: *Medical & biological engineering & computing* 54 (2016), pp. 1025–1035.

[5] Malcolm S Thaler. *The only EKG book you'll ever need*. Lippincott Williams & Wilkins, 1999.

[6] Kirina van der Bijl, Mohamed Elgendi, and Carlo Menon. “Automatic ECG quality assessment techniques: A systematic review”. In: *Diagnostics* 12.11 (2022), p. 2578.

[7] Udit Satija, Barathram Ramkumar, and M Sabarimalai Manikandan. “A review of signal processing techniques for electrocardiogram signal quality assessment”. In: *IEEE reviews in biomedical engineering* 11 (2018), pp. 36–52.

[8] Ikaro Silva, George B Moody, and Leo Celi. “Improving the quality of ECGs collected using mobile phones: The Physionet/Computing in Cardiology Challenge 2011”. In: *2011 computing in Cardiology*. IEEE. 2011, pp. 273–276.

[9] George B Moody and Roger G Mark. “The impact of the MIT-BIH arrhythmia database”. In: *IEEE engineering in medicine and biology magazine* 20.3 (2001), pp. 45–50.

[10] Stephen J Redmond et al. “Electrocardiogram signal quality measures for unsupervised telehealth environments”. In: *Physiological Measurement* 33.9 (2012), p. 1517.

[11] Aron Syversen et al. “Assessment of ECG signal quality index algorithms using synthetic ECG data”. In: (2024).

[12] Alvaro Huerta et al. “Comparative study of convolutional neural networks for ECG quality assessment”. In: *2020 Computing in Cardiology*. IEEE. 2020, pp. 1–4.

[13] Gari 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.

[14] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model inversion attacks that exploit confidence information and basic countermeasures”. In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 2015, pp. 1322–1333.

[15] Cynthia Dwork, Aaron Roth, et al. "The algorithmic foundations of differential privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.

[16] Zoher Orabe et al. "Securing Deep Learning Models with Differential Privacy for Cardiovascular Disease Prediction". Manuscript submitted for publication. 2024.

[17] Tuija Leinonen et al. "Empirical investigation of multi-source cross-validation in clinical ECG classification". In: *Computers in Biology and Medicine* 183 (2024), p. 109271.

[18] Dingfan Chen et al. "Gan-leaks: A taxonomy of membership inference attacks against generative models". In: *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 2020, pp. 343–362.

[19] Josh Tobin et al. "Domain randomization for transferring deep neural networks from simulation to the real world". In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 2017, pp. 23–30.

[20] Matti Kaisti et al. "Domain randomization using synthetic electrocardiograms for training neural networks". In: *Artificial Intelligence in Medicine* 143 (2023), p. 102583.

[21] Katri Karhinoja. *Framework for synthetic biosignals*. 2024. URL: `https://github.com/UTU-Health-Research/framework_for_synthetic_biosignals`.

[22] George B Moody, WE Muldrow, and Roger G Mark. "A noise stress test for arrhythmia detectors". In: *Computers in cardiology* 11.3 (1984), pp. 381–384.

[23] Jonathan Tremblay et al. "Training deep networks with synthetic data: Bridging the reality gap by domain randomization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018, pp. 969–977.

[24] Heikki Huttunen. *Signaalinkäsittelyn perusteet*. Lecture notes, Tampere University of Technology, Department of Signal Processing. 2014.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016. Chap. 5–9, pp. 96–366.

[26] Lu Lu et al. "Dying relu and initialization: Theory and numerical examples". In: *arXiv preprint arXiv:1903.06733* (2019).

[27] Djork-Arné Clevert. "Fast and accurate deep network learning by exponential linear units (elus)". In: *arXiv preprint arXiv:1511.07289* (2015).

[28] GE Hinton. "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv preprint arXiv:1207.0580* (2012).

[29] Diederik P Kingma. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).