

Anna Esposito
Marcos Faundez-Zanuy
Francesco Carlo Morabito
Eros Pasero *Editors*



Applications of Artificial Intelligence and Neural Systems to Data Science

Smart Innovation, Systems and Technologies

Volume 360

Series Editors

Robert J. Howlett, KES International Research, Shoreham-by-Sea, UK

Lakhmi C. Jain, KES International, Shoreham-by-Sea, UK

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.


Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST), SCImago, DBLP.


All books published in the series are submitted for consideration in Web of Science.


Anna Esposito · Marcos Faundez-Zanuy ·
Francesco Carlo Morabito · Eros Pasero
Editors

Applications of Artificial Intelligence and Neural Systems to Data Science

Editors

Anna Esposito 
Department of Psychology
International Institute for Advanced
Scientific Studies (IIASS)
University of Campania “Luigi Vanvitelli”
Caserta, Italy

Francesco Carlo Morabito 
Department of Civil, Energy,
Environmental and Materials Engineering
University Mediterranea of Reggio Calabria
Reggio Calabria, Italy

Marcos Faundez-Zanuy 
Fundació Tecnocampus
Pompeu Fabra University
Barcelona, Spain

Eros Pasero 
Dipartimento di Elettronica e
Telecomunicazioni
Politecnico di Torino
Turin, Italy

ISSN 2190-3018

ISSN 2190-3026 (electronic)

Smart Innovation, Systems and Technologies

ISBN 978-981-99-3591-8

ISBN 978-981-99-3592-5 (eBook)

<https://doi.org/10.1007/978-981-99-3592-5>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Program Committee

Amorese Terry, Università della Campania “Luigi Vanvitelli” and IIASS
Brunetti Antonio, Politecnico di Bari
Callejas Zoraida, University of Granada
Cordasco Gennaro, Università della Campania “Luigi Vanvitelli” and IIASS
Cuciniello Marialucia, Università della Campania “Luigi Vanvitelli” and IIASS
Damasevicius Robertas, Kaunas University of Technology
Esposito Anna, Università della Campania “Luigi Vanvitelli” and IIASS
Esposito Antonietta Maria, Osservatorio Vesuviano sezione di Napoli
Esposito Marilena, International Institute for Advanced Scientific Studies (IIASS)
Faundez-Zanuy Marcos, Tecnocampus Universitat Pompeu Fabra
Garzia Fabio, Università di Roma “La Sapienza”
Greco Claudia, Università della Campania “Luigi Vanvitelli” and IIASS
Griol David, University of Granada
Koutsombogera Maria, Trinity College Dublin
Ieracitano Cosimo, Università degli Studi Mediterranea Reggio Calabria
Morabito Francesco Carlo, Università degli Studi Mediterranea Reggio Calabria
Mekyska Jiri, Brno University
Prinzi Francesco, Università degli Studi di Palermo
Scarpiniti Michele, Università di Roma “La Sapienza”
Schaust Jonathan, University of Applied Science in Koblenz
Senese Vincenzo Paolo, Università degli Studi della Campania “Luigi Vanvitelli”
Squartini Stefano, Università Politecnica delle Marche
Uncini Aurelio, Università di Roma “La Sapienza”
Vitabile Salvatore, Università degli Studi di Palermo
Vogel Carl, Trinity College Dublin.

Sponsoring Institutions

International Institute for Advanced Scientific Studies (IIASS), <http://www.iiassvietri.it/it/>)

Department of Psychology, Università della Campania “Luigi Vanvitelli”, IT, <https://www.psicologia.unicampania.it/home-becogsy>

Provincia di Salerno, <https://www.provincia.salerno.it/>

Comune di Vietri sul Mare, <https://www.comune.vietri-sul-mare.sa.it/kweb/sito/vietrisulmare>

International Neural Network Society (INNS), <https://www.inns.org/>

Università Mediterranea di Reggio Calabria, <http://neurolab.ing.unirc.it/>

Società Italiana Reti Neuroniche (SIREN), <https://www.siren-neural-net.it/>.

Preface

This book provides an overview on the current progresses and applications exploiting Artificial Intelligence and Neural Systems for Data Science.

The contributions reported in the book cover different scientific areas. These areas are closely connected in the themes they afford and provide fundamental insights for the cross-fertilization of different disciplines.

This book provides an overview on the current progresses in Artificial Intelligence and Neural Nets in Data Science reporting on intelligent algorithms and applications modeling, prediction, and recognition tasks and in many other application areas supporting complex multimodal systems to enhance and improve human-machine or human-human interaction.

This field is broadly addressed by the scientific communities and has a strong commercial impact to the extent that it provides sophisticated computational intelligence tools for supporting multidisciplinary aspects of data mining and data processing and characterizing appropriate system reaction to human interactional exchanges in interactive scenarios.

The emotional issue has recently gained increasing attention for such complex systems due to its relevance in helping in the most common human tasks (like cognitive processes, perception, learning, communication, and even “rational” decision-making) and therefore improving the quality of life of the end users.

The book account of interdisciplinary aspects in data science research involving different fields among those mathematics, computer vision, speech analysis and synthesis, signal processing, psychology, sociology, and advanced sensing. It provides contributions on their most recent trends, innovative approaches, and future challenges.

The chapters composing this book were first discussed in regular and special sessions at the international workshop on neural networks (WIRN 2022) held in Vietri Sul Mare from 7 to 9 of September 2022. The workshop hosted the special session on “**Dynamics of Signal Exchanges and Empathic System**”, organized by Anna Esposito, Zoraida Callejas, Marialucia Cuciniello, Antonietta M. Esposito, Gennaro Cordasco, Nelson Mauro Maldonato, and Francesco Carlo Morabito. The session

emphasized contributions devoted to the implementation of Empathic Systems, considering that empathy is central for successful social interactional exchanges.

The session was sponsored by two H2020 funded projects “**Empathic**” (empathic-project.eu/) and “**Menhir**” (menhir-project.eu/), aiming to implement socially and emotionally believable automatic systems, the Italian Government funded project SIROBOTICS (<https://www.exprivia.it/it-tile-6009-si-robotics/>) aiming to implement social robot assistants for supporting elderly everyday independent living, and the ANDROIDS project (<https://www.psicologia.unicampania.it/android-project>) funded by the program V: ALERE 2019 Università della Campania “L. Vanvitelli”, D. R. 906 del 4/10/2019, prot. n. 157264, 17/10/2019.

This particular edition of WIRN was important for all the communities since it was the first after 2 years of silence due to the pandemic COVID-2019. It was then a very stimulating gathering and represented the end of the pandemic situation.

The book is divided into two parts: one dedicated to neural networks and their practical applications and the other dedicated to features deriving from dynamics of signal exchanges for implementing empathic AI systems.

The scientists contributing to this book are specialists in their respective disciplines and through their contributions have made this volume a significant scientific effort. The coordination and production of this book has been brilliantly conducted by the Springer Project Coordinator Mr. **Ramesh Kumaran**, the contact the Publishing Editor Aninda Bose, the Springer Executive Editor Dr. **Thomas Ditzinger**, and the Editor Assistant Mr. **Holger Schaepe**. They are the recipients of our deepest appreciation. This initiative has been skillfully supported by the Editors-in-Chief of the Springer series *Smart Innovation, Systems and Technologies*, Profs. **Lakhmi C. Jain**, and **Howlett Robert James**, to whom goes our deepest gratitude.

Caserta, Italy
Mataró, Spain
Reggio Calabria, Italy
Turin, Italy

Anna Esposito
Marcos Faundez-Zanuy
Francesco Carlo Morabito
Eros Pasero

Contents

Part I Neural Networks and Related Applications

1	Generating New Sounds by Vector Arithmetic in the Latent Space of the MelGAN Architecture	3
	Michele Scarpiniti, Edoardo Massaro, Danilo Comminiello, and Aurelio Uncini	
2	Graph Neural Networks for Topological Feature Extraction in ECG Classification	17
	Kamyar Zeinalipour and Marco Gori	
3	Manifold Learning by a Deep Gaussian Process Variational Autoencoder	29
	Francesco Camastra, Angelo Casolaro, and Gennaro Iannuzzo	
4	Analysis of Sensors for Movement Analysis	37
	Marcos Faundez-Zanuy, Anna Faura-Pujol, Hector Montalvo-Ruiz, Alexia Losada-Fors, Pablo Genovese, and Pilar Sanz-Cartagena	
5	Dual Deep Clustering	51
	Giansalvo Cirrincione, Vincenzo Randazzo, Pietro Barbiero, Gabriele Ciravegna, and Eros Pasero	
6	Learning-Based Approach to Predict Fatal Events in Brugada Syndrome	63
	Vincenzo Randazzo, Gaia Marchetti, Carla Giustetto, Erica Gugliermi, Rahul Kumar, Giansalvo Cirrincione, Fiorenzo Gaita, and Eros Pasero	
7	Breast Cancer Localization and Classification in Mammograms Using YoloV5	73
	Francesco Prinzi, Marco Insalaco, Salvatore Gaglio, and Salvatore Vitabile	

8	Deep Acoustic Emission Detection Trained on Seismic Signals	83
	Jonathan Melchiorre, Marco M. Rosso, Raffaele Cucuzza, Emanuela D’Alto, Amedeo Manuello, and Giuseppe C. Marano	
9	A Deep Learning Framework for the Classification of Pre-prodromal and Prodromal Alzheimer’s Disease Using Resting-State EEG Signals	93
	Elena Sibilano, Michael Lassi, Alberto Mazzoni, Vitoantonio Bevilacqua, and Antonio Brunetti	
10	Imitation Learning Through Prior Injection in Markov Decision Processes	103
	Giovanni Di Gennaro, Amedeo Buonanno, Francesco Verolla, Giovanni Fioretti, Francesco A. N. Palmieri, and Krishna R. Pattipati	
11	Vision-Based Human Activity Recognition Methods Using Pose Estimation	115
	Giovanni Di Gennaro, Amedeo Buonanno, Marilena Baldi, Enzo Capoluongo, and Francesco A. N. Palmieri	
12	Identifying Exoplanets in TESS Data by Deep Learning	127
	Stefano Fiscale, Laura Inno, Angelo Ciamarella, Alessio Ferone, Alessandra Rotundi, Pasquale De Luca, Ardelio Galletti, Livia Marcellino, and Giovanni Covone	
13	Computational Intelligence for Marine Litter Recovery	137
	Vincenzo Bevilacqua, Antonio Di Marino, Angelo Ciamarella, Anastasia Angela Biancardi, Giorgio Budillon, Paola de Ruggiero, Emanuele Della Volpe, Luigi Gifuni, Danilo Mascolo, Stefano Pierini, and Enrico Zambianchi	
14	A Synthetic Dataset for Learning Optical Flow in Underwater Environment	147
	Alessio Ferone, Marco Lazzaro, Vincenzo Mariano Scarrica, Angelo Ciamarella, and Antonino Staiano	
15	BERT Classifies SARS-CoV-2 Variants	157
	Giorgia Ghione, Marta Lovino, Elisa Ficarra, and Giansalvo Cirrincione	
16	Competence-Based Coalition Choice, a Non-additive Approach	165
	Michele Fedrizzi and Silvio Giove	
17	Forecasting Mortality with Autoencoders: An Application to Italian Mortality Data	173
	Michele La Rocca, Cira Perna, Marilena Sibillo, and Antonio Vignes	

18	Leaky Echo State Network for Audio Classification in Construction Sites	183
	Michele Scarpiniti, Edoardo Bini, Marco Ferraro, Alessandro Giannetti, Danilo Comminiello, Yong-Cheol Lee, and Aurelio Uncini	
19	ECG Signal Classification Using Long Short-Term Memory Neural Networks	195
	Sidhant Kumar, Vijayeshkar Kumar, Krishnil Ram, Daniel Wood, Giansalvo Cirrincione, and Rahul R. Kumar	
20	A Convolutional Neural Network Approach for the Classification of Subjects with Epileptic Seizures Versus Psychogenic Non-epileptic Seizures and Control, Based on Automatic Feature Extraction from Empirical Mode Decomposition of Interictal EEG Recordings	207
	Michele Lo Giudice, Nadia Mammone, Cosimo Ieracitano, Umberto Aguglia, Danilo Mandic, and Francesco Carlo Morabito	
21	Commerce Districts: Conditions for Customer Overall Satisfaction in a Multi-attribute Framework	215
	Nicola Camatti, Andrea Ellero, and Paola Ferretti	
22	Problematic Merging and Cartels: A Collusion Risk Factors Analysis	227
	Andrea Ellero, Paola Ferretti, and Elena Zocchia	
Part II Dynamics of Signal Exchanges and Empathic Systems		
23	Conversational Ontologies for Human–Machine Interaction: Application for Cultural Heritage	239
	Rosalba Mosca and Anna Esposito	
24	On Statistical Prediction of Geometric Features of Three-Dimensional Configurations of Proteins I: Theoretical Description of an Inference Method	249
	Federica Vitale	
25	Emotion Recognition in Preschool Children: The Role of Age, Gender and Emotional Categories	267
	Claudia Greco, Marialucia Cuciniello, Terry Amorese, Gennaro Raimo, Gennaro Cordasco, and Anna Esposito	
26	A Multilevel Approach on the Investigation of the Association Between Responses to Infant Cues and Caregiving Propensity	279
	Carla Nasti, Francesca Parisi, and Vincenzo Paolo Senese	

27	What Would Happen if Hackers Attacked the Railways? Consideration of the Need for Ethical Codes in the Railway Transport Systems	289
	Lidia Marassi and Stefano Marrone	
28	Home Automation and Applied Behavior Analysis: Mand's Development in the Natural Environment	297
	Flavia Morfini, Simona Durante, Angela Ammendola, Enrico Moretto, Roberta Stanzione, Ottavio Ragozzino, Lucia Luciana Mosca, Valeria Cioffi, Nelson Mauro Maldonato, Benedetta Muzii, Natascia De Lucia, and Raffaele Sperandeo	
29	The Role of HEXACO Personality Traits on Predicting Problematic and Risky Behaviors in Adolescents	303
	Francesca Mottola, Vincenzo Paolo Senese, Marco Perugini, Augusto Gnisci, and Ida Sergi	
30	Evolving Aggregation Behaviors in Swarms from an Evolutionary Algorithms Point of View	317
	Paolo Pagliuca and Alessandra Vitanza	
31	A Review of the Use of Neural Models of Language and Conversation to Support Mental Health	329
	Zoraida Callejas, Fernando Fernández-Martínez, Anna Esposito, and David Griol	
32	Generative Adversarial Networks in Federated Learning	341
	Miao Wei and Carl Vogel	
33	Identifying Key Physical and Natural Environmental Correlates of Child Development: An Exploratory Study Using Machine Learning on Data from Pakistan	351
	Andrea Bizzego and Gianluca Esposito	

About the Editors

Anna Esposito received her Laurea degree *summa cum laude* in Information Technology and Computer Science from Salerno University (1989) and the Ph.D. degree in Applied Mathematics and Computer Science from Napoli University Federico II (1995) with a thesis developed at MIT, Boston, USA. She was Postdoc at IIASS, Lecturer at Salerno University in Department of Physics (1996–2000), and Research Professor (2000–2002) at WSU in Department of Computer Science and Engineering, OH, USA. She is Full Professor at Campania University “L. Vanvitelli”. She is author of 300+ peer-reviewed publications in journals, books, and conference proceedings and editor-coeditor of 30+ books in the Springer series SIST, ISRL, LNCS and LNAI.

Marcos Faundez-Zanuy received the B.Sc. degree (1993) and the Ph.D. degree (1998), both from the Polytechnic University of Catalunya. He is Full Professor at ESUP Tecnocampus Mataro and heads the Signal Processing Group. His research interests lie in biometrics applied to security and health. He was Initiator and Chairman of the EU COST action 277 “Nonlinear speech processing” and Secretary of COST action 2102 “Cross-Modal Analysis of Verbal and Non-Verbal Communication”. He is author of 50+ papers indexed in ISI Journal citation report, 100+ conference papers, 10+ books, and PI of 10 national and EU funded projects.

Francesco Carlo Morabito joined the University of Reggio Calabria, Italy, in 1989 where he is a Full Professor (2001) of Electrical Engineering. He served as President of the Electronic Engineering Course, as Member of the University’s Inner Evaluation Committee, as Dean of the Faculty of Engineering, as Deputy Rector, and as Vice Rector for Internationalization. He is Member of the Steering Committee of the Italian Society of Electrical Engineering.

He served as President of Siren (2008–2014), as a Governor of INNS (2000–2012, 2022–) and is Vice-President of INNS.

Eros Pasero is Professor of Electronics at Politecnico of Turin since 1991. He was Visiting Professor at ICSI Berkeley (1991), Tongji University, Shanghai (2011,

2015), and Tashkent Politechic University, Uzbekistan. His interests are in Artificial Neural Networks and Electronic Sensors. He heads the Neuronica Lab (1990) where wired and wireless sensors are developed for biomedical, environmental, automotive applications, and sensor signals that are processed by neural networks. Professor Pasero is President of the Italian Society for Neural Networks (SIREN) and was General Chair of IJCNN2000, SIRWEC2006, and WIRN 2015. He received several awards and holds five international patents. He supervised 10 international Ph.D. and 100 master's theses and is Author of 100+ international publications.

Part I
Neural Networks and Related Applications

Chapter 1

Generating New Sounds by Vector Arithmetic in the Latent Space of the MelGAN Architecture



Michele Scarpiniti, Edoardo Massaro, Danilo Comminiello,
and Aurelio Uncini

Abstract In this paper, we investigate the exploitation of the latent space of a MelGAN architecture by the vector arithmetic for the generation of new sounds that may be appealing for musicians, similar to what has already been done in the case of words and images. Specifically, since the MelGAN uses directly the spectrogram as input to its generator, we focus our attention on the linear combination of two or three instrumental sounds. This combination is then fed to the MelGAN generator, and the produced output will be the new sound with innovative sonority. Some simulations, performed over different sounds and different combination coefficients, show the effectiveness of the proposed idea.

1.1 Introduction

In the last few years, researchers have heavily investigated deep learning techniques for audio and music generation [3]. Although the generation of high-quality audio samples is a very challenging problem, due to the high temporal resolution and dependencies at different timescales, recent approaches based on the Generative Adversarial Networks (GANs) [7] provide excellent results, overcoming those of more traditional approaches based on recurrent networks, like WaveNet [15].

M. Scarpiniti (✉) · E. Massaro · D. Comminiello · A. Uncini
Department of Information Engineering, Electronics and Telecommunications (DIET),
“Sapienza” University of Rome, Rome, Italy
e-mail: michele.scarpiniti@uniroma1.it

D. Comminiello
e-mail: daniilo.comminiello@uniroma1.it

A. Uncini
e-mail: aurelio.uncini@uniroma1.it

A GAN is an architecture composed of two sub-networks: a generator, which tries to generate high-quality audio samples, and a discriminator that should distinguish the generator's fake data from the real input data [7]. GANs have been successfully used to generate music by using both information from the time domain (like the WaveGAN) and the time-frequency domain (like the SpecGAN) [4]. Specifically, this latter is an architecture where the generator and the discriminator are composed of deep convolutional neural networks (CNNs), known as DCGAN [17], but using spectrogram inputs instead of normal images.

Other GAN architectures, which have been successfully used for audio generation, are the GANSynth [5] and the MelGAN [9]. The first architecture models log-magnitude and phases directly in the spectral domain, by using a sufficient frequency resolution, and produces high-fidelity and locally coherent audio [5]. The second one uses convolutional architectures in a GAN setup to perform audio waveform generation by exploiting the mel spectrogram inversion [9]. Different from other similar GANs, the MelGAN generator is fed by the mel spectrogram rather than a random input sampled by a Gaussian distribution. Very recently, MelGAN has also been extended to a multi-band approach for high-quality text-to-speech conversion [21]. A variant of the MelGAN is the MelGAN-VC [16] originally devoted to the voice conversion problem. This architecture is capable of generating audio frames longer than the other GANs, since spectrograms are split in shorter segments and then concatenated before the discriminator. Despite its good results, MelGAN-VC is quite complex and the training time is very high.

Based on the excellent results obtained by the MelGAN on the effective generation of audio samples, also evaluated by subjective evaluation metrics [9], and the not prohibitive computational cost, in this paper we adopt the MelGAN to generate new music samples.

The generator of a GAN picks its input by sampling from a latent space and creates a relationship between the latent space and the output. Usually, each variable is drawn from a Gaussian distribution; however, in the case of the MelGAN, the mel spectrogram is used as the input. The latent space is meaningful, since it contains all the information needed to represent the original data. Different data points in the latent space will produce a different output. It could be interesting to navigate such a latent space. For example, a series of points can be created in the latent space between the ends of a segment. These points can be used to generate a series of outputs that show a transition between the two generated ones [17].

Many efforts have been done in the exploitation of the latent space [14, 19, 20] in different applicative scenarios, like image processing [11], or medical imaging [2]. Some works are also addressed toward the audio generation [1, 8, 18]. In particular, [1] exploits the latent space in order to detect fake audio samples from the real ones, by extending the research conducted in [8] on the latent vector recovery by investigating the inverse mapping of GANs. The authors in [6] exploit the latent space produced by the WaveNet encoder-decoder architecture. Finally, the work in [18] proposes a hybrid GAN architecture that allows musicians to explore the GAN latent space in a controlled manner, and giving an opportunity to specify particular audio features to be present or absent in the generated audio samples.

In an interesting way, the points in the latent space can be combined by using a sort of vector arithmetic in order to obtain new points in the latent space that, in turn, can be used to generate new data. This is an interesting idea, and it was used in an intuitive manner for words and face images. Specifically, [13] show that combining the latent vector associated with some words will obtain a word conceptually related to a given simple arithmetic rule. As an example, the authors of [13] show that the output of the combination “King – Man + Woman” is “Queen”. Similarly, the authors of [17] provide an example with images: the output of a latent vector obtained as “smiling woman – neutral woman + neutral man” was an image related to a “smiling man”.

Motivated by these considerations and examples, in this paper we propose the first investigation on the possibility of exploiting the vector arithmetic of latent space to generate new sounds. After the training of a MelGAN architecture, we create new input vectors as the linear combination of two or three basic sounds, then we put these combinations in the MelGAN generator input in order to generate new sounds. We expect that these sonorities could be appreciated by musicians. Although this is a simple and preliminary idea, the use of vector arithmetic of latent space with the MelGAN has not been investigated.

The rest of the paper is organized as follows. Section 1.2 describes the proposed idea in terms of both the used architecture and vector arithmetic. Section 1.3 introduces the experimental setup, while Sect. 1.4 shows the obtained numerical results by depicting the spectrograms of the generated sounds. Finally, Sect. 1.5 concludes the paper and outlines some future works.

1.2 Proposed Idea

1.2.1 GAN: Generative Adversarial Network

Introduced in 2014 by Goodfellow et al. [7], GANs are based on the use of two (deep) competing neural networks. The first network is called the *generator* and has the task of generating sufficiently realistic samples. The latter are subsequently placed in the input to the second network, called the *discriminator*, which has the task of comparing the samples obtained from the generator with the real data coming from the dataset. The output of the discriminator will return a probability indicating whether the input data is real or not. The training is performed in order that the generator is able to generate realistic enough data to fool the discriminator, while the latter has to recognize, in the best possible way, the fake samples from the real ones (i.e., belonging to the dataset). This process can be summarized in a *minimax* game between the two networks. The learning of these networks is therefore unsupervised, i.e., without the use of a set of real data to control the estimation error. Both networks are usually implemented by deep neural architectures.

In the classic GAN model, the input data z to the generator is noisy and belong to a $p_z(z)$ distribution. The generator $G(z; \theta_g)$ is a differentiable function represented by a deep network with parameters θ_g . The discriminator $D(x; \theta_d)$ is also a deep network that returns a scalar: $D(x)$ represents the probability that x comes from the dataset rather than $p_z(z)$. The work done by the generator in generating samples that are as realistic as possible is equivalent to learning the distribution of the data in the dataset, that is, making sure that the latter is equivalent to the distribution of the generator. In other words, it aims to minimize the likelihood that the discriminator will recognize a false sample generated: $\log(1 - D(G(z)))$. The discriminator can be seen as a classifier that must maximize the probability of assigning the false class to the samples from the generator and the true class to the samples from the dataset. Therefore, the total cost function on which to apply the *minimax* game is the following [7]:

$$\min_G \max_D \mathcal{L}(D, G) = E_{x \sim p_{data}(x)} \{\log D(x)\} + E_{z \sim p_z(z)} \{\log(1 - D(G(z)))\}. \quad (1.1)$$

However, Goodfellow et al. [7] recommend to use the following non-saturating loss for the generator, which provides better results:

$$\min_G \mathcal{L}(G) = -E_{z \sim p_z(z)} \{\log(D(G(z)))\}. \quad (1.2)$$

Although GANs work well in many situations, they show some limitations, like the vanishing gradient problem and the mode collapse.

In order to avoid these problems, the loss function is often used as proposed in [10] for the Least Squares GANs (LSGANs). The cost functions that train the LSGANs are the following:

$$\begin{aligned} \min_D \mathcal{L}(D) &= \frac{1}{2} E_{x \sim p_{data}(x)} \{(D(x) - 1)^2\} + \frac{1}{2} E_{z \sim p_z(z)} \{(D(G(z)) + 1)^2\}, \\ \min_G \mathcal{L}(G) &= \frac{1}{2} E_{z \sim p_z(z)} \{(D(G(z)))^2\}. \end{aligned} \quad (1.3)$$

1.2.2 The Used MelGAN

A family of GANs used to generate high-quality audio samples is the MelGAN, which produces a synthetic audio signal by inverting in the time domain of the spectrogram generated in the mel scale [9, 21]. This network has a very different structure from the classic GAN. First of all, the generator input is not noisy but comes directly from a realistic distribution, i.e., the distribution of the mel spectrograms of the class to be generated. In addition, three discriminators are adopted. The architecture is equivalent for each of them; what varies is the type of input: the first discriminator

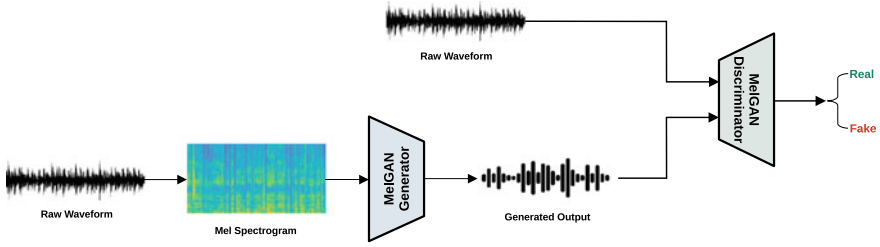


Fig. 1.1 General architecture of a MelGAN

works directly on the waveform generated, while the other two work on downsampled versions. In this way, each discriminator learns features at different frequency ranges.

Figure 1.1 shows the general architecture of a MelGAN. A complete description and explanation of the MelGAN can be found in the original paper of Kumar et al. [9]. In the following, we briefly describe the salient features.

The MelGAN generator is a fully convolutional network, whose input is the mel spectrogram and whose output is a raw waveform in the time domain. It uses a stack of transposed convolutional layers to upsample the input sequence, since the mel spectrogram has a temporal resolution lower than the original audio. Each transposed convolutional layer is, in turn, followed by a stack of residual blocks with dilated convolution. The use of a spectrogram input is justified in order to ensure the temporal coherence among adjacent inputs. To this purpose, the residual blocks after the oversampling are used for ensuring that there is a significant overlap between the inputs. Furthermore, the kernel size should be chosen as a multiple of the stride value, in order to avoid artifacts that would compromise the quality of the reconstruction. Finally, an important role is played by the weight normalization in order to speed up the convergence of the gradient descent algorithm.

As said before, the MelGAN is implemented by three parallel discriminators with identical architectures but working at different scales. Specifically, the first discriminator D_1 operates directly on the input audio in the time domain, while D_2 and D_3 operate on sub-sampled versions by a factor of 2 and 4, respectively. The downsampling is performed by the strided average pooling layers. The use of multiple discriminators is justified by the need to analyze audio at different scales, so that each discriminator can learn features at different frequency intervals. The discriminators' input is windowed in small frames, enough overlapped, in order to maintain the consistency among them. As for the generator, also the discriminators use the weight normalization. The detailed model architectures of the MelGAN generator and discriminator can be found in Fig. 1 of [9].

The training of the MelGAN is performed by the use of a modified version of the LSGAN loss in (1.3). Specifically, the discriminators are trained by minimizing the following loss functions that use the hinge version of (1.3):

$$\min_{D_k} \mathcal{L}(D_k) = E_x \{ \min\{0, 1 - D_k(x)\} \} + E_{s,z} \{ \min\{0, 1 + D_k(G(s, z))\} \}, \quad (1.4)$$

$\forall k = 1, 2, 3$, where x is the raw audio input in the time domain, s is the spectrogram used as input to the generator, and z is an additional Gaussian noise. The generator is trained by minimizing the following loss function, which is the sum of the single discriminators' losses:

$$\min_G \mathcal{L}(G) = E_{s,z} \left\{ - \sum_{k=1}^3 D_k(G(s, z)) \right\}. \quad (1.5)$$

Moreover, to further improve the quality of the generator output, an additional feature matching objective function has been added. This objective minimizes the L_1 distance between the discriminator outputs of the real audio and the generated audio, respectively. It is, therefore, a sort of similarity measure on the discriminator ability to distinguish real data from fake data:

$$\mathcal{L}_{FM}(G, D_k) = E_{x,s} \left\{ \sum_{i=1}^T \frac{1}{N_i} \left\| D_k^{(i)}(x) - D_k^{(i)}(G(s)) \right\|_1 \right\}, \quad (1.6)$$

where $D_k^{(i)}$ represents the i th layer feature map output of the k th discriminator block, and N_i denotes the number of units in each layer.

Hence, the objective function to be minimized for training of the generator is the following:

$$\min_G \left(\mathcal{L}(G) + \lambda \sum_{k=1}^3 \mathcal{L}_{FM}(G, D_k) \right), \quad (1.7)$$

where λ is a regularization parameter, usually set to $\lambda = 10$.

Overall, the MelGAN uses about 4.26 million of parameters, which is lower than other architectures used for the same purpose, such as the WaveNet.

1.2.3 The Vector Arithmetic in the Latent Space

The aim of this paper is to investigate the vector arithmetic in latent space of the MelGAN for the generation of new sounds, which may result very appealing for musicians.

To this purpose, we focus our attention on the linear combination of two or three basic sounds:

$$x = \sum_{i=1}^N \alpha_i x_i, \quad (1.8)$$

where x_i is the i th chosen input signal, N is 2 or 3 and represents the number of input signals (i.e., latent vectors), and $-1 \leq \alpha_i \leq 1$ is the i th coefficient of the linear

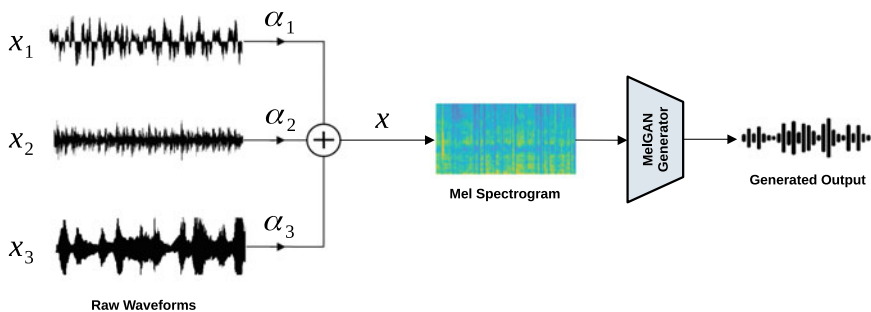


Fig. 1.2 An overview of the generation of a new sound by exploiting the latent space of a MelGAN by the vector arithmetic

combination. The combined input x is then fed to the MelGAN generator and the produced output is the new sound with innovative sonority. An overview of the generation process of new sounds by exploiting the latent space of a MelGAN by the vector arithmetic is shown in Fig. 1.2.

1.3 Experimental Setup

In this paper, we have used the large-scale and high-quality NSynth dataset¹ introduced in [6]. The NSynth dataset contains 305,979 musical notes, each with a unique pitch, timbre, and envelope. Each track consists of 4 seconds of monophonic audio belonging to 11 instrumental classes (generated from commercial sample libraries), sampled at 16 kHz. The sounds have been generated by ranging over every pitch of a standard MIDI Piano, as well as five different velocities. Moreover, the note was held for the first three seconds and allowed to decay for the final second. The NSynth dataset is already split in the training (289,205 instances), validation (12,678 instances), and testing (4,096 instances) sub-sets.

From the NSynth dataset, we have extracted a training set composed of four classes (keyboard, mallet, organ, and string), for a total of 88,616 files, corresponding to more than 98 hours of sounds. Table 1.1 summarizes the used classes, the number of available files for each class, and the corresponding duration. The related files of the test set have been used, after the training, as input for the generation process of the new sounds by exploiting the vector arithmetic in the latent space.

The MelGAN has been implemented in Python by using the PyTorch library. The audio waveform extraction and the mel spectrogram computation have been performed by using the librosa library² [12].

¹ Available at <https://magenta.tensorflow.org/datasets/nsynth>.

² Available at <https://librosa.github.io/librosa/>.

Table 1.1 Used classes, number of available files, and time duration (in [hr:min:sec]) for the training set

Class	Files	Duration
Keyboard	8,068	08:57:52
Mallet	26,857	29:50:28
Organ	34,301	38:06:44
String	19,390	21:32:40
Total	88,616	98:27:44

All the audio data has a sampling frequency of 16 kHz. The mel spectrogram, with 80 mel bands, has been computed by using frames of 1024 samples with a hop size of 256 samples, and a number of FFT points equal to 1024.

Simulations have been carried out by using a computer equipped with an Intel® Xeon 4110 CPU @ 2.10 GHz, with an NVIDIA Tesla V100 SXM2 32GB GPU.

The MelGAN has been trained by using the Adam algorithm with a learning rate $\eta = 10^{-4}$, and the other parameters set to $\beta_1 = 0.5$ and $\beta_2 = 0.9$, respectively. The mini-batch size is set to $B = 16$, while a total of 1,800 epochs has run.

1.4 Simulation Results

In this section, we show some simulation results obtained by exploiting the vector arithmetic in the latent space of the used MelGAN (refer to Eq. (1.8)). In order to provide a visual representation of the obtained audio signals and to evaluate the differences with respect to the original sounds, we show the spectrograms of the involved signals.³

In order to investigate our idea, we randomly extract from the test set four files from each class, denoted as the name of the class followed by an integer number (e.g., *Organ 1*, *Organ 2*, *Organ 3*, *Organ 4*, *String 1*, *String 2*, etc.). Similar experiments have been performed in the latent space of an autoencoder in [6].

The first set of simulations aims at verifying the quality of the reconstruction (i.e., the spectrogram inversion performed by the MelGAN generator). To this purpose, the selected test sounds have been converted into the mel spectrogram representations and then used as input to the generator. Figure 1.3 shows an example of spectrograms of two original sounds and the corresponding generated ones. Specifically, this figure shows the spectrograms related to the *String 1* and *Mallet 2* sound files. Figure 1.3 highlights that the reconstructed signals are very similar to the original ones, since the main spectral lines are preserved in the generated spectrograms. This fact could be also confirmed by a listening test. However, it should be noted that the *Mallet 2* sound has a slightly lower quality, since it presents a sort of noisy distortion in

³ Spectrograms and sound files of all the simulated cases are available at the following link: <https://github.com/mscarpiniti/LatentSound>.

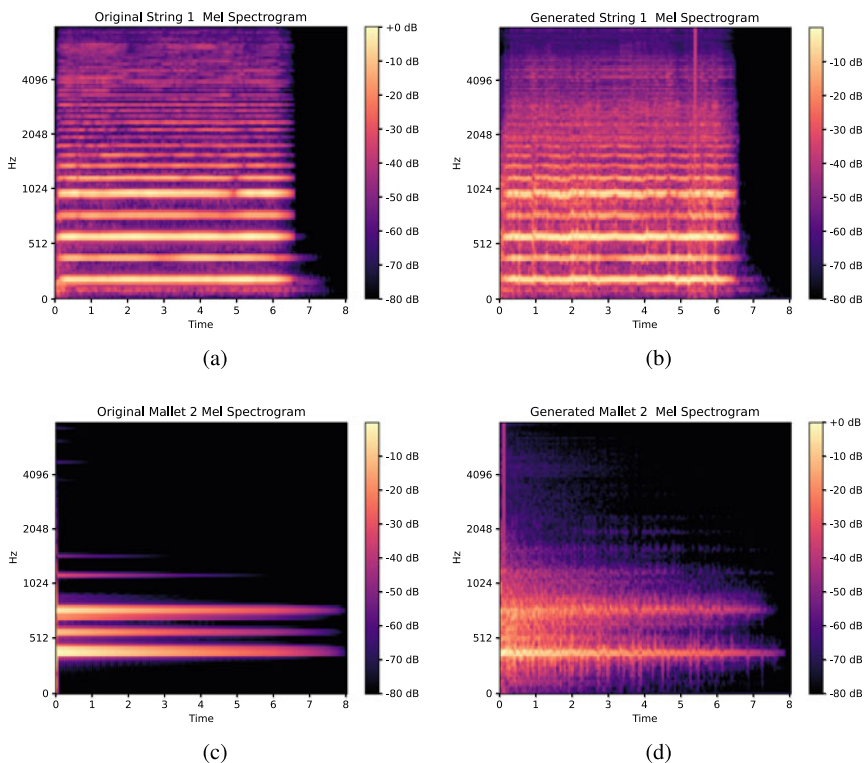


Fig. 1.3 Comparison between the spectrograms of **a** original “*String 1*” sound, **b** generated “*String 1*” sound, **c** original “*Mallet 2*” sound, and **d** generated “*Mallet 2*” sound

the background, although acceptable. Good results have been obtained also for the Keyboard and Organ classes.

In the second set of simulations, we explore the latent space of the MelGAN by the vector arithmetic. Specifically, we perform the linear combination of three test files and use this combination, after the mel spectrogram extraction, as input to the MelGAN. Due to length constraints, in this paper we show only three cases.

The first case is regarding the linear combination: “*Mallet 2* – *String 1* + *Keyboard 1*”. The resulting spectrogram is shown in Fig. 1.4a. Different from the case of words or images, it is quite difficult to have an intuition on what this output should sound like. However, we can argue that, since both the spectra of *String 1* and *Keyboard 1* (which is an acoustic Piano sound) present similar spectral lines, the output spectrogram should appear “similar” to that of the *Mallet 2* sound, at least at the lower frequencies due to the increasing differences between Piano and strings sounds at the higher ones. Moreover, the output should present more variability than the *Mallet 2*, due to an oscillating characteristic of *String 1*. This intuition is really observable in the corresponding spectrogram in Fig. 1.4a.

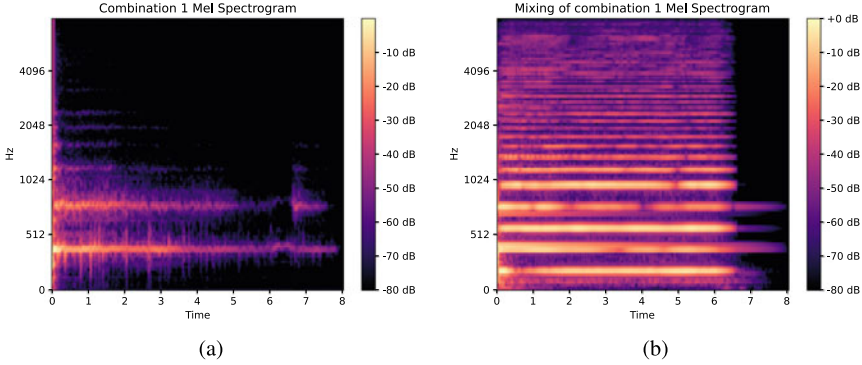


Fig. 1.4 **a** Spectrogram of the first combination: *Mallet 2 – String 1 + Keyboard 1*; **b** Differences between the simple mixing of the three considered sounds and the related sound generated by the MelGAN

In order to highlight the differences between such a generated output and the simple mixing of the same sounds, Fig. 1.4b shows the spectrogram of the mixed signal (i.e., the spectrogram evaluated on the input space). As can be seen from Fig. 1.4b, the two spectra are very different, and also the produced sound is different, highlighting the effect of the MelGAN generator. In fact, while the input sounds like a pure combination (i.e., like the three instruments playing at the same time), the output presents very peculiar features, and it sounds like a new and unheard sonority, as a “synthetic” sound produced by a sort of electronic device.

For the second combination, we adopt the idea in (1.8), and we use some parameters α_i set to values different from the unit. Specifically, we exploit the following combination: “*Keyboard 1 + 0.4 × Organ 2 – 0.4 × String 4*”. That is, we use the following combination parameters: $\alpha_1 = 1$, $\alpha_2 = 0.4$, and $\alpha_3 = -0.4$. In this case, as shown by the spectrogram in Fig. 1.5a, the sound is more rich. This behavior is due to the fact that each of the three considered sounds is composed of many spectral lines with different densities. Exploiting this combination, the generator will produce a sound with a spectral content that takes into account all the frequencies in each single signal.

Finally, the third combination is “ $0.4 \times \text{Organ 3} - 0.4 \times \text{String 1} + 0.9 \times \text{Organ 1}$ ”. That is, we use the following combination parameters: $\alpha_1 = 0.4$, $\alpha_2 = -0.4$, and $\alpha_3 = 0.9$. In this case, all the coefficients are different from the unit. The spectrogram obtained by the MelGAN generator exploiting such a combination is shown in Fig. 1.5b. Since *Organ 1* and *Organ 3* are two organ sounds with different frequencies, their combination should produce a richer spectral content. However, the subtraction of the *String 1* sound, in some way, tends to cancel the contribution in certain bands. Figure 1.5b confirms this simple intuition.

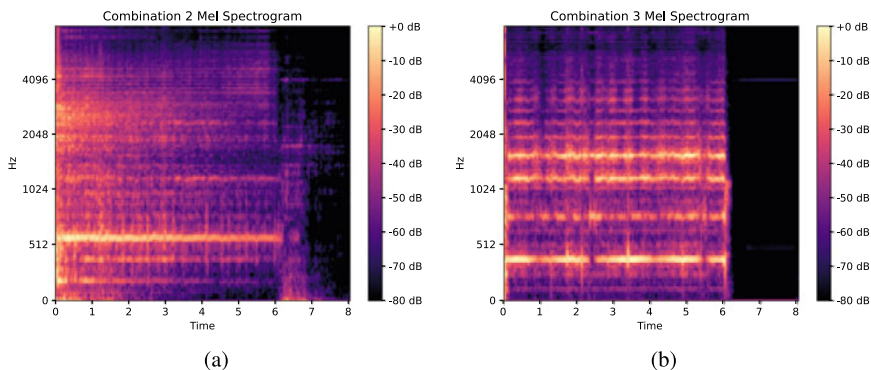


Fig. 1.5 **a** Spectrogram of the second combination: $\text{Keyboard } 1 + 0.4 \times \text{Organ } 2 - 0.4 \times \text{String } 4$; **b** Spectrogram of the third combination: $0.4 \times \text{Organ } 3 - 0.4 \times \text{String } 1 + 0.9 \times \text{Organ } 1$

Results obtained with many other combinations yield similar considerations. From this simple set of simulations, we can conclude that the exploitation of the vector arithmetic in the latent space of the MelGAN, similar to the case of words and images, can produce interestingly new sounds that could be considered appealing from musicians.

1.5 Conclusion

In this paper, we have proposed a preliminary investigation on the possibility to exploit the latent space of the MelGAN by the vector arithmetic, similar to what has already been done in the case of words and images, in order to generate new interesting sonorities, which may sound appealing to musicians. Specifically, since the MelGAN uses directly the spectrogram as its generator input, we construct a linear combination of two or three basic sounds. The output produced by the MelGAN to this input represents the new sound. Different cases, by using different sounds and different combination coefficients, show the effectiveness of the proposed idea. The obtained log mel scale spectrograms are also shown. In future work, we will analyze the effect of different types of (nonlinear) combination and the use of a greater number of basic sounds. Moreover, we will exploit the results that may be obtained by other GAN architectures, which will be released in the next future.

Acknowledgements This work has been supported by the project: “End-to-End Learning for 3D Acoustic Scene Analysis (ELeSA)” funded by Sapienza University of Rome Bando Acquisizione di medie e grandi attrezzature scientifiche 2018.

References

1. Bayat, N., Khazaie, V.R., Keyes, A., Mohsenzadeh, Y.: Latent vector recovery of audio GANs with application in deepfake audio detection. In: Proceedings of the 34th Canadian Conference on Artificial Intelligence, pp. 1–6 (2021). <https://doi.org/10.21428/594757db.1ee3922d>
2. Blanco, R.F., Rosado, P., Vegas, E., Reverter, F.: Medical image editing in the latent space of generative adversarial networks. *Intell.-Based Med.* **5**, 100040 (2021). <https://doi.org/10.1016/j.ibmed.2021.100040>
3. Briot, J.P., Hadjeres, G., Pachet, F.D.: *Deep Learning Techniques for Music Generation*. Springer, Cham, Switzerland (2020)
4. Donahue, C., McAuley, J., Puckette, M.: Adversarial audio synthesis. In: Seventh International Conference on Learning Representations (ICLR 2019), pp. 1–16. New Orleans, LA, USA (2019)
5. Engel, J., Agrawal, K.K., Chen, S., Gulrajani, I., Donahue, C., Roberts, A.: GANSynth: Adversarial neural audio synthesis. In: Seventh International Conference on Learning Representations (ICLR 2019), pp. 1–17. New Orleans, LA, USA (2019)
6. Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., Simonyan, K.: Neural audio synthesis of musical notes with WaveNet autoencoders. In: Proceedings of the 34th International Conference on Machine Learning (ICML 2017), vol. PMLR 70, pp. 1068–1077 (2017)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014), pp. 2672–2680 (2014)
8. Keyes, A., Bayat, N., Khazaie, V.R., Mohsenzadeh, Y.: Latent vector recovery of audio GANs (2020). [arXiv:2010.08534v1](https://arxiv.org/abs/2010.08534v1). <https://arxiv.org/abs/2010.08534>
9. Kumar, K., Kumar, R., de Boissiere, T., Geste, L., Teoh, W.Z., Sotelo, J., de Brébisson, A., Bengio, Y., Courville, A.C.: MelGAN: Generative adversarial networks for conditional waveform synthesis. In: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), pp. 1–12. Vancouver, Canada (2019)
10. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), pp. 2794–2802. Venice, Italy (2017). <https://doi.org/10.1109/ICCV.2017.304>
11. Marin, I., Gotovac, S., Russo, M., Božić-Štulić, D.: The effect of latent space dimension on the quality of synthesized human face images. *J. Commun. Softw. Syst.* **17**(2), 124–133 (2021). <https://doi.org/10.24138/jcomss-2021-0035>
12. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference (SciPy 2015), vol. 8, pp. 18–24 (2015). <https://doi.org/10.25080/Majors-7b98e3ed-003>
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (NIPS 2013), pp. 3111–3119 (2013)
14. Offert, F.: Latent deep space: generative adversarial networks (GANs) in the sciences. *Media+Environment* **3**(2), 1–20 (2021). <https://doi.org/10.1525/001c.29905>
15. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: WaveNet: A generative model for raw audio (2016). [arXiv:1609.03499v2](https://arxiv.org/abs/1609.03499v2), <https://arxiv.org/abs/1609.03499>
16. Pasini, M.: MelGAN-VC: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms (2019). [arXiv:1910.03713v2](https://arxiv.org/abs/1910.03713v2), <https://arxiv.org/abs/1910.03713>
17. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: 4th International Conference on Learning Representations (ICLR 2016). San Juan, Puerto Rico (2016)

18. Tahiroğlu, K., Kastemaa, M., Koli, O.: GANSpaceSynth: A hybrid generative adversarial network architecture for organising the latent space using a dimensionality reduction for real-time audio synthesis. In: Conference on AI Music Creativity, pp. 1–10. Graz, Austria (2021). <https://doi.org/10.5281/zenodo.5137902>
19. Van, T.P., Nguyen, T.M., Tran, N.N., Nguyen, H.V., Doan, L.B., Dao, H.Q., Minh, T.T.: Interpreting the latent space of generative adversarial networks using supervised learning. In: 2020 International Conference on Advanced Computing and Applications (ACOMP 2020), pp. 49–54. Quy Nhon, Vietnam (2020). <https://doi.org/10.1109/ACOMP50827.2020.00015>
20. Voynov, A., Babenko, A.: Unsupervised discovery of interpretable directions in the GAN latent space. In: 37th International Conference on Machine Learning (ICML 2020), pp. 9786–9796 (2020)
21. Yang, G., Yang, S., Liu, K., Fang, P., Chen, W., Xie, L.: Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech. In: 2021 IEEE Spoken Language Technology Workshop (SLT 2021), pp. 492–498. Shenzhen, China (2021). <https://doi.org/10.1109/SLT48900.2021.9383551>

Chapter 2

Graph Neural Networks for Topological Feature Extraction in ECG Classification



Kamyar Zeinalipour and Marco Gori

Abstract The electrocardiogram (ECG) is a dependable instrument for assessing the function of the cardiovascular system. There has recently been much emphasis on precisely classifying ECGs. While ECG situations have numerous similarities, little attention has been paid to categorizing ECGs using graph neural networks. In this study, we offer three distinct techniques for classifying heartbeats using deep graph neural networks to classify the ECG signals accurately. We suggest using different methods to extract topological features from the ECG signal and then using a branch of the graph neural network named graph isomorphism network for classifying the ECGs. On the PTB Diagnostics data set, we tested the three proposed techniques. According to the findings, the three proposed techniques are capable of making arrhythmia classification predictions with the accuracy of 99.38, 98.76, and 91.93%, respectively.

2.1 Introduction

One of the biophysical signals that may be monitored using special equipment from the human body is electrocardiography (ECG). It stores crucial information about how the heart functions and whether it is affected by aberrant conditions. Cardiologists and medical practitioners frequently utilize ECG to check heart health. The difficulty in recognizing and classifying diverse waveforms and morphologies in ECG signals, like with many other time-series data, is the fundamental issue with manual analysis. This task would take a human a long time to complete and is prone

K. Zeinalipour (✉) · M. Gori

Department of Information engineering and mathematics, University of Siena, Via Roma, 56, Siena 53100, Italy

e-mail: kamyar.zeinalipour2@unisi.it

M. Gori

e-mail: marco.gori@unisi.it

M. Gori

31A Côte d'Azur, Université Côte d'Azur, 28 Avenue de Valrose, Nice 06000, France