INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Data Augmentation with GANs applied to Healthcare

Carlos Daniel Costa Faria

Master Degree in Data Science

Supervisor:
PhD Diana E. Aldea Mendes, Associate Professor,
ISCTE – University Institute of Lisbon

October, 2024

Data Augmentation with GANs applied to Healthcare

Carlos Daniel Costa Faria

Master Degree in Data Science

Supervisor:
PhD Diana E. Aldea Mendes, Associate Professor,
ISCTE – University Institute of Lisbon

October, 2024

Aos meus pais e ao meu irmão, pela força silenciosa que sustentou cada passo desta jornada.

Acknowledgment

Um agradecimento especial à orientadora de dissertação, a Professora Diana Mendes, pelo auxílio constante durante todos os processos até ao momento de entrega da dissertação. Demonstrando sempre disponibilidade e rapidez na resposta, durante todo o desenvolvimento.

Com muita gratidão, dedico esta dissertação aos meus pais, Manuel e Paula, e ao meu irmão, Dinis, cuja presença e apoio constante me inspiraram ao longo de todo o percurso. Em cada desafio e conquista, estiveram ao meu lado, oferecendo força e motivação e aceitando com generosidade cada obstáculo que enfrentei. Obrigado por tornarem este caminho mais leve e significativo.

Aos amigos que me acompanharam neste percurso – João, Duarte, Pedro, Jhonata, Bárbara, Beatriz e Jaciara – deixo meu sincero agradecimento. Com vocês, cada etapa tornou-se mais leve, mais divertida e cheia de bons momentos. Obrigado por fazerem parte desta jornada.

Por fim, um agradecimento especial ao "Regresso dos que nunca foram", uma base sólida de amizade que tem sido um pilar na minha vida. São tantos amigos importantes que optei por me referir a vocês desta forma, pois sei que todos que fazem parte deste grupo entenderão a profundidade do meu carinho e gratidão. Obrigado por estarem sempre presentes e tornarem cada momento inesquecível.

Resumo

Esta dissertação explora a aplicação de Redes Adversariais Generativas (GANs) para gerar dados de séries temporais, com foco particular em sinais de eletrocardiograma (ECG) usados para detecção de arritmias. A escassez de dados nas áreas médicas é agravada pelas regulamentações de privacidade, pelas complexidades técnicas da recolha de dados e pela raridade de certas patologias, que limitam o acesso a conjuntos de dados abrangentes.

Recorrendo à base de dados de arritmia do MIT-BIH, este estudo aproveita uma arquitetura Wasserstein GAN com Gradient Penalty (WGAN-GP) e altera a estrutura do modelo adicionando camadas Long Short-Term Memory (LSTM) bidirecionais para gerar sinais de ECG sintéticos realistas. Esses sinais sintéticos visam equilibrar conjuntos de dados para classificação de arritmia, melhorando o desempenho do classificador onde os métodos tradicionais de aumento de dados são insuficientes devido a restrições de privacidade, raridade e complexidade em dados médicos.

O processo de treino do modelo GAN foi avaliado usando uma combinação de métricas quantitativas, como *Euclidean Distance* e *Dynamic Time Warping*, juntamente com técnicas visuais como PCA e t-SNE. Além disso, um modelo de classificação treinado com dados de ECG aumentados demonstrou potencial na abordagem de desequilíbrios no conjunto de dados e no aumento da precisão na deteção de eventos arrítmicos, demonstrando a eficácia do GAN na melhoria do desempenho do modelo.

Este trabalho contribui para o campo da ciência de dados em saúde. Destaca o potencial das GANs para superar desafios significativos, fornecendo conjuntos de dados diversos que preservam a privacidade e melhoram a precisão do modelo de diagnóstico. Através desta abordagem, os GANs oferecem uma ferramenta para a investigação médica, facilitando o desenvolvimento de modelos preditivos robustos, mantendo ao mesmo tempo, a integridade e a confidencialidade dos dados. Os resultados realçam o potencial de impacto dos GANs, onde a maior acessibilidade e diversidade dos dados podem melhorar significativamente os resultados dos pacientes na deteção de arritmia e muito mais.

Palavras-Chave: Redes Adversariais Generativas, Electrocardiograma, Aumento de Dados, Séries Temporais

Abstract

This dissertation explores the application of Generative Adversarial Networks (GANs) to generate time-series data, with a particular focus on Electrocardiogram (ECG) signals used for arrhythmia detection. Data scarcity in medical fields is compounded by privacy regulations, the technical complexities of data collection, and the rarity of certain pathologies, all of which limit access to comprehensive datasets.

With a foundation in the MIT-BIH Arrhythmia Database, this study leverages a Wasserstein GAN with Gradient Penalty (WGAN-GP) architecture and changes the model's structure by adding bidirectional Long Short-Term Memory (LSTM) layers to generate realistic synthetic ECG signals. These synthetic signals aim to balance datasets for arrhythmia classification, improving classifier performance where traditional Data Augmentation (DA) methods fall short due to privacy, rarity, and complexity constraints in medical data.

The GAN model's training was evaluated using a combination of quantitative metrics such as Euclidean Distance and Dynamic Time Warping (DTW), alongside visual techniques like Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). Additionally, a classification model trained on augmented ECG data demonstrated potential in addressing dataset imbalances and enhancing accuracy in detecting arrhythmic events, demonstrating the GAN's effectiveness in enhancing model performance.

This work contributes to the broader field of healthcare data science. It highlights the potential of GANs to overcome significant challenges by providing privacy-preserving, diverse datasets that improve diagnostic model accuracy. Through this approach, GANs offer a tool for medical research, facilitating the development of robust predictive models while maintaining data integrity and confidentiality. The results underscore the potential for GANs to impact, where enhanced data accessibility and diversity can significantly improve patient outcomes in arrhythmia detection and beyond.

Keywords: Generative Adversarial Networks, Electrocardiogram, Data Augmentation, Time-Series

Contents

List of Tables

List of Figures

List of Acronyms

AAMI: Association for the Advancement of Medical Instrumentation.
AE: Autoencoder.
AUC: Area Under the ROC Curve.

bpm: Beats per minute.

CNN: Convolutional Neural Network.

DA: Data Augmentation.
DTW: Dynamic Time Warping.

ECG: Electrocardiogram.
EEG: Electroencephalogram.
EMG: Electromyography.

GAN: Generative Adversarial Network.
GASF: Gramian Angular Summation Field.

ICU: Intensive Care Unit.

KLD: Kullback–Leibler divergence.

LSTM: Long Short-Term Memory.

ML: Machine Learning.

NN: Neural Network.
NYC: New York City.

PCA: Principal Component Analysis.
PCC: Pearson correlation coefficient.
PeMS: Caltrans Performance Management System.

RNN: Recurrent Neural Network.

SEN: Sensitivity.
SMOTE: Synthetic Minority Oversampling Technique.

t-SNE: t-distributed Stochastic Neighbor Embedding.
TRTS: Train on Real, Test on Synthetic.
TSTR: Train on Synthetic, Test on Real.

VAE: Variational Autoencoder.

CHAPTER 1

Introduction

1.1. Background and Motivation

In the evolving landscape of Data Science, time-series data has emerged as a crucial component of analysis across various domains. From finance and healthcare (Yoon et al., 2019) to energy (Demir et al., 2021) and transportation (Huang et al., 2020), the analysis of sequential data points collected over time has become instrumental in discovering patterns, making predictions, and helping people in decision-making tasks, that would be too complex to analyze. However, as the demand for time-series analysis grows, so does the challenge of gathering sufficient and diverse datasets to support deep learning model development, as these models require large amounts of data to achieve effective training and performance.

Data scarcity and limited diversity has been a big challenge in many areas of Data Science, especially in sensitive fields like healthcare. Acquiring healthcare data can present significant challenges, primarily due to strict privacy regulations, technical complexities of data collection, and the rarity of specific medical conditions. Privacy is important, as patient data is highly sensitive, requiring protocols to protect confidentiality, which is key for keeping the rights of the patients intact. Although this often limits the availability of comprehensive datasets, especially in the cases where sharing data is legally restricted. Additional, recording physiological data, such as ECG or Electroencephalogram (EEG) signals, requires specialized equipment and controlled environments, making it a demanding process logistically and financially. Furthermore, certain pathologies are rare, resulting in limited case examples that cannot easily be expanded through conventional data collection. All these factors contributes to obstacles for research and development such as Machine Learning (ML).

These challenges have increased the demand for innovative DA techniques to overcome such limitations. DA have long been employed to artificially enhance the size and variability of datasets, since 1998 by LeCun et al. (1998), when a technique for enhancing the dataset size and variability without gathering more real data was firstly introduced.

GANs, first introduced by Goodfellow et al. (2014), have revolutionized the field of synthetic data generation. Initially gaining attention for their applications in image creation, GANs have since demonstrated their potential across various domains, potentially transforming data-intensive fields by enhancing dataset diversity while preserving privacy. The adversarial training process of GANs, where a generator network competes against a discriminator network, allows for the creation of highly realistic synthetic data that resembles the statistical properties and temporal dependencies of real-world samples.

However, the use of GANs is not without challenges, specially in sequential data. GANs present challenges such as training instability and evaluation difficulties, necessitating continued research to optimize their efficacy and reliability. Unlike image-based GANs, where visual inspection can provide some insight into the quality of generated samples, assessing the fidelity of synthetic time-series data is less straightforward. The lack of standardized, reliable metrics makes it difficult to compare different models and evaluate improvements accurately.

1.2. Objectives and Research Questions

This dissertation dives into the world of time-series DA, with a particular focus on the application of GANs in healthcare DA. The primary objective of this research is to address the persistent issues of data scarcity and limited diversity in datasets, especially in sensitive domains where privacy concerns and data collection challenges are pushing back in the ability of getting large amounts of quality data.

This research aims to provide a comprehensive exploration of the application of GANs in time-series DA, with a specific focus on healthcare applications. The research begins with a literature review, examining the evolution of DA techniques, the fundamentals of GANs, and their adaptations for time-series. We then delve into the various GAN architectures and models that have been proposed for time-series generation, analyzing their strengths, limitations, and applicability to different types of data.

The application of GANs across various domains are also explored, including energy, traffic data, finance, and audio processing, before focusing more deeply on healthcare applications. This broad investigation provides insights into the versatility of GANs and the unique challenges posed by different types of data.

Building on the insights gained from chapter 2, this dissertation presents a novel GAN-based approach for time-series DA in healthcare. The objective is to design and implement a GAN model specifically adapted to generate synthetic ECG signals, contributing to enhanced data availability for arrhythmia classification and analysis. A *WGAN-GP* architecture is employed, optimized to capture the temporal characteristics of ECG signals and produce samples that reflect the diversity and complexity of real data. Section 4.1 outlines the detailed architecture of the model, highlighting the use of a gradient penalty to enhance training stability and issues such as mode collapse.

A critical aspect of this research is the evaluation of GAN performance in generating time-series data. Through quantitative metrics such as Euclidean Distance and DTW, Table 4.5, as well as visualizations like PCA and t-SNE, Figures 4.15 and 4.16, the quality and realism of the synthetic signals are evaluated. Additionally, the study explores the application of synthetic data in classification tasks, assessing its potential to address class imbalance and improve model performance, a common challenge in healthcare datasets where normal heartbeats significantly outnumber arrhythmic events.

To validate the effectiveness of the proposed approach, an experiment using a classification model is also produced. In section 4.4, the utility of synthetic ECG data for enhancing model performance is investigated in classification tasks by training a simple classification model on an augmented dataset, containing both original and synthetic data. This process assesses whether synthetic data can effectively mitigate class imbalance issues inherent in the dataset.

Ultimately, this research addresses critical questions about the effectiveness of GAN-based DA in healthcare and its ability to enhance model performance in classification applications. By experimenting methods to generate realistic synthetic data, this contributes to the broader field of healthcare data science, offering solutions to issues related to data scarcity, privacy, and the ethical sharing of medical information. Through these efforts, the potential for GANs to transform data availability in healthcare is underscored, opening new avenues for innovation and improving patient outcomes through enhanced predictive modeling and diagnostics.

1.3. Methodology and Organization of the Dissertation

This dissertation is organized to present a clear and concise overview of the development and validation of a GAN-based model designed to generate synthetic ECG data, addressing healthcare data scarcity and class imbalance in arrhythmia classification. Each chapter builds on the previous one, creating a narrative from theoretical foundation to practical implementation.

Chapter 2 delivers a comprehensive examination of DA techniques, focusing on GANs and their adaptations for time-series data. Subsections explore the evolution of GAN architectures, with attention to time-series applications in healthcare, finance, and other fields. The chapter discusses the unique challenges GANs face, such as training instability and mode collapse.

Chapter 3 describes the approach to developing a WGAN-GP architecture specifically for generating ECG signals. Section 3.1 covers the MIT-BIH Arrhythmia Database as the data source, detailing preprocessing steps like data normalization and segmentation.

Chapter 4 presents the outcomes of the GAN model's training on ECG data, assessing the quality of synthetic signals. Subsections analyze quantitative metrics (Euclidean Distance, DTW, PCC, and Kullback–Leibler divergence (KLD)) and visual methods (PCA, t-SNE) for evaluating the similarity between real and synthetic data. This chapter also examines the impact of synthetic data on classification performance, particularly in addressing dataset imbalances, providing a discussion of the results in relation to the research questions.

CHAPTER 2

Literature Review

In recent years, Data Science has seen a significant surge in the use of time-series data across various domains. However, a persistent challenge in this area is the scarcity and limited diversity of data. This issue is particularly common in fields like healthcare, where privacy concerns restrict access to sensitive information and collecting disease-related data is even more challenging due to the rarity of these events when those diseases are uncommon. GANs offer a promising solution to these challenges, especially through DA.

DA is a methodology used to artificially enhance the size and variability of datasets. For time-series, DA is not just a tool for enriching datasets but a crucial element in ensuring the robustness and accuracy of classification or predictive models, particularly those based on Neural Networks (NNs), as these require large amounts of data to learn effectively. GANs, initially getting widespread attention for their groundbreaking applications in image generation, have since also revolutionized the field of time-series. From their introduction by Goodfellow et al. (2014), GANs have quickly demonstrated their potential beyond image creation, becoming a powerful tool for DA across various domains.

This literature review aims to provide a comprehensive analysis of the application of GANs in DA for time-series. We begin with a foundational understanding of DA, exploring its purpose and importance in enhancing datasets. We then delve into GANs, examining their various adaptations and models. Each of these architectures offer unique features and benefits, tailored to specific types of data and application requirements.

The review extends its focus to investigate the application of GANs in several key domains: energy, traffic data, finance, audio (music and speech), and healthcare. Healthcare data often involves complex time-series, such as physiological signals (ECG[1], EEG[2], EMG[3], respiratory rate, etc.), some of them being multidimensional, which require sophisticated methods to learn the distribution and generate realistic synthetic data. This exploration is driven by the need to understand how GANs can effectively overcome the limitations inherent in time-series data in these areas. The review examines various GAN adaptations and models, including their features and benefits for specific types of data.

A deeper dive into healthcare developments in this context is conducted. After the decision to focus on healthcare data, this more targeted review was necessary to further understand the latest developments and challenges in this field of study. This deep dive

[1]Electrocardiogram
[2]Electroencephalogram
[3]Electromyography

provides insights into how GANs are being applied to address the specific needs and constraints of healthcare data.

2.1. Literature Review Methodology

This section outlines the systematic approach taken to review the literature on time-series DA. It describes the process of search, selecting and analyzing relevant studies, understanding the intricacies of GAN models and evaluation methods, and determining their applicability in various domains, in order to identify the optimal area of application for this dissertation.

2.1.1. Research Objectives

This literature review has two main goals. First, the motivation for this study comes from the growing importance of time-series analysis in various fields and the ongoing challenges of data scarcity and diversity in existing datasets of this nature. DA is crucial in addressing these challenges by increasing both the amount and variety of data in time-series datasets. Having large amounts of information is essential to effectively train NN models. Using enhanced datasets, researchers can improve the learning ability of the models, leading to better robustness and performance. Thus, this review will explore and investigate DA techniques specifically applied to time-series, examining their effectiveness and challenges in multiple domains.

Second, is to examine how effective the different architectures of GANs are in creating artificial data, and finding the area of application where this technique is going to be implemented and even improved. Other modern methods for generating data, like Variational Autoencoders (VAEs) (D. P. Kingma & Welling, 2013), will also be explored for comparison purposes.

2.1.2. Search Strategy

To initiate the literature review, a comprehensive understanding of DA techniques for time-series was necessary. This included exploring both modern and traditional approaches. The focus was on finding the most up-to-date articles, with a preference for any papers that provided a comprehensive summary of this particular field. Such a summary would serve as an ideal starting point for this review. A systematic review article by Iglesias et al. (2023) on the current state of DA techniques for time-series was recommended by the dissertation advisor. This review covers a wide range of techniques used to date, including traditional approaches and those based on VAEs and GANs.

Furthermore, this systematic review introduces evaluation metrics to compare and measure the performance of classification, forecasting, generation and anomaly detection models with DA in comparison to those without this technique. These metrics enable an objective assessment of the advantage provided by DA and the comparison of different approaches for certain application areas. This can also be valuable for comparing different DA techniques within the same application.

The review also mentions multiple application areas where the DA methods were used. Being those, audio processing and generation (Donahue et al., 2018), anomaly detection in ECG (G. Zhu et al., 2019), taxi traffic data (Huang et al., 2020), power demand forecasting (Ramponi et al., 2018), augmenting healthcare data like ECG and lung cancer events datasets for emotional classification (Chen et al., 2019) and event prediction (Yoon et al., 2019) respectively.

Following a broad investigation into DA for time series, the study narrowed its focus to articles that exclusively applied GAN models for time-series data generation. This focus was chosen because, following the initial review conclusions, GANs demonstrated superior results for this task compared to other architectures. While VAEs offer better control over the variability of generated data and are commonly used for anomaly detection, GANs are capable of much better generalization and can produce more diverse samples. However, they are also the most complex and difficult to train.

First, the articles using GANs for DA in time series referenced in the systematic review paper already mentioned were analyzed. Then, all relevant papers on GANs for DA in time series were gathered by conducting searches in databases like *Scopus*[4], *Science Direct*[5], *Research Gate*[6], *IEEE Xplore*[7], *Springer*[8], and *Google Scholar*[9]. This approach provides a comprehensive understanding of current developments in this specific field and helps identify research gaps that could lead to future studies. The following combination of words was applied in research:

(**"Generative Adversarial Networks"** OR **"GANs"**)
AND (**"Time Series"** OR **"Temporal Data"** OR **"Sequence Data"**)
AND (**"Data Augmentation"** OR **"DA"** OR **"Generation"** OR **"Imputation"** OR **"Synthetic"**)
AND (**"Review"** OR **"Systematic Review"** OR **"Survey"**)

This search strategy was applied across all mentioned research databases, with a time filter applied for the period from 2018 to 2024. The search yielded numerous papers, with some databases returning more results than others due to differences in advanced search capabilities. To manage the volume of results, in some cases, only the first 100 papers were screened. This approach was based on the assumption that the most relevant papers, which best matched the search criteria, would appear first in the search results. Beyond this point, the relevance of the papers to the research topic tended to decrease significantly.

[4]https://www.scopus.com/
[5]https://www.sciencedirect.com/
[6]https://www.researchgate.net/
[7]https://ieeexplore.ieee.org/
[8]https://link.springer.com/
[9]https://scholar.google.com/

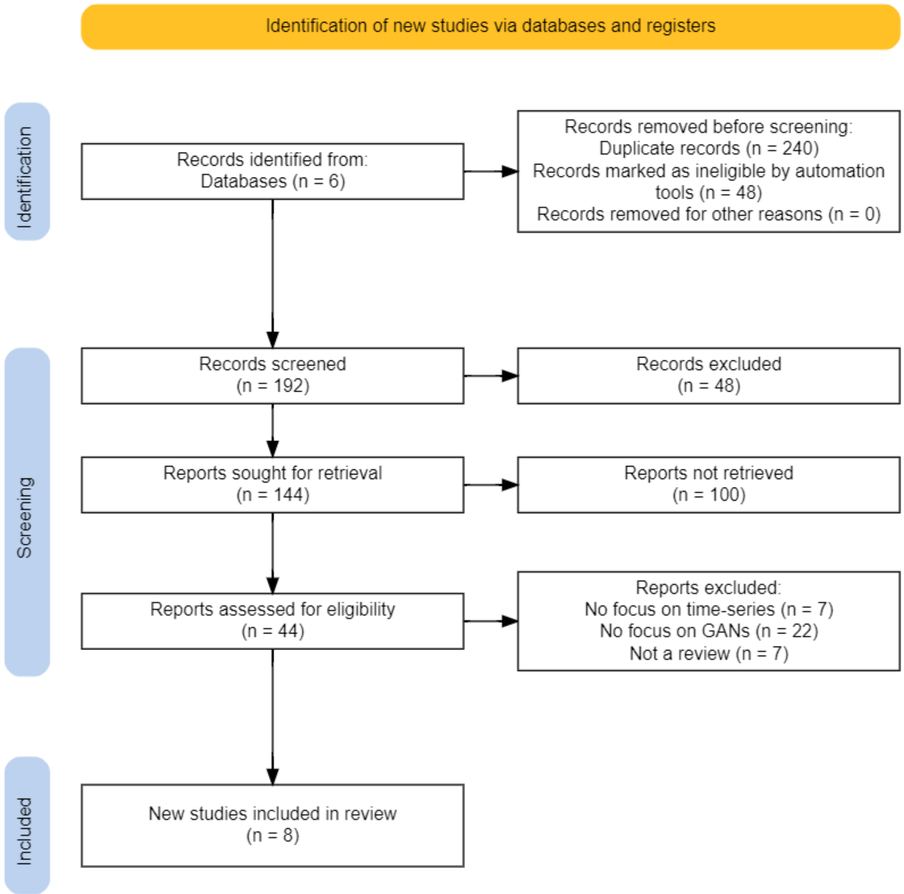

FIGURE 2.1. PRISMA flow diagram — GANs in time-series DA

After the initial screening, 144 papers remained. These papers focused on GANs, time-series, DA, or were systematic reviews related to Machine Learning. Further filtering was necessary to identify the most relevant papers for the research at hand. After analysis, three papers were selected as directly relevant to the search main focus. An additional five papers were identified for future review, although they addressed specific topics, which are not currently central to the study. The remaining papers were excluded for various reasons. Figure 2.1 shows the PRISMA[10] flow diagram for this search, created with the help of the PRISMA diagram tool by Haddaway et al. (2022).

After analyzing the selected papers, a comprehensive understanding of the current state of GANs applied to time-series data across various application areas and techniques was gained. This analysis will be elaborated upon in 2.4.

The study then focused on a more detailed investigation on the GAN models themselves. Some GAN technology focused articles were already gathered from the previously analyzed papers that mentioned multiple variations of GANs, especially applied to time-series, with some providing a brief history of this ML model and the evolution of this technology to address its initial challenges and limitations, others were sourced from review articles on the history and variations of GANs, searched in the same databases

[10] https://www.prisma-statement.org/

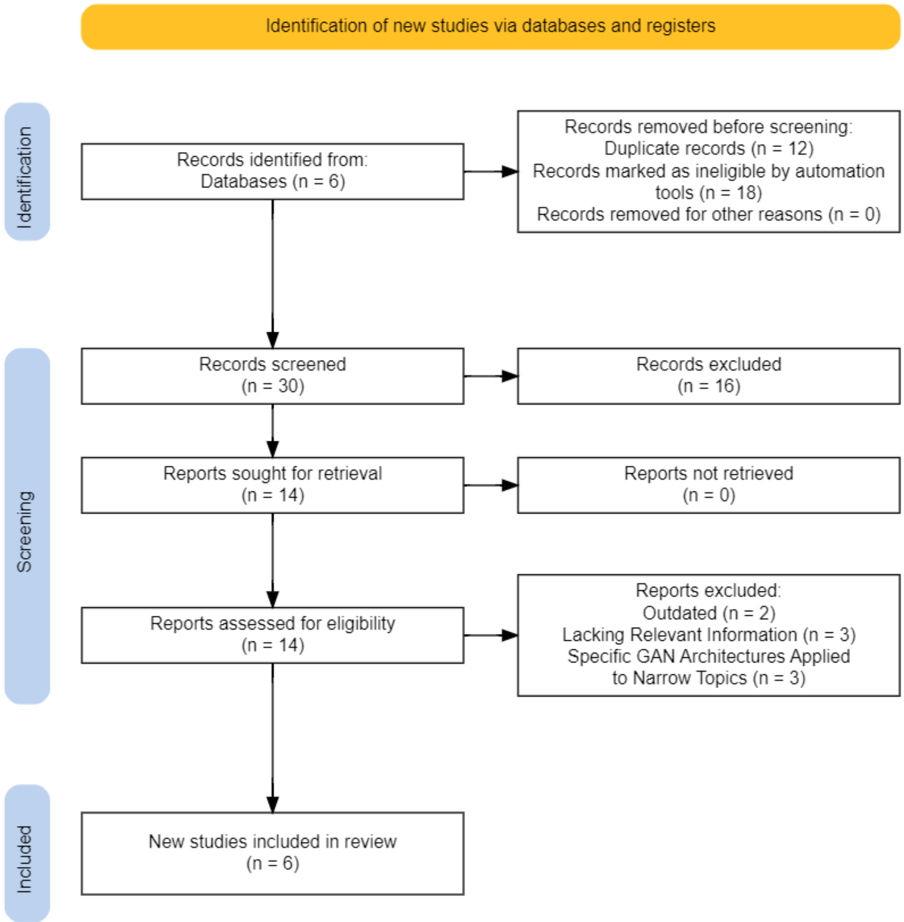

FIGURE 2.2. PRISMA flow diagram - GANs variations reviews

mentioned before. After an initial screening, only 6 out of 14 articles were selected for deeper analysis. The remaining were excluded for either being outdated, lacking relevant information, or focusing solely on specific GAN architectures applied to narrow topics. Figure 2.2 shows the PRISMA flow diagram for this search.

Gonog and Zhou (2019) discusses the basic theory, structure, advantages, and disadvantages of GANs, highlighting their ability to generate high-quality samples and approximate arbitrary probability distributions. Alqahtani et al. (2021) focuses on fundamental concepts, various architectures, and provides some information on applications in time-series areas like speech and music generation. Krichen (2023) covers architecture, loss functions, training challenges, applications, evaluation metrics, and future directions, emphasizing GANs' potential for generating high-quality synthetic data. Gui et al. (2023) explores algorithms and applications across various domains, including image synthesis, natural language processing, and sequential data tasks. Chakraborty et al. (2024) emphasises GANs' rapid development, recent advancements in architectures and training methods, and future research directions including privacy concerns and ethical considerations. Finally, Sharma et al. (2024) provides a taxonomy of GAN variants, discusses applications in fields like healthcare and computer vision, and explores potential solutions to improve GAN performance and stability. All these papers mentioned address common challenges in GAN technology, such as training instability and mode collapse, that will also be discussed in the section 2.3.

After gaining a solid understanding on the history and variations of these models, research was focused on five specific application areas in time-series, that were mentioned in the literature reviewed until this point: audio (music or speech generation), car traffic (anomaly detection), healthcare (anomaly detection, denoising and DA), finance (capturing long range dependencies) and energy (power demand DA) (Iglesias et al., 2023), (Brophy et al., 2023), (D. Zhang et al., 2022). A summary of this applications areas can be found in section 2.5.1.

The investigation then shifted its focus to healthcare data after making the decision to develop the model for this dissertation on this field of study, due to its critical impact on patient outcomes, and ability to address significant challenges in healthcare research stemming from data scarcity and privacy concerns. Specifically, the search concentrated on articles addressing health-related time series applications using GAN models. The following combinations of keywords were used in the previously mentioned research paper databases. Additionally, the *PubMed*[11] database was also included in this search, as it focuses on life sciences, biomedical, and healthcare-related topics—a relevant addition given the current focus drift.

> (**"generative adversarial networks"** OR **"GAN"** OR **"GANs"**)
> AND (**"time series"** OR **"temporal data"** OR **"sequential data"**)
> AND (**"healthcare"** OR **"medical"** OR **"clinical"** OR **"biomedical"**)

The search was also limited to papers published from 2018 to 2024. This timeframe was chosen to ensure that the most recent and relevant research in the field was captured. From the research datasets results, 33 papers were considered. After screening, 26 papers were selected for further analysis, comprising 4 reviews and 22 novel implementations. An additional 8 articles referenced in previously reviewed literature were also included, resulting in a total of 41 analyzed papers. Figure 2.3 shows the PRISMA diagram for this search.

Building on this focus, the investigation delved into the critical issue of privacy in time-series data generation, particularly in the medical field. Some datasets can be highly confidential, with access strictly controlled due to privacy issues or legal regulations. Synthetic data generated can help address these concerns by providing a means to share and publish data without compromising privacy, since the generated data is artificial. However, this synthetic data must also be protected against "membership inference attacks" (Monachino et al., 2023), where an attacker attempts to determine if an individual's data was used in the training set that led to the generated data. Differential privacy (Xie

[11] https://pubmed.ncbi.nlm.nih.gov/

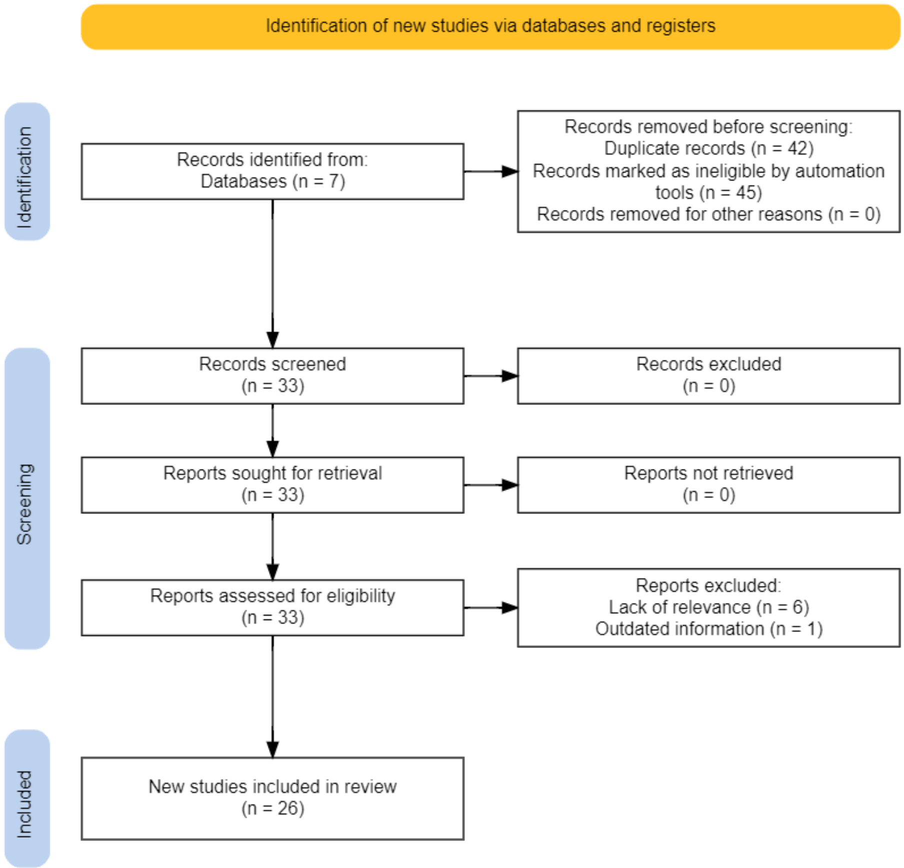

FIGURE 2.3. PRISMA flow diagram - GANs for time-series healthcare data

et al., 2018) and other privacy-preserving techniques can help mitigate this risk. Figure 2.4 provides a visual representation of this literature search progression.

2.2. Data Augmentation

Data Augmentation was originally used in image classification by LeCun et al. (1998), increasing image data with some techniques such as rotating, translating, scaling, adding noise, etc. Since then, DA is used to artificially enhance the size and variability of datasets in many applications. It is particularly useful in fields where data privacy and data scarcity is a significant challenge, such as healthcare (Hyland et al., 2017), (Y. Zhang et al., 2023) and time-series analysis. By enhancing datasets used to train predictive, classification or other types of models, DA improves both the robustness and accuracy of these models. Addressing challenges related to limited data availability, for multiple reasons like privacy or difficulty in gathering certain information.

Over the years, new DA approaches have emerged to enhance the quality of generated data and create entirely new information, rather than merely modifying existing data points. This evolution has led to innovative concepts in model engineering, enabling the generation of synthetic data. These advanced techniques are used to expand datasets, which in turn are utilized to training NN models. By providing more training data with

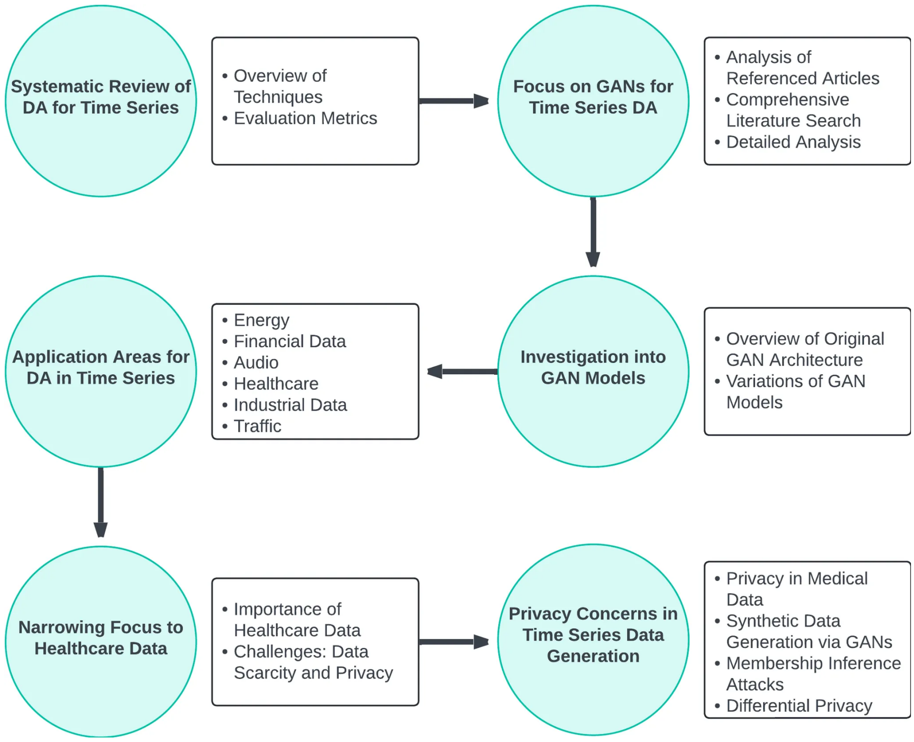

FIGURE 2.4. Flowchart of Search Strategy for Literature Review

greater variety, these models achieve improved performance and become more prepared to classify new data or completing any other task in a broader range of scenarios.

2.2.1. Modern Data Augmentation techniques in time-series

DA techniques for time-series have evolved significantly, utilizing advanced ML models to generate synthetic data that closely resemble real-world patterns. These techniques go beyond simple transformations, offering more sophisticated ways to enhance and expand datasets. Two prominent approaches in this category are VAEs and GANs, both of which have shown remarkable potential in creating high-quality synthetic time-series data (Brophy et al., 2023).

VAEs, introduced by D. P. Kingma and Welling (2013), are generative models that learn to encode data into a latent space and then decode it back, effectively learning the underlying distribution of the data. They consist of two main components, an encoder and a decoder (Iglesias et al., 2023):

- The **encoder** reduces the input data's dimensionality to a latent space representation.
- The **decoder** reconstructs the original data from this latent representation.

They have been successfully applied to time-series DA, as shown by Nazi et al. (2019) where a VAE is used to generate data for anomaly detection problems with LSTMs, and by Alawneh et al. (2021), in which they use a dataset augmented with VAE to improve the recognition of human activity also with LSTMs.

GANs, on the other hand, have gained significant attention due to their ability to generate highly realistic synthetic data. Originally proposed by Goodfellow et al. (2014) for image generation, GANs have since been adapted for time-series data with impressive results. Their adversarial training process allows them to capture complex temporal dependencies and produce diverse, high-quality synthetic time series. Continuous recurrent GAN (C-RNN-GAN) (Mogren, 2016) is one of the first GAN architectures proposed specifically for time-series, in particular to learn and synthesize classical music tracks.

Based on the review by Iglesias et al. (2023), there is no single most effective model for DA in time-series that applies universally. The effectiveness of a model depends on the specific application and characteristics of the data. Both GANs and Variational VAEs have shown promise in this field, each with their own strengths and challenges.

GANs are described as one of the main algorithms for DA, capable of producing more diverse samples compared to previous approaches. They are noted for their ability to learn the distribution of the data by extracting main features, without directly copying the distribution. This makes GANs particularly effective in generating high-quality, diverse synthetic data that closely resembles the original dataset distribution. However, they are also described as the most complex models available, with significant challenges in training and optimization.

VAEs, on the other hand, excel in their ability to learn a compact latent representation of the data, which can be useful for dimensionality reduction and feature extraction. They are often more stable during training compared to GANs and can provide a probabilistic interpretation of the latent space. This makes VAEs particularly useful in scenarios where interpretability of the generated data is crucial, or when the goal is to learn a meaningful representation of the data for downstream tasks. Although it may produce less sharp or detailed outputs compared to GANs.

The choice between these two approaches often depends on the specific requirements of the task at hand, the nature of the data, and the trade-offs between generation quality, training stability, and interpretability. Given the potential of GANs in generating diverse and high-quality synthetic time-series data, despite their complexity.

2.3. Generative Adversarial Networks

GANs have emerged as a powerful and versatile architecture of ML models, capable of generating high-quality synthetic data across various domains. This section provides an overview of GANs, exploring their history, structure, and the main challenges associated with their implementation and training.

2.3.1. History of GANs

The evolution of GANs began with the foundational **Standard GAN** (Goodfellow et al., 2014), the model that introduced the concept and laid the foundation for subsequent variants and was focused on computer vision tasks, like generating synthetic images. However, early GANs suffered from instability during training. The development of **Deep**

Convolutional GAN (DCGAN) (Radford et al., 2016) addressed this challenge by integrating convolutional layers to the NNs, improving the generation of realistic images. **Conditional GAN (cGAN)** (Mirza & Osindero, 2014), then brought in conditional control, allowing GANs to generate data based on specific labels, giving the user more control on the outputs.

In 2017, **Wasserstein GAN (WGAN)** (Arjovsky et al., 2017) revolutionized GAN training, by redefining the loss function with the Wasserstein distance, reducing training instability and resolving issues with vanishing gradients — a major disadvantage of GANs until this point. Building on this, **WGAN with Gradient Penalty (WGAN-GP)** (Gulrajani et al., 2017), further enhanced training stability with gradient regularization.

Until this point, GAN variants focused on improving the core architecture and training stability. Subsequent developments don't just improve the training performance but also explore novel applications and functionalities. **CycleGAN** (J. Y. Zhu et al., 2017) enabled image-to-image translation tasks even without the need for paired datasets. Paired datasets would normally contain matching images from two different domains, such as a photo of a landscape during day and its corresponding night version, or a horse image paired with a zebra image in the same posse. Transforming day images to night or horses to zebras with this model doesn't require this dataset structure, making it more practical to produce these outcomes. **ProGAN** (Karras et al., 2017) improves GAN generated image quality by progressively growing the layers in high-resolution outputs, a method later essential for achieving high realism synthetic images. **StyleGAN** (Karras et al., 2019) then introduced a new approach to generating highly detailed images by separating style and content, allowing the users precise control over features like hair and face shape for human synthetic portrait images for example, which was specially impactful in creative industries.

It's important to note that these models represent only a selection of the main GAN variants developed over time. Numerous other architectures and implementations exist and continue to be created, each with unique features and applications. Figure 2.5 represents the timeline of the main GANs variations and evolution.

2.3.2. GAN Architecture

The concept of a GAN is relatively straightforward but powerful. It consists of two key components, the *generator* and the *discriminator*, which are implemented as NNs.

- **Generator** (G): The generator's role is to produce synthetic data samples. It takes a random noise vector z as input and generates synthetic data samples $G(z)$ that resembles the real data distribution. The generator continuously improves its ability to create data as it learns through the adversarial process.
- **Discriminator** (D): The discriminator acts as a critic or judge. It receives either real data samples x from the training dataset or synthetic samples $G(z)$ from the generator. Its task is to distinguish between the two, identifying whether a given sample is real or fake.

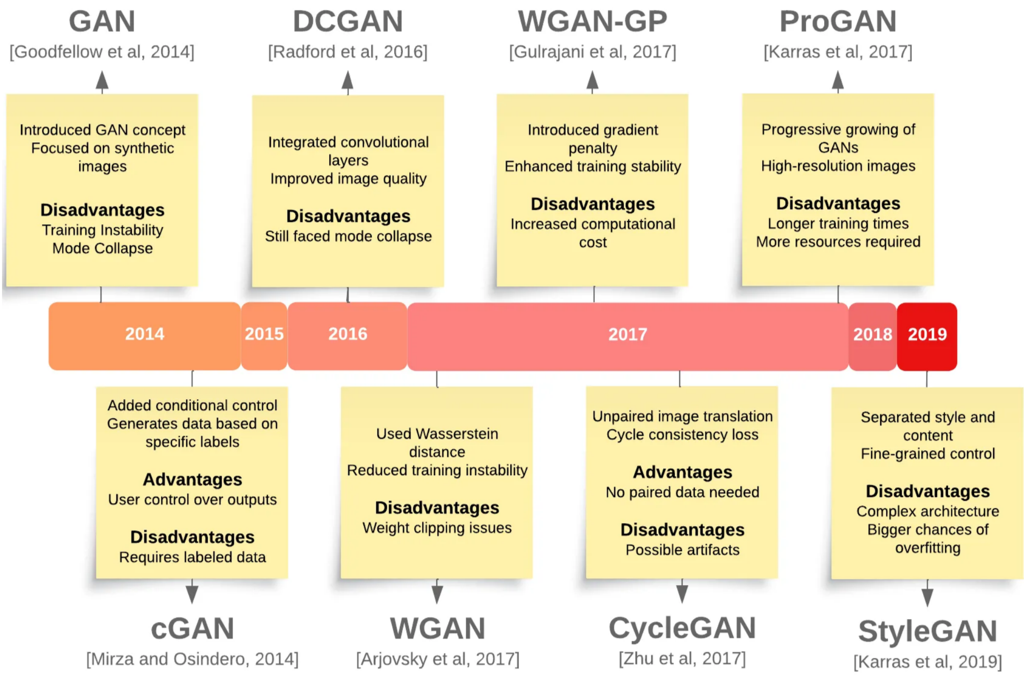

FIGURE 2.5. Timeline of GANs evolution

The training of a standard GAN is a dynamic process that involves both the generator and discriminator improving their performances simultaneously through a min-max game, described in equation (2.1), representing the value function:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.1}$$

where G stands for the generator network and D for the discriminator network. $V(D, G)$ is the value function to be minimized by G and maximized by D. x represents the real data samples and $p_{\text{data}}(x)$ the real data distribution, while z is the random noise input to the generator, being $p_z(z)$ the input noise distribution. $D(x)$ is the discriminator's output, representing the probability that x is real and $G(z)$ is the generator's output representing the synthetic data.

The generator aims to produce data so realistic that the discriminator cannot distinguish it from actual data. Conversely, the discriminator strives to become better at identifying the synthetic data. The discriminator provides feedback to the generator about how distinguishable its output is from real data. This feedback is used by the generator to improve its data generation. Both networks are updated through backpropagation given the loss functions outputs, in order to improve their respective performances. This process continues iteratively. Ideally, this competition results in a generator that produces data virtually indistinguishable from real data (Chakraborty et al., 2024).

2.3.3. Main Challenges of GANs

Although GANs are innovative models and have gained considerable attention from the scientific community, they also present challenges. These can significantly impact the performance and reliability of GANs in various applications. Some of the main issues include:

- **Training Instability**: GANs can be difficult to train, often causing the generator and discriminator losses to go up and down or move away from the desired result.
  - Krichen (2023) analyses different approaches to address this issue including adjusting learning rates, using different optimization algorithms, adding noise to the training process, and implementing more advanced architectures like Wasserstein GANs.
- **Mode Collapse**: Occurs when the generator produces only a limited set of outputs, therefore lacking diversity on the generated data. Mode collapse can happen when the discriminator becomes too good at spotting fake data. This causes the generator to make very similar samples that can trick the discriminator, resulting in those samples being either identical or showing minimal variation.
  - Krichen (2023) also explores some solutions include adding regularization terms to the loss function, modifying network architectures, and using advanced optimization methods.
- **Vanishing Gradients**: This problem occurs when the discriminator becomes too successful, preventing the generator from learning (Brophy et al., 2023).
  - Solutions include modifying the architecture or the loss function, these are reviewed in more detail by Wang et al. (2021).

The lack of consensus in evaluation metrics is also one of the main challenges of GANs. There is no widely accepted metric to evaluate the performance of these models, particularly for time series data. This makes it difficult to compare different implementations objectively. As this is a model more often applied in the imaging area, where initially, most of the tests to evaluate the models' performance were carried out by the human eye. Sharma et al. (2024) describes some essential performance measures for GANs, along with their mathematical basis.

Despite these challenges, the growing relevance of GANs in various areas and their rapid performance improvement motivate continued research into their specific application in DA for time-series. The main reason in studying GANs for time-series data is to develop a model that generates synthetic data that not only preserves the complex temporal dynamics of the original dataset but also outperforms or provides advantages over existing implementations.

2.4. Generative Adversarial Networks in Time-Series

Despite being originally designed for image generation, GANs have also been adapted to capture the complex temporal dynamics of time-series. To adapt the GAN architecture, some modifications are typically made. The use of Recurrent Neural Networks (RNNs) both in the generator and discriminator, particularly LSTMs allows the model to capture temporal dependencies in the data (Iglesias et al., 2023).

LSTMs have a unique structure that includes a memory cell, input gate, output gate, and forget gate. This architecture allows the model to selectively remember or forget information over long sequences, making it particularly well-suited for time-series. The memory cell can maintain information over long periods, while the gates control the flow of information into and out of the cell. This combination results in GANs that can generate more realistic and temporally consistent time-series data.

Models like **TS-GAN** (Yang et al., 2023) also incorporate attention mechanisms to better capture the relationship between dimensions at each time step, in order to understand how different features or measurements are related to each other at any given moment in time.

Some approaches, like **TimeGAN** (Yoon et al., 2019), incorporate Autoencoders (AEs) to enhance the model's capabilities. AEs help in two key ways:

- **Dimensionality Reduction**: They compress the input data into a lower-dimensional latent space, making it easier for the model to learn and process complex time-series data.
- **Feature Extraction**: By learning to reconstruct the original data from this compressed representation, AEs help capture the underlying structure and important features of the time-series.

This combination allows more accurate and realistic synthetic time-series generation (Gatta et al., 2022).

2.4.1. History of GANs in Time-Series

The evolution of this architecture for sequential data has been marked by some milestones that address the unique challenges of temporal information. This is only a brief historic overview of the use of GANs in time-series. Only a few examples are mentioned here, representing the main developments through the years, but many more exist.

SeqGAN (Yu et al., 2017) was among the first to extend GANs to sequence data by using policy gradients and Monte Carlo Search to handle discrete sequential outputs. The generator uses RNNs with LSTM cells, and the discriminator employs a Convolutional Neural Network (CNN). This model is applied to text generation, achieving realistic and diverse output by maximizing a reinforcement learning goal. However, SeqGAN faced difficulties in capturing long-term dependencies and suffered from issues like *mode collapse*. To overcome these limitations, **RGAN (Recurrent GAN)** (Hyland et al., 2017) integrated RNNs into both the generator and discriminator and focused on continuous outputs, since it is possible to effectively model time-dependent data without resorting to reinforcement learning. This made the training process more straightforward and improved the ability to capture long-term dependencies. **WaveGAN** (Donahue et al., 2018) made advancements in the field by generating raw audio waveforms, showing versatility in

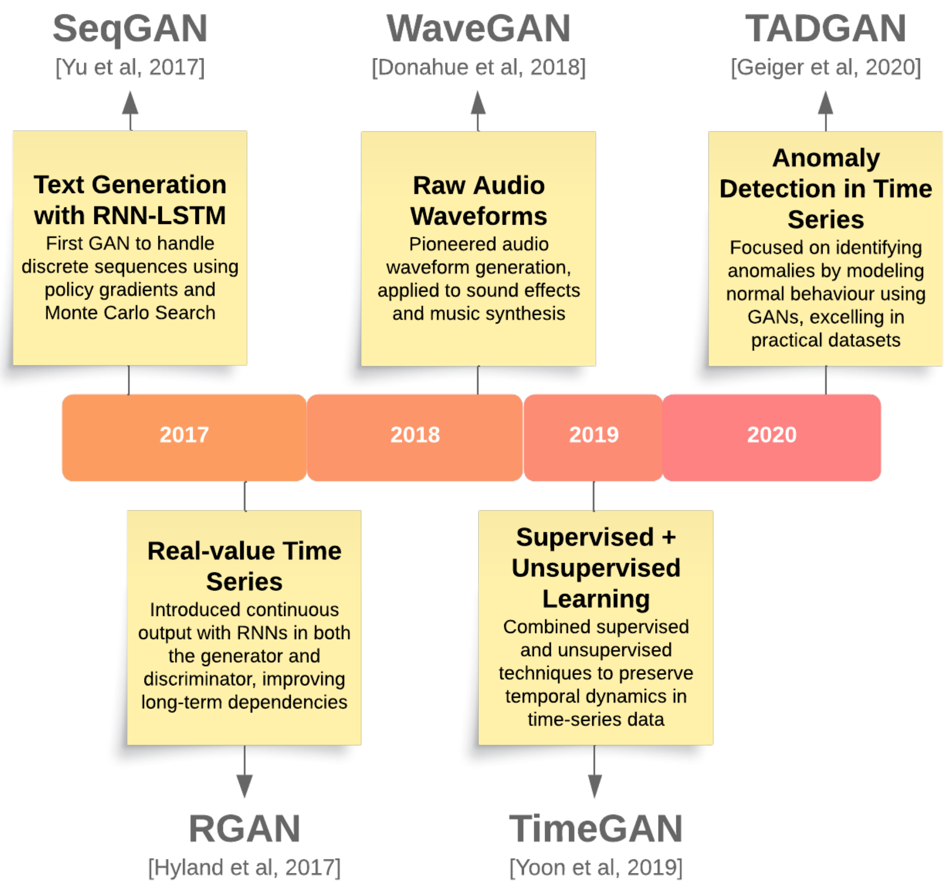

FIGURE 2.6. Timeline of Time-Series GANs evolution

various applications, notably in creating short sound effects and also developing a procedural drum machine for electronic musicians. Despite these improvements, earlier models still struggled with preserving the temporal dependencies in certain types of time-series data. **TimeGAN** (Yoon et al., 2019) addressed this challenge by combining supervised and unsupervised learning within the GAN framework, to effectively generate time-series data with preserved temporal dynamics. The authors of ClinicalGAN (Chandra et al., 2024) mention TimeGAN as a state-of-the-art approach for multi-variate time-series generation and uses it as the foundation for their model. Finally, **TADGAN** (Geiger et al., 2020) focused on anomaly detection within time-series, using GANs to model normal behavior and identify deviations, thus solving practical challenges that previous models couldn't specifically target. It was tested on 11 datasets from various sources including NASA, Yahoo, Numenta, Amazon and Twitter, outperforming baseline methods in 6 of those datasets, while also achieving the highest averaged F1 score across all datasets. Figure 2.6 shows the timeline evolution with the mentioned time-series GANs.

2.5. Results and Discussion

In this section, the outcomes of the comprehensive literature review and analysis are expressed, with a particular lens on healthcare. Here, the insights gathered from various studies are summarized, comparing the effectiveness, challenges, and potential of GANs across different applications. The transformative impact of GAN models in healthcare is evaluated, discussing the significant improvements in data quality and model performance they offer in this domain, compared to the other explored areas.

2.5.1. Application Areas

The use of DA in time series, using GANs, has proven to be a very useful tool in many areas of study. This technique has been explored in diverse contexts, such as energy data, car traffic, finance, audio, and healthcare.

In each of these areas, a variety of studies have already adopted this approach, using different variations of GANs to increase the size of the datasets. Each of these domains presents unique challenges in development. Two studies will be presented for each category. These will show the specific methods used, the problems faced, and the important discoveries made in each area.

By the end of this section, a comparison of the different areas of application will be conducted, and the most suitable area for further research will be determined.

2.5.1.1. *Energy Data:* C. Zhang et al. (2018) introduces an innovative method for generating synthetic data for smart grids using GANs. This research tackles the challenge of limited availability and access to detailed data in smart grids, especially at the distribution system level. These restrictions are usually due to issues with gathering data, ensuring security, and protecting privacy. The main goal of the study is to generate artificial datasets that closely resemble real smart grid time-series data. The team utilizes the Pecan Street Dataset for their experiments, which comprises energy consumption and solar generation data from 25 users with photovoltaic panels, gathered from 2013 to 2016. The results reveal that the artificial datasets produced by their GAN model are statistically indistinguishable from the real datasets in various ML tasks and statistical tests. This shows that synthetic data generated by GANs can help solve data availability problems in smart grid research. It enables the creation of new models based on data while keeping the actual datasets confidential by not using them for training.

In Demir et al. (2021), they focus on improving the accuracy of forecasting electricity market prices by employing advanced DA techniques. This research shows that the size of a training dataset affects how well a model can predict new data, especially when dealing with multiple variables over time. Traditional DA methods become too complex to employ in such cases and don't provide good results. A key challenge identified is the constraints imposed by limited sample sizes, which can potentially lead to overfitting in regression models. The researchers propose using AEs, VAEs, and WGAN-GPs (Gulrajani et al., 2017) to generate new data. The study looks at predicting prices for the next day in the electricity markets of Belgium and the Netherlands. It uses data from January 1, 2016, to December 31, 2018. The findings suggest that the application of AEs, VAEs, and WGAN-GPs significantly enhances regression accuracies, with these methods decreasing the mean absolute errors of benchmark models by 2.23%, 2.73%, and 2.97%, respectively. Notably, the GAN approach exhibited the most substantial improvement in the study.

2.5.1.2. *Car Traffic Data:* **TSDIGAN (Traffic Sensor Data Imputation GAN)** (Huang et al., 2020) focuses on developing a DCGAN framework to address missing data in traffic sensor systems. Traffic data is critical for intelligent transportation applications such as congestion prediction and traffic flow analysis, most recently used in apps like Uber to predict the duration of a trip, but sensor malfunctions and communication issues often lead to incomplete datasets. The proposed framework uses GANs to generate realistic synthetic traffic data to fill in these missing gaps, employing a novel technique called the Gramian Angular Summation Field (GASF) to convert traffic data into image-like structures, so, in this approach, the GAN architecture doesn't require significant changes, such as incorporating LSTMs to handle raw time-series data, since the GAN effectively trains on image-like representations. The model is then tested on the Caltrans Performance Management System (PeMS) dataset, showing that it performs well even with high rates of missing data, demonstrating robustness and efficiency in traffic data reconstruction.

G. Zhu et al. (2019) presents an LSTM-GAN algorithm for anomaly detection in time-series, focusing on handling non-stationary datasets that exhibit significant temporal fluctuations. The approach is applied to two primary datasets: traffic data from the New York City (NYC) taxi dataset, representing passenger flow anomalies during events like holidays and snowstorms, and ECG data for detecting heartbeat anomalies. During training, only normal (non-anomalous) data is used. This approach allows the generator to learn the characteristics of normal data. In the testing phase, both normal and anomalous data are passed through the model. If the generator fails to accurately reconstruct the input, it means the data point is likely an anomaly. The results demonstrate superior performance compared to traditional algorithms like isolation forest and one-class support vector machines, detecting traffic anomalies with greater accuracy. The framework is also promising for applications in system health monitoring, an area of application that will also be explored in this section.

2.5.1.3. *Financial Data:* Liao et al. (2020) introduces an innovative conditional Sig-WGAN framework. This framework effectively integrates WGANs (Arjovsky et al., 2017) with a sophisticated and efficient path feature extraction method known as the "signature of a path". A significant advancement is the conditional Sig-W1 metric, created to capture the conditional joint law of time-series models, which acts as a powerful discriminator. This method makes it easier to represent the suggested discriminators, which greatly reduces the requirement for expensive training procedures. To test this method, there were used various datasets, including the SPX and DJI index datasets, along with the Bitcoin-USD dataset. The results were impressive, demonstrating that the conditional Sig-WGAN framework consistently performs better than current benchmarks in terms of similarity and predictive accuracy.

The study by Yoon et al. (2019), known as **TimeGAN**, presents a novel framework for generating realistic time-series data. Its importance lies in its capability to preserve the complex temporal patterns found in time-series, a task that was quite challenging.

TimeGAN uniquely combines the unsupervised nature of the GAN framework with the precision of supervised training, resulting in a versatile and effective model. **A critical aspect of TimeGAN's evaluation is its performance on a variety of real and synthetic datasets**, particularly in the financial domain. The model was evaluated using daily historical Google stock data from 2004 to 2019, which included different aspects like volume, high and low prices, opening and closing prices, and adjusted closing prices. TimeGAN's proficiency in generating sequences that accurately reflect the complex, aperiodic, and correlated nature of financial data underscores its value. It shows an impressive ability to generate sequences that maintain the original relationships between variables over time, making it a powerful tool for financial analysis and modelling.

2.5.1.4. *Audio Data:* Mogren (2016) presents an innovative generative adversarial model designed for continuous sequential data, which represents a significant departure from conventional approaches in sequence modeling. Prior models, particularly RNNs, were predominantly focused on predicting the next token in a sequence, such as in language processing. However, Mogren's model represents a breakthrough by being capable of processing fully continuous sequence data. This includes elements of music such as tone lengths, frequencies, intensities, and timing. The model's ability to handle these continuous data elements allows for a more sophisticated and accurate representation of music sequences, surpassing previous models that depended on symbolic representations. The model was trained and evaluated on a dataset consisting of 3,697 MIDI files from 160 different composers, encompassing a wide range of classical music pieces. Each tone in the dataset was represented by its corresponding sound frequency, and the data was normalized to a tick resolution of 384 per quarter note. The study concluded that the model's capability to generate music improved progressively with further training, showcasing its potential in music synthesis.

Donahue et al. (2018) presented **WaveGAN**, a significant innovation in audio generation technology. This model tackles the difficult task of creating audio signals, which requires capturing detailed structures over different time periods. WaveGAN has been shown to effectively synthesize one-second slices of audio waveforms, making it suitable for sound effect generation in diverse applications. WaveGAN takes a different approach by flattening the DCGAN architecture to work in a single dimension, thereby preserving the same scale of parameters and operations as its two-dimensional counterpart. The versatility of WaveGAN is evident in its wide range of potential applications, notably in creating short sound effects for the music and film industries. The researchers also trained WaveGAN on drum sounds, leading to the development of a procedural drum machine intended to aid electronic musicians. Furthermore, the model was also tested through a speech benchmark, allowing for straightforward assessment by human annotators.

2.5.1.5. *Medical Data:* Y. Zhang et al. (2023) introduces an innovative approach with the WGAN-GP algorithm, specifically designed for generating radiomics data, a one-dimensional dataset extracted from radiotherapy (RT) and computed tomography (CT) images, crucial in modern medical analysis. Zhang's work stands out for its ability to create synthetic medical data samples that replicate the distribution of real data. This is particularly vital in addressing the challenges associated with collecting annotated medical data, where privacy, ethical considerations, and the rarity of certain medical conditions pose significant hurdles. The study employed two key datasets for validation: the widely-used public Heart Disease Cleveland dataset, known for its role in heart disease prediction, and a private dataset focusing on Radiation Pneumonitis. The former contains 76 attributes, with 14 being primarily used in the experiments to assess heart disease risk in patients. The latter includes data from 300 patients, with radiomics features from CT images and radiotherapy planning dose files. The research demonstrated that the WGAN-GP model improves classification performance, specifically in terms of Area Under the ROC Curve (AUC) and Sensitivity (SEN), outperforming traditional methods like Synthetic Minority Oversampling Technique (SMOTE) and common GANs, especially in scenarios with limited samples.

Hyland et al. (2017) study proposes two models: RGAN and RCGAN. These models implement RNNs in both the generator and discriminator components. The RCGAN, in particular, conditions these RNNs on auxiliary information, enhancing the specificity and context of the generated data. This approach proved effective in creating realistic data suitable for supervised training, with minimal loss in performance compared to real test data. The medical dataset had data from the Philips eICU database, including about 200,000 patients from 208 care units across the United States, holding over 224 million entries across 33 tables. The potential of these models in the Intensive Care Unit (ICU) context is significant, where they could assist doctors in making rapid and critical decisions, a common challenge in such high-pressure environments.

2.5.2. Overview

Each domain benefits uniquely from this technology, but when evaluating the importance and urgency of these applications, with the objective of finding the area to study in the dissertation, medical data stands out as the most critical area, and the most interesting one to delve into. As seen in studies by Y. Zhang et al. (2023) and Hyland et al. (2017), GANs can generate medical datasets that maintain the complexity and variability of real patient data. This is vital in a field where data scarcity and/or privacy concerns pose significant obstacles to research and training. The ability to generate realistic, diverse medical data can revolutionize areas like disease diagnosis, treatment planning, and medical training. In scenarios where rapid and accurate decision-making can be life-saving, such as in ICUs, this technology can have real practical impact. It directly contributes to improving patient outcomes and probably saving lives, addressing some of the most pressing challenges in healthcare today.

In conclusion, while GANs' DA is transformative across various sectors, its application in medical data is arguably the most consequential. The ability to generate accurate, diverse, and specially privacy-compliant medical data addresses a fundamental challenge in healthcare, offering the potential to improve patient care and outcomes, making it arguably the most critical and impactful application of this technology. Consequently, the focus now turns to a more in-depth investigation of how GANs can be applied specifically in medical time-series data, and what different approaches already exist in this field of study. Tables 2.1 and 2.2 present the summary of the analyzed papers for the areas of application comparison.

Study/Author	**Focus**	**GAN Variant**	**Type of Data**
(C. Zhang et al., 2018)	Smart Grids	Standard GAN	Energy consumption data
(Demir et al., 2021)	Electricity Market Forecasting	AEs, VAEs, WGAN-GPs	Energy prices data
(Huang et al., 2020)	Traffic Data Imputation	DCGAN (TSDI-GAN)	Traffic sensor data
(G. Zhu et al., 2019)	Anomaly Detection in Time Series	LSTM-GAN	Traffic, ECG data
(Liao et al., 2020)	Time Series Data	Sig-WGAN	Financial indices, Bitcoin
(Yoon et al., 2019)	Time-Series Data	TimeGAN	Financial stocks data
(Mogren, 2016)	Audio Data	Continuous GAN	Music (MIDI)
(Donahue et al., 2018)	Audio Synthesis	WaveGAN, Spec-GAN	Audio waveforms
(Y. Zhang et al., 2023)	Medical Data	WGAN-GP	Radiomics data
(Hyland et al., 2017)	Medical Time-Series Data	RGAN, RC-GAN	ICU patient data

Table 2.1. Comparative Analysis of GAN Variants in Different Application Areas (Part 1)

2.6. Literature Review Conclusions

This literature review underscores the significant potential and challenges of applying GANs to increase datasets volume and variability. Starting from understanding the fundamentals of DA, from the traditional algorithms to modern ML techniques, such as VAEs and GANs. The review then delved into the foundational aspects of GANs to the more refined variations like DCGAN, cGAN, WGAN, and WGAN-GP, and also the variations created specifically to handle time-series data like SeqGAN, RGAN, WaveGAN, TimeGAN and TADGAN. It is evident that GANs offer a versatile and powerful tool for enhancing dataset quality and diversity, essential for effective model training and analysis.

Furthermore, the comparative analysis of GAN variants across different application areas reveals the versatility of these models. From energy and traffic data to financial and audio applications, GANs have proven their adaptability. However, their application in healthcare stands out due to its potential to directly impact patient outcomes and address critical data scarcity issues in medical research.

In healthcare, models like RGAN, RCGAN for example, have shown promising results in generating realistic ICU patient data. Their ability to generate realistic, synthetic data

Study/Author	Key Applications	Notable Findings
(C. Zhang et al., 2018)	Synthetic data generation for smart grids	Synthetic datasets indistinguishable from real datasets
(Demir et al., 2021)	Forecasting electricity market prices	Enhanced regression accuracies
(Huang et al., 2020)	Traffic data imputation	Robust traffic data reconstruction even with high missing data rates
(G. Zhu et al., 2019)	Anomaly detection in traffic and ECG data	Detected anomalies with superior accuracy compared to traditional methods
(Liao et al., 2020)	Financial data analysis	Outperformed benchmarks in similarity and predictability
(Yoon et al., 2019)	Financial market analysis	Realistic sequences respecting temporal dynamics
(Mogren, 2016)	Music generation	Improved music generation with training
(Donahue et al., 2018)	Sound effect generation, electronic music	Coherent audio waveform synthesis
(Y. Zhang et al., 2023)	Medical DA, disease classification	Improved classification performance
(Hyland et al., 2017)	Early warning systems in ICUs, medical training	Effective in generating realistic time-series data

Table 2.2. Comparative Analysis of GAN Variants in Different Application Areas (Part 2)

could revolutionize aspects of patient care, from diagnosis to treatment planning or even medical training. However, the journey towards fully realizing this potential is not without obstacles. Issues such as mode collapse, training instability, vanishing gradients, and the development of robust evaluation metrics remain significant challenges. Additionally, ethical considerations, especially around the generation and use of synthetic healthcare data, demand careful attention to ensure privacy, and trustworthiness.

Looking forward, the field of GAN-based DA in healthcare presents opportunities for further research. Areas of focus might include improving the interpretability of generated data, developing more robust evaluation metrics specific to medical applications, and exploring ways to integrate domain expertise more easily into the model development.

CHAPTER 3

Methodology

3.1. Database and Exploratory Data Analysis

The MIT-BIH Arrhythmia Database (Goldberger et al., 2000) is the database used in this study for the creation of synthetic heartbeats using a GAN. It was created to evaluate arrhythmia detection algorithms. It contains 48 half-hour, two-channel ECG recordings from 47 subjects, captured between 1975-1979. It includes both common and rare arrhythmias, with 25 recordings including less common but clinically significant arrhythmias that would not be well-represented in a small random sample, while the other 23 primarily consists of more typical heart rhythms. The records are digitized at 360 samples per second (360 Hz) over a 10 mV range. Cardiologists have annotated each recording, enhancing its value for building a classification model capable of detecting heartbeat types.

An **ECG** is used to record the electrical activity of the heart from different angles. ECGs are recorded by placing electrodes on a patient. An **ECG lead** is a graphical representation of the heart's electrical activity, generated from several electrodes. This study will use the MLII (Modified Limb II) lead from the ECG signals.[1]

The normal ECG heartbeat signal has distinct characteristics identifiable in a graph visualization. Figure 3.7 displays a segment of a normal heartbeat from the database, exemplifying the main components present in a typical ECG heartbeat representation.

Arrhythmia is defined as an abnormal rhythm of the heart, where the heart may beat too fast (tachycardia), too slow (bradycardia), or irregularly, disrupting the heart's ability to pump blood efficiently. (J. Kingma et al., 2023)

Figure 3.8 represents a segment of the first 10 seconds of an ECG signal from the database. The amplitude in the signals is not always the same, varying from record to record and even from within the same record.

Each record in the database has detailed annotations. These annotations provide users with deeper insights into signal behavior and various heartbeat types. Figure 3.9 demonstrates annotations for a 10-second ECG segment. Each annotation in the image represents a segment with normal beats (N), premature ventricular contractions (V) and rhythm variations (+), though various other annotations exist for different heartbeat types and rhythm variations (possible Arrhythmias). The figure also illustrates the signal's amplitude variation, with some instances showing lower or higher amplitudes.

[1]https://geekymedics.com/understanding-an-ecg/

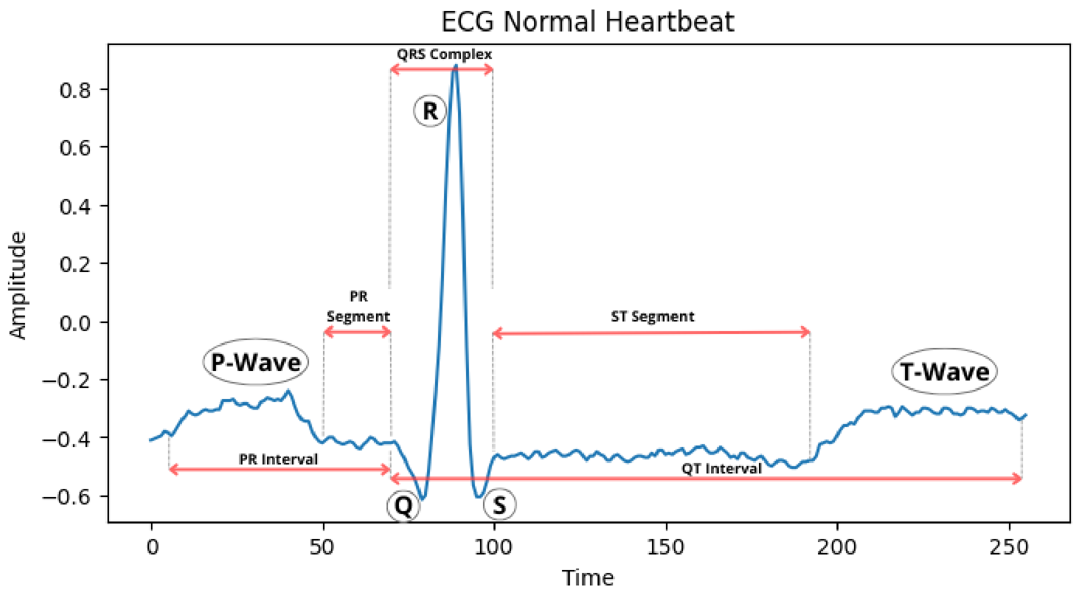

FIGURE 3.7. Representation of the different parts of a normal heartbeat represented through an ECG

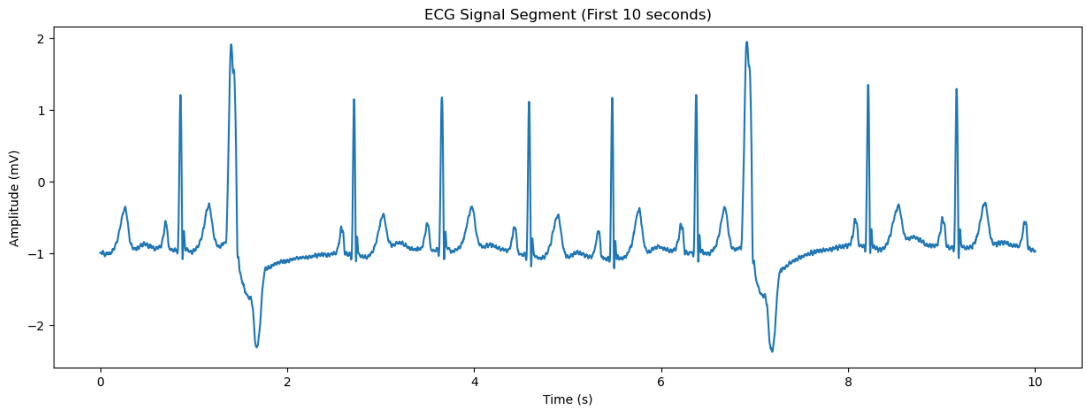

FIGURE 3.8. 10-second segment of MLII ECG signal

Each 30-minute record allows for the calculation of instantaneous heart rate from the signal and the calculation of average heart rate for any record length. This is achieved by measuring the time intervals between consecutive heartbeats (RR Intervals). The instantaneous heart rate is calculated as the reciprocal of RR intervals multiplied by 60 to convert to Beats per minute (bpm). Figure 3.10 shows the instantaneous heart rate over a full 30-minute record. The mean heart rate in this record is approximately 75 bpm, occasionally dipping below 60 bpm and rising above 100 bpm—the range typically considered healthy for adults (Mason et al., 2007).

Like mentioned before, the MIT-BIH Arrhythmia database is composed of two main groups, 23 recordings with normal rhythms and 25 with more uncommon rhythms and heartbeats. By analyzing the heart rate distributions across records from both groups, we can discern key differences between these two sets of recordings. Figure A.21 available in the appendix illustrates the heart rate distribution of the records from common signals,

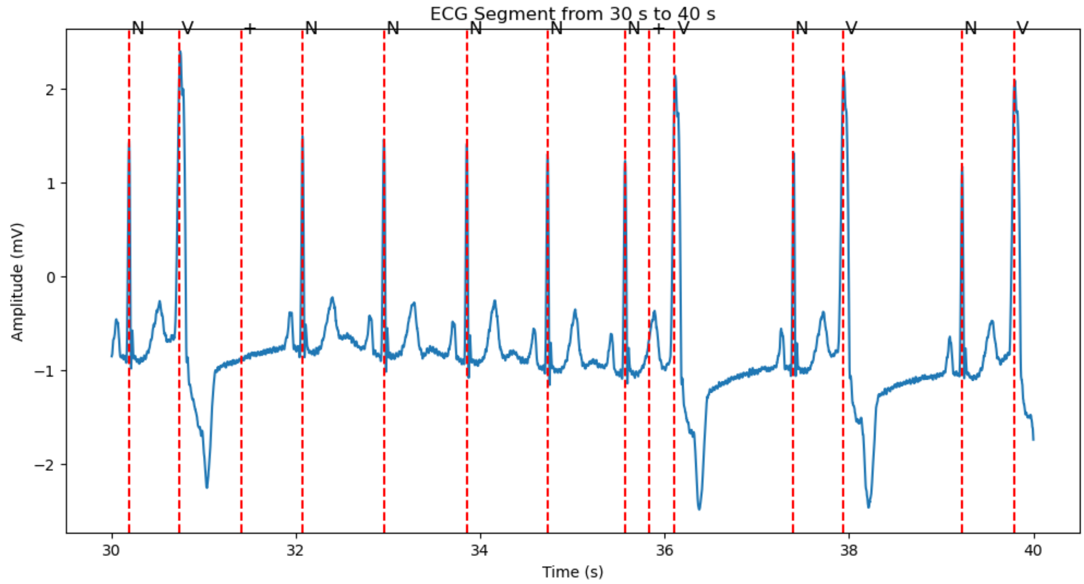

FIGURE 3.9. 10-second segment of MLII ECG Signal with annotations

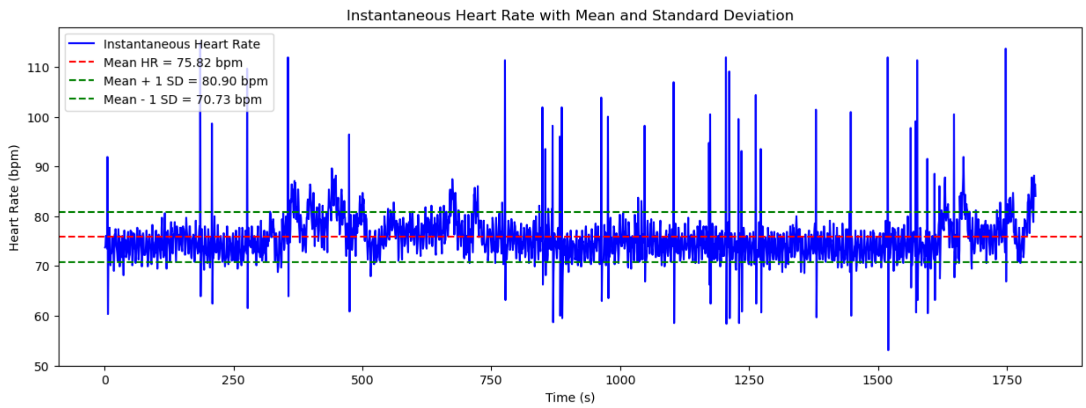

FIGURE 3.10. Instantaneous Heart Rate of a 30-minute ECG with mean and standard deviation calculations

while Figure A.22 also available in the appendix depicts the distribution for uncommon signals. The common group predominantly exhibits heart rates within the healthy range of 60–100 bpm, with some instances dipping to around 40 bpm and others reaching approximately 140 bpm. In contrast, the uncommon group displays notably higher heart rates in most records, with some exceeding 150 bpm. Figures A.23 and A.24 present in the appendix, show some distributions of heart beats in some records.

3.2. Data Preprocessing

Four records in the database were generated from pacemakers: numbers 102, 104, 107, and 217. In these records, the normal beats are actually considered paced beats, with the corresponding annotation symbol "/". Golany et al. (2020) removed the paced beats from their dataset, citing this as a recommended practice by the Association for the Advancement of Medical Instrumentation (AAMI). A pacemaker is a small, battery-powered device that prevents the heart from beating too slowly. These recordings are

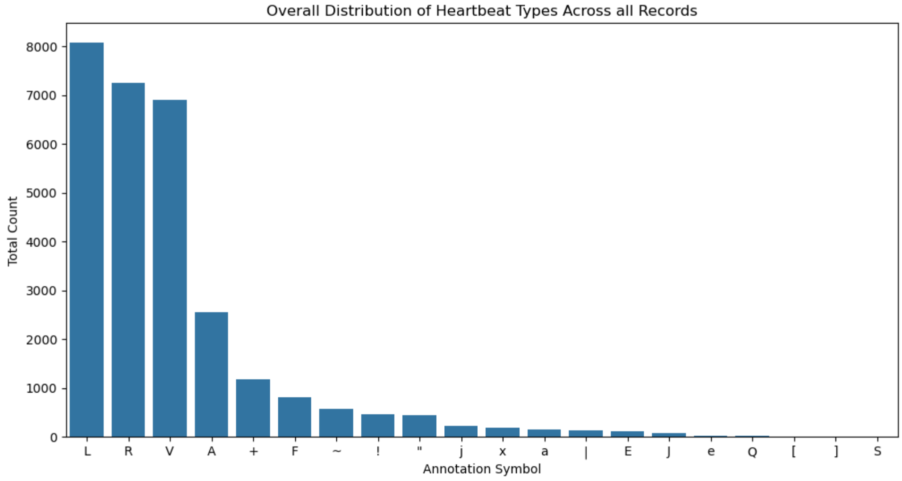

FIGURE 3.11. Overall Distribution of Heartbeat Types Across all Records

removed from further examination because they are inconsistent with the rest of the database. Including them could introduce uncertainty to the model when training on signals from these records alongside the rest of the database. Therefore, these recordings were also removed from this study.

To select heartbeat types for GAN training with the objective of augmentation, understanding the frequency of each type in the database was needed. A visualization was created to display the quantities in descending order, excluding type N (the most common) due to its overwhelming prevalence with 75052 occurrences. Figure 3.11 presents a bar graph showing the number of samples for each symbol across all recordings.

Looking at the graph, besides N, there are 3 main annotations present in the database, L (8075), R(7259) and V(6903). Beyond these top three annotations, the frequency drops significantly for all other types. In (Yang et al., 2021), besides these main 4 heartbeats, the authors also include type A (2546) segments in training. As the GAN model's purpose is to augment data, it should be effective even with limited input data, producing high-quality synthetic heartbeats for future use. Therefore, this less frequent heartbeat type is also included in the study. Here is a description of each of the mention heartbeat types [2]:

- **Type N (Normal Beat)**: Signal segment representing a normal heartbeat.
- **Type L (Left bundle branch block beat)**: Occurs when something blocks or disrupts the electrical impulse that causes the heart to beat.
- **Type R (Right bundle branch block beat)**: An obstacle in the right bundle branch that makes the heartbeat signal late and out of sync with the left bundle branch.
- **Type V (Premature ventricular contraction)**: Extra heartbeats that begin in one of the heart's two lower pumping chambers (ventricles)

[2]https://archive.physionet.org/physiobank/annotations.shtml

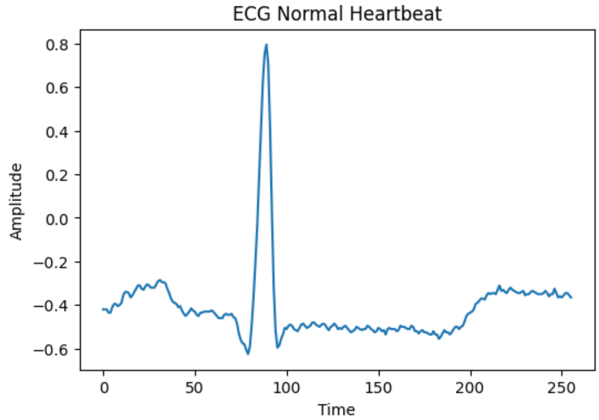

Figure 3.12. Heartbeat segment with 256 data points (R Peak around 88 points)

- **Type A (Atrial premature beat)**: An extra beat caused by electrical activation of the atria (upper chambers of the heart) from an abnormal site before a normal heartbeat would occur.

To prepare data for training the GAN model, we needed to segment the records into heartbeat intervals of each type. First, we looped through all records, identifying every sample annotated with one of the five heartbeat types mentioned above. Then, following Yang et al. (2021), we created segments consisting of 88 points before the beat and 168 points after it (256 points total), as a heartbeat typically lasts 0.6–0.8 seconds (216–288 points). Figure 3.12 shows an example of a normal (N) heartbeat segment extracted using this method.

Studies by Wulan et al. (2020), Xia et al. (2023), and Shaker et al. (2020) eliminate some noise from the signals before training to improve model accuracy and achieve smoother training. However, in this case, it was chosen not to perform any signal filtering. Instead, the model trains on raw original data, making it more robust for generating synthetic data that resembles real data with natural noise.

For heartbeat types N, L, R, and V, 5000 samples of each were used in training the GAN models. Type A was trained using all, 2546 available segments. As a dedicated model was trained for each heartbeat type, there were no issues with data imbalance.

Finally, the data is normalized to the range $[-1, 1]$ to match the "tanh" activation used in the Generator and reshaped to (num_samples, 256, 1) suitable for the *Conv1D* layers in the discriminator.

CHAPTER 4

Results and Discussion

In this chapter, the outcomes of training a GAN to produce synthetic ECG heartbeat signals of five types: Normal (N), Left Bundle Branch Block (L), Right Bundle Branch Block (R), Premature ventricular contraction (V), and Atrial premature contraction (A). The objective was to generate high-quality synthetic heartbeat signals using a limited amount of training data, addressing data scarcity issues in medical datasets. Additionally, the quality of the generated data was evaluated using a very simple classification model.

4.1. GAN Model Implementation

As mentioned in section 2.3.1 in Literature Review, Gulrajani et al. (2017) tackled challenges in training Wasserstein GANs. They proposed an enhanced approach using gradient penalty, which led to more stable training and improved performance. By penalizing the gradient norm to be close to 1, this method better enforces the Lipschitz constraint without weight clipping, like it was done formerly, resulting in a more stable training process. With the gradient penalty, the model achieves better stability in training and avoids the main issues of vanishing and exploding gradients.

Considering this, the plan is to implement this training method in our GAN model, with necessary adaptations to handle sequential data like ECG signals. All the code regarding the development of the GAN model and testing is available in GitHub Repository URL provided in the footnote[1]. The GAN consists of two primary components: the **Generator** and the **Discriminator**.

4.1.1. Generator

The generator aims to produce synthetic ECG heartbeat signals that resemble the real data.

It takes a latent vector z sampled from a standard normal distribution, $z \sim N(0, I)$ where I is the identity matrix. The output of the first dense layer is reshaped into a 2D tensor suitable for sequence modeling. The model then implements a bidirectional LSTM that captures temporal dependencies in both directions. This bidirectional approach is crucial, as a heartbeat's characteristics depend on both past and future signal intervals. Table 4.3 illustrates the full architecture of the Generator.

4.1.2. Discriminator

The discriminator's role is to distinguish between real ECG signals and those generated by the generator. It takes an ECG heartbeat segment as input and outputs a single

[1] https://github.com/CarlosISCTE/Masters_Code_GAN

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 8192)	827,392
reshape (Reshape)	(None, 64, 128)	0
bidirectional (Bidirectional)	(None, 64, 128)	98,816
conv1d_4 (Conv1D)	(None, 64, 128)	82,048
leaky_re_lu_4 (LeakyReLU)	(None, 64, 128)	0
conv1d_5 (Conv1D)	(None, 64, 64)	41,024
leaky_re_lu_5 (LeakyReLU)	(None, 64, 64)	0
up_sampling1d (UpSampling1D)	(None, 128, 64)	0
conv1d_6 (Conv1D)	(None, 128, 32)	10,272
leaky_re_lu_6 (LeakyReLU)	(None, 128, 32)	0
conv1d_7 (Conv1D)	(None, 128, 16)	2,576
leaky_re_lu_7 (LeakyReLU)	(None, 128, 16)	0
up_sampling1d_1 (UpSampling1D)	(None, 256, 16)	0
conv1d_8 (Conv1D)	(None, 256, 1)	81

TABLE 4.3. Generator Model Architecture

scalar value representing the critic score—an assessment of whether the segment is real or generated. Table 4.4 illustrates the full architecture of the Discriminator.

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 256, 32)	544
leaky_re_lu (LeakyReLU)	(None, 256, 32)	0
conv1d_1 (Conv1D)	(None, 256, 64)	32,832
leaky_re_lu_1 (LeakyReLU)	(None, 256, 64)	0
max_pooling1d (MaxPooling1D)	(None, 128, 64)	0
conv1d_2 (Conv1D)	(None, 128, 128)	131,200
leaky_re_lu_2 (LeakyReLU)	(None, 128, 128)	0
conv1d_3 (Conv1D)	(None, 128, 256)	524,544
leaky_re_lu_3 (LeakyReLU)	(None, 128, 256)	0
max_pooling1d_1 (MaxPooling1D)	(None, 64, 256)	0
flatten (Flatten)	(None, 16384)	0
dense (Dense)	(None, 1)	16,385

TABLE 4.4. Discriminator Model Architecture

4.1.3. Framework

This architecture implements the WGAN-GP framework by modifying the loss function and enforcing a Lipschitz constraint on the discriminator via a gradient penalty.

Unlike traditional GANs that use the Jensen-Shannon divergence, WGAN uses the Earth-Mover (Wasserstein-1) distance, which provides smoother gradients and better training dynamics. The discriminator aims to maximize the difference between its outputs on real and fake data while adhering to the Lipschitz constraint. The discriminator's loss function is defined as:

$$L_D = \mathbb{E}_{x_{\text{fake}} \sim P_g} \left[D(x_{\text{fake}}) \right] - \mathbb{E}_{x_{\text{real}} \sim P_r} \left[D(x_{\text{real}}) \right] + \lambda \cdot \text{GP}$$

λ is the gradient penalty weight, which is set to 10 in this implementation. GP is the gradient penalty term that ensures the discriminator satisfies the Lipschitz constraint by

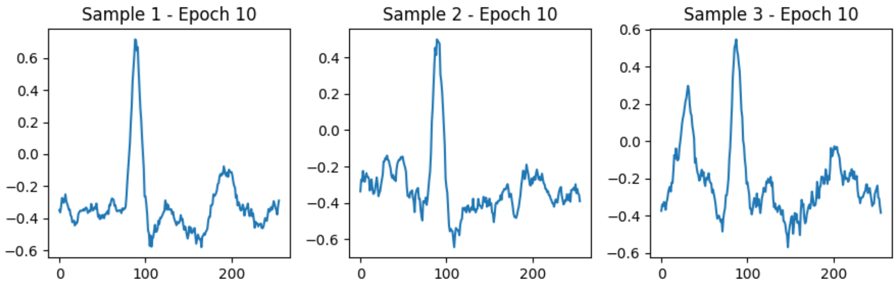

FIGURE 4.13. Samples for Normal Heartbeat generated at epoch 10

penalizing deviations from a gradient norm of 1. The Gradient penalty term is defined as:

$$\mathrm{GP} = \mathbb{E}_{\hat{x} \sim P_{\hat{x}}} \left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1 \right)^2 \right]$$

- $\hat{x}$ are samples interpolated between real and fake data.
- $\nabla_{\hat{x}} D(\hat{x})$ is the gradient of the discriminator output with respect to $\hat{x}$.

The generator aims to produce data that the discriminator rates highly, effectively minimizing the discriminator's output on fake data. The Generator's loss is defined as:

$$L_G = -\mathbb{E}_{x_{\text{fake}} \sim P_g} \left[D(x_{\text{fake}}) \right]$$

4.2. Training Procedure

For each training step, a batch of real ECG segments $\{\mathbf{x}_{\text{real}}\}$ is sampled from the dataset. Then, random latent vectors $\{\mathbf{z}\}$ are sampled, and the generator produces fake ECG signals $\{\mathbf{x}_{\text{fake}} = G(\mathbf{z})\}$.

The discriminator then computes its outputs on real and on fake data. The gradient penalty is calculated, and the discriminator loss is returned, updating the weights on the discriminator in order to minimize this loss function. The discriminator is updated 5 times before updating the generator to ensure the critic remains optimal during training.

Based on the results from the discriminator, the generator loss is then calculated, updating the generator weights in order to minimize its loss function.

Both the generator and discriminator use the Adam optimizer with a learning rate of 1×10^{-4} and $\beta_1 = 0.5$. β_1 controls the exponential decay rate for the first moment estimates of the gradients, managing the momentum aspect of the optimizer.

Every 10 epochs, a custom callback generates and plots three synthetic ECG samples, offering a snapshot of the generator's performance during training. Figure 4.13 illustrates an example of generated signals for normal heartbeats produced at epoch 10.

When the training is over, the generator model is saved, and 2000 samples of artificial heartbeat segments are generated and saved for each heartbeat type.

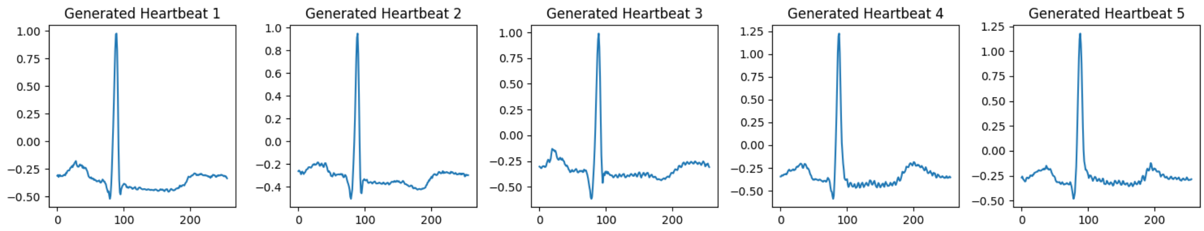

FIGURE 4.14. Generated Heartbeat from generator. Normal Beats.

4.3. GAN Training Results

For training the GAN, a subset of heartbeat segments was selected from the original data to demonstrate the model's ability to generate quality synthetic data that resembles real samples even with limited input. Specifically, 5,000 segments were used for each heartbeat type N, L, R, and V, while only 2240 segments were used for type A, given that this was the total of heartbeats of this types present in the dataset. Figures from appendix B show some examples of heartbeat segments of every type.

4.3.1. Training Process and Signal Generation

The GAN was trained separately on each heartbeat type. Before training, we normalized the data to a range of -1 to 1, enhancing the GAN's performance by standardizing the signal amplitudes. This normalization step is crucial as it ensures that all input features are on a similar scale, which can help the model converge faster and perform more effectively. It also aligns well with the *tanh* activation function used in the generator, which outputs values in the range of -1 to 1, like mentioned in section 3.2.

The training process involved 40 epochs for each type, with the generator and discriminator alternating updates in the min-max game. During training, the loss functions were monitored of both networks to ensure stable convergence and to observe if either model was learning much faster than the other. Significant learning disparities could lead to problems such as mode collapse or vanishing gradients. Every 10 epochs, sample synthetic ECG signals were generated to visually assess the quality of the generated data. This allowed to track the model's progress in capturing the characteristics of each heartbeat type and understand whether the generator was learning the distribution or not.

At the end of the 40 epochs, five heartbeat signals were generated and plotted for each type, to visually assess the consistency and quality of the GAN outputs after training. Figure 4.14 shows 5 generated heartbeats for the heartbeat type N. The characteristics of the signals vary significantly, even within the distribution of one type of heartbeat. This variability is desirable, as it mirrors the original dataset. Heartbeats of each type don't always look similar and can have different characteristics depending on the patient and the recording device.

4.3.2. GAN Model Evaluation

To evaluate the similarity between the real and synthetic data distributions, PCA and t-SNE were visualized.

These visual evaluation techniques were applied to both real and synthetic datasets for each heartbeat type, allowing for a visual comparison of their distributions in lower-dimensional spaces. PCA was used to reduce the dimensionality of the data to two principal components, providing insights into the overall structure and variability of the datasets. t-SNE, on the other hand, was employed to capture local structures and reveal potential clusters or patterns within the data. By analyzing both these methods, it is possible to understand if the distribution of the synthetic data correctly follows the original data. However, if the distributions match too closely, it may indicate overfitting by the model. Figures 4.15 and 4.16 shows the PCA and t-SNE methods respectively employed to the generated and original segment for the heartbeat type R, more examples can be found in the appendix.

The PCA plot from figure 4.15 shows the original data (orange) and the generated data (blue) plotted along the first two principal components. This reduction in dimensionality highlights the primary sources of variance within the dataset, comparing the global structure between original and synthetic heartbeats. The overlap between the original and generated data indicates that the generative model has successfully learned and replicated the overall variability and structure present in the original dataset. The presence of different clusters in the plot suggests that the model has captured key patterns or groupings within the heartbeat segments, which likely correspond to specific physiological characteristics or variations in the data. However, a closer inspection reveals some areas where the synthetic data distribution diverges from the original, representing potential opportunities to refine the model for even better accuracy. The use of PCA is especially valuable in generative model evaluation because it provides a high-level summary of the data's variability and structure, making it easier to identify similarities or discrepancies between the original and synthetic datasets. By capturing global trends, PCA helps assess whether the synthetic data maintains the large-scale relationships between features, which is important for tasks requiring realistic data distributions.

The t-SNE plot presented in Figure 4.16 focuses on preserving the local similarities and relationships within the data, adding to the perspective shown in the PCA plot. Unlike PCA, which highlights global variance and large-scale structure, t-SNE is particularly valuable for capturing fine-grained, nonlinear patterns in the data. In this plot, there is significant overlap between the original data (orange) and the generated data (blue), suggesting that the generative model has successfully learned many of the features of the original dataset. However, deviations exist, particularly in the larger clusters, where the alignment of generated and original data points is not perfect. This discrepancy highlights areas where the model could be improved to better replicate subtle details within the data. The importance of the t-SNE plot lies in its ability to reveal local structures that might be obscured in other dimensionality reduction techniques like PCA. For example, the clustering observed in the t-SNE visualization may represent different physiological or morphological characteristics of specific heartbeat patterns. By comparing the overlap and distribution of these clusters between the original and synthetic datasets, we can assess how well the generative model captures the underlying relationships and variability within the data.

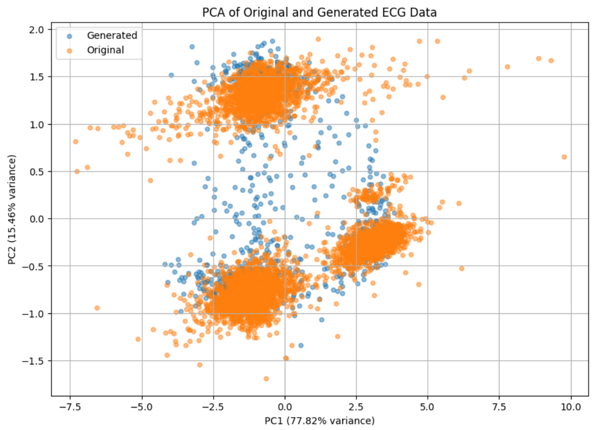

FIGURE 4.15. PCA test of generated heartbeat type R segments

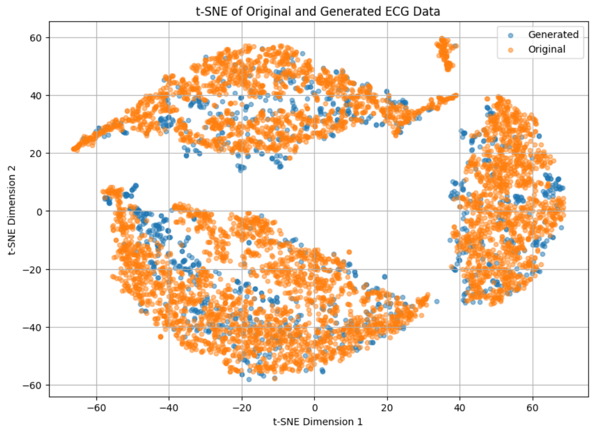

FIGURE 4.16. t-SNE test of generated heartbeat type R segments

Quantitative metrics were also calculated to evaluate the quality of the synthetic signals. Values for Euclidean Distance, DTW, PCC, and KLD were calculated between the original and generated values for every heartbeat type. These metrics provide a comprehensive assessment of the similarity between the original and synthetic data distributions. The Euclidean Distance measures the overall similarity in shape between the signals, while DTW accounts for temporal variations. The PCC evaluates the linear relationship between the datasets, and KLD quantifies the difference between probability distributions. Together, these metrics offer a multi-faceted evaluation of the GAN's performance in generating realistic ECG signals for each heartbeat type, with lower Euclidean Distance and DTW values, higher PCC, and lower KLD values suggesting better quality of the generated data. Table 4.5 shows all the values for these metrics for every heartbeat type.

Heartbeat Type	**Euclidean Distance**	**DTW**	**PCC**	**KLD**
N	3.018	2.009	0.976	0.027
L	3.554	2.533	0.967	0.018
R	4.772	3.299	0.971	0.032
V	7.820	4.699	0.941	0.009
A	6.129	4.210	0.898	0.031

TABLE 4.5. Calculated values for Euclidean Distance, DTW, PCC, and KLD for each heartbeat type

The relatively low Euclidean Distance values across heartbeat types indicate that the GAN effectively generated signals with similar overall shapes to the original data. DTW values, while slightly higher for more complex arrhythmias like V and A, suggest that the synthetic data preserved temporal dependencies, an essential feature for ECG signal realism. Higher PCC values reflect a strong linear relationship between the synthetic and real data, especially for the more frequently occurring heartbeat types (N and R), indicating that the model captures essential features of these patterns.

The KLD metric further supports the GAN model's ability to capture the probability distribution of the real data, with lower KLD values showing closer alignment in data distribution. The slightly higher KLD for the rarer arrhythmia types (R and A) indicates that the GAN might require more specific tuning for generating these classes accurately. Overall, these metrics collectively affirm that the WGAN-GP architecture is effective in generating realistic synthetic ECG data, with the potential for further refinement in classes with greater structural variability.

4.4. Classification Model Testing

A simple classification model was developed to classify ECG heartbeat signals into their respective types. The dataset was divided into 80% training, 10% validation, and 10% test sets. The training data was used to fit the model, the validation set to monitor for overfitting and checking the model performance during training, and the test set provided the evaluation target for the model's performance.

The model was first trained and tested on original data to establish a baseline performance. It was designed to be as simple as possible, making it easier to understand the impacts that the augmented data might have on the model. If the model were too powerful, these impacts would be less noticeable.

The dataset is highly imbalanced, with significantly more heartbeats of type N than the others and a low number of type A heartbeats. This imbalance makes it an ideal scenario for testing DA.

The model itself is a simple NN with an input layer that get a flattened 256-sample heartbeat segment and an output layer with five neurons corresponding to the five classes, using the *softmax* activation function to classify the data. The model is compiled with the Adam optimizer and categorical cross-entropy is used as loss function, suitable for multiclass classification with one-hot encoded labels. Accuracy is tracked to monitor the model's performance. It uses the early stopping callback function to monitor the validation loss and stop training immediately when the model stops improving, preventing overfitting. The model is trained for a maximum of 10 epochs. Table 4.6 shows the architecture of the model.

Layer (type)	**Output Shape**	**Param #**
flatten_4 (Flatten)	(None, 256)	0
dense_4 (Dense)	(None, 5)	1,285

TABLE 4.6. Model Architecture for Classification Model

4.4.1. Training and Testing on Original Data

Using the original dataset only, the model achieves an overall accuracy of 89%. The confusion matrix in Figure 4.17 illustrates the classification performance across different heartbeat types. Table 4.7 presents the precision, recall, and F1-score values for this classification test.

Heartbeat Type	Precision	Recall	F1-Score
A	0.52	0.21	0.30
L	0.84	0.54	0.66
N	0.90	0.98	0.94
R	0.94	0.93	0.93
V	0.79	0.63	0.70

TABLE 4.7. Classification Model: Trained on Original - Tested on Original

The confusion matrix provides insights into the model's performance for each heartbeat type, highlighting areas where the classifier excels and where it struggles. Notably, the model shows strong performance in identifying normal (N) heartbeats, but faces challenges with less common types like atrial premature beats (A). This baseline performance sets a benchmark for comparison when evaluating the impact of DA using synthetic ECG signals.

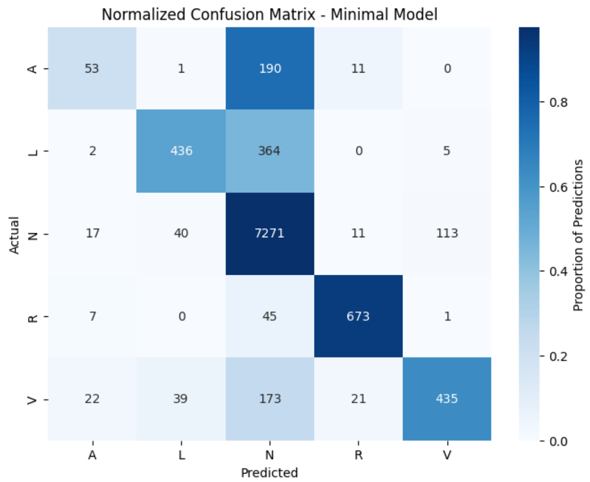

FIGURE 4.17. Confusion matrix for the classification model, trained and tested with original data

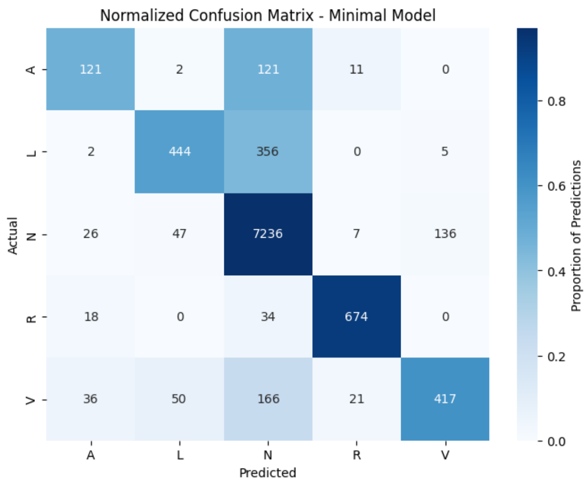

FIGURE 4.18. Confusion matrix for the classification model, trained with both original and synthetic data and tested with original data

4.4.2. Data Augmentation with Synthetic Data

To address the class imbalance in the dataset, synthetic data generated by the GAN was used to augment the underrepresented classes (A, V, L and R). Considering the differences in quantity among the various heartbeat types in the database, the number of synthetic segments added for each class accounts for this discrepancy. More segments are added to classes with fewer values and fewer to those with more values. The number of synthetic samples added for each class is as follows:

- **Class N**: 0 synthetic samples needed
- **Class A**: 1,999 synthetic samples added
- **Class V**: 1,878 synthetic samples added
- **Class L**: 1,846 synthetic samples added
- **Class R**: 1,869 synthetic samples added

After augmentation, the classification model was retrained. The augmented dataset led to a small improvement in the model's performance, particularly for classes that were unrepresented, increasing the overall accuracy by 1% to 90%. Notably, the precision for heartbeat type A improved from 52% to 69%, as shown in Table 4.8. The confusion matrix for this model is shown in Figure 4.18.

Heartbeat Type	**Precision**	**Recall**	**F1-Score**
A	0.69	0.47	0.53
L	0.82	0.55	0.66
N	0.91	0.97	0.94
R	0.95	0.93	0.94
V	0.75	0.60	0.67

TABLE 4.8. Classification model: Trained on Original + Synthetic — Tested on Original

The addition of GAN-generated synthetic samples mainly improved the classification model's performance on underrepresented heartbeat types. The precision and recall metrics for rarer classes, such as A, L and R, showed a slight improvement, highlighting that the synthetic data helped alleviate the class imbalance issue, though it has not been completely resolved. As shown in Table 4.8, adding synthetic data improved the model's sensitivity to less common arrhythmias, enhancing recall for the A and L heartbeat types without reducing the performance on more common classes like N.

Comparing F1-scores before and after DA reveals that the augmented dataset helped the classifier achieve a more balanced performance across classes, reducing bias toward the majority class. This result underscores the practical utility of synthetic data in enhancing model robustness and accuracy, particularly in healthcare applications where underrepresented classes can have critical diagnostic implications.

The addition of the synthetic data helped in balancing the dataset, which in turn improved the model's ability to correctly classify previously underrepresented heartbeat types. This demonstrates the practical utility of the GAN-generated data in enhancing classification tasks like this one.

4.4.3. Train on Synthetic, Test on Real & Train on Real, Test on Synthetic

Hyland et al. (2017) proposed two evaluation methods to assess the quality and diversity of generated data. The purpose of **Train on Synthetic, Test on Real (TSTR)** approach is to analyze how well the synthetic data captures the characteristics of real data. The classification model is trained using synthetic data generated by the GAN, and it is then tested on a separate set of real data. If the model performs well in this method, it suggests that the synthetic data effectively captures important features and patterns of the original data. Conversely, the purpose of **Train on Real, Test on Synthetic (TRTS)** is to evaluate how diverse the generated data is. In this method, the model is trained on real data and tested on synthetic data, assessing how well the synthetic data spans the original data's feature space.

These two methods were applied to the classification model, achieving an accuracy of **60% in TSTR** and an accuracy of **76% in TRTS**. This indicates that while the synthetic data captures some characteristics of the real data, there are discrepancies that affect classification performance. However, the synthetic data also contains variations that are recognized by the model trained on real data but includes novel patterns not fully represented in the original data.

Figure 4.19 presents the confusion matrix for the TRTS scenario. The high diagonal values, especially for categories such as "N" and "R", indicate that the synthetic data captures major patterns and clusters of the original data. However, the misclassifications in classes like "A", "V" and "L" suggest that the generative model still struggles to replicate some minority classes accurately. This limitation represents the imbalanced distribution of these classes in the training data.

Figure 4.20 illustrates the confusion matrix for the TSTR scenario. The prediction of correct diagonal values looks rather similar to the earlier scenario, while also showing a better performance in classifying the heartbeat of class "A", even though it still struggles to effectively predict it, mistaking several cases with the heartbeat type "N", which can also represent a big similarity between these two classes, giving the classification model a hard time.

The combined insights from these testing scenarios reinforce the importance of these evaluation methods. The higher accuracy in the TRTS scenario suggests that the synthetic data introduces diversity, potentially adding new patterns that complement the original data. However, the lower accuracy in the TSTR scenario points to limitations in the generative model's ability to replicate subtle and complex structures inherent to the real data.

4.5. Discussion

The results indicate that the GAN successfully generated synthetic ECG heartbeat signals that are similar to real signals, as evidenced by the visualizations and quantitative metrics. The use of synthetic data for DA proved beneficial, particularly for underrepresented classes, improving both the overall accuracy and the precision for heartbeat type A (the less represented type).

The additional testing methods revealed that while the synthetic data is of reasonable quality, there is room for improvement in fully capturing the complexity of real ECG

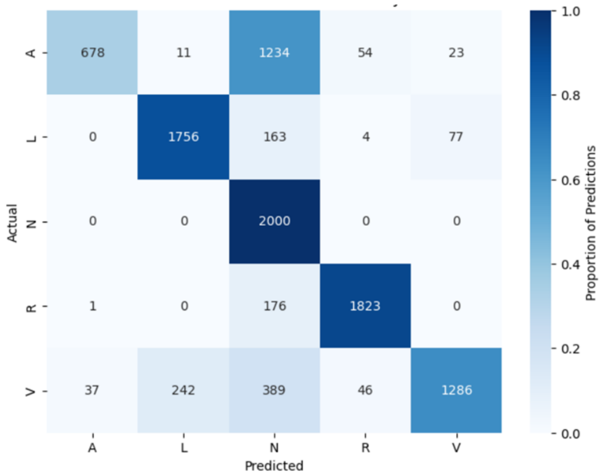

FIGURE 4.19. TRTS Confusion Matrix

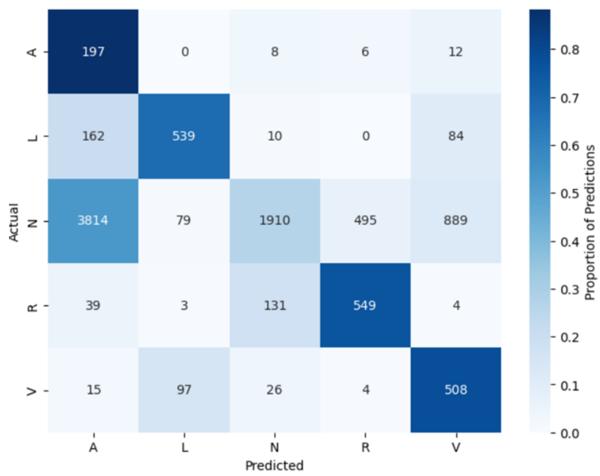

FIGURE 4.20. TSTR Confusion Matrix

signals. The lower accuracy in the TSTR scenario underscores the limitations in how well the synthetic data represents real ECG signals.

These finding demonstrate the potential of GANs in medical signal processing, particularly in generating synthetic data to augment limited datasets.This can enhance the performance of classification models, leading to better diagnostic tools.

CHAPTER 5

Conclusions

This dissertation has explored the potential of GANs to generate synthetic ECG signals, specifically addressing the challenges of data scarcity, privacy constraints, and class imbalance in the classification of arrhythmias. A WGAN-GP model was developed and implemented, tailored to generate realistic synthetic ECG data, thereby augmenting existing datasets and supporting deep learning models in healthcare.

The main findings of this study underscore the effectiveness of GAN-generated data in healthcare. The generated synthetic ECG signals maintained essential characteristics and variability of real signals, as demonstrated through quantitative metrics (Euclidean Distance, DTW, PCC, and KLD) and visual analysis methods (PCA and t-SNE). These evaluations revealed that the GAN model successfully captured the underlying distribution of the ECG data, producing high-quality synthetic signals that mirrored real patient data.

The application of TSTR and TRTS methodologies provided valuable insights into the quality and diversity of the generated data. While the results indicated room for improvement, especially in the quality area, they also highlighted the GAN's ability to capture important characteristics of real ECG signals and generate diverse synthetic samples.

Another key outcome is the positive impact of synthetic data on classification performance. By augmenting the training dataset with GAN-generated samples, the model effectively addressed class imbalance by improving performance on the classification of less present classes. This was most notable for the least represented class (type A), where precision increased from 52% to 69%. This improvement underscores the potential of synthetic data in addressing class imbalance issues, a common challenge in medical datasets. The synthetic data also slightly improved the model's classification accuracy, indicating that GANs can play a significant role in enhancing dataset diversity.

These findings have significant implications for medical research and clinical practice. The ability to generate high-quality synthetic ECG data can potentially accelerate cardiology research, allowing the development of more robust diagnostic tools and ML models. By demonstrating the generation of realistic synthetic data , GANs offer a solution to some challenges in healthcare data science, paving the way for improved diagnostic models and more personalized patient treatment options.

However, limitations were also revealed, pointing to areas for future research. The focus on generating individual heartbeat segments, while successful, leaves room for exploring the generation of longer ECG sequences with multiple heartbeat types and signal variations. Future work could involve training the GAN to generate ECG signal intervals with greater complexity and variability. Experimenting with different GAN architectures and applying advanced evaluation metrics could further validate the quality of the synthetic data. Investigating the potential of transfer learning techniques could enhance the GAN's ability to generate more diverse and accurate ECG signals. Additionally, the performance differences observed in TSTR and TRTS scenarios suggest that there is potential for further refinement of the GAN architecture to fully capture the complexities of real ECG signals.

Future research could also explore the application of this methodology to other types of medical signals, such as EEG or Electromyography (EMG), to expand the utility of DA in various medical fields. By extending this work to other domains, the methodology demonstrated here could have a broader impact on medical research and practice.

In conclusion, this study demonstrates that GANs are a promising tool for addressing data scarcity and enhancing classification accuracy in healthcare, with significant implications for patient care and clinical research. By demonstrating the generation of realistic synthetic data, GANs offer a solution to some challenges in healthcare data science, paving the way for improved diagnostic models and more personalized patient treatment options.

References

Alawneh, L., Alsarhan, T., Al-Zinati, M., Al-Ayyoub, M., Jararweh, Y., & Lu, H. (2021). Enhancing human activity recognition using deep learning and time series augmented data. *Journal of Ambient Intelligence and Humanized Computing*, *12*, 10565–10580. https://doi.org/10.1007/S12652-020-02865-4

Alqahtani, H., Kavakli-Thorne, M., & Kumar, G. (2021). Applications of generative adversarial networks (gans): An updated review. *Archives of Computational Methods in Engineering*, *28*, 525–552. https://doi.org/10.1007/S11831-019-09388-Y/TABLES/2

Arjovsky, M., Chintala, S., conference on machine …, L. B. .-., & undefined 2017. (2017). Wasserstein generative adversarial networks. *M Arjovsky, S Chintala, L BottouInternational conference on machine learning, 2017*. https://proceedings.mlr.press/v70/arjovsky17a.html

Brophy, E., Wang, Z., She, Q., & Ward, T. (2023). Generative adversarial networks in time series: A systematic literature review. *ACM Computing Surveys*, *55*, 31. https://doi.org/10.1145/3559540

Chakraborty, T., S, U. R. K., Naik, S. M., Panja, M., & Manvitha, B. (2024). Ten years of generative adversarial nets (gans): A survey of the state-of-the-art. *Machine Learning: Science and Technology*, *5*, 011001. https://doi.org/10.1088/2632-2153/ad1f77

Chandra, S., Prakash, P. K., Samanta, S., & Chilukuri, S. (2024). Clinicalgan: Powering patient monitoring in clinical trials with patient digital twins. *Scientific reports*, *14*. https://doi.org/10.1038/S41598-024-62567-1

Chen, G., Zhu, Y., Yang, Z., & Hong, Z. (2019). Emotionalgan: Generating ecg to enhance emotion state classification. *ACM International Conference Proceeding Series*, 309–313. https://doi.org/10.1145/3349341.3349422

Demir, S., Mincev, K., Kok, K., & Paterakis, N. G. (2021). Data augmentation for time series regression: Applying transformations, autoencoders and adversarial networks to electricity price forecasting. *Applied Energy*, *304*, 117695. https://doi.org/10.1016/J.APENERGY.2021.117695

Donahue, C., McAuley, J., & Puckette, M. (2018). Adversarial audio synthesis. *7th International Conference on Learning Representations, ICLR 2019*. https://arxiv.org/abs/1802.04208v3

Gatta, F., Giampaolo, F., Prezioso, E., Mei, G., Cuomo, S., & Piccialli, F. (2022). Neural networks generative models for time series. *Journal of King Saud University -*

Computer and Information Sciences, *34*, 7920–7939. https://doi.org/10.1016/j.jksuci.2022.07.010

Geiger, A., Liu, D., Alnegheimish, S., Cuesta-Infante, A., & Veeramachaneni, K. (2020). Tadgan: Time series anomaly detection using generative adversarial networks. *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, 33–43. https://doi.org/10.1109/BIGDATA50022.2020.9378139

Golany, T., Lavee, G., Yarden, S. T., & Radinsky, K. (2020). Improving ecg classification using generative adversarial networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 13280–13285. https://doi.org/10.1609/AAAI.V34I08.7037

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

Gonog, L., & Zhou, Y. (2019). A review: Generative adversarial networks. *Proceedings of the 14th IEEE Conference on Industrial Electronics and Applications, ICIEA 2019*, 505–510. https://doi.org/10.1109/ICIEA.2019.8833686

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *dl.acm.org*, *63*, 139–144. https://doi.org/10.1145/3422622

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2023). A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, *35*, 3313–3332. https://doi.org/10.1109/TKDE.2021.3130191

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of wasserstein gans [arXiv preprint]. https://doi.org/10.48550/arXiv.1704.00028

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Systematic Reviews*, *18*, e1230. https://doi.org/10.1002/cl2.1230

Huang, T., Chakraborty, P., & Sharma, A. (2020). Deep convolutional generative adversarial networks for traffic data imputation encoding time series as images. *International Journal of Transportation Science and Technology*, *12*, 1–18. https://doi.org/10.1016/j.ijtst.2021.10.007

Hyland, S. L., Zurich, E., Esteban, C., & Zurich, G. R. E. (2017). Real-valued (medical) time series generation with recurrent conditional gans [arXiv preprint]. https://doi.org/10.48550/arXiv.1706.02633

Iglesias, G., Talavera, E., González-Prieto, Á., Mozo, A., & Gómez-Canaval, S. (2023). Data augmentation techniques in time series domain: A survey and taxonomy. *Neural Computing and Applications*, *35*, 10123–10145. https://doi.org/10.1007/S00521-023-08459-3

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. https://arxiv.org/abs/1710.10196v3

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 4396–4405. https://doi.org/10.1109/CVPR.2019.00453

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. https://doi.org/10.61603/ceas.v2i1.33

Kingma, J., Simard, C., & Drolet, B. (2023). Overview of cardiac arrhythmias and treatment strategies. *Pharmaceuticals 2023, Vol. 16, Page 844*, *16*, 844. https://doi.org/10.3390/PH16060844

Krichen, M. (2023). Generative adversarial networks. *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*. https://doi.org/10.1109/ICCCNT56998.2023.10306417

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278–2323. https://doi.org/10.1109/5.726791

Liao, S., Ni, H., Sabate-Vidales, M., Wiese, M., Xiao, B., & Ni, S.-V. (2020). Conditional sig-wasserstein gans for time series generation. *2019 Prognostics and System Health Management Conference, PHM-Qingdao 2019*. http://arxiv.org/abs/2006.05421

Mason, J. W., Ramseth, D. J., Chanter, D. O., Moon, T. E., Goodman, D. B., & Mendzelevski, B. (2007). Electrocardiographic reference ranges derived from 79,743 ambulatory subjects. *Journal of Electrocardiology*, *40*, 228–234.e8. https://doi.org/10.1016/J.JELECTROCARD.2006.09.003

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets [arXiv preprint]. https://doi.org/10.48550/arXiv.1411.1784

Mogren, O. (2016). Data augmentation using synthetic data for time series classification with deep residual networks [arXiv preprint]. https://doi.org/10.48550/arXiv.1611.09904

Monachino, G., Zanchi, B., Fiorillo, L., Conte, G., Auricchio, A., Tzovara, A., & Faraci, F. D. (2023). Deep generative models: The winning key for large and easily accessible ecg datasets? *Computers in Biology and Medicine*, *167*, 107655. https://doi.org/10.1016/J.COMPBIOMED.2023.107655

Nazi, Z. A., Biswas, A., Rayhan, M. A., & Abir, T. A. (2019). Classification of ecg signals by dot residual lstm network with data augmentation for anomaly detection. *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. https://doi.org/10.1109/ICCIT48885.2019.9038287

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

Ramponi, G., Protopapas, P., Brambilla, M., & Janssen, R. (2018). T-cgan: Conditional generative adversarial network for data augmentation in noisy time series with irregular sampling. *International Journal of Transportation Science and Technology*. https://arxiv.org/abs/1811.08295v2

Shaker, A. M., Tantawi, M., Shedeed, H. A., & Tolba, M. F. (2020). Generalization of convolutional neural networks for ecg classification using generative adversarial networks. *IEEE Access*, *8*, 35592–35605. https://doi.org/10.1109/ACCESS.2020.2974712

Sharma, P., Kumar, M., Sharma, H. K., & Biju, S. M. (2024). Generative adversarial networks (gans): Introduction, taxonomy, variants, limitations, and applications. *Multimedia Tools and Applications*, 1–48. https://doi.org/10.1007/S11042-024-18767-Y/TABLES/6

Wang, Z., She, Q., & Ward, T. E. (2021). Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys*, *54*, 37. https://doi.org/10.1145/3439723/SUPPL_FILE/WANG.ZIP

Wulan, N., Wang, W., Sun, P., Wang, K., Xia, Y., & Zhang, H. (2020). Generating electrocardiogram signals by deep learning. *Neurocomputing*, *404*, 122–136. https://doi.org/10.1016/J.NEUCOM.2020.04.076

Xia, Y., Wang, W., & Wang, K. (2023). Ecg signal generation based on conditional generative models. *Biomedical Signal Processing and Control*, *82*, 104587. https://doi.org/10.1016/J.BSPC.2023.104587

Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018). Differentially private generative adversarial network. *Neural Computing and Applications*. https://arxiv.org/abs/1802.06739v1

Yang, Z., Li, Y., & Zhou, G. (2023). Ts-gan: Time-series gan for sensor-based health data augmentation. *ACM Transactions on Computing for Healthcare*, *4*, 12. https://doi.org/10.1145/3583593/ASSET/AF303717-63A7-4508-A4CA-844C10D7C6BB/ASSETS/GRAPHIC/HEALTH-2022-0057-F12.JPG

Yoon, J., Jarrett, D., & Schaar, M. V. D. (2019). Time-series generative adversarial networks. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. https://doi.org/10.5555/3454287.3454781

Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). Seqgan: Sequence generative adversarial nets with policy gradient. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*, 2852–2858. https://doi.org/10.1609/AAAI.V31I1.10804

Zhang, C., Kuppannagari, S. R., Kannan, R., & Prasanna, V. K. (2018). Generative adversarial network for synthetic time series data generation in smart grids. *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, SmartGridComm 2018*. https://doi.org/10.1109/SMARTGRIDCOMM.2018.8587464

Zhang, D., Ma, M., & Xia, L. (2022). A comprehensive review on gans for time-series signals. *Neural Computing and Applications*, *34*, 3551–3571. https://doi.org/10.1007/S00521-022-06888-0/FIGURES/9

Zhang, Y., Wang, Z., Zhang, Z., Liu, J., Feng, Y., Wee, L., Dekker, A., Chen, Q., & Traverso, A. (2023). Gan-based one dimensional medical data augmentation. *Soft Computing*, *27*, 10481–10491. https://doi.org/10.1007/s00500-023-08345-z

Zhu, G., Zhao, H., Liu, H., & Sun, H. (2019). A novel lstm-gan algorithm for time series anomaly detection. *2019 Prognostics and System Health Management Conference, PHM-Qingdao 2019*. https://doi.org/10.1109/PHM-QINGDAO46334.2019.8942842

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-October*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244

APPENDIX A

Distribution of Heart Beats Images

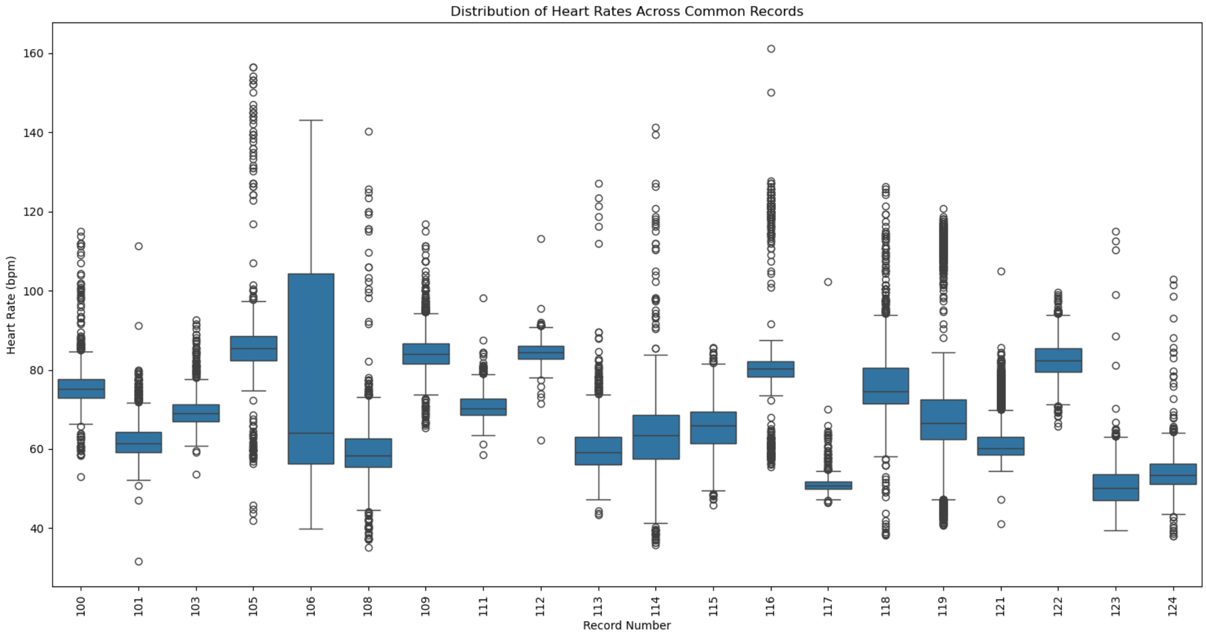

FIGURE A.21. Distribution of Heart Beats across common Records

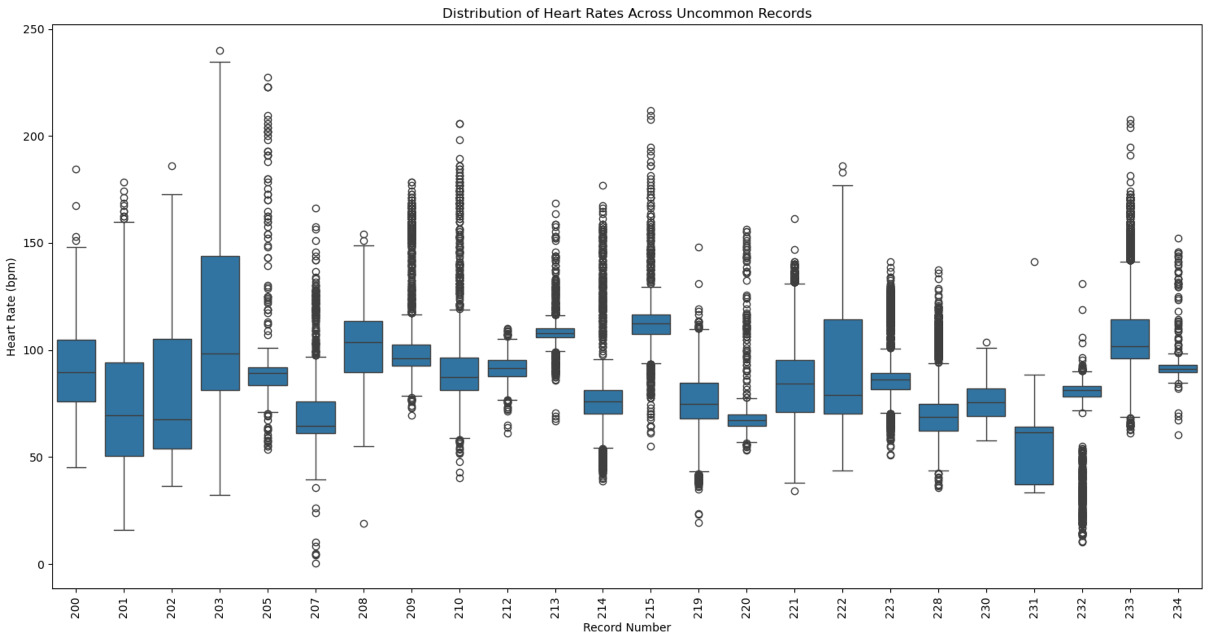

FIGURE A.22. Distribution of Heart Beats across uncommon Records

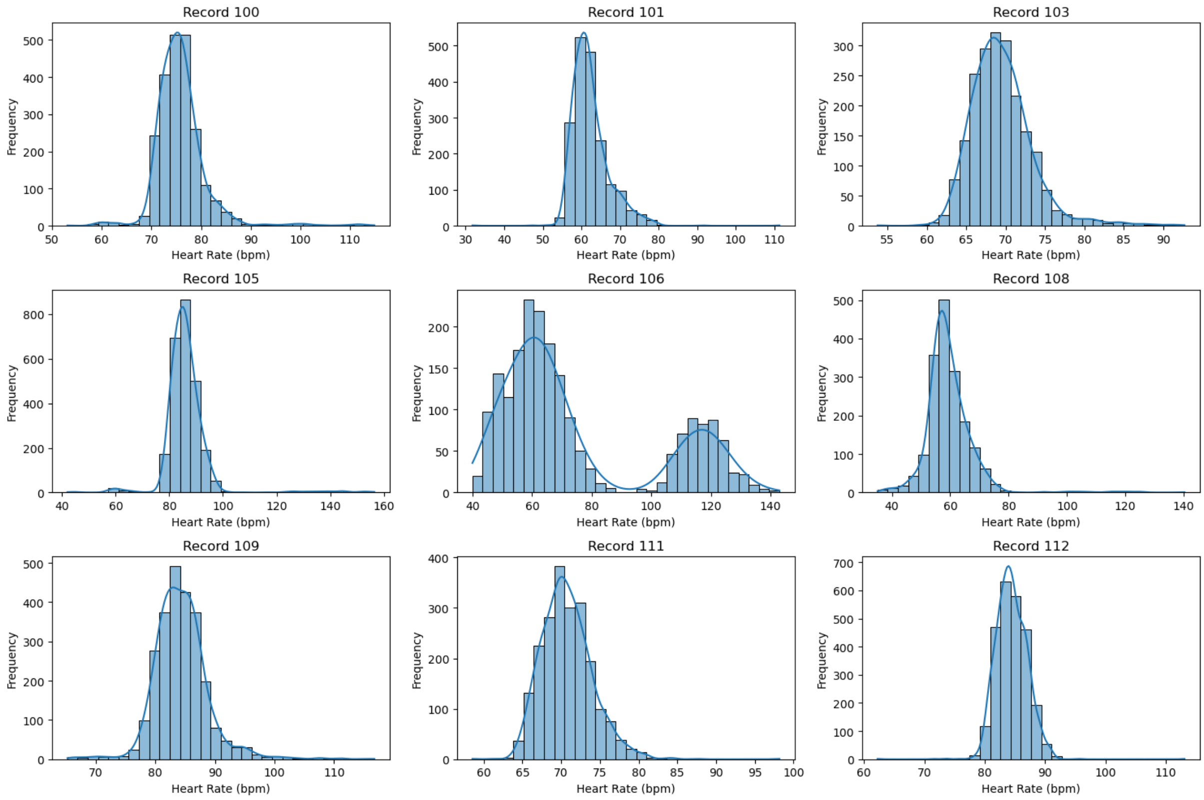

FIGURE A.23. Distribution of Heart Beats in some common records ECGs

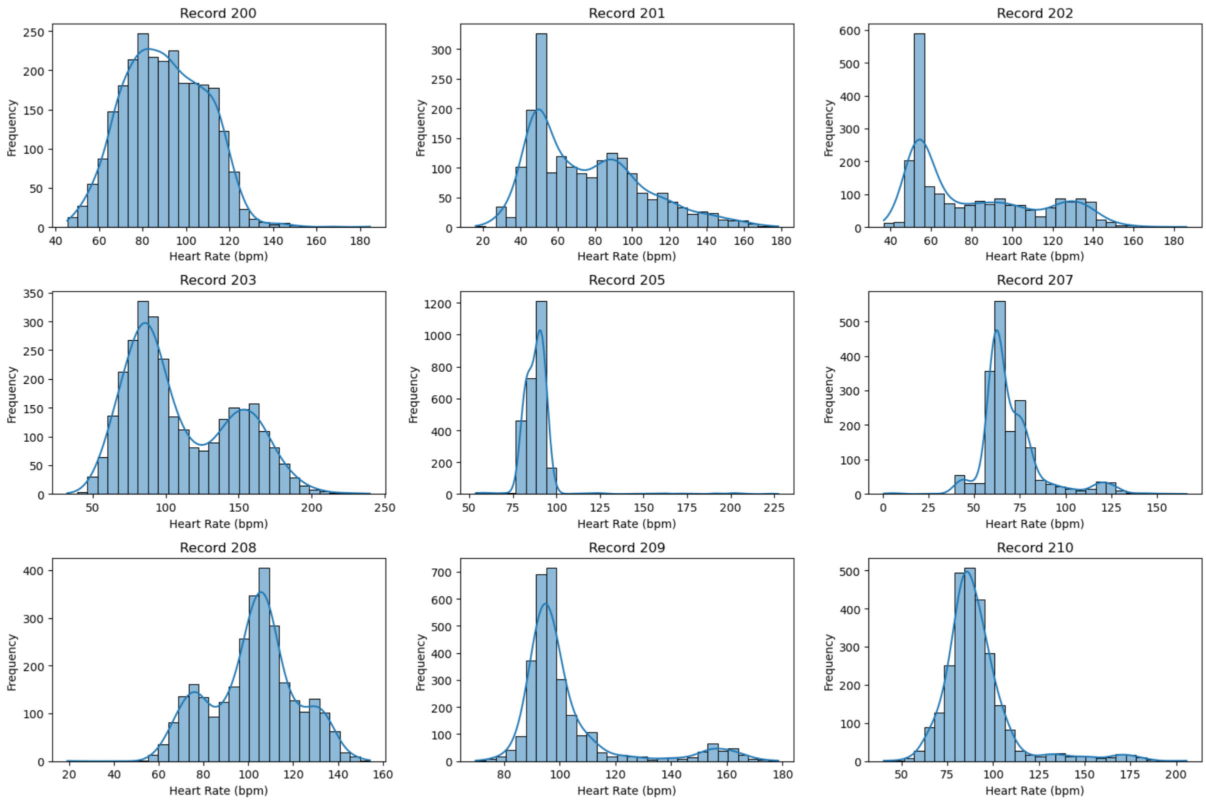

FIGURE A.24. Distribution of Heart Beats in some uncommon records ECGs

ECG Intervals Segmented from Original Data

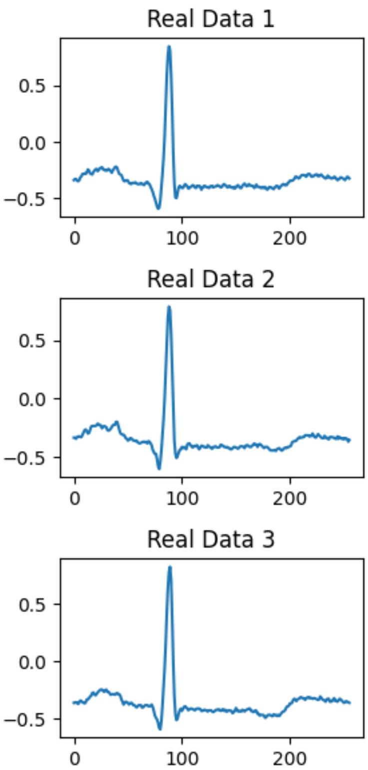

FIGURE B.25. Heartbeat signals of type N taken from the Database

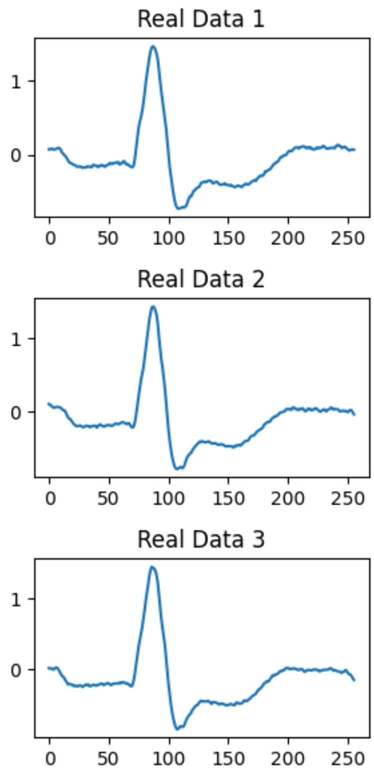

FIGURE B.26. Heartbeat signals of type N taken from the Database

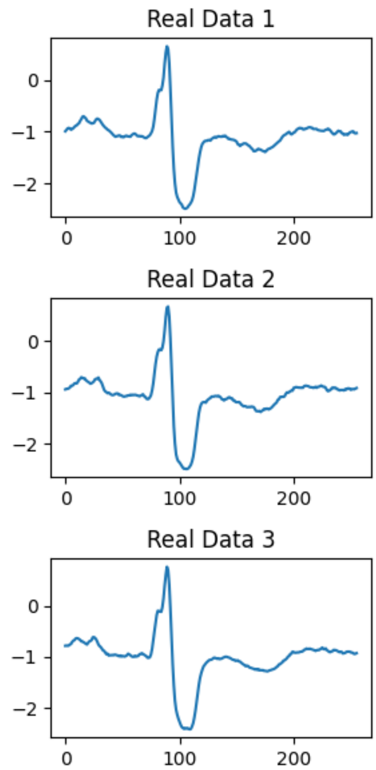

FIGURE B.27. Heartbeat signals of type N taken from the Database

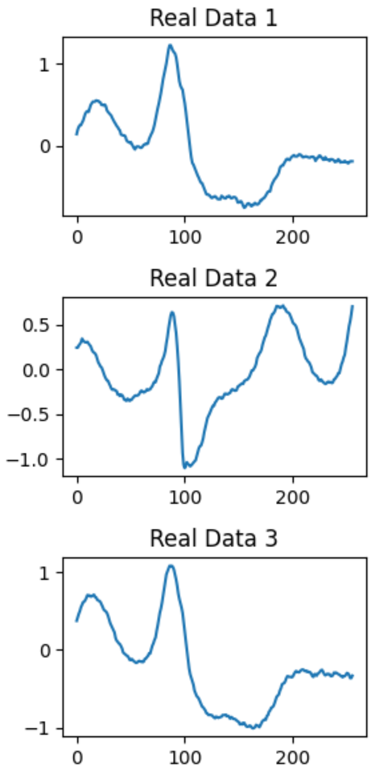

Figure B.28. Heartbeat signals of type V taken from the Database

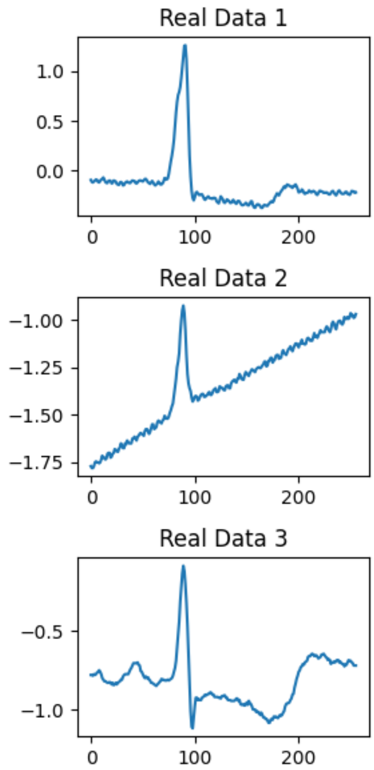

Figure B.29. Heartbeat signals of type A taken from the Database

APPENDIX C

PCA and t-SNE tests across records

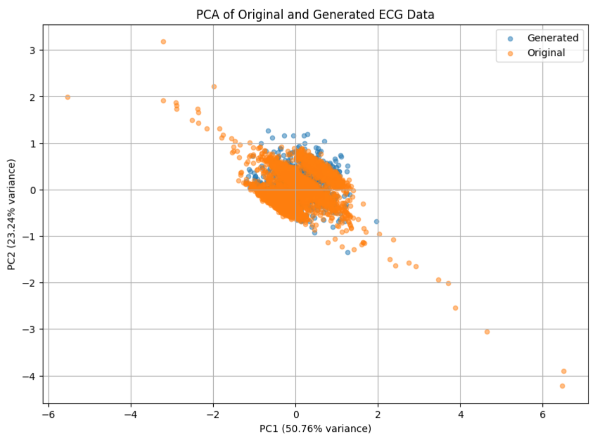

FIGURE C.30. PCA test of generated heartbeat type N segments

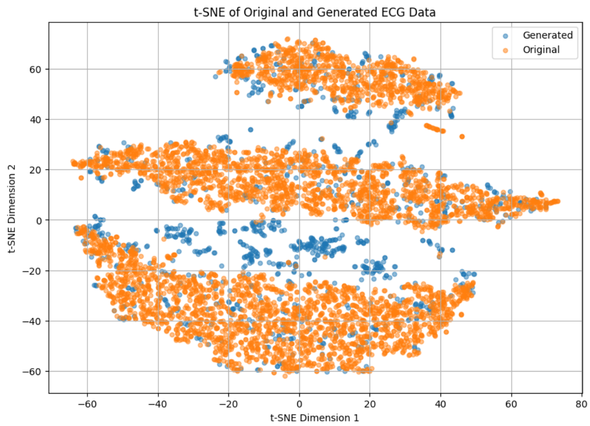

FIGURE C.31. t-SNE test of generated heartbeat type N segments

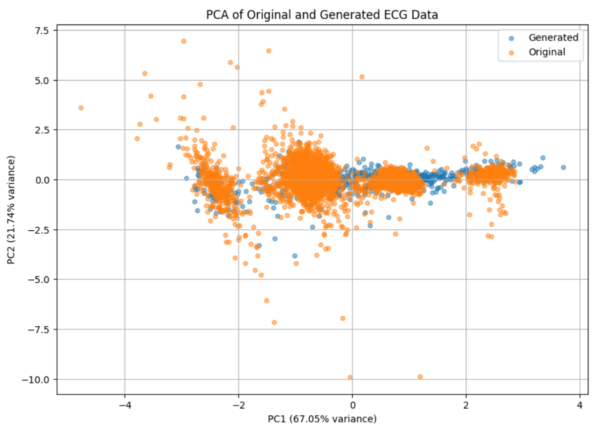

FIGURE C.32. PCA test of generated heartbeat type L segments

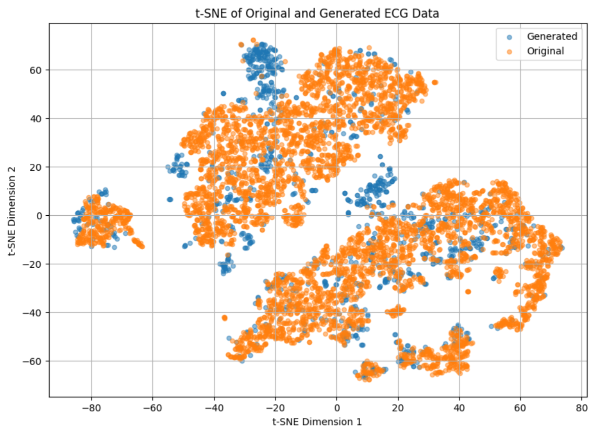

FIGURE C.33. t-SNE test of generated heartbeat type L segments

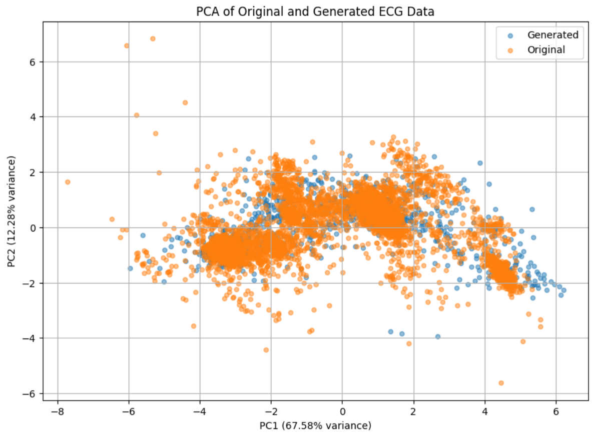

FIGURE C.34. PCA test of generated heartbeat type V segments

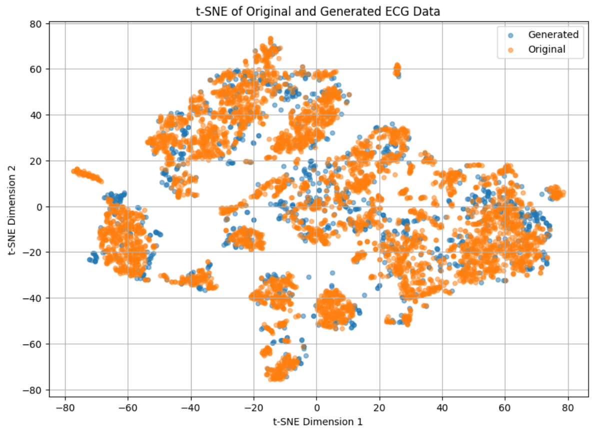

FIGURE C.35. t-SNE test of generated heartbeat type V segments

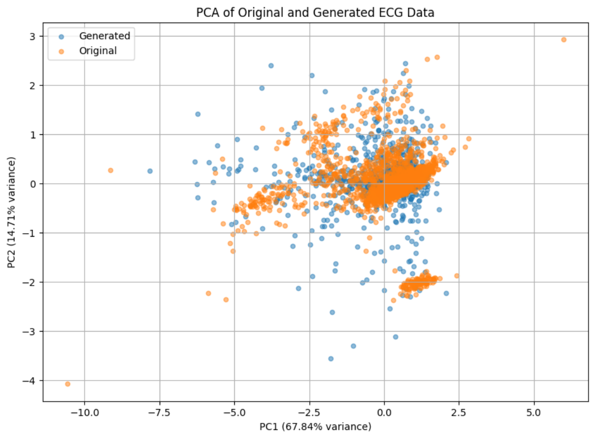

FIGURE C.36. PCA test of generated heartbeat type A segments

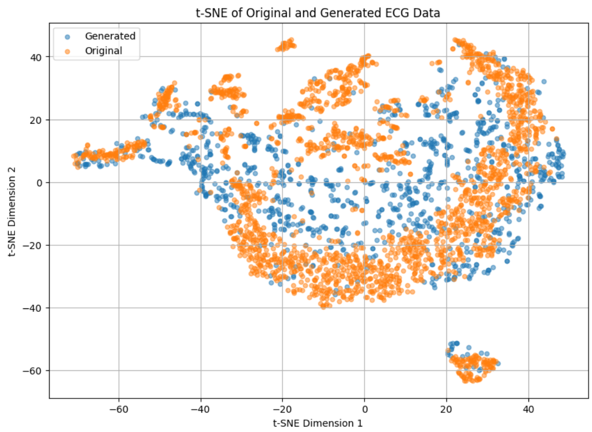

FIGURE C.37. t-SNE test of generated heartbeat type A segments