NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

MARIA MOURÃO GONÇALVES RUSSO

Bachelor of Science in Biomedical Engineering

TOWARDS HIGH-FIDELITY ECG GENERATION: EVALUATION VIA QUALITY METRICS AND HUMAN FEEDBACK

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon

December, 2024

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

TOWARDS HIGH-FIDELITY ECG GENERATION: EVALUATION VIA QUALITY METRICS AND HUMAN FEEDBACK

MARIA MOURÃO GONÇALVES RUSSO

Bachelor of Science in Biomedical Engineering

Adviser: Dr. Hugo Filipe Silveira Gamboa
Full Professor, NOVA University of Lisbon

Examination Committee

Chair: Dr. Célia Maria Reis Henriques
Associate Professor, NOVA University of Lisbon

Rapporteur: Dr. Pedro Manuel Cardoso Vieira
Assistant Professor, NOVA University of Lisbon

Adviser: Dr. Hugo Filipe Silveira Gamboa
Full Professor, NOVA University of Lisbon

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
December, 2024

Towards High-Fidelity ECG Generation: Evaluation via Quality Metrics and Human Feedback

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.18) [1].

Acknowledgements

This dissertation marks the culmination of a challenging journey, filled with moments of doubt but also perseverance and growth. Throughout this path, I was fortunate to have the guidance and support of remarkable people who helped me reach this milestone.

First, I would like to express my gratitude to my advisor, Professor Hugo Gamboa, for introducing me to the world of data science through his inspiring classes. His invaluable knowledge and guidance were instrumental in shaping this work.

I extend my appreciation to Fraunhofer AICOS, especially the Lisbon team, for providing such a supportive and collaborative environment. Special thanks to Joana and Nuno, whose guidance and daily advice were indispensable, especially during challenging times. I have learned so much from their mentorship, for which I am profoundly grateful.

To my lifelong friends, whose companionship and support kept me grounded. To Inês, for your patience in listening to my countless rants of stress, and for offering words of comfort that meant more than I can express. And to the wonderful friends I made along the way, Margarida, Catarina, and Matilde, your friendship made this journey incredibly rewarding.

To Miguel, my lobster, thank you for standing by my side over the past six years. Your laughter, hugs, and endless encouragement lifted me up during the darkest times. Your love, dedication, and constant motivation gave me the strength to keep moving forward.

Finally, I am eternally grateful to my family for their support and unconditional love. You have been my greatest sponsors, in every sense. To my mom, for the love and care you put into everything you do for me, for always showing interest in my work, and for supporting me even from afar with countless acts of love. To my dad, for believing in me and never letting me give up on my dreams, though your greatest wish was always for me to be happy. To my sister, my role model, thank you for the laughter, the wisdom, and the best advice. I strive every day to be as amazing as you. And finally, to my niece, Francisca, the happiest and most wonderful surprise of this year. Your smile at the end of a long day reminded me of what I was fighting for, so that one day, you can be proud of your engineer auntie. This journey was an incredible ride, and it would not have been possible without my amazing family.

"It matters not how strait the gate,
How charged with punishments the scroll,
I am the master of my fate,
I am the captain of my soul."

William Ernest Henley, *Invictus*, 1875

Abstract

The Electrocardiogram (ECG) is a powerful diagnostic tool that offers valuable insights into the complex functioning of the heart. Despite its widespread use, researchers face significant challenges in accessing openly shared ECG datasets due to privacy concerns and data scarcity. In response, synthetic data generated through Deep Learning (DL) models has emerged as a promising solution. However, evaluating synthetic medical data solely with general quality metrics may not be sufficient, as medical data must be both realistic and clinically relevant.

This dissertation aims to overcome such challenges by advancing high-fidelity ECG data generation and evaluation, presenting an approach for generating realistic ECG signals using a diffusion model and introducing a novel evaluation metric based on a DL evaluator model. The initial phase of this research focused on refining the state-of-the-art Structured State Space Diffusion (SSSD-ECG) model through hyperparameter optimization. The fidelity of the generated synthetic dataset was assessed using both quantitative metrics and feedback from medical experts. Additionally, the diversity and utility of the synthetic signals were assessed to ensure comprehensive evaluation. The evaluator model was developed to classify individual synthetic ECG signals into four quality levels and was trained on a custom-developed quality dataset specifically designed for the generation of 12-lead ECG signals.

Regarding the goal of generating high-fidelity ECG data, both the evaluation metrics and expert feedback indicate that the SSSD-ECG model was successful. In correlation studies, the evaluator model demonstrated alignment with the fidelity metrics. Overall, this research addresses key challenges in the generation and evaluation of synthetic ECG data, contributing with a novel approach for generating realistic ECG signals, the development of a quality dataset, and a new metric for classifying synthetic ECG data into distinct quality levels.

Keywords: ECG synthesis, Deep Generative Models, Synthetic Data Evaluation, Human Feedback

Resumo

O Eletrocardiograma (ECG) é uma ferramenta de diagnóstico que oferece informações valiosas sobre o funcionamento do coração. Apesar da sua ampla utilização, os investigadores enfrentam desafios no acesso a conjuntos de dados de ECG, devido a preocupações de privacidade e à escassez de dados. Como solução surgem os dados sintéticos gerados através de modelos de Aprendizagem Profunda (AP). Contudo, avaliar dados médicos sintéticos apenas com métricas gerais de qualidade é insuficiente, já que estes devem ser tanto realistas como clinicamente relevantes.

Esta dissertação aborda os desafios na geração e avaliação de dados de ECG realistas, apresentando uma abordagem de geração que utiliza um modelo de difusão e introduzindo uma nova métrica de avaliação baseada num modelo avaliador de AP. A fase inicial desta pesquisa centrou-se no aprimoramento do modelo da literatura *Structured State Space Diffusion (SSSD-ECG)* através da otimização de hiperparâmetros. A fidelidade do conjunto de dados sintéticos gerado foi avaliada utilizando métricas quantitativas e feedback de médicos. Adicionalmente, a diversidade e a utilidade dos sinais sintéticos foram avaliadas para garantir uma avaliação abrangente. O modelo avaliador foi desenvolvido para classificar individualmente sinais de ECG sintéticos em quatro níveis de qualidade, sendo treinado com um conjunto de dados de qualidade desenvolvido especificamente para a geração de sinais de ECG de 12 derivações.

Relativamente ao objetivo de gerar dados de ECG realistas, tanto as métricas como o feedback dos especialistas indicam que o modelo SSSD-ECG foi bem-sucedido. Em estudos de correlação, o modelo avaliador demonstrou alinhamento com as métricas de fidelidade. Concluindo, este projeto aborda desafios relevantes na geração e avaliação de dados de ECG sintéticos, contribuindo com uma abordagem inovadora para gerar sinais realistas, com o desenvolvimento de um conjunto de dados de qualidade e uma nova métrica para classificar dados sintéticos de ECG em níveis de qualidade distintos.

Palavras-chave: Síntese de ECG, Modelos generativos profundos, Avaliação de dados sintéticos, Feedback Humano

Contents

List of Figures

List of Tables

Abbreviations

Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
AISym4Med	Synthetic and Scalable Data Platform for Medical Empowered AI
ANN	Artificial Neural Network
BUT QDB	Brno University of Technology ECG Quality Database
CD	Conduction Disturbance
CNN	Convolutional Neural Network
DC-GAN	Deep Convolutional Generative Adversarial Network
DL	Deep Learning
DM	Diffusion Model
DNN	Deep Neural Network
DTW	Dynamic Time Warping
ECG	Electrocardiogram
FN	False Negatives
FP	False Positives
GAN	Generative Adversarial Network
HYP	Hypertrophy
MI	Myocardial Infarction
ML	Machine Learning
MLP	Multi-layer Perceptron
MMD	Maximum Mean Discrepancy

NORM	Normal
ReLU	Rectified Linear Unit
RF	Random Forest
S4	Structured State Space Sequence Model
SSSD-ECG	Structured State Space Diffusion ECG
STTC	ST/T change
TN	True Negatives
TP	True Positives
TRTS	Train on Real, Test on Synthetic
TSFEL	Times Series Feature Extraction Library
TSTR	Train on Synthetic, Test on Real
TSTS	Train on Synthetic, Test on Synthetic
VAE	Variational Autoencoder

1 Introduction

1.1 Context and Motivation

Electrocardiograms (ECGs) are the cornerstone of cardiovascular diagnostics, offering vital insights into the electrical activity of the heart. As a non-invasive and widely used diagnostic tool, ECGs are crucial in detecting a broad spectrum of heart conditions [2]. The accuracy and reliability of these diagnoses depend heavily on the availability of high-quality ECG data. However, the acquisition of large and diverse datasets of real ECG recordings is often hindered by privacy concerns and data scarcity [3].

ECG data is highly sensitive, often compared to a human fingerprint due to its unique patterns that can reveal individual-specific information when analyzed in both frequency and time domains [3]. This sensitivity poses challenges in data sharing, limiting the ability to leverage such data for research and the development of advanced diagnostic tools. Additionally, the scarcity of data, due to difficulties in acquisition, presents another obstacle. Data scientists, tasked with developing and training Machine Learning (ML) models, often struggle with the need for substantial volumes of data that are both heterogeneous and representative of the population of interest [3]. Unfortunately, access to such datasets, especially those with specific pathologies or rare conditions, remains a rarity.

To address these challenges, deep generative models have emerged as a promising solution, capable of generating synthetic data that exhibits the same structural and statistical characteristics of real data. However, in the healthcare domain, the generation of synthetic data is not sufficient. It is imperative that medical data, such as the ECG, is both realistic and clinically relevant [4].

Evaluating the quality of synthetic data is therefore a critical step. Current evaluation metrics, which rely on the statistical distributions between synthetic and real datasets, have contributed to valuable advances in the field. However, these metrics might overlook complex signal features essential for accurate medical interpretation. Additionally, clinicians often struggle to contextualize statistical criteria in clinical context. This highlights the need for more sophisticated evaluation methods, potentially ones that assess data at the sample level rather than collectively, and that offer more intuitive interpretation [4].

In addition to quantitative assessments, researchers have increasingly emphasized the importance of qualitative evaluation by medical experts as a necessary step in identifying discrepancies in synthetic samples [4]. Involving clinical professionals in the evaluation process could provide valuable insights into the realism and clinical utility of synthetic ECG data, as human experts offer the most reliable "ground truth" in determining realism [5].

In response to these needs, this work was developed as part of the Synthetic and Scalable Data Platform for Medical Empowered AI (AISym4Med) project. The primary goal of the project is "developing a platform that provides healthcare data engineers, practitioners and researchers access to a trustworthy dataset system augmented with controlled data synthesis for experimentation and modeling purposes" [6].

1.2 Objectives

The aim of this dissertation is to evaluate and enhance the generation of synthetic ECG data using deep generative models with an emphasis on achieving high realism. Another key objective is to develop a robust metric that can accurately capture the complexity and subtle nuances of individual ECG patterns. To accomplish this, three specific goals have been set:

- **Evaluate a State-of-the-Art Generative Model:** Refine the chosen model to generate synthetic ECG signals that are closely resembling real data in both appearance and diagnostic relevance, and assess their quality through metrics that evaluate fidelity, diversity, and utility.
- **Develop an Individualized Evaluation Metric:** Create a specialized metric for evaluating each synthetic ECG sample individually, focusing on generation rather than merely detecting artifacts and noise.
- **Validate Generated Data Quality:** Conduct comprehensive evaluations of the synthetic ECG data using the newly developed metric and gather expert human feedback to validate the fidelity of the generated signals.

By achieving these objectives, this dissertation aims to make a significant contribution to the AISym4Med project and the broader field of Artificial Intelligence (AI) in healthcare.

1.3 Document Structure

This dissertation is organized into seven chapters. Chapter one introduces the motivation behind the work and outlines its main objectives. Chapter two covers the foundational concepts and theoretical knowledge that support the development of this research. In chapter three, a review of the relevant literature is provided, focusing on deep generative

models and current evaluation techniques. Chapter four details the dataset used for training the generative models. Chapter five explains the methodology, including data preprocessing, model implementation, design of a quality dataset, development of an evaluation metric and the assessment of synthetic data. Chapter six presents a thorough analysis of the results obtained throughout the study. Finally, chapter seven concludes the dissertation by summarizing the key contributions, discussing limitations, and offering suggestions for future research. Additionally, Appendix A provides supplementary information about the human evaluation approach used in the study, while Annex I includes descriptions of the diagnostic classes used for model training.

1.4 Declaration of Originality

The research work described in this dissertation was carried out in accordance with the norms established in the ethics code of Universidade Nova de Lisboa. The work described and the material presented in this dissertation, with the exceptions clearly indicated, constitute original work carried out by the author.

2

Theoretical Background

2.1 Electrocardiogram

The ECG is a graphical representation of the heart's electrical activity arising from the processes of depolarization and repolarization within the cardiac muscle. In the ECG, distinct characteristics are evident, each revealing specific aspects of cardiac function [7]. The initial P wave delineates the propagation of electrical impulses that spreads across the atria. Then the QRS complex becomes prominent, indicating ventricular excitation, while the T wave corresponds to the subsequent ventricular repolarization. The analysis of these characteristic entities enables clinicians to extract valuable information, aiding in the calculation of heart rate and the identification of rhythm abnormalities [7].

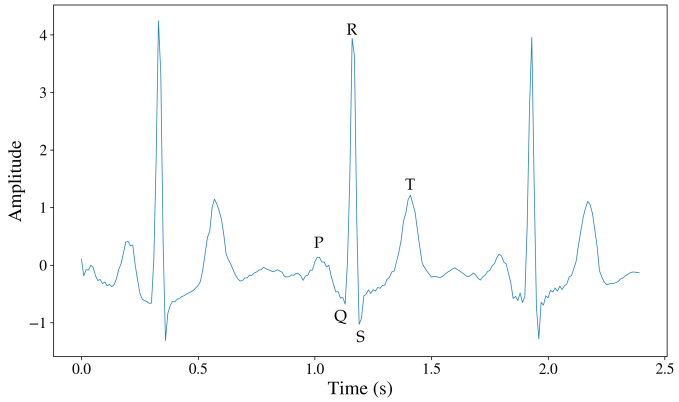


Figure 2.1: Example of normal ECG signal from the PTB-XL dataset, with labeled P wave, QRS complex, and T wave. The x-axis represents time in seconds, and the y-axis shows normalized amplitude using z-score normalization.

2.2 Machine Learning

In the domain of AI, ML stands as a subfield dedicated to investigating and developing algorithms with the capacity to learn and predict outcomes from data. These algorithms diverge from traditional static program instructions, as they derive predictions or decisions

from the data itself [8]. The classification of ML algorithms depends on the level of human supervision during training, resulting in distinct types such as supervised, unsupervised, semi-supervised, and reinforcement learning.

- Supervised learning algorithms are trained with labeled data. During this process, the model learns the mapping from inputs to outputs, adjusting its parameters to minimize the disparity between predictions and actual labels. Employing techniques such as classification and regression, supervised learning extrapolates patterns to forecast label values on additional unlabeled data [8].
- Unsupervised learning algorithms operate on unlabeled data, where models aim to analyze and cluster the data to uncover inherent patterns based on the features of the dataset [9].
- Semi-supervised learning uses both labeled and unlabeled data during training, typically incorporating a small amount of labeled data alongside a substantial amount of unlabeled data[9].
- Reinforcement Learning operates by enabling a model to learn through trial and error which actions yield the greatest rewards [8].

2.2.1 Traditional Machine Learning

Traditional ML algorithms have long been the foundation of predictive modeling, providing powerful tools for analyzing structured data and making predictions based on learned patterns. These algorithms are typically divided into two categories: classification and regression. Classification methods predict discrete labels or classes, while regression methods forecast continuous values. The choice of algorithm depends on the specific application and the characteristics of the dataset being used [10].

One of the most widely employed traditional ML algorithms is Random Forest (RF), which can be applied to both classification and regression tasks. RF is an ensemble learning method that constructs multiple decision trees and combines their outputs to enhance predictive accuracy and reduce overfitting [10]. In classification tasks, the final prediction is determined by majority voting among the decision trees, while in regression tasks, the result is the average of the trees' outputs [11].

When evaluating the performance of ML models, it is essential to use appropriate evaluation metrics to ensure the model generalizes well to unseen data. For binary classification, a valuable evaluation tool is the confusion matrix. This matrix compares the actual and predicted classes, offering a comprehensive view of the model's performance. Each row in a confusion matrix represents an actual class, while each column represents a predicted class [9]. The main components of a confusion matrix are

- **True Positives (TP):** Correctly predicted positive instances.
- **True Negatives (TN):** Correctly predicted negative instances.
- **False Positives (FP):** Incorrectly predicted positive instances
- **False Negatives (FN):** Incorrectly predicted negative instances

Table 2.1: Representation of a confusion matrix for a binary classification problem.

	Predicted Positive Class	Predicted Negative Class
Actual Positive Class	TP	FN
Actual Negative Class	FP	TN

To evaluate the performance of a model, various metrics are derived from the confusion matrix. In this work, accuracy and F1-Score are the primary metrics used. Since the F1-Score is the harmonic mean of precision and recall, both metrics are also presented.

- **Accuracy:** The ratio of correct samples to the total number of instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** The proportion of correctly classified positive samples among all predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** The proportion of correctly predicted positive samples identified out of all actual positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** The harmonic mean of precision and recall.

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

2.2.2 Deep Learning

Deep Learning (DL) is a part of a broader family of ML methods based on learning data representations. It aims to replace handcrafted features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction [8]. The distinctiveness from traditional ML is prominent with large datasets, showcasing the ability of DL to process a vast array of features effectively [12]. The term "Deep" refers to

the incorporation of multiple layers in the model architecture, facilitating the construction of a data-driven model [13].

DL is based on the concept of Artificial Neural Networks (ANNs), which draw inspiration from the organization of biological nervous systems. In an ANN, numerous interconnected processing elements, or neurons, form the structure, with each neuron generating a series of real-valued activations for the desired outcome.

In Sarker et al. [12], a taxonomy for DL is introduced, categorizing techniques into three major groups: deep networks for supervised or discriminative learning, deep networks for unsupervised or generative learning, and deep networks for hybrid learning, combining both approaches.

In the first category, architectures like Multi-layer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs) are used to classify data by modeling the posterior distributions of classes conditioned on observed data. In contrast, deep networks in the second category focus on analyzing higher-order correlation features for pattern recognition or synthesis. Deep generative models, such as Generative Adversarial Networks (GANs) and Diffusion Models (DMs), are often employed for feature learning, data generation, and representation [12].

2.2.2.1 Multi-layer Perceptron

A MLP is an ANN, with information flowing unidirectionally from input to output. The typical MLP is fully connected, comprising an input layer for data reception, an output layer for decisions, and hidden layers for computation [12]. Within these hidden layers, transformations to input data are applied and each neuron in these layers has an associated weight and bias. The output of an MLP is determined by activation functions like Rectified Linear Unit (ReLU), Tanh, Sigmoid, or Softmax, introducing non-linearity for complex pattern learning.

Training an MLP involves adapting the connection weights to minimize the difference between the network output and the desired output. The most used algorithm for this purpose is backpropagation, that calculates the derivatives of the loss function with respect to the previous weights, aiding optimization. In the forward pass, the input data propagates through the network layers to generate an output. The error, computed by comparing this output to the actual target output using a designated loss function, serves as a measure of the disparity between the predicted and desired outcomes. The backward pass is initiated by computing the gradients of the error with respect to the weights. This is achieved through the application of the chain rule from calculus, enabling the algorithm to trace how changes in each weight contribute to the overall error. These gradients signify the direction and magnitude of adjustments required to minimize the error. Subsequently, the weights are updated in the opposite direction of the computed gradients, effectively reducing the error. The learning rate, a hyperparameter, regulates the size of these weight updates, ensuring a balance between convergence speed and stability.

This iterative process of forward and backward passes, weight updates, and error minimization is repeated for multiple epochs until the neural network converges to a state where the difference between its predictions and the actual outputs is minimized.

2.2.2.2 Convolutional Neural Network

Inspired by the human visual cortex, a CNN is a feedforward ANN, designed to reduce the need for extensive preprocessing by automatically learning spatial hierarchies in data. It excels in feature extraction from multidimensional data like images and videos, demonstrating remarkable efficiency in tasks such as computer vision, pattern recognition, and image processing [8]. The architecture of CNNs comprises three main types of layers: convolutional, pooling, and fully connected layers. The stacking of these layers forms the overall structure of the network.

- The Convolutional layer captures local patterns and hierarchies through convolution operations. A kernel slides over input data, conducting element-wise multiplications and summing the results for crucial feature extraction. The output feature map is formed by convolving feature maps from the preceding layer with learnable kernels and applying an activation function [14].

- The pooling layer aims to reduce the dimensionality of the representation simply performing downsampling along the spatial dimensionality of the given data and thus further reduce the number of parameters and the computational complexity of the model. A common pooling operation is max pooling, that operates in local regions of the input feature map. It slides a window, selecting the maximum value in each region, which is then used to form the output feature map [14].

- Fully connected layers directly link neurons from two adjacent layers. They make predictions based on learned features. Preceding these layers, data is flattened from the current dimension to 1D, as fully connected layers operate on one-dimensional vectors [14].

2.2.2.3 Generative Adversarial Networks

GANs are a class of DL models crafted to generate artificial samples that closely resemble real ones. Their primary objective is to autonomously discern and understand patterns within input data, enabling the generation of new examples from the original dataset. Comprising two neural networks, namely the generator and discriminator, these networks function as antagonists. The generator strives to create synthetic data closely resembling the real data from the training set, attempting to deceive the discriminator. On the other hand, the discriminator assesses both genuine and fake data, aiming to distinguish between them. Both networks undergo training in a competitive manner: the generator adjusts its parameters to generate increasingly realistic data, while the discriminator

improves its ability to differentiate between genuine and synthetic data [3]. Figure 2.2 illustrates the representation of the GAN framework.

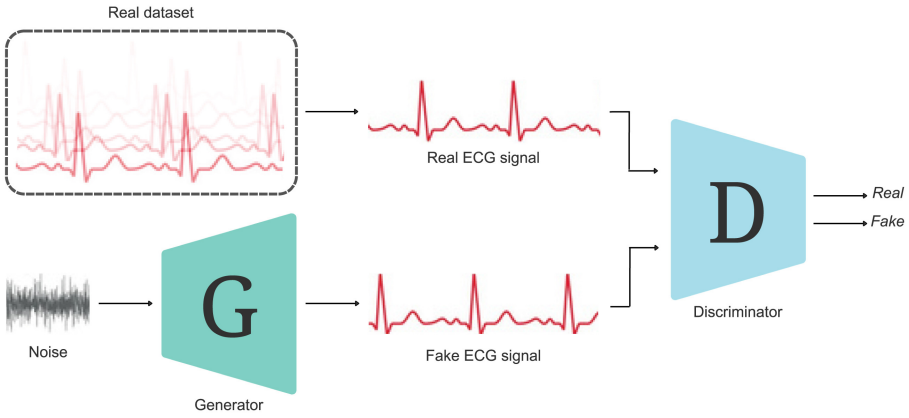


Figure 2.2: Representation of the GAN framework, adapted from [3].

2.2.2.4 Diffusion Models

DMs are generative models with a probabilistic nature, drawing inspiration from the thermodynamic diffusion process in physics. These models are designed to reverse the effects of data destruction or corruption, such as the introduction of noise, thereby enabling the generation of realistic and clean data samples. Consequently, DMs can be employed for generating new samples from a specific distribution, and they can implicitly learn the underlying distribution from which real samples are drawn [15]. The process involves two sequential steps: forward and reverse diffusion, as illustrated in Figure 2.3. The forward diffusion process entails gradually introducing noise to the input signal until it transforms into noise. On the other hand, the reverse diffusion process utilizes a trainable neural network to systematically eliminate the noise and reconstruct the original signal. Synthetic signals are subsequently generated by applying the trained neural network to simple noise [3].

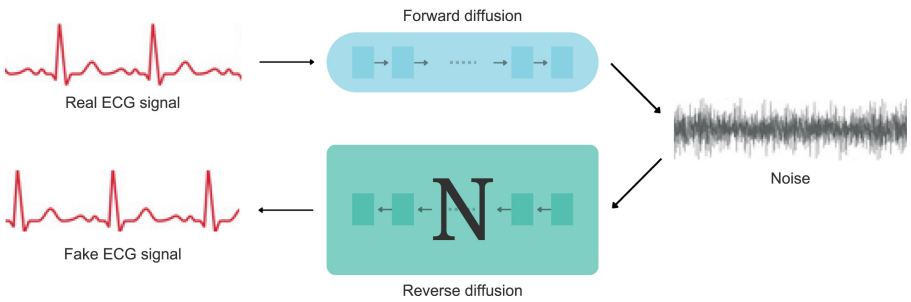


Figure 2.3: Illustration of the DM framework, adapted from [3].

Structured State Space Diffusion ECG

In this section, the focus shifts to the specific DM applied in this dissertation: the Structured State Space Diffusion ECG (SSSD-ECG), proposed by Alcaraz and Strodthoff [16]. As illustrated in Figure 2.4, this model stands out among generative models for 12-lead ECG generation by leveraging two key components: the Structured State Space Sequence Model (S4) and the conditional diffusion process.

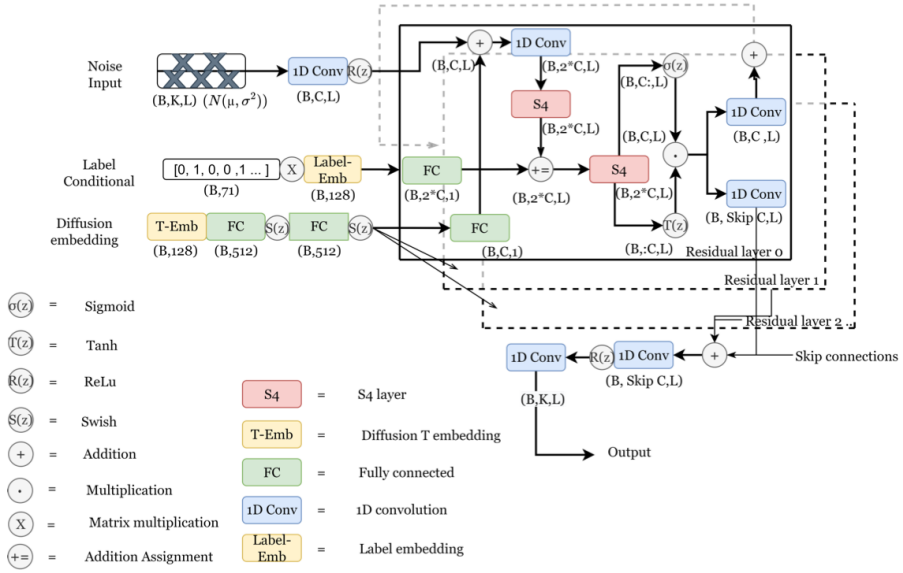


Figure 2.4: Illustration of the SSSD-ECG model architecture, adapted from [16].

SSSD-ECG is an adaptation of the previously proposed SSSD^{S4} model by the same authors. SSSD stands for Structured State Space Diffusion and was originally designed for time series imputation and forecasting [17]. The SSSD^{S4} model itself is built on the DiffWave architecture developed by Kong et al. [18], a diffusion probabilistic model proposed for audio synthesis, which employs a bidirectional dilated convolution architecture. In SSSD^{S4}, the dilated convolutions are replaced by two S4 layers, which are more effective at capturing long-term dependencies in time series data.

The S4 layer forms the core of this architecture, enabling the model to capture both short- and long-term dependencies in sequential data efficiently. It improves upon traditional state space models by introducing a carefully structured design that allows for highly efficient computations, even with very long sequences. This efficiency is achieved through a specialized mathematical technique called the High-Order Polynomial Projection Operator framework, which enables the model to represent and process long-term dependencies in a compact form [19]. By replacing the dilated convolutions with S4 layers, the SSSD^{S4} enhances its ability to model complex temporal relationships, making it particularly suited for time series tasks like ECG signal generation.

The key difference between SSSD^{S4} and SSSD-ECG lies in the conditional information used. While SSSD^{S4} receives an imputation mask and remaining input signals as conditional information, SSSD-ECG is conditioned on a set of annotated ECG statements,

represented as a binary vector of length 71. This vector is transformed into a continuous representation by multiplying it with a learnable weight matrix, which is then passed through a fully connected layer and used as conditional information in different $SSSD^{S4}$ layers [16].

In addition, another modification was made to the SSSD-ECG model. It was adapted to generate only 8 leads: leads I, aVF, and V1-V6, with the remaining 4 leads reconstructed from these primary leads. This approach was based on the observation that, in a standard 12-lead ECG, only two of the six limb leads are independent. Consequently, any set of limb leads can be reconstructed from two given limb leads using the following relationships: III = II - I, aVL = (I - III)/2, aVF = (II + III)/2, and -aVR = (I + II)/2 [16].

2.3 Evaluation Metrics for Synthetic Time Series

Many generative models and evaluation metrics for synthetic time series data have been proposed. However, there is still no universally accepted approach or consensus among researchers on how to quantify the quality of generated data. Assessing synthesis quality is a complex task that encompasses fidelity, diversity, generalizability, and privacy considerations [20]. In this work, the evaluation metrics selected to assess the quality of synthetic ECG data were primarily chosen to evaluate fidelity, with diversity and utility considered afterward.

2.3.1 Improved precision and recall

In the research by Sajjadi et al. [21], the authors proposed the precision and recall metrics to express the quality of generated data in two different components: fidelity and diversity. However, some practical limitations raised concerns, particularly the ambiguity in interpreting relative density differences, the sensitive initialization of the k-means algorithm, and the difficulty in reliably estimating extremes in situations such as mode collapse or truncation.

To address these challenges, improved precision and recall metrics were proposed by Kynkäänniemi et al. [22]. The improved metrics overcome these limitations by employing non-parametric methods to better estimate the manifolds of real and generated data distributions. The key innovation involves constructing a more accurate representation of these distributions in a high-dimensional feature space, utilizing pairwise Euclidean distances and k-nearest neighbors [22].

Precision measures the proportion of generated samples that are realistic. In the improved version, precision is calculated by first embedding both real and generated samples into a high-dimensional feature space using a pre-trained classifier. Next, the manifold of the real data is estimated by calculating the Euclidean distance between the feature vectors of real samples and constructing hyperspheres around each sample, with the radius determined by the distance to its k-th nearest neighbor. Improved precision is

then quantified by determining how many generated samples lie within this estimated manifold of real data. This refined approach now accounts for the true geometric structure of the data distribution [22].

On the other hand, recall measures the proportion of generated samples that cover the diversity of real data. In the improved recall metric, the same non-parametric manifold estimation is applied but in reverse. Recall assesses how many real samples fall within the manifold of generated data. This manifold of synthetic samples is estimated in the same way, by embedding the generated samples into the feature space, calculating pairwise distances, and constructing hyperspheres. This method allows for a more accurate assessment of how well the generated samples cover the variety of the real data distribution [22].

2.3.2 Density and Coverage

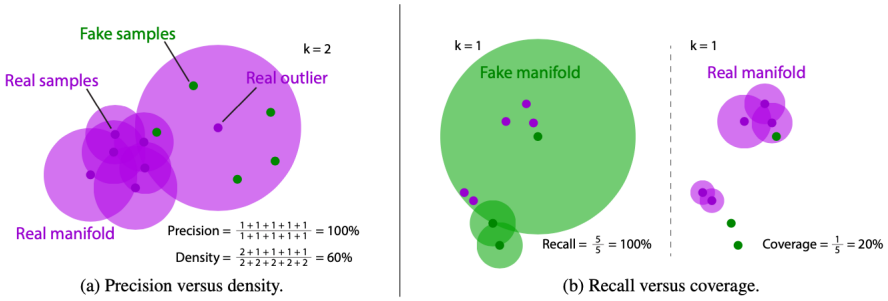
Naeem et al. [23] highlighted in their work that the effectiveness of the improved precision and recall metrics is limited due to their sensitivity to outliers, hyperparameters, and computational inefficiency. To address these issues, the authors proposed density and coverage metrics as practical solutions to remedy the mentioned problems.

Density improves the precision metric by correcting the overestimation of the manifold around real outliers. Instead of simply checking whether the generated samples fall within a neighborhood sphere of real data, as precision does, the density metric counts how many real-sample neighborhood spheres contain a synthetic sample. This approach balances the classic precision metric with the Parzen window estimate, providing a more reliable assessment by rewarding samples in densely packed regions of real data while mitigating the impact of outliers. Unlike precision, which is bounded by 1, density can exceed 1, depending on the density of real samples surrounding the generated samples [23].

On the other hand, coverage is an enhancement of the recall metric, designed to better capture the extent to which the diversity of real data is represented by the generated samples. Instead of building manifolds around the generated samples, which are more susceptible to outliers, coverage constructs nearest-neighbor manifolds around the real samples. This approach is also computationally favorable, as the manifold can be computed per dataset rather than per model. Coverage measures how many of these real samples have neighborhoods containing at least one generated sample, and it is bounded between 0 and 1 [23]. Figure 2.5 demonstrates the comparison between precision, recall, density, and coverage, underscoring the benefits of the newer metrics.

2.3.3 Train on Real, Test on Synthetic and Train on Synthetic, Test on Real

Evaluating the utility of synthetic data involves assessing how useful the generated data is for practical applications compared to real-world data. One of the most common ways to assess utility is through downstream classification tasks. Esteban et al. [24] proposed



(a) Precision versus density. (b) Recall versus coverage.

Figure 2.5: Comparison of precision, recall, density, and coverage metrics across two example scenarios. The figure highlights how density provides a more robust alternative to precision, while coverage enhances the recall metric. Adapted from [23].

two novel measures for evaluating the usefulness of synthetic data: Train on Synthetic, Test on Real (TSTR) and Train on Real, Test on Synthetic (TRTS)

For TSTR, the dataset generated by the generative model is used to train a classifier, which is then tested on an unseen set of true samples. Since the task is supervised, the generated labels must be provided. According to the authors, this evaluation metric is ideal because it demonstrates the practical value of the generated data, which could potentially replace the real data without compromising model quality [24].

In TRTS, the approach is reversed: real data is used to train the supervised model on a set of tasks, and then a test set of generated samples is used for evaluation. If the classifier achieves high accuracy, one can assume that the generated samples have features sufficiently similar to those in the real data [24].

3

Literature Review

3.1 Generative Models

According to the World Health Organization, cardiovascular diseases are the leading cause of death, underscoring the importance of accurate ECG analysis for timely detection [25]. Given its crucial role in diagnosing heart conditions, the ECG is one of the most synthesized biomedical signals.

In the initial stages of synthetic ECG signal generation, some of the first approaches relied on mathematical models consisting of three coupled ordinary differential equations, to mirror heart rate signals [26]. Alternatively, autoregressive models were employed to capture temporal dependencies and intricate patterns in the ECG by predicting the next value based on previous ones. Despite their ability to produce realistic signals in terms of ECG morphology, these models often resulted in synthetic heartbeats that were too standardized.

Recent DL advancements have greatly influenced the biomedical field, surpassing traditional methods. Wulan et al. [27] employed a Deep Convolutional Generative Adversarial Network (DC-GAN) to generate realistic ECG signals, encompassing different types of heartbeats. However, challenges were identified, such as the need for R-peak-centered data segments and limited extensibility to longer ECG signals and other biosignal modalities, necessitating annotated data for R-peak location.

To address these challenges, the work presented by Dissanayake et al. [28] focuses on 1D biosignal data, using adversarial models like GANs, an adversarial autoencoder, and a modality transfer GAN. The results indicate improved capabilities, including the generation of independent peak annotations and longer synthetic signals with multiple R-peaks compared to the DC-GAN approach.

In another notable contribution, Belo et al. [29] proposed a Deep Neural Network (DNN) model that learns and synthesizes biosignals, demonstrating morphological equivalence to real signals, including Electromyogram, ECG, and respiratory signals. The model, consisting of an embedded matrix, three layers of Gated Recurrent Units, and a softmax function, successfully captured the basic morphological and cyclical characteristics of the

original signals. For ECG, the model showcased the ability to discern distinct traits of each subject during training. Furthermore, Nishikimi et al. [30] introduced a DNN-based method, utilizing the conditional Variational Autoencoder (VAE), to synthesize ECG signals from cardiac parameters. The model, conditioned by the parameters, efficiently generated accurate ECG signals and demonstrated feasibility in terms of computational cost for creating a compilation of ECG signals.

Recent advancements in the DM domain have introduced remarkable approaches for time series modeling, demonstrating outcomes that surpass GANs and autoregressive models. Diffusion methods offer several advantages, including training stability and the ability to generate diverse synthetic samples. Alcaraz and Strodthoff [16] proposed the SSSD-ECG framework, which combines a conditional DM with structured state space sequences to synthesize short 12-lead ECG signals based on over 71 statements in a multi-label setting. Their approach excels in quantitative, qualitative, and human evaluations. Inspired by SSSD-ECG, Zama and Schwenker [31] developed the Diffusion State Space Augmented Transformer model, which also generates conditional 12-lead ECG data, replacing S4 layers with State Space Augmented Transformer layers. This novel approach also demonstrated strong performance in qualitative and authenticity assessments.

Additionally, Neifar et al. [32] developed a versatile framework based on Diffusion Denoising Probabilistic Models for ECG signal generation, imputation, and forecasting. Their approach integrates a simple yet efficient conditioning encoding, allowing seamless transitions between tasks, and showed promising results across various evaluations.

3.2 Quality Metrics for Synthetic Data Evaluation

Several measures have been proposed to assess synthetic data quality, but the choice of metrics depends on the specific problem and domain, leading to diverse evaluation approaches. As a result, no universal metric or general approach exists for evaluating synthetic data, particularly for time series, making it an active research area [20].

Stenger et al. [20] suggest categorizing the extensive list of proposed metrics for synthetic time series evaluation into two groups: distribution-level metrics, which assess all samples collectively, and sample-level metrics, which evaluate individual samples.

At the distribution level, common methods involve calculating the distance between synthetic and real samples, such as Average Euclidean Distance, Jensen-Shannon Distance, and Maximum Mean Discrepancy (MMD). Building upon these traditional metrics, Sajjadi et al. [21] proposed a novel definition of precision and recall for distributions, based on the estimated supports of real and synthetic data. Their approach moves beyond the one-dimensional scores, offering a more nuanced assessment by separately evaluating quality (precision) and diversity (recall) of synthetic datasets relative to real data. To demonstrate the practical utility of these metrics, they applied them across several variants of GANs and VAEs.

Kynkäänniemi et al. [22] addressed limitations in the previous metrics by introducing improved precision and improved recall, which better estimate real and synthetic data distributions using non-parametric methods, pairwise Euclidean distances, and k-nearest neighbors in a high-dimensional feature space. More recently, Naeem et al. [23] highlighted the unreliability of newer precision and recall metrics, introducing density and coverage metrics as alternative approaches designed to be less vulnerable to outliers and more computationally efficient, further advancing the evaluation of synthetic data. Both articles demonstrated the effectiveness of the metrics using GANs.

Shifting focus to sample-level metrics, Dynamic Time Warping (DTW) is a widely used distance measure for time series, as it captures flexible similarities under time distortions [33]. This metric measures the similarity between two sequences by aligning them point-to-point under certain constraints. Delaney et al. [34] employed DTW and MMD to evaluate synthetic ECG signals generated by GANs, finding DTW more effective for ECG data due to its sensitivity and robustness. However, a limitation of DTW is its sensitivity to noise and outliers, which can distort the alignment and lead to inaccurate similarity assessments. Another sample-based metric introduced by Alaa et al. [35] is α-Precision and β-Recall, which builds on Sajjadi et al. metrics by using a refined soft-boundary classification. These metrics assess how well synthetic samples cover typical regions of the real data distribution, offering a more detailed evaluation that detects mode collapse or distribution mismatches. However, the authors of the SSSD-ECG framework have raised concerns about these metrics, citing issues with instability during the training of one-class embeddings, which significantly affected the results.

Turning the attention to ECG quality assessment, several studies have utilized ML and DL techniques for evaluating ECG signals. Notable works [36], [37], and [38] train ML classifiers to assess various quality aspects, including background noise, beat consistency (detecting unexpected events), amplitude range, and the identification of signals with missing leads. In contrast, [39] and [40] employ non-feature-based approaches.

Despite their differing methodologies, all these studies rely on the PhysioNet/Computing in Cardiology Challenge 2011 dataset, indicating a shared focus on artifact and noise detection. Considering these findings, the proposed metric in this dissertation sets itself apart by specifically evaluating the realism of individual ECG samples, concentrating on the quality of the generated signals rather than merely identifying noise and artifacts.

4

Dataset

To accomplish the objectives of this thesis, the PTB-XL dataset was employed for training and evaluating the generative models. Additionally, it was used to develop one of the classes in the quality dataset, which will be explained in the following chapter.

4.1 PTB-XL

The PTB-XL dataset used in this dissertation originates from the "Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021" [41]. This annual competition promotes the development of open-source solutions for complex physiological signal processing and medical classification challenges. The goal of the 2021 challenge was to identify clinical diagnoses using ECG data from various lead configurations [41].

The dataset provided for this competition consists of annotated 12-lead ECG recordings sourced from nine different databases. These recordings have been curated and standardized to meet the specific needs of the competition, making the dataset both scalable and adaptable for use in multiple algorithms. Consequently, this dissertation utilizes data from this challenge, specifically focusing on the third source, the Physikalisch-Technische Bundesanstalt (PTB), for training and evaluating generative models.

The PTB-XL dataset was selected due to its large number of 12-lead ECG samples and its diversity in signal quality and pathology coverage. It comprises 21,837 clinical 12-lead ECG records, each 10 seconds in length, collected from 18,885 patients. It is gender-balanced, with 52% male and 48% female participants, and covers a wide age range from 0 to 95 years. Each ECG record was annotated by one or two cardiologists, who assigned multiple ECG statements based on the SCP-ECG standard, covering form, rhythm, and diagnostic categories. These annotations span 71 distinct categories [42].

The diagnostic labels, which were the primary focus of this research, are organized hierarchically into 5 broad superclasses and 24 more specific subclasses. For the challenge, however, all the labels were mapped to SNOMED-CT codes. A description of the classes can be found in Annex I.

5 Methodology

5.1 Overview

To accomplish the objectives outlined in Section 1.2, this dissertation was structured around several key stages: data preprocessing, generative model implementation, synthetic ECG signal generation, quality dataset construction, evaluator model development, and finally, a comprehensive evaluation using quality metrics and expert feedback.

The first step involved preprocessing the ECG signals from the publicly available PTB-XL dataset, which were used for model training and as a reference for evaluating the synthetic signals through various metrics. The generative approaches implemented included a GAN model and a DM from the literature, named SSSD-ECG. As anticipated, in preliminary experiments, the DM outperformed his competitor in generating realistic synthetic ECG samples, and therefore, only the signals generated by the SSSD-ECG were used for subsequent evaluations. However, both models played a crucial role in developing the custom quality dataset of 12-lead ECG signals, which was then used to train the evaluator model. This model was specifically designed to classify synthetic data into four distinct quality levels.

In the final stage, a thorough evaluation of the generated signals was conducted to assess their fidelity, diversity and utility. This was achieved by employing state-of-the-art evaluation metrics, along with a downstream classification task. To complement these evaluations, the synthetic signals generated by the SSSD-ECG model were presented to clinical experts via a questionnaire for qualitative assessment. This approach aimed to ensure that the synthetic samples retained the essential characteristics of real ECG signals, making them useful for future applications.

5.2 Data preprocessing

Both generative models were conditioned on the diagnostic labels associated with the ECG signals, therefore data preprocessing was performed in two stages: one for the signals and one for the labels.

5.2.1 Real Signal Preprocessing

The ECG signals from the PTB-XL dataset were first resampled from 500Hz to 100Hz for each lead. This resampling reduced the data size while preserving essential information, making the data more manageable. A moving average filter was then applied to remove baseline wander - a low-frequency noise component - by smoothing the signals and subtracting the baseline wander from the original ECG signal.

Subsequently, each ECG channel was normalized using z-score normalization, which standardizes the data by centering it around a mean of zero and scaling it to have a standard deviation of one.

Finally, since the PTB-XL dataset is characterized by its diversity and contains many co-occuring pathologies, signals with more than one diagnostic class label were excluded from the data. This ensured that each signal in the new processed dataset belonged to only one diagnostic 'superclass', to reduce complexity in model training. The distribution of signals in each class after preprocessing is shown in Table 5.1.

Table 5.1: Distribution of ECG signals across different categories in the training dataset after preprocessing.

Diagnostic Classes	Number of signals
Conduction Disturbance (CD)	404
Myocardial Infarction (MI)	573
Hypertrophy (HYP)	91
ST/T change (STTC)	758
Normal (NORM)	8773

5.2.2 Diagnostic Label Preprocessing

As outlined in Section 4.1, the dataset contains 71 unique statements categorized into three groups: diagnostic, form, and rhythm. Specifically, there are 44 diagnostic statements, 19 form statements describing the shape of the ECG signal, and 12 statements describing the cardiac rhythm [42]. For the 2021 challenge, these statements were standardized and mapped to SNOMED-CT codes to ensure a consistent representation of cardiac abnormalities.

However, this dissertation narrows the focus to diagnostic classes, particularly the five broad diagnostic 'superclasses'. This decision was made because, for the validation phase of the pipeline, it was impractical to present 71 examples of cardiac pathologies to the clinical experts for evaluation. Additionally, consolidating the labels into fewer, more general categories enhances model performance and generation robustness by providing the model with a sufficient number of examples from each class, allowing it to learn meaningful patterns.

The preprocessing of labels involved two main steps. First, the labels were mapped from SNOMED-CT codes back to the corresponding diagnostic 'superclasses'. Second, the labels were converted into one-hot encoded vectors, ensuring a consistent and structured format for model input.

5.3 Generative Models

In this chapter, several generative models were explored for synthesizing ECG signals. Among the approaches tested, the GAN and the SSSD-ECG models ultimately contributed to the final work. While GANs have demonstrated significant potential in image generation and time series applications, recent research indicates that DMs can outperform them [43], [32]. Indeed, in preliminary experiences, the GAN model developed for this study produced inferior results compared to the SSSD-ECG model, primarily due to higher noise levels in the generated signals. Regardless, both models were essential in developing the custom quality dataset.

5.3.1 GAN

As previously mentioned, the initial approach for generating realistic ECG samples involved a GAN model. However, the performance was inferior compared to the SSSD-ECG model. Even so, the GAN demonstrated the ability to produce signals with discernible R peaks, while other waves were obscured by noise. These characteristics seemed ideal for populating Class 2 of the quality dataset. Given that early-stage results from the SSSD-ECG model did not meet the specific needs for Class 2, the decision was to adapt the GAN architecture to generate the lower-quality signals required for this class.

The model consists of two CNNs: a discriminator and a generator. The discriminator comprises four convolutional layers, each progressively increasing in the number of filters, starting at 64 and doubling with each subsequent layer. These layers utilize a kernel size of 4, a stride of 2, and padding of 1, followed by Leaky ReLU activations (with a negative slope of 0.2) to introduce non-linearity. The final convolutional layer compresses the output to a single channel, which is reshaped and passed through a sigmoid activation function to produce a binary classification output indicating whether the input is real or generated.

The generator synthesizes 1D signals from a latent vector of size 128. It employs five transposed convolutional layers to upsample the latent vector to the desired length and dimensionality. Each layer uses the same kernel size, stride and padding as the discriminator, as also the ReLU activations. Once the signal is upsampled, the output is passed through 12 fully connected layers, each producing a 1000-dimensional sample, corresponding to one signal channel. This allows the generator to produce 12-lead synthetic ECG signals. A high-level schematic of the described GAN architecture is shown in Figure 5.1.

For training, the preprocessed 12-lead ECG dataset from PTB-XL was used. The discriminator and generator were alternately updated using the Adaptive Moment Estimation (Adam) optimizer, with a batch size of 64, over 35 epochs. The learning rates were set to 1×10^{-5} for the discriminator and 2×10^{-5} for the generator. Both networks used binary cross-entropy loss.

For signal generation, outputs from epochs 25 to 35 were used to generate signals for Class 2, as these outputs met the quality requirements. It is important to mention that the conditional component was not used in generating signals for Class 2, as the focus was on producing signals for all classes indiscriminately.

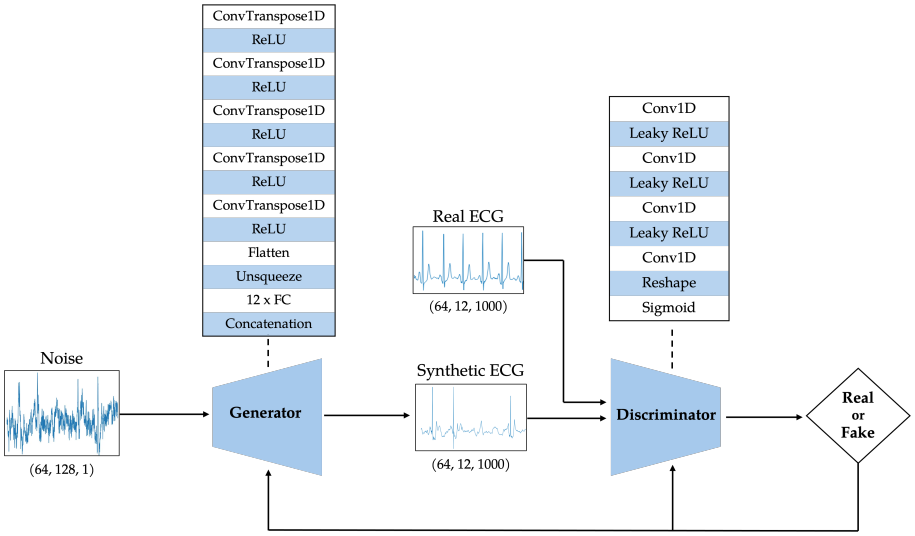


Figure 5.1: Schematic representation of the GAN architecture used in this work. 'Conv1D' denotes one-dimensional convolutional layers, 'ConvTranspose1D' represents transposed convolutional layers, 'ReLU', 'Leaky ReLU', and 'Sigmoid' are activation functions, 'FC' represents fully connected layers, and 'Unsqueeze' adds a dimension to the tensor.

5.3.2 SSSD-ECG

Since the developed GAN model was unsuccessful in generating realistic samples and therefore unsuitable for further analysis, a novel approach, the SSSD-ECG model, emerged as a promising solution for generating 12-lead ECG data. In the original paper, this model excelled in various evaluation contexts, including qualitative, quantitative, and human expert assessments [16].

The SSSD-ECG model, developed by Alcaraz and Strodthoff [16], represents a state-of-the-art framework for ECG generation by leveraging conditional DMs and structured state-space dynamics, as explained in detail in Section 2.2.2.4. In this thesis, two objectives were defined for this model: to produce signals that resemble ECGs with visible periodic R waves and most other waves observable, though containing conceptual errors (intended for Class 3 in the quality dataset), and to generate highly realistic ECG samples for further quality assessment by clinical experts.

The first step in adapting this model for generation involved modifying its conditional input. Instead of using the original 71 statements, the model was conditioned on five broad annotated ECG statements, encoded as a binary vector of length 5, corresponding to the PTB-XL database classes. Additionally, to accomplish the desired results for both objectives, several hyperparameters were modified and tested across different configurations. Table 5.2 presents the experiments conducted with the tested hyperparameters, including the variable names used for reference and their corresponding values.

Table 5.2: Hyperparameters tested in the experiments, detailing the experiment name, the adjusted hyperparameter, its abbreviation for reference, and the respective values applied.

Experiment	Hyperparameter (variable name)	Value
Residual Layers 48	Number of Residual Layers (res_layers)	48
Residual Layers 24	Number of Residual Layers (res_layers)	24
Label Embedding Dimension	Dimension of Label Embedding (label_embed_dim)	256
T Steps	Number of Diffusion Time Steps (T_steps)	300
Batch	Batch size (batch_size)	4
Diffusion Step Embedding In	Dimension of Diffusion Step Embedding (diffusion_step_embed_dim_in)	256
S4 State Dimension	State Dimension for S4 Layer (s4_d_state)	128
S4 Dropout	Dropout Rate for S4 Layer (s4_dropout)	0.2
S4 Layer Normalization	Layer Normalization for S4 Layer (s4_layernorm)	0 (disable)
S4 Bidirectional	Bidirectionality for S4 Layer (s4_bidirectional)	0 (disable)
Hyperparameter Combination	res_layers, label_embed_dim, T_steps	48, 256, 300
Best Hyperparameter Combination	**res_layers, label_embed_dim, T_steps**	**48, 256, 1000**

For the goal of generating realistic synthetic ECG signals, these experiments helped fine-tune the model for improved performance. Only one hyperparameter was adjusted at a time to assess its impact on the generation process. Subsequently, synthetic ECG samples were generated for each configuration and evaluated using the improved precision, improved recall, density, and coverage metrics.

The first effective configuration was achieved by combining the hyperparameters that yielded the best results. The key hyperparameters were the number of diffusion time steps (T_steps = 300), the number of residual layers (res_layers = 48), and the dimension of the label embedding (label_embed_dim = 256). This configuration is referred to as the "Hyperparameter Combination". While these three hyperparameters significantly improved the generation capacity of the model, further optimization was achieved by increasing the number of diffusion time steps to 1000. This refined configuration is referred to as the "Best Hyperparameter Combination."

For generating Class 3 signals, the model used was the one with 24 residual layers. This model was chosen based on extensive testing and visual inspection of generated signals, selecting the configuration that best fit the criteria to populate the class.

It is important to highlight that the model selected for quality assessment was the one with the highest performance as of July, when the evaluation questionnaire for the clinical experts was developed. Even after this, experiments continued to improve model performance, ultimately reaching a better hyperparameter combination.

5.4 Quality Dataset

To develop an evaluator model capable of assessing the quality of synthetic ECG data on a sample-by-sample basis, it was essential to find a dataset that met several specific criteria: it needed to include 12-lead ECG signals, provide a large number of samples suitable for training a DNN, and offer clear, detailed descriptions of quality levels. During this search, two databases were considered: the dataset from the PhysioNet/Computing in Cardiology Challenge 2011 and the Brno University of Technology ECG Quality Database (BUT QDB)

The PhysioNet dataset provided standard 12-lead ECG recordings, each graded by multiple annotators for signal quality [44]. However, it lacked detailed criteria explaining the factors leading to classifications of acceptable, indeterminate, or unacceptable signals, making it unsuitable for the specific needs of this project.

The BUT QDB, on the other hand, was designed specifically for evaluating ECG quality, with annotations categorizing signals into three quality classes based on the clarity and detectability of significant waveforms [45]. Despite its relevance, this dataset was not chosen due to its limited number of signals and focus on single-lead ECGs.

Given that neither dataset fully met the criteria, there was a need to construct a custom quality dataset from scratch. The classification system from the BUT QDB inspired this new dataset, ensuring that the evaluator model could effectively assess the realism and quality of synthetic ECG signals.

The custom quality dataset was constructed using a mix of basic wave functions, synthetic signals generated by a GAN model, synthetic signals from the SSSD-ECG model, and real signals from the PTB-XL dataset. It was categorized into four distinct classes, with examples of each class illustrated in Figure 5.2. The description of each class was based on the ECG characteristic waves, particularly by the R wave, which is typically the first feature a generative model learns when synthesizing ECG signals. The four classes are defined as follows:

- **Class 1:** Signals that do not resemble ECGs.
- **Class 2:** Signals similar to ECGs, but only show discernible R peaks, with noise that obscures other waves.
- **Class 3:** Signals that resemble ECGs with visible periodic R waves and most other waves observable, but containing conceptual errors that result in highly improbable ECG patterns.
- **Class 4:** Real ECG signals.

For Class 1, signals were created using basic wave functions, such as sine, triangular, rectangular, and sawtooth waves, each with varying levels of noise. Class 2 samples were generated using the GAN model described in Section 5.3.1, chosen for their higher noise

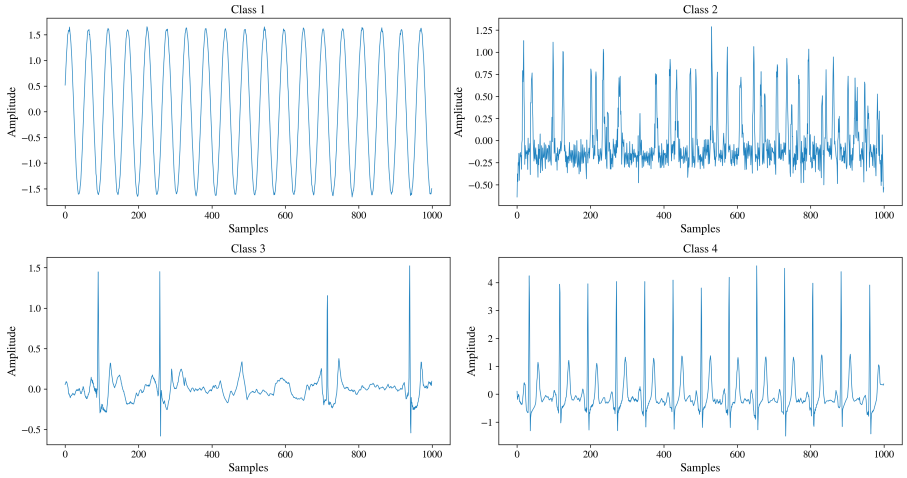


Figure 5.2: Representative signal examples for each class in the Quality Dataset.

levels. Class 3 was produced with specific parameters from the SSSD-ECG model to ensure higher fidelity, as detailed in Section 5.3.2. Class 4 consisted of real signals from the PTB-XL database. The distribution of the number of signals generated for each class is summarized in Table 5.3, providing a detailed breakdown of the dataset composition. It is observable that Class 3 has fewer samples than the other classes, which is due to the manual selection of samples that met the required characteristics, out of the initially 10,000 samples generated.

Table 5.3: Signal distribution across each class of the Quality Dataset.

Class	Number of signals
1	10000
2	8960
3	176
4	10599

5.5 Evaluator Model

The proposed evaluator model serves as a novel evaluation metric for assessing the quality of synthetic ECG data at the sample level, classifying each signal into one of four predefined classes from the custom quality dataset. The necessity of such an evaluator model arises from the limitations of conventional time series generation metrics, which typically rely on statistical comparisons between real and synthetic datasets. These metrics often require compressed representations of the signals, such as extracted features, and depend on analyzing multiple samples from both datasets, which can compromise subject privacy.

In contrast, the evaluator model focuses on evaluating each synthetic ECG sample individually. It was designed using ensemble DL techniques, selected for their demonstrated success in various classification tasks, particularly with DNN [46]. These architectures not

only learn complex relationships between variables but also automatically extract features, eliminating the need for traditional preprocessing.

The evaluator consists of five neural networks, each initialized with different random seeds while maintaining a consistent architecture. Each network includes five convolutional layers, followed by Leaky ReLU activations and dropout layers to prevent overfitting. During training, signals and their corresponding class labels are passed through the networks. The output of the model is then compared to the target class label, and cross entropy loss is calculated. The Adam optimizer is used to adjust the weights and biases, minimizing this loss.

The training data used to develop this model was sourced from the custom quality dataset described earlier, which was unbalanced. Notably, Class 3 contained fewer signals than the other classes. To address this, a function to estimate class weights for the unbalanced data was incorporated. This ensured that the underrepresented classes received more weight during the training process, allowing the model to learn better from the fewer signals available and reducing bias toward the more frequent classes.

Since ensemble learning enhances prediction performance by combining multiple models, it was essential to define an effective strategy for aggregating predictions. Therefore, soft voting was implemented, where each model predicts the probability for each class label, and these probabilities are averaged across all models. The class with the highest average probability is then selected as the final prediction, effectively considering the confidence levels of all model predictions [47].

As mentioned, the training data was obtained from the quality dataset, which includes both synthetic and real data. However, when this work is made publicly available, researchers will have access to the model's weights rather than the raw features of the real and synthetic data. This approach introduces an additional layer of privacy protection, compared to traditional assessment methods.

Additionally, because the quality dataset was specifically designed with diverse waveform characteristics, the evaluator is focused on classifying the quality of signal generation rather than merely detecting noise and artifacts.

5.6 Evaluation

Assessing the quality of time-series generation has demonstrated to be a multidimensional task, covering various aspects such as fidelity, diversity, and utility [20]. As outlined in Section 1.2, the main goal of this work was to produce realistic synthetic ECG samples using the SSSD-ECG model. Therefore, the focus was primarily on fidelity, by evaluating how closely the generated samples resemble real ECG signals. In addition, the diversity of synthetic signals was also evaluated to ensure that the samples represent the full variability of the real data. Moreover, the utility of the synthetic data was assessed through a downstream classification task to measure its usefulness for ML applications.

To complement the quantitative metrics, the generated signals were also subjected to qualitative evaluation by clinical experts through a questionnaire, providing expert feedback on the realism of the data.

Finally, the evaluator model, designed to assess individual synthetic ECG samples, was evaluated using performance metrics such as accuracy, F1-score, and a confusion matrix. The correlation between the evaluation metrics for synthetic data and the evaluator was analyzed to determine if the model aligns with the state-of-the-art metrics. Additionally, the relationship between the human evaluation and the evaluator was also studied.

5.6.1 Fidelity and Diversity of Synthetic ECG Signals

The metrics used to assess the fidelity of the generated data were improved precision and density, while improved recall and coverage metrics were used to evaluate diversity, as proposed by Naeem et al. [23]. For simplicity, throughout this work, improved precision and improved recall will be referred to as precision and recall, respectively. The implementation was adapted to use 5 nearest neighbors (k=5) and 200 samples from each diagnostic class for both real and synthetic data, ensuring a balanced dataset. However, it is worth mentioning that the HYP category in the processed PTB-XL dataset had fewer samples, as shown in Table 5.1. Consequently, the shape of the real dataset used was 891 samples, compared to 1000 synthetic samples.

During the 12 experiments conducted to optimize the performance of the SSSD-ECG model, each experiment produced corresponding fidelity and diversity results for the synthetic data generated. The real dataset used for comparison remained consistent across all experiments.

To compute the metrics, features from multiple domains, such as statistical, spectral, and temporal, were extracted using the Times Series Feature Extraction Library (TSFEL), a Python package optimized for automatic feature extraction from time series data [48].

5.6.2 Utility of the Synthetic Dataset

The utility of the synthetic dataset was evaluated through a downstream classification task using the TRTS and TSTR metrics proposed by Esteban et al. [24], as well as the additional Train on Synthetic, Test on Synthetic (TSTS) metric, introduced in the work of Fekri et al. [49].

The supervised classification task was carried out using a RF classifier (configured with n_estimators=100 and random_state=42), and features were extracted from both the real and synthetic datasets using the TSFEL library. For baseline comparisons, the classifier was trained and tested on real data, with the dataset divided into training and test sets. A test size of 30% was consistently used across all classification tasks.

From these evaluations, three performance measures were derived and analyzed:

- **TSTR:** This metric assesses the capacity of synthetic data to replace real data by evaluating how well a model trained on generated samples performs when tested with real ones. This is essential for assessing the practical applicability of a synthetic dataset in real-world scenarios.
- **TRTS:** This metric measures the realism of synthetic samples by training the classifier on real data and evaluating its performance on synthetic data. It evaluates the ability of the generative model to produce realistic samples.
- **TSTS:** This metric evaluates the internal consistency of the synthetic dataset by measuring how well a model trained on synthetic samples generalizes to unseen synthetic data.

5.6.3 Human Evaluation

The human evaluation specifically targeted the realism of the synthetic dataset, as realism is a property for which humans can provide an unequivocal "ground truth" [5]. Unlike diversity or utility, which may require more quantitative assessments, realism can be directly validated by expert judgment.

To validate the realism of synthetic ECG samples generated by the SSSD-ECG model, a structured questionnaire was developed using Microsoft Forms. The questionnaire featured 20 images of ECG tracings – 10 synthetic signals from the generative model and 10 real signals from the PTB-XL database. This study was conducted in July and involved the evaluations of three clinical experts: one cardiologist, one internist, and a final-year medical student.

Each tracing was paired with a set of questions, beginning with an inquiry about the nature of the signal. The respondents were asked to indicate whether they believed the tracing to be an ECG or not. If uncertain, they could select the 'not sure' option, which allowed them to proceed to the next image. For tracings identified as ECGs, participants were then asked to classify the tracing into one of several diagnostic categories: NORM, MI, STTC, HYP, or CD. These categories correspond to the five 'superclasses' used to classify the PTB-XL data in terms of disease diagnosis.

If a tracing was not recognized as an ECG, the clinical experts were asked to evaluate its quality by selecting one of the following options, which correlate to the quality levels defined in the quality dataset:

- **Noise (Class 1):** the tracing does not resemble an ECG, and the R waves are not reliably observable.
- **Clearly not an ECG (Class 2):** periodic R waves are visible in some leads, but other ECG waves are not clearly identifiable.
- **Almost an ECG (Class 3):** periodic R waves are visible, and most of the waves can be observed, but there are conceptual errors resulting in highly unlikely ECG patterns.

The signals selected for the questionnaire were chosen to represent the diversity within the dataset. To achieve this, a method using a nearest neighbors model was employed to determine how different each sample was from the others, ensuring that the selected samples were as unique as possible. Each selected sample was then reviewed to ensure it accurately reflected the diverse characteristics of the dataset. Ultimately, 10 samples were carefully chosen to ensure a well-rounded representation of the dataset's diversity.

The images of two example questions from the questionnaire used for the human evaluation are provided in Appendix A.

6

Results and Discussion

This chapter presents and discusses the results of this dissertation in accordance with the goals established in Section 1.2. First, the fidelity, diversity, and utility of the generative model are evaluated. Then, the realism of the synthetic signals is validated through expert feedback. Lastly, the performance of the custom evaluator model is assessed, along with its alignment to both quality metrics and expert insights.

6.1 SSSD-ECG Synthetic Signal Evaluation

The hyperparameters from the original paper were used in initial experiments to generate synthetic ECG signals, which were provided to clinical experts for qualitative assessment. While awaiting feedback, further experimentation was conducted to enhance the realism of the synthetic ECG signals, leading to the development of an improved set of hyperparameters detailed in Section 5.3.2.

To assess the fidelity and diversity of the generated samples, a quantitative comparison was made between the two hyperparameter configurations using precision, density, recall, and coverage metrics. The results are summarized in Table 6.1 for each configuration across the five diagnostic classes. Additionally, the utility of the synthetic dataset was evaluated through downstream classification tasks, detailed in Table 6.2.

Table 6.1: Comparison of precision, recall, density, and coverage values across the diagnostic classes for two hyperparameter configurations: the original and the best-performing.

Diagnostic Class	**Original Hyperparameters**				**Best Hyperparameters**			
	Precision	Density	Recall	Coverage	Precision	Density	Recall	Coverage
CD	0.65	0.99	0.00	0.25	0.97	4.15	0.01	0.83
HYP	0.03	0.01	0.01	0.03	0.95	3.13	0.02	0.95
MI	0.95	1.41	0.00	0.32	0.89	2.73	0.03	0.79
NORM	0.28	0.17	0.00	0.06	0.90	3.99	0.04	0.91
STTC	0.95	1.43	0.00	0.27	0.98	5.25	0.06	0.99
Mean	**0.57**	**0.80**	**0.00**	**0.19**	**0.94**	**3.85**	**0.03**	**0.89**

6.1.1 Fidelity and Diversity

As shown in Table 6.1, the average precision of the synthetic ECG signals increased substantially from 0.57 with the original hyperparameters to 0.94 with the best configuration. The most significant improvement was observed in the HYP class, where precision dramatically increased from 0.03 to 0.95. Additionally, the density metric improved across all diagnostic classes, with several exhibiting values greater than 1. Consequently, the overall average density increased significantly from 0.80 to 3.85. These values suggest that the model is generating more synthetic samples in proximity to real data points.

While these improvements in fidelity are significant, it is essential to examine the diversity of the generated signals to gain a holistic comprehension of the performance of the model. Although recall improved with the best set of hyperparameters, it remained low, particularly for the CD and HYP classes. These results suggest that several synthetic samples fall outside the real dataset space, especially for these diagnostic categories. In contrast, the coverage metric showed notable improvements across all categories. Even though some synthetic samples lie outside the real data space (as reflected by low recall), the model is still capable of generating a diverse set of samples that cover the majority of the data space. This increase in coverage may be attributed to the model building the manifold using real samples rather than synthetic ones, as discussed in Section 2.3.2. Therefore, coverage appears more effective in avoiding outliers in distributions compared to recall.

After analyzing both the fidelity and diversity results, it is evident that the best hyperparameters configuration have successfully achieved the goal of generating synthetic ECG signals that exhibit statistical characteristics similar to those of real ones. As confirmed by high precision and density values. However, the lower recall and higher coverage scores indicate that while the model generates a broad array of signals (high coverage), many real points are still not represented in the synthetic dataset (low recall). This limitation, highlights the need for future work to enhance the diversity of the synthetic signals to better capture the full range of characteristics present in real data.

6.1.2 Utility

Synthetic datasets are often used for specific ML applications, and their usefulness can be assessed by evaluating how well they support these applications. The evaluation procedure for assessing the utility of synthetic data involved performing several classification tasks with a RF classifier, as detailed in Section 5.6.2 and shown in Table 6.2

The classifier trained on real data has nearly identical performance when tested on both real (56.58%) and synthetic data (57.03%). These results indicate that the synthetic dataset seems to preserve the characteristics of the real one, confirming the realism of the generated samples.

The model trained on synthetic data performed significantly better on synthetic data

(78.00%) compared to real data (40.84%). This suggests that while data conditioning produces consistent results, it may lack generalization when applied to real-world scenarios. This limitation may be due to the lower values of the diversity metrics. Nevertheless, the synthetic data still displays some quality, despite of not being able to fully replace real data in practical applications

Examining the entire scope, the high similarity between the performance on real and synthetic data suggests that the synthetic dataset replicates many patterns from the real dataset. This is a positive indication of its quality and aligns with the main goal of this dissertation. However, its utility is more limited for training models intended for real-world applications. Nonetheless, this limitation provides an opportunity to enhance synthetic data in future work.

Table 6.2: Macro Average F1-score for RF classification on real and synthetic datasets.

	Test on Real	Test on Synthetic
Train on Real	56.58%	57.03%
Train on Synthetic	40.84%	78.00%

6.2 Human Evaluation

To complement the quantitative metrics, three clinical experts evaluated the synthetic signals to assess their realism. This evaluation was conducted through a questionnaire, detailed in Section 5.6.3, where the primary objective was for the experts to classify the samples as either real or synthetic. A total of 20 ECG tracings were used, consisting of 10 real and 10 synthetic samples generated with the original hyperparameter configuration.

After assessing each ECG tracing, different follow-up questions were presented depending on the answer. If an ECG was considered real, the following question prompted the expert to categorize the signal into one of five diagnostic classes. This step allowed the evaluation of whether the model correctly generated signals corresponding to the intended diagnostic categories. On the other hand, if an ECG was considered synthetic, the expert was asked to indicate the quality level the signal belongs to. In turn, this provides insights into understanding where the generative process might have failed when synthetic signals are accurately identified as such.The responses from all the questions are compiled in Table A.1, which presents a detailed breakdown of the evaluations provided by each expert.

The responses about the nature of the signals from each expert were first analyzed individually, as illustrated in Figure 6.1, where classification outcomes are categorized into distinct groups: real signals identified as real, real signals misclassified as synthetic, synthetic signals misidentified as real, and synthetic signals correctly classified. Following the individual assessments, the responses were then collectively evaluated through majority voting. It is crucial to acknowledge that the first question for each ECG tracing allowed

the evaluator to select the option "Not sure" regarding the nature of the signals. Although only one expert used this option, for the statistical analysis, "Not sure" was treated as a positive classification, indicating that the signal had sufficiently realistic characteristics to cause indecision and was therefore considered real.

Examining individual cases, medical expert A classified all 20 signals as real, without considering any as synthetic. Clinician B correctly identified 8 real signals but also classified 8 synthetic signals as real. The final evaluator classified 5 real ECG tracings as real but labeled the other 5 as synthetic, and 4 synthetic ECGs were classified as real. These results suggest that the synthetic signals possess realistic characteristics, as most were perceived as real. Additionally, the misclassification of some real ECG signals as synthetic further supports the realism of the generated data.

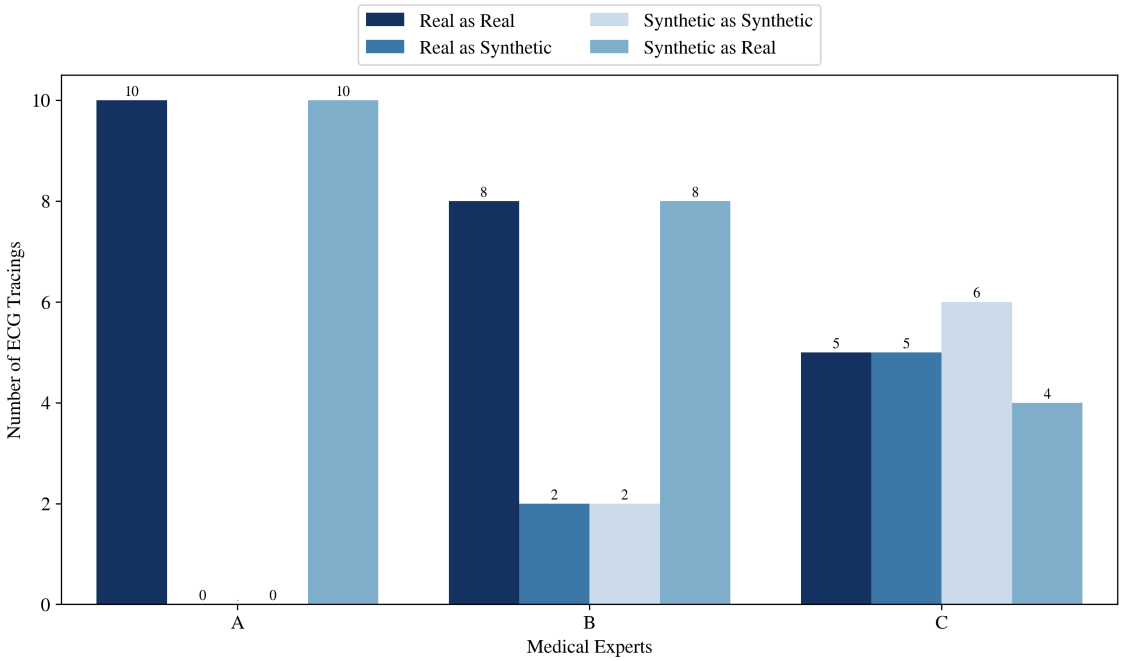


Figure 6.1: Classification of real and synthetic ECG signals by clinical experts.

Taking a holistic view, the majority of the three clinicians identified 16 out of 20 signals as real and 4 as synthetic. Notably, 80% of the synthetic signals (8 out of 10) were misclassified as real, while 20% of the real signals (2 out of 10) were misclassified as synthetic. These results suggest a high degree of realism in the synthetic ECG signals, as the generated data closely resembles genuine ECG tracings in terms of visual characteristics and rhythmic patterns. This aligns with the previously evaluated metrics of precision and density. Additionally, the 20% of real signals misclassified as synthetic highlights the challenge medical experts face in distinguishing between real and generated data.

Figure 6.2 illustrates the realism achieved by presenting two normal 12-lead ECGs evaluated by medical experts. Image 6.2a depicts a real signal from the PTB-XL database, while image 6.2b shows a synthetic signal generated by the SSSD-ECG model using the original hyperparameters. Notably, all evaluators identified the synthetic signal as real.

For the analysis of the second set of follow-up questions, only the feedback from two medical experts was considered, since there was no information on this scenario from one clinician. As mentioned, 8 out of 10 synthetic signals were mistaken for real ones, while the remaining two were correctly classified as synthetic. According to the evaluators, these two synthetic samples fell into the 'Noise' quality level (Class 1), characterized by the absence of observable R waves. This finding suggests that, while the majority of synthetic signals captured the characteristics of real signals, the two that were accurately identified lack resemblance to authentic ECG tracings. Furthermore, the analysis of the synthetic signals classified as real revealed a lack of consensus among the clinicians regarding the assigned diagnostic categories. This inconsistency suggests that the conditional aspect of the generative model may not be functioning as intended, indicating room for improvement.

In conclusion, human evaluation validates the effectiveness of the SSSD-ECG model in generating highly realistic ECG signals. While the evaluation was based on just three clinicians, the results indicate that the synthetic data is considered to possess sufficient quality to merit further exploration.

6.3 Evaluation of the Evaluator Model

6.3.1 Performance Assessment

The performance of the evaluator model was assessed using accuracy and F1-score as evaluation metrics, utilizing the test set from the quality dataset. Since the evaluator is an ensemble of five CNNs, these metrics were averaged across all networks, resulting in a mean accuracy of 99.9884% (±0.0045%) and an average F1-score of 99.6953% (±0.1171%). The confusion matrix, presented in Figure 6.3, further demonstrates the performance of the model, showing that the evaluator correctly classifies signals across all classes. The only misclassification involves a Class 4 signal being incorrectly labeled as Class 3.

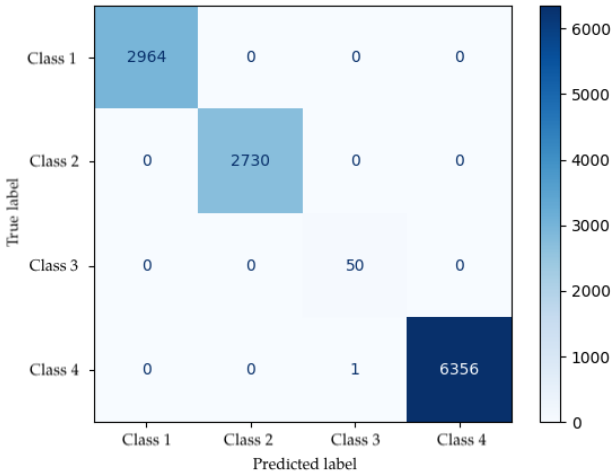
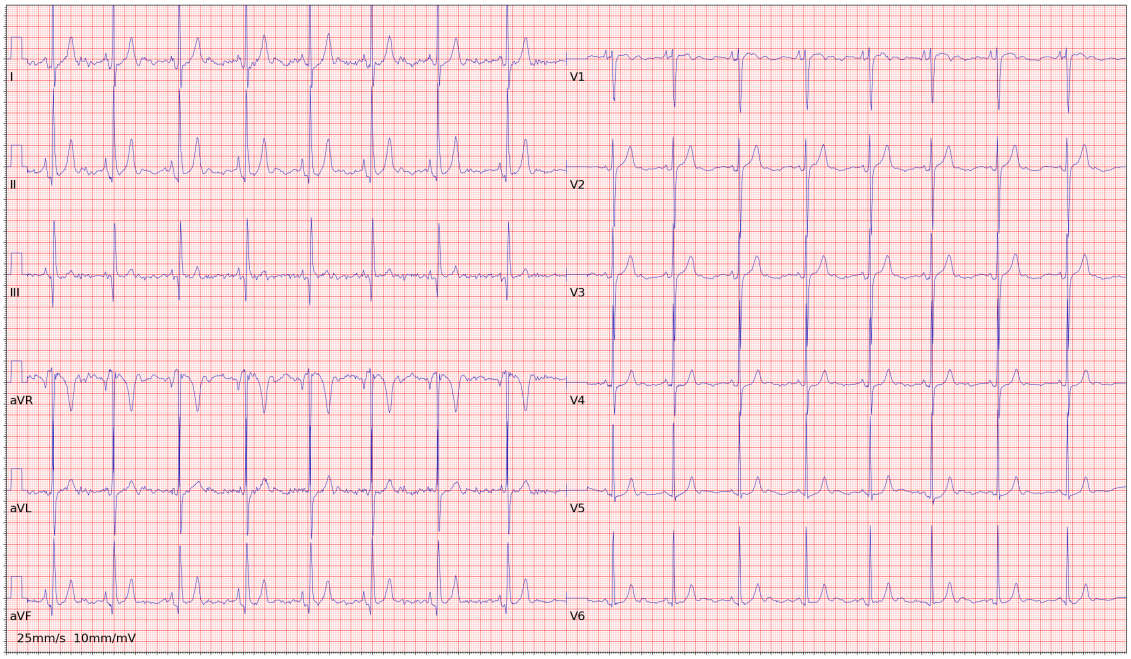
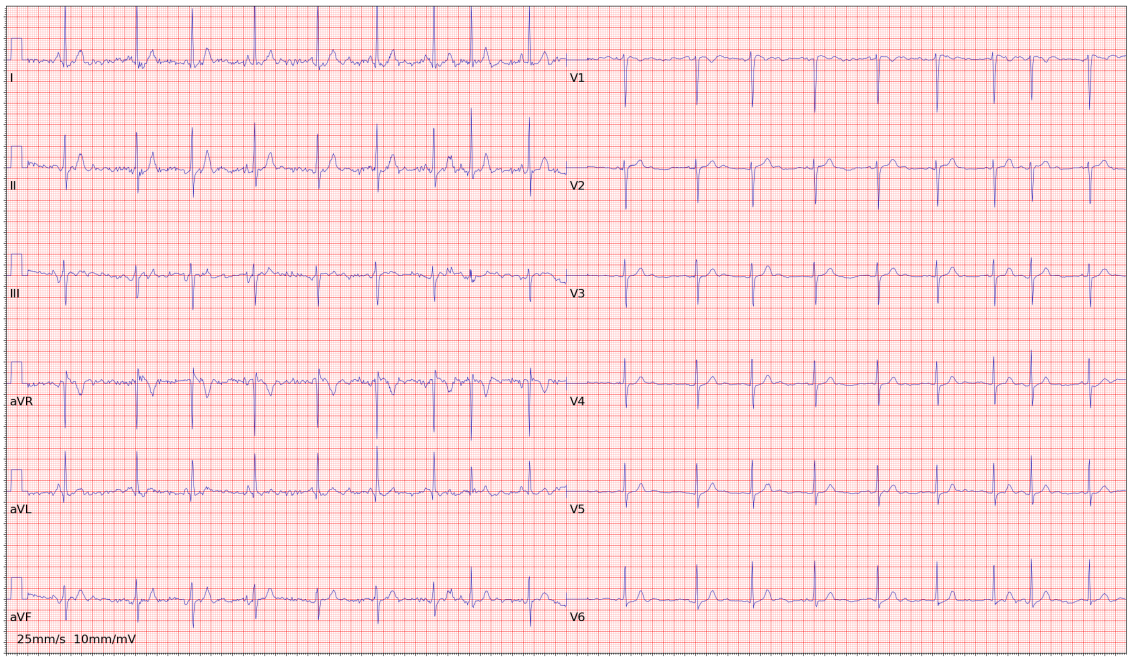


Figure 6.3: Confusion matrix for evaluator model performance.



(a) Real ECG sourced from the PTB-XL database.



(b) Synthetic ECG generated by the SSSD-ECG.

Figure 6.2: Example of two Normal 12-lead ECG tracings presented in the human evaluation questionnaire. Image (a) shows a real ECG from the PTB-XL database, and image (b) represents a synthetic ECG generated by the SSSD-ECG model.

During model evaluation, the class imbalance present in the data, particularly in class 3, was carefully considered as it could impact the model's ability to generalize across different datasets. Therefore, to mitigate this issue, more weight was assigned to the underrepresented class during training, as explained in Section 5.5. Overall, these results demonstrate the effectiveness of the model in distinguishing signals between different quality levels.

6.3.2 Correlation Analysis Between Synthetic Evaluation Metrics and the Evaluator Model

Another approach to assess the performance of the evaluator model involved exploring its relationship with several key evaluation metrics: precision, density, recall, and coverage. By examining these relationships, the objective was to determine whether the evaluator model aligns with the established state-of-the-art metrics.

Synthetic datasets generated from hyperparameter experiments were passed through the evaluator model to predict the percentage of signals classified as Class 4, which were interpreted as real signals. The results for each experiment, including the values of precision, density, recall, and coverage, are detailed in Table 6.3. The final row emphasizes the values of the best hyperparameter configuration. Here, the evaluator model classified 96% of the signals as real, a finding that is corroborated by the high values of precision and density, previously analyzed in this chapter.

Table 6.3: Synthetic datasets generated from each experiment, evaluated by the evaluator model (signals classified as Class 4) and the evaluation metrics (precision, recall, density, and coverage). These values were used to compute the correlations.

Hyperparameter Experiment	Predicted Class 4	Precision	Density	Recall	Coverage
Original	0.44	0.57	0.80	0.00	0.18
Residual Layers	0.71	0.58	0.49	0.00	0.14
Label Embedding Dimension	0.57	0.82	1.61	0.00	0.25
T Steps	0.61	0.62	0.46	0.01	0.15
Batch	0.10	0.34	0.36	0.00	0.09
Diffusion Step Embedding In	0.43	0.66	0.75	0.00	0.19
S4 State Dimension	0.13	0.35	0.39	0.00	0.12
S4 Dropout	0.24	0.59	0.87	0.00	0.19
S4 Layer Normalization	0.02	0.27	0.21	0.00	0.08
S4 Bidirectional	0.05	0.10	0.09	0.00	0.09
Hyperparameter Combination	0.52	0.70	0.69	0.01	0.20
Best Hyperparameter Combination	**0.96**	**0.94**	**3.85**	**0.03**	**0.89**

The correlation values, presented in Figure 6.4, shows that the evaluator model exhibits strong correlations with all evaluation metrics, with coefficients calculated using the Pearson correlation method. Particularly, the model achieves a correlation of 0.88 with precision and a correlation of 0.72 with density. These strong correlations serve as a positive indicator, since both precision and density metrics are focused on fidelity, similar to the evaluator model.

It is worth mentioning that the evaluator model operates at a sample level, whereas the other metrics evaluate at a distribution level. Therefore the predictions of the model were averaged across datasets in order to do this comparison. The significant correlation between the evaluator model and these metrics confirms that the evaluator is a valid tool for quality assessment.

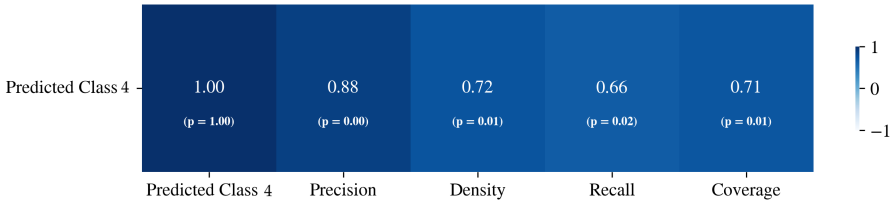


Figure 6.4: Correlation values between evaluator model and evaluation metrics.

6.3.3 Correlation Analysis Between Human Evaluation and the Evaluator Model

Another interesting perspective emerged from analyzing the relationship between the evaluator model and the medical experts. The evaluator model was designed to assess the quality of each ECG signal at a sample level, similar to how the experts performed their evaluation. Thus, it was logical to examine the fidelity of the synthetic ECG signals by comparing the performance of the evaluator model with that of the human evaluators, using the same classification task.

The results, illustrated in Figure 6.5, reveal that the evaluator model correctly identified 7 real signals and classified 7 synthetic signals as real. This performance demonstrates a notable degree of similarity with the medical experts, who also misclassified 8 synthetic signals as real. Additionally, both the evaluator model and the experts exhibited some difficulty in distinguishing certain real signals as real. The alignment in performance between the evaluator model and the human evaluators supports the conclusion that the synthetic data closely resembles genuine ECG tracings, reinforcing their fidelity.

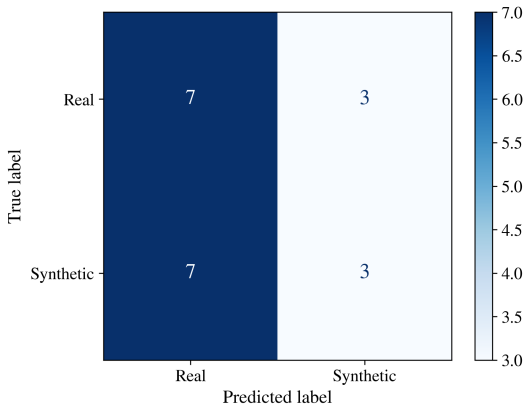


Figure 6.5: Confusion matrix for the classification of signals by the evaluator model.

In summary, the results from the evaluator model and the human evaluators exhibit significant similarity, which validates the realism of the ECG samples generated by the SSSD-ECG model. This also suggests that the evaluator model has the potential to approximate human-like classification performance. However, further experimentation and additional studies are necessary to validate this potential.

7

Conclusions and Future Work

7.1 Conclusion

Synthetic data has emerged as a promising solution to address the challenges of data scarcity and privacy concerns. In healthcare, it has proven to be a tool with potential to improve patient care by supporting clinical research and advancing the development and training of ML models, that can serve as diagnostic support systems. However, in the medical domain, generating data is not sufficient. Medical data must be of high quality and have clinical relevance, as it can significantly impact patient outcomes. Therefore, evaluating the generated data is a critical, yet ambiguous, step since there is no standard procedure for assessing the quality of synthetic data.

Considering the above challenges, this dissertation introduces an approach for generating highly realistic ECG signals through the refinement of a state-of-the-art DM. This was followed by a study conducted to validate the realism of the synthetic samples which gathered insights from medical experts. Additionally, a novel evaluation metric based on a DL evaluator model was proposed to assess the quality of synthetic ECG data at sample level, focusing on generation quality rather than noise and artifacts. To evaluate how this metric aligns with state-of-the-art evaluation methods, a comprehensive correlation analysis was performed.

Regarding the goal of generating highly realistic ECG data, the present work successfully produced synthetic samples that closely resemble real ones. This was accomplished through the adaptation of the SSSD-ECG model, which involved conducting various experiments by modifying one hyperparameter at a time. The fidelity and diversity of each generated ECG dataset were assessed using evaluation metrics from the literature. This process led to the identification of the best model, achieved by combining the best hyperparameters tested and refining them further. The evaluation results indicated that the generated samples displayed statistical characteristics similar to those of real data. The utility of the generated ECG dataset was further examined. The performance analysis in classification tasks confirmed the realism of the synthetic data but also revealed its limitations for training models intended for real-world applications.

Despite these limitations, the research focused on achieving realism, and several criteria support the synthetic ECG data as sufficiently realistic. Additionally, human expert feedback was incorporated to validate the realism of the synthetic dataset. Three medical experts evaluated the generated ECG signals through a questionnaire, where they were presented with an equal number of real and synthetic signals and tasked with identifying their authenticity. While the evaluations conducted by the clinicians indicated that the conditional aspect of the generation did not perform as expected, the consensus among them was that many of the synthetic signals were indistinguishable from genuine ones, further validating the realism of the generated samples. Therefore, the human evaluation supports the effectiveness of the SSSD-ECG model in generating highly realistic ECG signals and demonstrates that the synthetic data is considered to possess sufficient quality to be explored further.

Building on the diverse challenges in evaluating synthetic ECG data, a novel metric for assessing generation quality has been developed. This metric is an evaluator model leveraging a deep ensemble classifier, which categorizes synthetic ECG samples into four levels of quality. Unlike other evaluation metrics in the literature, which operate at a distribution level by assessing the statistical characteristics of synthetic and real data distributions, this model focuses on evaluating the quality of individual ECG signals. Additionally, while other classifiers often emphasize detecting quality by identifying the absence of artifacts and noise, the developed evaluator model provides a more nuanced assessment of synthetic ECG data.

The correlation study performed confirmed an alignment between the state-of-the-art metrics and the evaluator model. The key advantage of the developed metric is that, while it performs equally well at the distribution level, it provides a more detailed analysis at the sample level compared to traditional metrics. Given that the evaluator model assesses synthetic ECG data on a per-sample basis, it provided an interesting opportunity to explore its relationship with human evaluators, who also evaluated individual ECG signals. The results showed a notable degree of similarity between the two, suggesting that the evaluator model might have potential to approximate human-like classification performance, with further analysis and experimentation.

The evaluator model was trained using a quality dataset also developed in this dissertation. This is a significant contribution, as existing databases lack datasets that meet the diverse criteria established in the custom quality dataset. It is categorized into four distinct classes, with each class description based on ECG characteristic waves, particularly the R wave, which is typically the first feature a generative model learns.

In conclusion, this dissertation addresses challenges in generating and evaluating synthetic ECG data, presenting several key contributions. While there are always areas for improvement, particularly in enhancing the diversity and utility of the synthetic dataset, high-quality medical data remains essential for research and the development of models for real-world applications. By taking significant steps towards high-fidelity ECG data generation and evaluation, this work paves the way for future innovations in the field.

7.2 Contributions Summary

- Development of a quality dataset for the generation of 12-lead ECG signals.
- Implementation of an evaluator model that classifies synthetic ECG data into four distinct quality levels.
- Correlation of the developed evaluator model with state-of-the-art metrics and human evaluations.
- Design of an online questionnaire for medical assessment of synthetic ECG signals.
- Validation of the synthetic dataset produced by the SSSD-ECG model through medical expert feedback.

7.3 Limitations and Future Work

The previous section discussed the main contributions of this work, but there are limitations to consider. While the SSSD-ECG model has proven to be an excellent tool for generating realistic ECG signals, the conditional part of the model still leaves room for improvement. The generated samples do not have the desired amount of diversity, as evidenced by significant low recall values, which indicate that the model still struggles with capturing complex features of real signals. Despite this limitation, the DM achieved significant results that support the fidelity of the synthetic ECG dataset, which was further validated by medical experts. These experts confirmed the realism of the samples, but this conclusion was based on a small number of clinical evaluators available at the time of the questionnaire implementation. Therefore, further research involving a larger number of evaluators would provide a more robust validation of the signals.

The evaluator model is a major contribution of this work, classifying synthetic ECG data into four quality levels, from random noise signals to real ECG signals. It would be valuable to enhance the model to not only distinguish between real and synthetic samples but also assess whether the diagnostic labels of real signals are correctly assigned. This would allow for a more comprehensive evaluation of the conditional part of the generative model. Additionally, since the evaluator model operates at the sample level, future projects could explore other sample-level metrics for synthetic data evaluation. These considerations and suggestions could provide valuable insights and contribute to further advancements in time series synthesis and evaluation.

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).

[2] A. Di Costanzo et al. "An artificial intelligence analysis of electrocardiograms for the clinical diagnosis of cardiovascular diseases: a narrative review". In: *Journal of Clinical Medicine* 13.4 (2024), p. 1033 (cit. on p. 1).

[3] G. Monachino et al. "Deep Generative Models: The winning key for large and easily accessible ECG datasets?" In: *Computers in biology and medicine* (2023), p. 107655 (cit. on pp. 1, 9).

[4] H. Murtaza et al. "Synthetic data generation: State of the art in health care domain". In: *Computer Science Review* 48 (2023), p. 100546 (cit. on pp. 1, 2).

[5] G. Stein et al. "Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models". In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on pp. 2, 27).

[6] *AISyM4Med Project Website*. https://aisym4med.eu/about-the-project/. Accessed: January 20, 2024 (cit. on p. 2).

[7] K. Antczak. "A generative adversarial approach to ECG synthesis and denoising". In: *arXiv preprint arXiv:2009.02700* (2020) (cit. on p. 4).

[8] P. Ongsulee. "Artificial intelligence, machine learning and deep learning". In: *2017 15th international conference on ICT and knowledge engineering (ICT&KE)*. IEEE. 2017, pp. 1–6 (cit. on pp. 5, 6, 8).

[9] A. Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022 (cit. on p. 5).

[10] A. F. Alnuaimi and T. H. Albaldawi. "An overview of machine learning classification techniques". In: *BIO Web of Conferences*. Vol. 97. EDP Sciences. 2024, p. 00133 (cit. on p. 5).

[11] P. A. A. Resende and A. C. Drummond. "A survey of random forest based methods for intrusion detection systems". In: *ACM Computing Surveys (CSUR)* 51.3 (2018), pp. 1–36 (cit. on p. 5).

[12] I. H. Sarker. "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions". In: *SN Computer Science* 2.6 (2021), p. 420 (cit. on pp. 6, 7).

[13] N. K. Chauhan and K. Singh. "A review on conventional machine learning vs deep learning". In: *2018 International conference on computing, power and communication technologies (GUCON)*. IEEE. 2018, pp. 347–352 (cit. on p. 7).

[14] K. O'Shea and R. Nash. "An introduction to convolutional neural networks". In: *arXiv preprint arXiv:1511.08458* (2015) (cit. on p. 8).

[15] Z. Guo et al. "Diffusion Models in Bioinformatics: A New Wave of Deep Learning Revolution in Action". In: *arXiv preprint arXiv:2302.10907* (2023) (cit. on p. 9).

[16] J. M. L. Alcaraz and N. Strodthoff. "Diffusion-based conditional ECG generation with structured state space models". In: *Computers in Biology and Medicine* (2023), p. 107115 (cit. on pp. 10, 11, 15, 21).

[17] J. M. L. Alcaraz and N. Strodthoff. "Diffusion-based time series imputation and forecasting with structured state space models". In: *arXiv preprint arXiv:2208.09399* (2022) (cit. on p. 10).

[18] Z. Kong et al. "Diffwave: A versatile diffusion model for audio synthesis". In: *arXiv preprint arXiv:2009.09761* (2020) (cit. on p. 10).

[19] A. Gu, K. Goel, and C. Ré. "Efficiently modeling long sequences with structured state spaces". In: *arXiv preprint arXiv:2111.00396* (2021) (cit. on p. 10).

[20] M. Stenger et al. "Evaluation is key: a survey on evaluation measures for synthetic time series". In: *Journal of Big Data* 11.1 (2024), p. 66 (cit. on pp. 11, 15, 25).

[21] M. S. Sajjadi et al. "Assessing generative models via precision and recall". In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 11, 15).

[22] T. Kynkäänniemi et al. "Improved precision and recall metric for assessing generative models". In: *Advances in neural information processing systems* 32 (2019) (cit. on pp. 11, 12, 16).

[23] M. F. Naeem et al. "Reliable fidelity and diversity metrics for generative models". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7176–7185 (cit. on pp. 12, 13, 16, 26).

[24] C. Esteban, S. L. Hyland, and G. Rätsch. "Real-valued (medical) time series generation with recurrent conditional gans". In: *arXiv preprint arXiv:1706.02633* (2017) (cit. on pp. 12, 13, 26).

[25] S. Aziz, S. Ahmed, and M.-S. Alouini. "ECG-based machine-learning algorithms for heartbeat classification". In: *Scientific reports* 11.1 (2021), p. 18738 (cit. on p. 14).

[26] P. E. McSharry et al. "A dynamical model for generating synthetic electrocardiogram signals". In: *IEEE transactions on biomedical engineering* 50.3 (2003), pp. 289–294 (cit. on p. 14).

[27] N. Wulan et al. "Generating electrocardiogram signals by deep learning". In: *Neurocomputing* 404 (2020), pp. 122–136 (cit. on p. 14).

[28] T. Dissanayake et al. "Generalized Generative Deep Learning Models for Biosignal Synthesis and Modality Transfer". In: *IEEE Journal of Biomedical and Health Informatics* 27.2 (2022), pp. 968–979 (cit. on p. 14).

[29] D. Belo et al. "Biosignals learning and synthesis using deep neural networks". In: *Biomedical engineering online* 16 (2017), pp. 1–17 (cit. on p. 14).

[30] R. Nishikimi et al. "Variational autoencoder–based neural electrocardiogram synthesis trained by FEM-based heart simulator". In: *Cardiovascular Digital Health Journal* (2023) (cit. on p. 15).

[31] M. H. Zama and F. Schwenker. "Ecg synthesis via diffusion-based state space augmented transformer". In: *Sensors* 23.19 (2023), p. 8328 (cit. on p. 15).

[32] N. Neifar et al. "DiffECG: A Versatile Probabilistic Diffusion Model for ECG Signals Synthesis". In: *arXiv preprint arXiv:2306.01875* (2023) (cit. on pp. 15, 20).

[33] Z. Zhang et al. "Dynamic time warping under limited warping path length". In: *Information Sciences* 393 (2017), pp. 91–107 (cit. on p. 16).

[34] A. M. Delaney, E. Brophy, and T. E. Ward. "Synthesis of realistic ECG using generative adversarial networks". In: *arXiv preprint arXiv:1909.09150* (2019) (cit. on p. 16).

[35] A. Alaa et al. "How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 290–306 (cit. on p. 16).

[36] C. Liu et al. "Signal quality assessment and lightweight QRS detection for wearable ECG SmartVest system". In: *IEEE Internet of Things Journal* 6.2 (2018), pp. 1363–1374 (cit. on p. 16).

[37] M. Athif and C. Daluwatte. "Combination of rule based classification and decision trees to identify low quality ECG". In: *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*. IEEE. 2017, pp. 1–4 (cit. on p. 16).

[38] E. Morgado et al. "Quality estimation of the electrocardiogram using cross-correlation among leads". In: *Biomedical engineering online* 14 (2015), pp. 1–19 (cit. on p. 16).

[39] G. Liu et al. "ECG quality assessment based on hand-crafted statistics and deep-learned S-transform spectrogram features". In: *Computer Methods and Programs in Biomedicine* 208 (2021), p. 106269 (cit. on p. 16).

[40] J. Zhang et al. "A signal quality assessment method for electrocardiography acquired by mobile device". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 1–3 (cit. on p. 16).

[41] M. A. Reyna et al. "Will two do? Varying dimensions in electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021". In: *2021 Computing in Cardiology (CinC)*. Vol. 48. IEEE. 2021, pp. 1–4 (cit. on p. 17).

[42] P. Wagner et al. "PTB-XL, a large publicly available electrocardiography dataset". In: *Scientific data* 7.1 (2020), pp. 1–15 (cit. on pp. 17, 19, 49).

[43] E. Adib et al. "Synthetic ecg signal generation using probabilistic diffusion models". In: *IEEe Access* (2023) (cit. on p. 20).

[44] I. Silva, G. B. Moody, and L. 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 (cit. on p. 23).

[45] A. Nemcova et al. "Brno university of technology ECG quality database (BUT QDB)". In: *PhysioNet* 101 (2020), e215–e220 (cit. on p. 23).

[46] A. Mohammed and R. Kora. "A comprehensive review on ensemble deep learning: Opportunities and challenges". In: *Journal of King Saud University-Computer and Information Sciences* 35.2 (2023), pp. 757–774 (cit. on p. 24).

[47] P. Mahajan et al. "Ensemble learning for disease prediction: A review". In: *Healthcare*. Vol. 11. 12. MDPI. 2023, p. 1808 (cit. on p. 25).

[48] M. Barandas et al. "TSFEL: Time series feature extraction library". In: *SoftwareX* 11 (2020), p. 100456 (cit. on p. 26).

[49] M. N. Fekri, A. M. Ghosh, and K. Grolinger. "Generating energy data for machine learning with recurrent generative adversarial networks". In: *Energies* 13.1 (2019), p. 130 (cit. on p. 26).

A

Human Evaluation Questionnaire

This appendix provides additional information about the human evaluation approach used to assess the realism of synthetic ECG signals generated by the SSSD-ECG model. It first includes two example questions from the structured questionnaire, followed by a table summarizing the responses from medical experts.

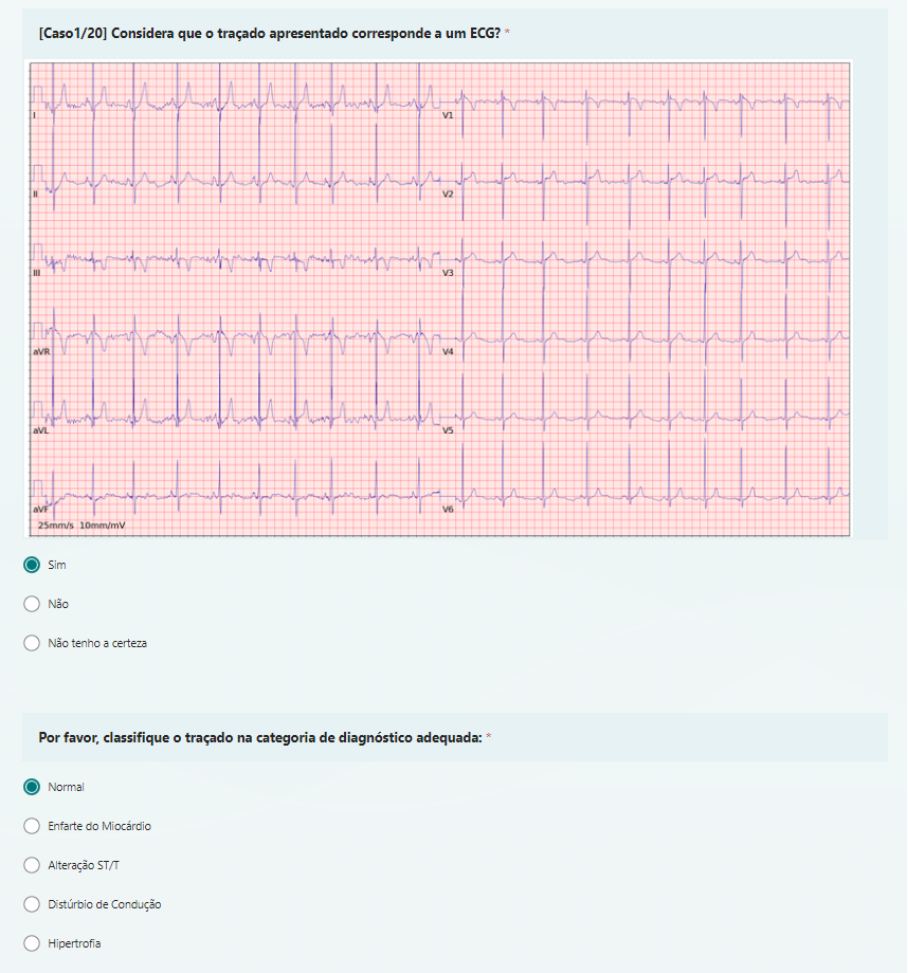


Figure A.1: Example of the questionnaire developed for human evaluation of ECG signals. The image illustrates the scenario where the ECG tracing is classified as real by the experts. It then displays the follow-up question, prompting the classification of the signal into one of five diagnostic categories.

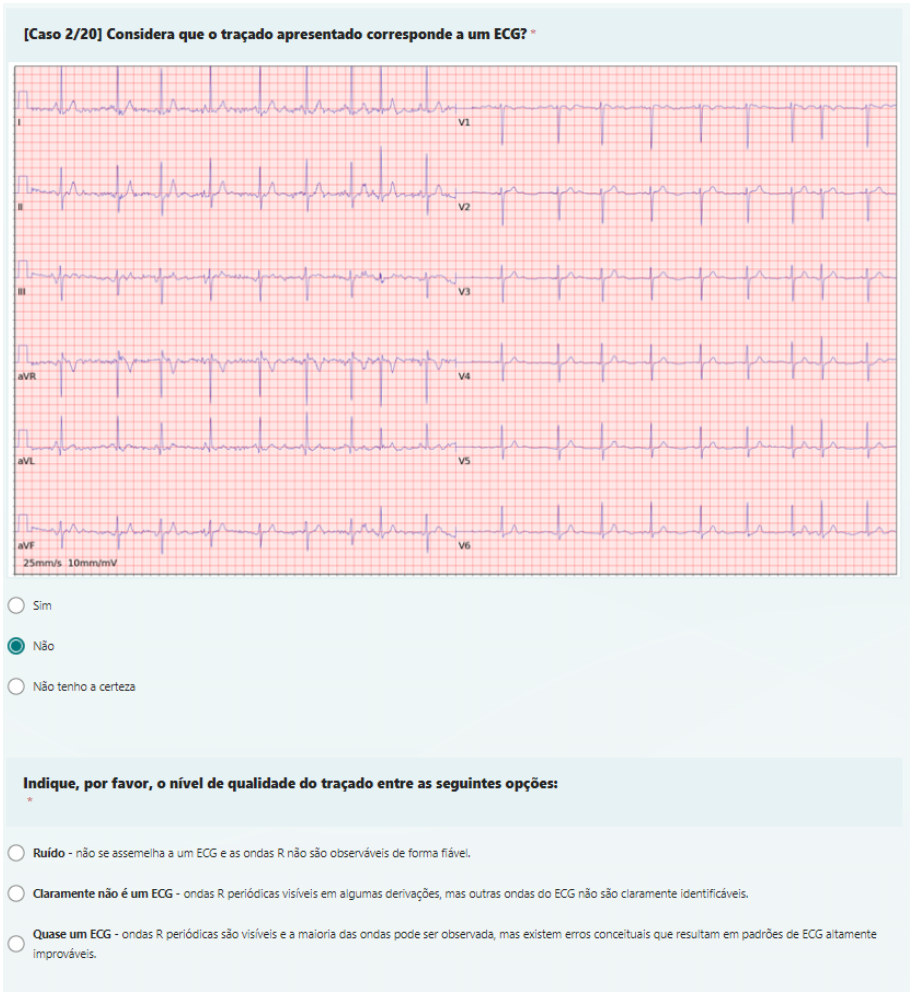


Figure A.2: Example of the questionnaire developed for human evaluation of ECG signals. This image depicts the scenario where the ECG tracing is perceived as synthetic by the experts. It also shows the follow-up question, which prompts the classification of the tracing into one of three quality levels.

Table A.1: Summary of the responses of medical experts to the questionnaire. The table presents data from the questionnaire answered by medical experts, detailing their evaluation of each ECG tracing. For each signal, the actual nature (real or synthetic) and diagnostic category are provided. The responses from each expert include the perceived nature of the signal (real or synthetic). If the signal is considered real, the diagnostic category is assessed, but if the signal is considered synthetic, only the quality level is evaluated. If an expert was "Not sure" about the nature of the signal, no diagnostic category or quality level was assigned.

	Actual		Medical Expert A			Medical Expert B			Medical Expert C		
Signal	Nature	Diagnostic Category	Nature	Diagnostic Category	Quality Level	Nature	Diagnostic Category	Quality Level	Nature	Diagnostic Category	Quality Level
1	Real	NORM	Real	HYP	-	Real	NORM	-	Real	MI	-
2	Synthetic	NORM	Real	CD	-	Real	CD	-	Real	CD	-
3	Synthetic	NORM	Real	CD	-	Real	CD	-	Synthetic	-	Almost an ECG
4	Real	CD	Real	HYP	-	Synthetic	-	Noise	Synthetic	-	Clearly not an ECG
5	Synthetic	CD	Real	NORM	-	Not sure	-	-	Synthetic	-	Clearly not an ECG
6	Synthetic	NORM	Real	NORM	-	Real	CD	-	Real	NORM	-
7	Real	NORM	Real	HYP	-	Not sure	-	-	Real	CD	-
8	Real	MI	Real	HYP	-	Not sure	-	-	Synthetic	-	Clearly not an ECG
9	Synthetic	NORM	Real	NORM	-	Not sure	-	-	Real	-	-
10	Real	STTC	Real	MI	-	Real	CD	-	Synthetic	-	Clearly not an ECG
11	Synthetic	HYP	Real	NORM	-	Synthetic	-	Noise	Synthetic	-	Noise
12	Synthetic	CD	Real	CD	-	Not sure	-	-	Synthetic	-	Almost an ECG
13	Real	NORM	Real	HYP	-	Real	NORM	-	Real	STTC	-
14	Real	NORM	Real	HYP	-	Real	NORM	-	Real	NORM	-
15	Real	HYP	Real	HYP	-	Synthetic	-	Almost an ECG	Synthetic	-	Almost an ECG
16	Synthetic	STTC	Real	NORM	-	Not sure	-	-	Synthetic	-	Clearly not an ECG
17	Synthetic	MI	Real	NORM	-	Synthetic	-	Noise	Synthetic	-	Noise
18	Real	MI	Real	MI	-	Not sure	-	-	Synthetic	-	Clearly not an ECG
19	Real	NORM	Real	HYP	-	Real	HYP	-	Real	CD	-
20	Synthetic	NORM	Real	NORM	-	Real	CD	-	Real	CD	-

I

PTB-XL Dataset

This annex presents a figure detailing the descriptions of the SCP-ECG acronyms from the PTB-XL dataset, which was used for training the generative models.

		Acronym	SCP statement Description
Superclasses		NORM	Normal ECG
		CD	Conduction Disturbance
		MI	Myocardial Infarction
		HYP	Hypertrophy
		STTC	ST/T change
Subclasses	NORM	NORM	Normal ECG
	CD	LAFB/LPFB	left anterior/left posterior fascicular block
		IRBBB	incomplete right bundle branch block
		ILBBB	incomplete left bundle branch block
		CLBBB	complete left bundle branch block
		CRBBB	complete right bundle branch block
		_AVB	AV block
		IVCB	non-specific intraventricular conduction disturbance (block)
		WPW	Wolff-Parkinson-White syndrome
	HYP	LVH	left ventricular hypertrophy
		RHV	right ventricular hypertrophy
		LAO/LAE	left atrial overload/enlargement
		RAO/RAE	right atrial overload/enlargement
		SEHYP	septal hypertrophy
	MI	AMI	anterior myocardial infarction
		IMI	inferior myocardial infarction
		LMI	lateral myocardial infarction
		PMI	posterior myocardial infarction
	STTC	ISCA	ischemic in anterior leads
		ISCI	ischemic in inferior leads
		ISC_	non-specific ischemic
		STTC	ST-T changes
		NST_	non-specific ST changes

Figure I.1: SCP-ECG acronym descriptions for super- and subclasses, adapted from [42].