POLITECNICO DI TORINO

MASTER'S DEGREE
in
MATHEMATICAL ENGINEERING

Politecnico
di Torino

MASTER'S DEGREE THESIS

Wavelet Scattering Transform. Mathematical Analysis and Applications to VIRGO Gravitational Waves Data

Supervisors	Candidate
Prof. LAMBERTO RONDONI	ALESSANDRO LICCIARDI
Dr. DAVIDE CARBONE	

JULY 2023

Abstract

In machine learning (ML) algorithms for image and signal classification it is fundamental to build architectures which enable to learn and extract discriminative features regardless to translation, rotation, small stochastic perturbations and stretches which affect real-world data. In this thesis we investigate the properties of the Wavelet Scattering Transform (WST), a mathematical technique recently proposed in the context of signal processing and ML. We explore the application of WST in distinguishing glitches that impact the accurate detection of gravitational waves (GW). The first detection of a GW took place in 2015 by the LIGO-Virgo collaboration. This thesis is done in collaboration with the Istituto Nazionale di Fisica Nucleare (INFN), in the *interTwin* project framework. In recent years the use of ML to improve analysis and manipulation of GW signals has increased significantly.

This work investigates the possibility of applying WST to extract discriminative features from glitches, i.e. noises captured by gravitational wave detectors not related to astrophysical phoenomena. Physical issues, e.g. non-homogeneity in centering the time-series over the energy spike, environmental noise and deformations caused by the high sensitivity of the interferometer are problems that critically plague the state-of-the-art techniques. As a side note, our analysis is developed on the data provided by INFN because of the privacy policy of LIGO-Virgo experiments. This work is organized in three sections.

The first section provides a formal definition of the wavelet scattering operator and a comparison to other standard signal representations, with particular regards to three fundamental properties: *translation invariance*, *stability to additive perturbations*, *local invariance to continuous stretches* (in terms of Lipschitz continuity to the action of C^2-diffeomorphisms). These properties are crucial for classification and discrimination tasks, as natural elements of the same class exhibit slight variations. A reliable representation should keep similar images close in the target space, even when they undergo minor geometric modifications. This proximity facilitates their effective use in ML applications. However, standard methods like Q-transform based on Short Time Fourier Transform (STFT) lack these properties, leading to distorted representations that are less accurate in capturing essential features.

The second part of the thesis focuses on the experiments and results on the Free Spoken Digits dataset, analyzing the effect on classification algorithms of WST, compared to the state-of-the-art representation technique based on STFT. On average the algorithms trained in the wavelet scattering domain outperformed the same classifiers trained in the Fourier domain, proving from an empirical point of view that WST enhances discriminative features across the different classes. These

preliminary studies and results on the Free Spoken Digits were crucial to test and evaluate WST based techniques for studying GW, due to the similarity of glitch signals and human voice recordings - both are one-dimensional and non-stationary. The third part focuses on applying WST to GW data collected by VIRGO interferometer, and comparing the results with the state-of-the-art method, based on Q-transform, which, being strictly related to STFT, does not provide translation invariant features. This work is the first that applies WST to gravitational data in order to classify and discriminate different kinds of glitches.
Trials are developed on the dataset provided by INFN, consisting of 855 samples of *scattered-light glitches*. Q-transform is known to perform bad on this class, since it is not able to discriminate the most significant features. An analysis of dispersion has been conducted to compare the behaviour of WST against Q-transform. We develop a technique to estimate dispersion for the two representations based on the Principal Component Analysis (PCA). We compute two different representative spectrograms, one for each operator, and then estimate the sample mean distance and the relative standard deviation. The results are compared in terms of *coefficient of variation*, since, in general, having less dispersed classes significantly enhance the performance of simple classification algorithms. In conclusion the results show that in WST space *scattered light glitches* appear to be much more clustered.
This thesis provides the first steps towards the application of WST for gravitational glitch discrimination. Promising future directions lie in understanding how to leverage WST patterns to whiten astrophysical signals and in developing efficient neural architectures with optimal parametrization to classify glitches belonging to different classes, exploiting invariance properties of WST. This study sets the stage for further advancements in the field, offering potential solutions to improve the accuracy and efficiency of gravitational glitch detection methods. Additionally, an important result worth noting is the impact on computational resources. Lighter architectures are consistently preferred due to their lower training costs.

Acknowledgements

I would like to thank all those who have contributed to the realization of this study. I express my gratitude to Professor Lamberto Rondoni for involving me in this project with INFN, and for always lending a wise and trustful hand throughout the realization of this work. I am sincerely grateful to Doctor Davide Carbone for his patient guidance and meticulousness provided in these months.

A heartfelt appreciation goes to my family who has always supported me throughout my academic path, guiding me with immeasurable love and patiently dispensing invaluable advice. I dedicate to my parents everything I have been able to harvest throughout these years, hoping to always benefit of their wisdom in times of failure and of their joy in times of success. I will always look upon them with infinite admiration.

Table of Contents

List of Tables

List of Figures

Acronyms

AI
Artificial Intelligence

BBH
Binary Black Hole

BNS
Binary Neutron Star

CNN
Convolutional Neural Network

CNRS
Centre National de la Recherche Scientifique

CV
Coefficient of Variation

FFNN
Feed Forward Neural Network

GW
Gravitational Wave

INFN
Istituto Nazionale di Fisica Nucleare

KNN
K Nearest Neighbours

MSE

Mean Squared Error

ML

Machine Learning

MLP

MultiLayer Perceptron

NSBH

Neutron Star-Black Hole

PCA

Principal Component Analysis

STFT

Short Time Fourier Transform

WST

Wavelet Scattering Transform

Introduction

The construction of mathematical representations that extract hidden patterns and similarities from data, e.g. auditory signals or images, is crucial to build reliable and accurate machine learning architectures, enhancing the performances and the robustness of the algorithms.
Signal representation has historically bridged theoretical mathematics, such as functional analysis, with various applied fields including engineering, physics, and more recently, computer science and artificial intelligence. *Joseph Fourier* (1768-1830) in his work *"Théorie Analytique de la Chaleur"* [1] introduced the revolutionary idea that decomposing signals into their frequency components allows the extraction of significant information. Fourier's contribution established the mathematical framework for modern signal processing techniques, such as the *Short Time Fourier Transform* (STFT), proposed in 1946 by D.Gabor [2], and the more recent *Q-transform* (1992) [3]. STFT is a common technique which provides a representation applying the Fourier transform on a signal over short, overlapping time windows. Similarly the Q-transform divides the frequency range into sub-bands with different widths, logarithmically spaced, allowing a more close correspondance to the perception of sound by human auditory system. Representation techniques in signal analysis applying wavelet decomposition started in 1982, when Jean Morlet proposed the *Continuous Wavelet Transform* [4] to analyze seismic signal propagation over different media. Continuous Wavelet Transform provides a representation of a signal applying a convolution with a family of wavelets that are scaled and translated across the time domain.
Before the emergence of machine learning, signal analysis primarily focused on developing analytical techniques and algorithms to process and analyze signals. Dealing with large amount of data led to a change in perspective and set new paradigms for signal representations. In fact deep learning architectures designed for auditory or image classification achieve higher performances when relying on stable representations of the data, in particular translation invariance, stability to continuous stretches and to additive noise. Standard methods based on the Fourier transform lacks of these properties, therefore new techniques have been proposed. Convolutional neural networks aim to acquire an invariant representation, yet

they heavily depend on computationally expensive operations, and it still lacks rigorous proofs that they satisfy the aforementioned properties [5]. In order to offer a solution to the problem of stable local invariant representations the wavelet scattering transform has recently been proposed [6]. The wavelet scattering transform is a non-linear integral operator that allows to extract patterns and similarity between signals, by cascading convolutions with *wavelet functions* and moduli operations.
In this thesis we discuss the mathematical properties of the wavelet scattering transform, and, throughout comprehensive tests and experiments, we introduce a novel approach based on the wavelet scattering operator to overcome the challenge of detecting and characterizing glitches that affect gravitational waves signals. This thesis is made in collaboration with the Istituto Nazionale di Fisica Nucleare (INFN), in the *interTwin* project framework. Our analysis has been conducted on portions of sensible gravitational waves data collected by VIRGO interferometer, kindly provided to us by INFN.
Albert Einstein (1879-1955) predicted gravitational waves in *"Die Feldgleichungen der Gravitation"* [7], while introducing the theory of general relativity. Gravitational waves are ripples in the fabric of space-time caused by the acceleration of massive objects. Einstein's remarkable prediction took several decades for being confirmed experimentally. The historic first detection of gravitational waves, announced by the LIGO-Virgo collaboration in 2015 [8], confirmed Einstein's theory and marked a fundamental milestone in mankind's understanding of the nature of the Universe. Gravitational waves are detected by extremely sensitive interferometers, therefore collected signals are heavily affected by glitches that can compromise the study and analysis of astrophysical signals. Classifying and characterizing these glitches is a primary interest in this research field, in particular INFN, and other leading research institutes, who take part to the *interTwin* project.
The first section of the thesis focuses on formally stating the definition and properties of the wavelet scattering transform, introducing a comprehensive mathematical framework, throughout examples and comparisons with standard techniques. We present the fundamental theoretical elements, providing detailed results that justify the reliability of this representation technique, making it a suitable instrument for machine learning based signal analysis. This section studies the wavelet scattering transform applied to deterministic squared-integrable signals, hinting a further generalization to stochastic processes.
We preliminary developed experiments to test the wavelet scattering transform on auditory data, more precisely for spoken digits classification on Free Spoken Digits dataset [9]. We compared the performances achieved with the wavelet scattering transform to the ones of the same classification algorithms trained with the state-of-the-art representation technique, i.e. STFT. In this experiments the wavelet scattering transform outperformed the STFT, in terms of classification

accuracy, achieving higher scores, also with simple distance based algorithms, e.g. K-Nearest Neighbours.
Analysis on the Free Spoken Digits dataset paved the way for applying the wavelet scattering transform on gravitational waves. In particular we present a comparative analysis with the state-of-the-art representation technique based on the Q-transform. Our experiments were developed on the single class dataset provided by INFN, which consisted of samples of *scattered light glitches*. Scattered light glitches represent unwanted noise appearing in gravitational wave interferometers. These glitches are caused by different sources, e.g. scattering of laser light within the instrument's optics. Scattered light glitches are among the most challenging glitches families to distinguish and characterize. We compare the dispersion of the spectrograms in the wavelet scattering domain with the one in the Q-transform space. Our analysis, developed with Principal Component Analysis (PCA), displayed way less dispersion in the wavelet scattering domain, providing a very promising outcome. All the experiments are developed in Python language. *Kymatio* [10] is a package that provides efficient tools for the wavelet scattering transform computations. As far as gravitational analysis is concerned, the package *GWpy* [11] offers fundamental functions and methods to deal with LIGO/Virgo data. All the codes developed for the experiments presented in this thesis are available at `https://github.com/alelicciardi99/wavelet_scattering_transform_thesis`.
In this work we show that the wavelet scattering transform offers a stable and reliable representation for one-dimensional data. We argue that the wavelet scattering transform can capture high-level discriminative features and fine-grained textures in gravitational wave data, without using over-parametrized architectures, like convolutional neural networks. This significantly reduces the computational effort in the training phase. Moreover due to the increasing interest on decentralized edge computing and low-powered devices, reducing the cost of training algorithms, without losing accuracy, is becoming crucial in machine learning research (Federated Learning [12], Tiny ML [13]). Technological advancements in computational hardware resources have enabled AI algorithms to achieve excellent results. However, as highlighted in *"Training a single AI model can emit as much carbon as five cars in their lifetimes"* published by MIT researchers [14], several studies show how over-parametrized complex networks consume a massive amount of energy, remarking the importance of developing new solutions and architectures that balance the trade-off between accuracy of predictions and energetical cost of training [15]. In this scenario the wavelet scattering transform may provide valuable improvements. The results of this thesis reinforce the idea that building a locally translation invariance representation that is not learnt from data, may significantly reduce training efforts.

Part I

Wavelet Scattering Transform. A Mathematical Analysis

Chapter 1

An Introduction to Signal Representation

In this first part a detailed introduction to the wavelet scattering transform and its properties is provided [6],[16]. In several machine learning tasks, such as classification, it is required to build architectures able to capture distinctive features, highlighting the patterns which characterize the different classes.
In mathematical terms one can see these architectures as suitable operators which map the input signal, a function x defined over a domain $\Omega \subset \mathbb{R}^d$, into another space where, hopefully, features of interests can be extracted. In this work such operator is generally denoted with Φ, therefore the transformed signal becomes $\Phi(x)$. In order to address properly the formulation of the problem it is fundamental to focus on the properties that the operator Φ should satisfy.
Let us consider the following example. Let the signal x be a bi-dimensional image of a cat. The information provided by Φ should ideally be the same for any image showing the same cat, regardless to any translation, rotation, small stretch or additive noise, but it should also be able to discriminate the picture of that same cat from the picture of any other different animal species, hence the representation should be informative. In this first chapter we provide a description from a mathematical point of view of such properties and, furthermore, we show how classical results, such as Fourier transform, fail to reach this strict target, leading to the second chapter, where a detailed description of the wavelet scattering transform is dispensed.

1.1 Signal Representation and Properties Formulation

In this framework we suppose that the signals are functions in the space $L^2(\mathbb{R}^d)$. It is important to denote that the representation Φ induces a metric d over $L^2(\mathbb{R}^d)$, i.e. for any couple of signals x, x' the metric d can be seen as the Euclidean distance among the two transformed signals

$$d(x, x') = \|\Phi(x) - \Phi(x')\|. \tag{1.1}$$

Before proceeding to the description of some well-known signal representations, we aim to point out and clarify the following notions:

- local translation invariance
- stability to additive perturbations
- stability to deformations.

The first property is quite trivial, and can be stated as follows.

Definition 1.1.1. *Let $x(u)$ be a signal, and $c \in \mathbb{R}^d$ be a constant. Being called $x_c(u) = x(u + c)$, we say that a representation Φ is* ***translation invariant*** *if for any $c \in \mathbb{R}^d$*

$$\Phi(x) = \Phi(x_c)\,.$$

Translation invariance is a desirable property in many mathematical and machine learning contexts. It implies that the operator Φ does not depend on the specific position or location of the function's argument but only on its relative differences or distances. This property often simplifies the analysis and computation of the operator, leading to more elegant mathematical formulations and efficient algorithms. Note that global and local translation invariance are slightly different properties, and have different implications in machine learning architectures. Global translation invariance refers to the property of a system where the output remains the same regardless of the input's position, while local translation invariance means that the output may vary depending on the input position within a specific region or context. Machine learning exploits local translation invariance to extract fine-grained features from input signals [5].
Formally defining an operator that is robust and stable to additive perturbations is the same as requiring that Φ is a (*Lipschitz*) contraction, i.e. for any couple of signals x, x' in $L^2(\mathbb{R}^d)$ it exists a positive constant $C < 1$ such that

$$\|\Phi(x) - \Phi(x')\| \leq C\|x - x'\|.$$

So if we consider a signal x and a perturbed version $x+\epsilon$, the representation Φ must ensure us that the noise does not propagate in the transformed space:

$$\|\Phi(x+\epsilon)-\Phi(x)\| \leq \|\epsilon\|.$$

A little more technical is the discussion of the notion of stability to deformations. Let us recall the definition of *diffeomorphism*.

Definition 1.1.2. *Let $\Omega \subset \mathbb{R}^d$ be an open set and let $\tau : \Omega \to \mathbb{R}^d$ be a bijection. Function τ is said to be a $\mathbf{C^r}$-**diffeomorphism** from Ω to $\tau(\Omega)$ if it is differentiable with continuity r-times and if its inverse τ^{-1} is also differentiable with continuity r-times.*

In order to obtain stability to deformations, we need to control the behaviour of the representation Φ under the action of the diffeomorphism

$$u \mapsto u - \tau(u)$$

where τ is an invertible field from $\mathbb{R}^d$ into itself. Let us introduce the following norm over $\mathbf{C}^2$:

$$\|\tau\| := \sup_{u\in\Omega}|\tau(u)| + \sup_{u\in\Omega}|\nabla\tau(u)| + \sup_{u\in\Omega}|H\tau(u)| \tag{1.2}$$

where $\Omega \subset \mathbb{R}^d$ is a compact set, and $H\tau$ is the Hessian matrix. Let us point out that the norm defined in (1.2) measures the amount of elastic deformation resulting from applying τ over Ω. This justifies the following requirement on Φ.

Definition 1.1.3. *The representation Φ is **Lipschitz continuous** to the action of C^2-diffeomorphisms and **locally translation invariant** at scale 2^J if for any compact set $\Omega \subset \mathbb{R}^d$ exists a positive constant C such that for all x in $L^2(\mathbb{R}^d)$ supported in Ω and all $\tau \in \boldsymbol{C}^2$*

$$\|\Phi(L_{[\tau]}x)-\Phi(x)\| \leq C\|x\|\left(2^{-J}\sup_{u\in\Omega}|\tau(u)| + \sup_{u\in\Omega}|\nabla\tau(u)| + \sup_{u\in\Omega}|H\tau(u)|\right)$$

where with the notation $L_{[\tau]}x$ the action of τ on x is denoted.

We observe that for a field τ having $\sup_{u\in\Omega}|\tau(u)|$ much smaller than 2^J the representation stability is controlled by the maximum elasticity of τ, i.e. the term $\sup_{u\in\Omega}|\nabla\tau(u)|$.

1.2 Classic Signal Representations

In this section some of the most common representations used in signal processing are discussed, focusing the attention in showing why such operators do not enjoy the previously stated properties.

1.2.1 Linear Signal Representations

Linear Signal Representation [17] is the simplest way to transform a function, however due to its extreme simplicity it does not provide any further information.

Definition 1.2.1. *For a signal $x \in L^2(\mathbb{R}^d)$ its linear representation is given by*

$$\Phi(x) = \sum_{j=1}^{N} c_j \phi_j \tag{1.3}$$

where $\{\phi_j\}_{j=1}^N \subset L^2(\mathbb{R}^d)$ is an orthonormal set, and for each $j = 1, \dots, N$ the coefficient c_j is the projection of x onto ϕ_j, namely $= c_j = \langle x, \phi_j \rangle$.

As previously mentioned such representation does not add any kind of information, furthermore the induced metric is not stable to local deformations. One can prove that for any couple of $L^2(\mathbb{R}^d)$ functions x, y the distance

$$d(\Phi(x), \Phi(y)) = \left\| \sum_{j=1}^{N} \langle x - y, \phi_j \rangle \phi_j \right\|$$

is equivalent to the Euclidean norm over $L^2(\mathbb{R}^d)$ which is not Lipshitz continuous to the action of C^2-diffeomorphism, therefore linear signal metric inherits the property.

Remark 1.2.1. *A special case of linear representations is provided by the Fourier series, in such case $N = \infty$, and the set $\{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis of $L^2(\mathbb{R}^d)$. Due to completeness properties of orthonormal basis of Hilbert spaces, i.e.*

$$\overline{span\{\phi_j : j = 1, \dots, \infty\}} = L^2(\mathbb{R}^d) \ ,$$

the Fourier series of any function $x \in L^2(\mathbb{R}^d)$ converges to x, therefore the equivalence of the metrics is trivial.

1.2.2 Non-Linear Signal Representations

As previously illustrated linear signal representations are not helpful to extract discriminative features and do not result robust to small deformations, in the sense that the norm of the error can not be controlled and may explode. The vast majority of methods used in signal processing and machine learning exploit non-linearity. Historically the most notorious examples are based on the Fourier transform, e.g. Fourier modulus and power spectrum [18], [19]. With the uprising of machine learning and the continuous increase in computational power have led

to the construction of more powerful (and costly) techniques, such as convolutional neural networks [20]. Kernel methods [21] constitute a common technique in machine learning. Kernel methods utilize kernel functions to map signals into a high-dimensional feature space, enabling effective signal representation and analysis [16]. However these methods do not enjoy of all the three fundamental properties which we previously required. Note that also for convolutional filters such properties are not proved to be true [5], yet those methods are more heuristically constructed [20]. In the following pages we analyze Fourier transform based methods.

Fourier Analysis and Autocorrelation

Fourier analysis [19], [18] consists in a series of well established techniques for signal analysis. Let us recall the definition of continuous Fourier transform.

Definition 1.2.2. *Let x be a signal in $L^2(\mathbb{R}^d)$. We define the Fourier transform of x, denoted by $\hat{x}(\omega)$, the function*

$$\hat{x}(\omega) = \int_{\mathbb{R}^d} x(u)e^{-i\omega\cdot u}\, du \tag{1.4}$$

To simplify the notation we will omit the subscript $\mathbb{R}^d$ under the integral operator.
Let us observe that Fourier transform is a linear operator, due to the linearity of the integral. However in order to show non-linear Fourier based methods, we report some notable results regarding the Fourier transform in this section.
Let us recall that Fourier transform enjoys the **isometric** property, i.e. for any signal in x in $L^2(\mathbb{R}^d)$ the equality $\|x\| = \|\hat{x}\|$ holds true [17]. This isometry is also referred in literature as *Plancherel identity* [19].
However, as far as our task is concerned, we immediately see that Fourier representation is not translation invariant, as shown by the following trivial example.

Example 1.2.1. Let us consider $x \in L^2(\mathbb{R}^d)$ and let us fix a shift $c \in \mathbb{R}^d$. If we denote the shifted version of x by $x_c = x(u - c)$, we observe that $\hat{x}_c$ differs from the non-shifted transform $\hat{x}$. Indeed

$$\hat{x}_c = \int x_c(u)e^{-i\omega\cdot u}\, du = \int x(u - c)e^{-i\omega\cdot u}\, du$$

and imposing $v := u - c$ we get

$$\int x(v)e^{-i\omega\cdot(v+c)}\, du = e^{-ic\cdot\omega} \int x_c(v)e^{-i\omega\cdot v}\, dv = e^{-ic\cdot\omega}\hat{x}(\omega).$$

However the previous example suggests that if we took the modulus of the Fourier transform we would get a signal representation which is invariant under any shift. Indeed, let $\Phi(x) = |\hat{x}|$ be the so called *Fourier modulus*, we easily see that

$$\Phi(x_c) = |\hat{x}_c| = |e^{-ic\cdot\omega}\hat{x}| = |\hat{x}| \ .$$

Nevertheless Fourier modulus does not enjoy the property of stability to small perturbations (in terms of Lipschitz strict contractions), as we portrait in the following example.

Example 1.2.2. For the sake of simplicity we assume that the domain of the signal x is one-dimensional, e.g. time domain. Let $\theta(u)$ be a symmetric lowpass filter centered in 0. Let us define $x(u) = e^{i\xi u}\theta(u)$ and $x_\tau(u) = x((1+s)u)$, i.e. a perturbed version of x by a small positive parameter s.
We want to show that as s tends towards zero, the representation Φ given by Fourier modulus is not a Lipschitz strict contraction, i.e. the following inequality does not hold true

$$\|\Phi(x_\tau) - \Phi(x)\| \leq \|x_\tau - x\| \,. \tag{1.5}$$

Let us compute the Fourier transform of the signals under analysis:

$$\hat{x}(\omega) = \int e^{i\xi u}\theta(u)e^{-i\omega u}\, du = \int \theta(u)e^{-i(\omega-\xi)u}\, du = \hat{\theta}(\omega - \xi) \tag{1.6}$$

where $\hat{\theta}$ is the Fourier transform of the lowpass filter, and, imposing $(1+s)u =: v$, we get

$$\begin{aligned}\hat{x}_\tau(\omega) = \int x((1+s)u)e^{-i\omega u}\, du &= \frac{1}{s+1}\int x(v)\exp\left\{-i\left(\frac{\omega}{s+1}\right)v\right\}\, dv \\ &= \frac{1}{s+1}\hat{x}\left(\frac{\omega}{s+1}\right)\end{aligned}$$

which can be written in terms of θ, due to relation (1.6) as

$$\hat{x}_\tau(\omega) = \frac{1}{s+1}\hat{\theta}\left(\frac{\omega - (s+1)\xi}{s+1}\right) \,.$$

So we got that $\hat{x}$ is a lowpass filter centered in $\omega_0 = \xi$, while $\hat{x_\tau}$ is centered in $\omega_0^\tau = (s+1)\xi$.
Since x_τ is the same signal as x, just displaced by the positiven parameter s, it is trivial to show that in the limit $s \to 0^+$, the right-hand side of equation (1.5) tends to zero. Let us focus on the left-hand term of (1.5). We know that the variance of a signal is a measure of how the energy is displaced around the central frequency, hence for the lowpass filter θ we get

$$\sigma_\theta^2 = \int |\omega|^2 |\hat{\theta}(\omega)|^2\, d\omega \,,$$

as for the signal x, indeed, with the help of a trivial change of variable

$$\sigma_x^2 = \int |\omega - \omega_0|^2 |\hat{x}(\omega)|^2\, d\omega \;= \int |\omega - \xi|^2 |\hat{\theta}(\omega - \xi)|^2\, d\omega \;= \int |\nu|^2 |\hat{\theta}(\nu)|^2\, d\nu = \sigma_\theta^2 \,.$$

The same thing is done on the displaced signal x_τ

$$\begin{aligned}\sigma_{x_\tau}^2 &= \int |\omega - \omega_0^\tau|^2 |\hat{x_\tau}(\omega)|^2 \, d\omega \; = \int |\omega - (s+1)\xi|^2 |(s+1)^{-1}\hat{x}((s+1)^{-1}\omega))|^2 \, d\omega \\ &= \int |(s+1)^{-1}\omega - \xi|^2 |\hat{x}((s+1)^{-1}\omega))|^2 \, d\omega = (s+1)\int |\nu - \xi|^2 |\hat{x}(\nu)| \, d\nu \\ &= (s+1)\sigma_x^2 = (s+1)\sigma_\theta^2 \, .\end{aligned}$$

We notice that the distance of the central frequencies $\omega_0^\tau - \omega_0 = s\xi$ is way much larger than the frequency spread $\sigma_{x_\tau} + \sigma_x = (\sqrt{1+s}+1)\sigma_\theta$. This means that in the limit the supports can be considered disjoint, and hence, as $\|x_\tau - x\|$ goes to zero, the term $\|\Phi(x_\tau) - \Phi(x)\|$ tends to a positive value, therefore inequality (1.5) does not hold. Hence we conclude that the Fourier modulus does not satisfy the non-expansive property required.

Beside Fourier transform and Fourier modulus, another classical tool for signal processing is the *auto-correlation function*. Due to its fame it may be worthwhile to discuss an example that shows why it is not suitable for our pre-processing task. Again this representation is not a Lipschitz strict contraction, since it induces a metric equivalent to the one of the Fourier modulus.

Definition 1.2.3. *Let $x \in L^2(\mathbb{R}^d)$ be a signal, the auto-correlation function of x is defined as*

$$R_x(v) = x \star x^*(-v) = \int x(u)x^*(u-v) \, du$$

where x^* is the conjugate complex of x and $\star$ denotes the continuous convolution operation.

In order to prove the metric equivalency we shall introduce a fundamental property relating the auto-correlation function of a signal x and its spectral density, i.e. the squared modulus of $\hat{x}$. The property is a special case of the *Wiener-Kinchin Theorem*[22], stated in the following proposition.

Proposition 1.2.1. *For any signal $x \in L^2(\mathbb{R}^d)$ and for any $\omega \in \mathbb{R}^d$*

$$\hat{R}_x(\omega) = |\hat{x}(\omega)|^2 \quad . \tag{1.7}$$

Here we provide the proof for the one-dimensional case.

Proof. The Fourier transform of the auto-correlation can be written as

$$\hat{R}_x(\omega) = \int R_x(v) e^{-i\omega v} \, dv = \int \left(\int x(u)x^*(u-v) \, du \right) e^{-i\omega v} \, dv$$

since the exponential term does not depend on u, it can be brought inside the inner integral and, adding and subtracting in $\pm i\omega u$, we get

$$\int \left(\int x(u)x^*(u-v)e^{-i\omega v}\,du \right)\, dv = \int\int x(u)x^*(u-v)e^{-i\omega(v-u+u)}\,du\,dv\ .$$

The above expression can easily be rearranged into

$$\int\int x(u)e^{-i\omega u}x^*(u-v)e^{-i\omega(v-u)}\,du\,dv$$

and setting $t := u - v$, it results to be equal to

$$\begin{aligned}\left(\int x(u)e^{-i\omega u}\,du \right) \left(-\int x^*(t)e^{i\omega t} \right) =& \hat{x}(\omega) \left(-\int x^*(t)[e^{-i\omega t}]^* \right) \\ =& \hat{x}(\omega) \left(\int x(t)e^{-i\omega t} \right)^* \\ =& \hat{x}(\omega)\hat{x}^*(\omega) = |\hat{x}(\omega)|^2\ .\end{aligned}$$

□

Example 1.2.3. Let us consider the signals x and x_τ introduced in Example 1.2.2. Auto-correlation function induces the metric $\|R_x - R_{x_\tau}\|$. Due to the isometric property and linearity of Fourier transform the following equivalences hold

$$\|R_x - R_{x_\tau}\| = \|\widehat{R_x - R_{x_\tau}}\| = \|\widehat{R_x} - \widehat{R_{x_\tau}}\|\,.$$

Applying relation (1.7) we get

$$\|\widehat{R_x} - \widehat{R_{x_\tau}}\| = \|\,|\hat{x}|^2 - |\hat{x_\tau}|^2\|\,.$$

Hence one can conclude the proof following the argument of Example 1.2.2 that also the auto-correlation function is not stable to small perturbations.

Convolutional Neural Networks

Convolutional Neural Networks [23][24] are one of the most used family of algorithms in deep learning. These architectures achieve their best performances in classification tasks involving structured complex data, such as images. A convolutional neural network consists of two main blocks, one extracts meaningful patterns via non-linear operators which are obtained by cascading trainable filter banks, pooling operators and *sigmoid* or *ReLU* activation functions [25], the other is a fully connected neural network outputting the posterior probability of the input data belonging to each class. The weights and the filters are learnt by minimizing the loss function via optimization methods, such as *Stochastic Gradient Descent*

[26] or *Adam* [27].
The convolutional layers aim to build (learning the filters from the data) a (locally) translation invariant representation of the data. Even though it has not been proved that this representation enjoys such property, empirical results have shown that the deeper the network the better it learns representations that get closer to be translation invariant [28]. However deeper network are heavy to train and suffer of over-parametrization (state-of-the-art GPipe has more than 500 milion parameters and reaches the top-accuracy of 84.3% on ImageNet dataset [29]).

Chapter 2

Introducing Wavelet Scattering Transform

In this chapter we explain the construction of the *wavelet scattering transform* [6]. After a review of wavelets and their properties, we delve into defining the wavelet scattering transform, and how, by the simple introduction of non-linearity, this operator overcomes the issues characterizing its predecessor, the *Littlewood-Paley wavelet transform* [30] . Many results proved for the Littlewood-Paley wavelet operator are helpful to demonstrate properties that characterizes the wavelet scattering transform, therefore a detailed description of the former is provided [16]. In the last section of the chapter important results about the wavelet scattering transform induced metric are shown [31].

2.1 Wavelet Functions

As suggested by its naming, wavelet functions play a central role in the in the computation of the wavelet scattering transform. Therefore choosing in a proper way wavelet function affects the behaviour and the properties of scattering operators. In many applications the most used are *Gaussian* and *Morlet* wavelets.
Let us introduce a signal $\psi \in L^2(\mathbb{R}^d)$, usually referred to as *mother wavelet* and let $\{a^j\}_{j\in\mathbb{Z}}$ be a sequence of scale factors, taking $a > 1$ (e.g. in audio signal processing a is often smaller than $2^{1/8}$). Furthermore, let $r \in G$ be a rotation in $\mathbb{R}^d$, where G denotes a discrete rotation group. Wavelets are obtained applying a rescaling a^j and a rotation $r \in G$, therefore we will use the following definition

$$\psi_{a^j r}(u) = a^{-dj}\psi(a^{-j}r^{-1}u). \tag{2.1}$$

The latter wavelets enjoy the following properties.

Proposition 2.1.1. *For any scale factor $a^j \in a^{\mathbb{Z}}$ and for any rotation $r \in G$:*

1. *the norm of the mother wavelet ψ is preserved in $L^1(\mathbb{R}^d)$, i.e.*

$$\|\psi_{a^j r}\|_1 = \|\psi\|_1 .$$

2. *the Fourier transform of a wavelet satsfies at any given frequency ω the following relation*

$$\widehat{\psi_{a^j r}}(\omega) = \hat{\psi}(a^j r\omega).$$

Proof. Let us prove the first statement.

1. Recalling (2.1) we can write L^1-norm of a wavelet $\psi_{a^j r}$ as follows

$$\|\psi_{a^j r}\|_1 = \int |\psi_{a^j r}(u)|\, du = \int a^{-dj} |\psi(a^{-j} r^{-1} u)|\, du .$$

Let us define an auxiliary variable $v := a^{-j} r^{-1} u$, which implies $u = r(a^j v)$. Since rotation matrices determinant equals ± 1, we may find that the modulus of the Jacobian determinant of the inverse mapping $a^j r$ equals a^{dj}. This substitution allow us to write the above integral as

$$\int a^{-dj} a^{dj} |\psi(v)|\, dv = \|\psi\|_1 .$$

2. Let us write the Fourier transform for a wavelet $\psi_{a^j r}$.

$$\widehat{\psi_{a^j r}}(\omega) = \int \psi_{a^j r}(u) e^{-i\omega\cdot u}\, du = \int a^{-dj} \psi(a^j r^{-1} u) e^{-i\omega\cdot u}\, du .$$

Applying the aforementioned change of variable, and exploiting linearity, we get

$$\widehat{\psi_{a^j r}}(\omega) = \int a^{-dj} a^{dj} \psi(v) e^{-i\omega\cdot r(a^j v)}\, dv = \int \psi(v) e^{-i(a^j r\omega)\cdot v}\, dv = \hat{\psi}(a^j r\omega) .$$

□

Let us introduce the following notation, that will be used throughout this work. We define a *scale-rotation* operator $\lambda \in a^{\mathbb{Z}} \times G$, which is defined as $\lambda = a^j r$. The scale factor will be referred to as the modulus of the latter operator, i.e. $|\lambda| = a^j$. Hence instead of writing $\psi_{a^j r}$, we will denote

$$\psi_\lambda(u) = a^{-dj} \psi(\lambda^{-1} u) .$$

A common and interesting choice for the mother wavelet ψ is to take $\psi(u) = e^{i\eta u}\theta(u)$, where θ is a lowpass filter centered in the null frequency with real Fourier transform whose support is in the order of magnitude of π. As seen in the previous section, multiplying a filter θ for a complex exponential function is the equivalent of doing a translation in the frequency domain, more precisely

$$\hat{\psi}(\omega) = \hat{\theta}(\omega - \eta)\,.$$

Hence, if we consider a wavelet ψ_λ, its Fourier transform, according to statement 2. of Proposition 2.1.1, would be

$$\hat{\psi_\lambda}(\omega) = \hat{\psi}(a^j r\omega) = \hat{\psi}(\lambda\omega) = \hat{\theta}(\lambda\omega - \eta)\,.$$

Let us observe that factorizing the operator λ we get $\hat{\psi_\lambda}(\omega) = \hat{\theta}(\lambda(\omega - \lambda^{-1}\eta))$, which implies that:

- it is centered in the frequency $\lambda^{-1}\eta$
- the bandwidth it is of the order $\pi|\lambda|^{-1} = a^{-j}\pi \simeq a^{-j}$.

Example 2.1.1. Morlet wavelets are widely used in signal processing. A Morlet wavelet ψ is given by

$$\psi(u) = \alpha(e^{iu\cdot\xi} - \beta)e^{-|u|^2/(2\sigma^2)}\,. \tag{2.2}$$

The parameters $\alpha, \xi, \beta, \sigma$ allow to shape the wavelet in order to reach different features and enjoy certain properties. For instance a common choice is $\beta \ll 1$ in order to get zero average, i.e. $\langle\psi\rangle = \int \psi(u)\,du = 0$. Common values used in image classification for the phase and the spread are respectively $\xi = 3\pi/4$ and the spread $\sigma = 0.85$ [16].

2.2 From The Littlewood-Paley Wavelet Transform to the Wavelet Scattering Transform

Littlewood-Paley transform is a mathematical operator that allows to study the localization properties of functions in the frequency domain and their relationship to various function spaces [30].

Definition 2.2.1. *Let $\psi \in L^2(\mathbb{R}^d)$ be a mother wavelet, and let $\lambda \in a^{\mathbb{Z}} \times G$ be a scale-rotation operator. For any $u \in \mathbb{R}^d$ we define the* ***Littlewood-Paley transform*** *of the signal $x \in L^2(\mathbb{R}^d)$ as*

$$W_\lambda(x)(u) := x \star \psi_\lambda(u) = \int x(v)\psi_\lambda(u - v)\,dv\,.$$

Remark 2.2.1. *Let x is a real signal, i.e. $x(u) = x^*(u)$ and suppose that $\hat{\psi}$ is real (if and only if ψ is an Hermitian symmetric function, i.e. $\psi(-u) = \psi^*(u)$), we can easily state that $(W_\lambda x)^* = W_{-\lambda} x$.*
Indeed one can write

$$\begin{aligned}(W_\lambda x)^* &= \left(\int x(v) a^{-dj} \psi(\lambda^{-1}(u-v))\, dv \right)^* \\ &= \int x^*(v) a^{-dj} \psi^*(\lambda^{-1}(u-v))\, dv \\ &= \int x(v) a^{-dj} \psi(-\lambda^{-1}(u-v))\, dv \\ &= W_{-\lambda} x\,.\end{aligned}$$

This allows to state that in the considered case it is possible to define the equivalency class of discrete rotations $G^+ := G/\{\pm 1\}$, which assimilates both r and $-r$.

We are now set to start talking about the wavelet transform. First thing to say is that, for a fixed $J \in \mathbb{Z}$, a wavelet transform takes only into account scale-rotations λ such that $|\lambda| < 2^J$, i.e. all $j \in \mathbb{Z}$ satisfying $a^j < 2^J$. Moreover, in order to recover the low frequencies not captured by the wavelets, it is introduced a lowpass filter ϕ_J having a spatial support proportional to 2^J, which can be written as $\phi_J(u) = 2^{-dJ}\phi(2^{-J}u)$.

Definition 2.2.2. *The **(Littlewood-Paley) wavelet transform** of a signal $x \in L^2(\mathbb{R}^d)$ at scale 2^J, for some $J \in \mathbb{Z}$ is defined as*

$$\mathcal{W}_J x := \{x \star \phi_J, (W_\lambda x)_{\lambda \in \Lambda_J}\}\,,$$

where $\Lambda_J = \{\lambda = a^j r,\ r \in G^+, |\lambda| < 2^J\}$. Moreover for such representation it is possible to define the following norm

$$\|\mathcal{W}_J x\|^2 := \|x \star \phi_J\|^2 + \sum_{\lambda \in \Lambda_J} \|W_\lambda x\|^2\,. \tag{2.3}$$

Due to linearity of integral convolution, Littlewood-Pailey wavelet transform is a linear operator, more specifically defines the following map

$$\mathcal{W}_J : x \in L^2(\mathbb{R}^d) \longmapsto \mathcal{W}_J x \in L^2(\mathbb{R}^d) \times \{L^2(R^d)\}^{|\Lambda_J|}$$

Mathematicians Littlewood and Paley proved a very interesting results, characterizing the boundaries for the norm introduced in relation (2.3), which we report in the following proposition.

Proposition 2.2.1. *(Littlewood-Paley Condition) If it exists $\epsilon \geq 0$ such that for a.e. $\omega \in \mathbb{R}^d$ and for all $J \in \mathbb{Z}$ satisfies*

$$(1-\epsilon) \leq |\hat{\phi}(2^J\omega)|^2 + \frac{1}{2}\sum_{j\leq J}\sum_{r\in G}|\hat{\psi}(a^j r\omega)|^2 \leq 1 \tag{2.4}$$

hence the square into which $\mathcal{W}_J$ maps x is bounded, more precisely the following inequality holds

$$(1-\epsilon)\|x\|^2 \leq \|\mathcal{W}_J x\|^2 \leq \|x\|^2 \,. \tag{2.5}$$

Proof. We notice that inequality (2.5) and condition (2.4) are much related. For the sake of simplicity, let us define the functional

$$\Xi(\omega) := |\hat{\phi}(2^J\omega)|^2 + \frac{1}{2}\sum_{j\leq J}\sum_{r\in G}|\hat{\psi}(a^j r\omega)|^2 \,.$$

In terms of Ξ, the Littlewood-Paley hypothesis can be written as

$$(1-\epsilon) \leq \Xi(\omega) \leq 1$$

by multiplying each term for $|\hat{x}(\omega)|^2$ and integrating over ω, we get

$$(1-\epsilon)\int |\hat{x}(\omega)|^2\, d\omega \leq \int \Xi(\omega)|\hat{x}(\omega)|^2\, d\omega \leq \int |\hat{x}(\omega)|^2\, d\omega \,,$$

since Fourier transform is an isometric operator the condition becomes

$$(1-\epsilon)\|x\|^2 \leq \int \Xi(\omega)|\hat{x}(\omega)|^2\, d\omega \leq \|x\|^2 \,.$$

Now it is left to prove that the integral $\int \Xi(\omega)|\hat{x}(\omega)|^2\, d\omega$ equals actually $\|\mathcal{W}_J x\|^2$. Due to linearity of integral operator, the following identity holds

$$\int \Xi(\omega)|\hat{x}(\omega)|^2\, d\omega = \int |\hat{\phi}(2^J\omega)\hat{x}|^2\, d\omega + \frac{1}{2}\sum_{j\leq J}\sum_{r\in G}\int |\hat{\psi}(a^j r\omega)\hat{x}(\omega)|^2\, d\omega \,.$$

If we focus on the first summand, noticing that for any $\omega \in \mathbb{R}^d$ the filter $\hat{\phi}_J(\omega) = \hat{\phi}(2^J\omega)$, where $\phi_J(u) = 2^{-dJ}\phi(2^{-J}u)$, we obtain

$$\int |\hat{\phi}(2^J\omega)\hat{x}|^2\, d\omega = \int |\hat{\phi}_J(\omega)\hat{x}|^2\, d\omega = \int |\phi_J \star x(u)|^2\, du = \|\phi_J \star x\|^2 \,.$$

Let us focus on the second summand. Recalling that $\Lambda_J = \{\lambda = a^j r \in a^{\mathbb{Z}} \times G^+ : |a^j| \leq 2^J\}$, we have that

$$\begin{aligned}\frac{1}{2}\sum_{j\leq J}\sum_{r\in G}\int |\hat{\psi}(a^j r\omega)\hat{x}(\omega)|^2\, d\omega &= \sum_{j\leq J}\sum_{r\in G^+}\int |\hat{\psi}(a^j r\omega)\hat{x}(\omega)|^2\, d\omega \\ &= \sum_{\lambda\in\Lambda_J}\int |\hat{\psi}_\lambda(\omega)\hat{x}(\omega)|^2\, d\omega\end{aligned}$$

which, using convolution property of Fourier transform, becomes

$$\sum_{\lambda\in\Lambda_J}\int |\psi_\lambda \star x(u)|^2\,du = \sum_{\lambda\in\Lambda_J}\|W_\lambda x\|^2.$$

This allows to conclude

$$\int \Xi(\omega)|\hat{x}(\omega)|^2\,d\omega = \|\mathcal{W}_J x\|^2\,,$$

hence (2.5) follows. □

Remark 2.2.2. *Morlet wavelet, defined in* (2.2)*, alongside a Gaussian low-pass filter, i.e.*

$$\phi(u) = \frac{1}{2\pi\sigma^2}\exp\left(\frac{-|u|^2}{2\sigma^2}\right)$$

with $\sigma = 0.7$*, satisfy Littlewood-Pailey condition, Proposition 2.2.1, with* $\epsilon = 0.25$*.*

Let us point out that if the wavelet ψ and the low-pass filter ϕ are chosen in a way such that

$$|\hat{\phi}(2^J\omega)|^2 + \frac{1}{2}\sum_{j\le J}\sum_{r\in G}|\hat{\psi}(a^j r\omega)|^2 = 1$$

the operator $\mathcal{W}$ is isometric, since $\epsilon = 0$, therefore the energy of the signal is preserved by the Littlewood-Pailey representation, namely

$$\|x\| = \|\mathcal{W}_J x\|\,.$$

Another interesting property is that Fourier transform of mother wavelet ψ computed in the null frequency $\omega = 0$, is zero, namely $\hat{\psi}(0) = 0$. This result implies that ψ has null average, indeed

$$\langle\psi\rangle = \int \psi(u)\,du = \int \psi(u)e^{-i\omega\cdot u}\,du\Big|_{\omega=0} = \hat{\psi}(0) = 0\,.$$

However Littlewood-Pailey wavelet transform is not translation invariant, as for any constant $c\in\mathbb{R}^d$, and for any $\lambda\in\Lambda_J$

$$W_\lambda x_c = \int x_c(v)\psi_\lambda(u-v)\,dv = \int x(c+v)\psi_\lambda(u-v)\,dv \neq \int x(v)\psi_\lambda(u-v)\,dv = W_\lambda x\,.$$

Furthermore averaging the wavelet coefficients $W_\lambda x$ does not produce further information since $\langle\psi_\lambda\rangle = 0$. However if we average the modulus of wavelets coefficients we get a non negative value, hence we do not lose information.
Such result leads us to consider for each λ a new wavelet decomposition, given by

$$\{|x\star\psi_\lambda|\star\psi_{\lambda'}\}_{\lambda'\in\Lambda_J}\,.$$

It is useful to introduce now the following operator $U[\lambda]$, acting onto a signal x for a given sub-band λ, defined such that

$$U[\lambda]x := |x \star \psi_\lambda| \,.$$

Moreover, given a path of m sub-bands $p = (\lambda_1, \dots, \lambda_m)$, and we define the operator $U[p]$, such that

$$U[p]x := U[\lambda_m] \dots U[\lambda_2]U[\lambda_1]x = |\dots||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \dots | \star \psi_{\lambda_m}|$$

where $U[\emptyset]x = x$.
Let ϕ be a lowpass filter and let us denote by $\phi_{2^J}(u) := 2^{-2J}\phi(2^{-J}u)$, we can finally define the *windowed scattering transform.*

Definition 2.2.3. *For each path $p = (\lambda_1, \dots, \lambda_m)$, where $\lambda_i \in \Lambda_J$ and $x \in L^1(\mathbb{R}^d)$, the **wavelet scattering transform** is defined as*

$$S_J[p]x(u) := U[p]x \star \phi_{2^J}(u) = \int U[p]x(v)\phi_{2^J}(u - v)\, dv \,.$$

Note that sometimes the wavelet scattering transform is also referred as windowed scattering transform [16].

2.3 Elementar Properties of the Wavelet Scattering Transform

Let us introduce the operator $\mathcal{U}_J$, called *propagator*, which operates on a signal x in the following way

$$\mathcal{U}_J x = \{x \star \phi_J, (U[\lambda]x)_{\lambda \in \Lambda_J}\} \,,$$

where we recall that $U[\lambda]x = |x \star \psi_\lambda|$. Moreover, given two paths $p = (\lambda_1, \dots, \lambda_m)$ and $p' = (\lambda'_1, \dots, \lambda'_{m'})$, let us introduce the operation of concatenation, which we denote as $p + p' = (\lambda_1, \dots, \lambda_m, \lambda'_1, \dots, \lambda'_{m'})$.
It trivial to show that for any path p and sub-band λ the following statements hold true:

- $U[\lambda]U[p] = U[p + \lambda]$
- $U[p]x \star \phi_J = S_J[p]x \,.$

Therefore we see that the windowed scattering representation is obtained by cascading the propagator $\mathcal{U}_J$ to each output $U[p]x$, as one easily can see that

$$\mathcal{U}_J(U[p]x) = \{S_J[p]x, (U[p + \lambda]x)_{\lambda \in \Lambda_J}\}$$

Let Λ_J^m be the set of all paths having length m, we define as

$$\mathcal{P}_J := \bigcup_m \Lambda_J^m$$

the set of paths of any length up to scale 2^J.

Proposition 2.3.1. *The propagator $\mathcal{U}_J$ is non-expansive i.e. for any $x, x' \in L^2(\mathbb{R}^d)$*

$$\|\mathcal{U}_J x - \mathcal{U}_J x'\| \leq \|x - x'\|.$$

Moreover if the wavelet decomposition is isometric, also the propagator is isometric.

Proof. The result is trivially implied by the Littlewood-Paley condition (Proposition 2.2.1)

$$\|\mathcal{U}_J x - \mathcal{U}_J x'\|^2 = \|x \star \phi_J - x' \star \phi_J\|^2 + \sum_{\lambda \in \Lambda_J} \| |W_\lambda x| - |W_\lambda x'| \|^2 \leq \|x - x'\|^2 .$$

If the wavelet decomposition preserves the L^2-norm, we get, by Proposition 2.2.1,

$$\|\mathcal{U}_J x\|^2 = \|x \star \phi_J\|^2 + \sum_{\lambda \in \Lambda_J} \| |W_\lambda x| \|^2 = \|x \star \phi_J\|^2 + \sum_{\lambda \in \Lambda_J} \|W_\lambda x\|^2 = \|x\|^2 .$$

□

Let Ω be a set of paths, we can define the norm induced by all the scattering coefficients of the paths p included in Ω, i.e.

$$\|S_J[\Omega]x\|^2 := \sum_{p \in \Omega} \|S_J[p]x\|^2 .$$

Let us consider, for instance, the path set $\mathcal{P}_J$, i.e. the set of paths of any length up to scale 2^J, we can construct the scattering operator $S_J[\mathcal{P}_J]$ by cascading propagetors $\mathcal{U}_J$ as seen before. Since $S_J[\mathcal{P}_J]$ is constructed by concatenating non-expansive Lipschitz operators, it still enjoys non-expansive property, as remarked by the following result.

Proposition 2.3.2. *Windowed scattering trasform operator $S_J[\mathcal{P}_J]$ is non-expansive i.e. for any couple of signals $x, x' \in L^2(\mathbb{R}^d)$*

$$\|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\| \leq \|x - x'\| .$$

Let us point out that since the operator is Lipschitz-continuous and, moreover, non-expansive it results stable to additive noise. The idea behind the proof is quite trivial since non-expansive property is inherited by cascading non-expansive propagator operators.
The following theorem is essential, since it sets necessary conditions that the wavelets must satisfy in order to make the operator $S_J[\mathcal{P}_J]$ an isometry. Let us introduce the following notion of admissibility for the wavelets.

Definition 2.3.1. *We say that a mother wavelet ψ is **admissible** if there exists a vector $\eta \in \mathbb{R}^d$ and a non-negative function $\rho \in L^2(\mathbb{R}^d)$, having $|\hat{\rho}(\omega)| \leq |\phi(\hat{2}\omega)|$ and $\rho(0) = 1$, such that for any $\omega \in \mathbb{R}^d$*

$$\hat{\Psi}(\omega) = |\hat{\rho}(\omega - \eta)|^2 - \sum_{k=1}^{\infty} k \left(1 - |\hat{\rho}(2^{-k}(\omega - \eta))|^2\right)$$

satisfies

$$\inf_{1 \leq |\omega| \leq 2} \sum_{j \in \mathbb{Z}} \sum_{r \in G} \hat{\Psi}(2^{-j} r^{-1} \omega) |\hat{\psi}(2^{-j} r^{-1} \omega)|^2 > 0 \tag{2.6}$$

Now all the elements are set to state the isometry theorem.

Theorem 2.3.1. *If an admissible wavelet ψ satisfies the Littlewood-Pailey condition* (2.2.1) *with $\epsilon = 0$, then for any $x \in L^2(\mathbb{R}^d)$*

$$\lim_{m \to \infty} \|U(\Lambda_J^m)x\|^2 = \lim_{m \to \infty} \sum_{n \geq m} \|S_J[\Lambda_J^n]x\|^2 = 0 \tag{2.7}$$

and

$$\|S_J[\mathcal{P}_J]x\| = \|x\| . \tag{2.8}$$

We do not provide a proof for Theorem 2.3.1, however we can remark some interesting facts. If we interpret $\|U(\Lambda_J^m)x\|^2$ as energy captured by the operator $U(\Lambda_J^m)$, relation (2.7) tells us that this goes to zero as the order of the paths goes to infinity. In other words we do not need to use long paths if we want to capture much information. Indeed we will see that in many applications m is chosen to be at most equal to 3. Furthermore identity (2.8) tells us that the windowed scattered transform, for a suitable choice of the wavelets, is an isometric operator, hence this representation preserves the energy of the signal. In conclusion just few layers (2 or 3) of convolutions are needed to capture high level features and patterns, as the energy of the signal fastly decays to zero as we go deeper in the scattering network [31].

Chapter 3

Stability Properties of the Wavelet Scattering Transform

As previously introduced the principal target of the analysis is to be able to construct a non-linear operator which induces a metric that results **stable to additive noise** and **invariant to local deformations and translations**, in terms of *Lipschitz continuity*. Proposition 2.3.2 states that the wavelet scattering metric is stable to additive noise, therefore the error between two signals, say x and a perturbed version $x' = x + \epsilon$, can be controlled in the transformed space, namely

$$\|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\| \leq \|\epsilon\| .$$

Intuitively this property gains interest in many audio classification problems: small stochastic perturbations of the same signals do not affect much its representations in the scattering transformed space. In the following chapter we try to walk across the principal stages that lead to the statement of the Lipschitz stability to deformations and translation invariance.

3.1 Local Translation Invariance

Proposition 3.1.1. *For any $x, x' \in L^2(\mathbb{R}^d)$ and for any $J \in \mathbb{Z}$*

$$\|S_{J+1}[\mathcal{P}_{J+1}]x - S_{J+1}[\mathcal{P}_{J+1}]x']\| \leq \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\| \tag{3.1}$$

Here we provide a proof in the case $J \in \mathbb{Z}^+$, however the result can be easily generalized for any $J \in \mathbb{Z}$.

Proof. Let J be any positive integer. We recall that the set $\mathcal{P}_J = \bigcup_m \Lambda_J^m$ and $\mathcal{P}_{J+1} = \bigcup_m \Lambda_{J+1}^m$; since $\Lambda_J^m = \{(\lambda_1, \dots, \lambda_m) : |\lambda_i| \leq 2^J\}$ and $\Lambda_{J+1}^m = \{(\lambda_1, \dots, \lambda_m) : |\lambda_i| \leq 2^{J+1}\}$ it is easy to see that for any length m, it holds the inclusion $\Lambda_{J+1}^m \supset \Lambda_J^m$, and, moreover, it follows that $\mathcal{P}_{J+1} \supset \mathcal{P}_J$. Let us work on the first term of the inequality

$$\begin{aligned}
\|S_{J+1}[\mathcal{P}_{J+1}]x - S_{J+1}[\mathcal{P}_{J+1}]x'\|^2 &= \left\| \sum_{p \in \mathcal{P}_{J+1}} S_{J+1}[p]x - \sum_{p \in \mathcal{P}_{J+1}} S_{J+1}[p]x' \right\|^2 \\
&= \left\| \sum_{p \in \mathcal{P}_{J+1}} U[p]x \star \phi_{J+1} - \sum_{p \in \mathcal{P}_{J+1}} U[p]x' \star \phi_{J+1} \right\|^2 \\
&= \left\| \left(\sum_{p \in \mathcal{P}_{J+1}} U[p]x - U[p]x' \right) \star \phi_{J+1} \right\|^2 .
\end{aligned}$$

In order to complete the proof we need to state the following claim.

Claim 3.1.1. *For any function $f \in L^2(\mathbb{R}^d)$ and for any scale J the following holds true: $\|f \star \phi_{J+1}\| \leq \|f \star \phi_J\|$.*

Such claim is quite immediate to prove, indeed, if we recall that $\phi_J(u) = 2^{-dJ}\phi(2^{-J}u)$, we easily see that for any ω its Fourier transform $\hat{\phi}_J(\omega) = \hat{\phi}(2^J\omega)$. Hence the scaling does not affect the amplitude of the lowpass filter in the frequency domain, but it just squeezes its support, indeed $Supp(\hat{\phi}_{J+1}) \subset Supp(\hat{\phi}_J)$, and it follows that $Supp(\hat{\phi}_{J+1}\hat{f}) \subset Supp(\hat{\phi}_J\hat{f})$. Since the norm of the convolution equals the norm of the product of the Fourier transforms, the previous inequality is proven. Applying the claim, and the inclusion of the paths set we get

$$\begin{aligned}
\left\| \left(\sum_{p \in \mathcal{P}_{J+1}} U[p]x - U[p]x' \right) \star \phi_{J+1} \right\|^2 &\leq \left\| \left(\sum_{p \in \mathcal{P}_J} U[p]x - U[p]x' \right) \star \phi_{J+1} \right\|^2 \\
&\leq \left\| \left(\sum_{p \in \mathcal{P}_J} U[p]x - U[p]x' \right) \star \phi_J \right\|^2 \\
&= \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\|^2 .
\end{aligned}$$

And such result proves the thesis. □

Let us point out that we just proved that the succession $\{\|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\|\}_J$ is decreasing, and, since its terms are non-negative, it converges to a non negative value l. We show that, provided that the wavelets are admissible, such limit is zero when x' is a translated version of x. Such result is properly stated by the following result.

Theorem 3.1.1. *Let us consider $x \in L^2(\mathbb{R}^d)$ and a translated version $x_c(u) = x(u-c)$, for any constant $c \in \mathbb{R}^d$. Then for any choice of admissible wavelets we have*

$$\lim_{J\to\infty} \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x_c\| = 0 \tag{3.2}$$

Such result is essential, indeed as the scale J grows, the distance between the scattering representation of x and the representation of any translation of it, x_c, goes to zero. We stress again as this property of the wavelet scattering representation is shared also other classical tools, such as the Fourier modulus.

3.2 Stability to The Action of Displacements

We discussed how the most difficult property to reach is the stability under the action of diffeomorphisms and displacements. In terms of pre-processing audio and image data it is fundamental to build a mathematical representation which is able to extract the more information, regardless of shape deformations. For instance in speech recognition the signal of a spoken word may change significantly when told by two different people, however mathematically we can say that the two signals are just stretched versions of each other.
Let us recall the notation we previously introduced. We denote by $\tau : \mathbb{R}^d \to \mathbb{R}^d$ the displacement field having $\|\nabla\tau\|_\infty < 1$, and by

$$\begin{aligned} L[\tau] : L^2(\mathbb{R}^d) &\longrightarrow L^2(\mathbb{R}^d) \\ x(u) &\longmapsto L[\tau]x(u) = x(u-\tau(u)) \end{aligned}$$

the action of the diffeomorphism on a signal $x \in L^2(\mathbb{R}^d)$. We are interested in bounding from above the quantity

$$\|S_J[\mathcal{P}_J]L[\tau]x - S_J[\mathcal{P}_J]x\|\,,$$

hence we need to define the following auxiliary norm

$$\|U[\mathcal{P}_J]x\|_1 = \sum_{m\geq 0} \|U[\Lambda_J^m]x\|\,.$$

Theorem 3.2.1. *There exists a constant C such that every $x \in L^2(\mathbb{R}^d)$ having $\|U[\mathcal{P}_J]x\|_1 < \infty$ and every $\tau \in C^2(\mathbb{R}^d)$ with $\|\nabla\tau\|_\infty \leq 1/2$ satisfy*

$$\|S_J[\mathcal{P}_J]L[\tau]x - S_J[\mathcal{P}_J]x\| \leq C\|U[\mathcal{P}_J]x\|_1 K(\tau)\,, \tag{3.3}$$

where

$$K(\tau) := 2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty \max\left(1, \log\frac{\sup_{u,u'}|\tau(u)-\tau(u')|}{\|\nabla\tau\|_\infty}\right) + \|H\tau\|_\infty\,,$$

and for all $m \geq 0$, *if* $\mathcal{P}_{J,m} = \bigcup_{n<m} \Lambda_J^n$, *then*

$$\|S_J[\mathcal{P}_{J,m}]L[\tau]x - S_J[\mathcal{P}_{J,m}]x\| \leq Cm\|x\|K(\tau)\,. \tag{3.4}$$

This result shows that the distance between the representation of the displaced signal $L[\tau]x$ and x is controlled by a term proportional to $2^{-J}\|\tau\|_\infty$ and by another term proportional to the maximal elastic deformation $\|\nabla\tau\|_\infty$. Moreover if x has a compact support the following result can be proven.

Corollary 3.2.1. *For any compact set* $\Omega \subset \mathbb{R}^d$, *there exists a constant* C *such that for any* $x \in L^2(\mathbb{R}^d)$ *supported in* Ω *having* $\|U[\mathcal{P}_J]x\|_1 < \infty$ *and for all* $\tau \in C^2(\mathbb{R}^d)$ *with* $\|\nabla\tau\|_\infty \leq 1/2$, *then*

$$\|S_J[\mathcal{P}_{J,m}]L[\tau]x - S_J[\mathcal{P}_{J,m}]x\| \leq C\|U[\mathcal{P}_J]x\|_1(2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty + \|H\tau\|_\infty)\,. \tag{3.5}$$

The proofs of Theorem 3.2.1 and Corollary 3.2.1 can be found on [6] and [31]. Basically this corollary states that when the hypothesis are met, the windowed scattering transform metric is Lipschitz-continuous to the action of diffeomorphisms.

Chapter 4

Scattering Transform and Stochastic Processes with Stationary Increments

Real world signals, such as images or auditory data, are realizations and trajectories of stochastic processes, since all events are subject to randomness, that might be caused by the measuring instrument or by the nature of the phenomenon [32]. Therefore to provide a more detailed and accurate mathematical description it is important to spend a few pages to describe the behaviour of the wavelet scattering transform when applied to random processes [6][16]. More specifically after a suitable definition of L^2 norm (re-adopted in terms of expected value), all the result shown in the deterministic case still holds for certain classes of stochastic processes. In this chapter we recall some necessary definitions and properties of stationary processes and, furthermore, we show how the wavelet scattering transform behaves for such kind of random events.

4.1 Stochastic Processes with Stationary Increments

Definition 4.1.1. *A random process X has stationary increments if for any time $t, t' \geq 0$ and any lag $h > 0$ the random variables $X(t+h)-X(t)$ and $X(t'+h)-X(t')$ have the same probability distribution.*

There is an interesting result showing that the convolution of a stationary process X with any deterministic function preserves the stationarity in the resulting process. We sum up such result in the following proposition.

Proposition 4.1.1. *Let X be a random process with stationary increments, then*

1. *$Y_1 = X \star \psi_\lambda$ has stationary increments;*
2. *$Y_2 = |X \star \psi_\lambda|$ has stationary increments.*

Proof. Let us point out the idea behind statement 1). For any time $t \geq 0$ the process Y_1 is defined as

$$Y_1(t) = \int X(\tau)\psi_\lambda(t-\tau)\, d\tau\,.$$

Therefore, if we consider any lag $h > 0$, the increment $Y_1(t+h) - Y_1(t)$, namely

$$\int X(\tau)(\psi_\lambda(t+h-\tau) - \psi_\lambda(t-\tau))\, d\tau\,,$$

depends on h only through the term $\psi_\lambda(t+h-\tau)-\psi_\lambda(t-\tau)$, which is deterministic, hence Y_1 still has stationary increments. Such argomentum may be extended to prove statement 2. □

4.2 Wavelet Scattering Transform for Stationary Processes

Let us introduce the completed path set $\bar{\mathcal{P}}_\infty^\infty := \Lambda_\infty \cup \mathcal{P}_\infty$. As a consequence of the previous proposition, this result directly follows.

Proposition 4.2.1. *If the random process X is stationary, or has stationary increments, then for any path $p = (\lambda_1, \dots, \lambda_m) \in \bar{\mathcal{P}}_\infty^\infty$ the process*

$$U[p]X = |\,\dots|\,|X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \cdots \star \psi_{\lambda_m}\,|$$

is stationary, or ,respectively, has stationary increments.

It is a fundamental result, as for any time t, the expected value $\mathbb{E}(U[p]X(t))$ does not depend on time t. In this terms we can denote such value disregarding time dependence, writing simply $\mathbb{E}(U[p]X)$.

Definition 4.2.1. *For any path $p \in \bar{\mathcal{P}}_\infty$ the expected scattering transform of X is defined as*

$$\bar{S}X(p) := \mathbb{E}(U[p]X)\,. \tag{4.1}$$

The expected scattering transform, induces a norm such that for any stationary random processes X and Y

$$\|\bar{S}X - \bar{S}Y\|^2 := \sum_{p \in \bar{\mathcal{P}}_\infty} |\bar{S}X(p) - \bar{S}Y(p)|^2\,.$$

In machine learning tasks one can estimate the scattering representation of a random signal X through the computation the windowed scattering transform of a set of observations. However the power of this tool is that to properly characterize X just a few realizations are required.
Let us introduce the set of scales smaller than J denoted by$\Lambda_J = \{\lambda = 2^j : 2^{-j} > 2^{-J}\}$ and, let $\mathcal{P}_J = \{p = (\lambda_1, \dots, \lambda_m) : \lambda_i \in \Lambda_J \, \forall i = 1, \dots, m\}$. If we consider a realization $x(t)$ of the random process X, we can compute the windowed scattering transform at scale J, which, as far as we have seen in the previous chapters, results

$$S_J[\mathcal{P}_J]x = \{U[p]x \star \phi_J, p \in \mathcal{P}_J\} .$$

An essential result is that if we used the operator S_J to estimate the scattering representation of a random process, we would get a pretty accurate outcome, since such estimator is unbiased. Let us state more properly this observation.

Proposition 4.2.2. *The operator S_J is an unbiased estimator of the scattering coefficients contained in $\mathcal{P}_J$.*

Proof. Let us recall that the low-pass filter ϕ_J is chosen to satisfy $\int \phi_J(u)\, du = 1$. As already seen this trivially implies that $\hat{\phi}_J(0) = 1$. To prove the thesis we need to show that for any path $p \in \mathcal{P}_J$ yields the equality

$$\mathbb{E}[S_J[\mathcal{P}_J]X] = \bar{S}X(p) .$$

Let us fix a path p in $\mathcal{P}_J$, we have that

$$\mathbb{E}[S_J[\mathcal{P}_J]X] = \mathbb{E}[U[p]X \star \phi_J] .$$

We know that if X is stationary, also the process $U[p]X \star \phi_J$ is stationary (i.e. its expectancy does not depend on the time). Without loss of generality we introduce a time variable u, and recalling the fundamental convolution properties of the Fourier transform we can say that for any frequency ω

$$\mathbb{E}[U[p]X(u) \star \phi_J(u)] = \mathbb{E}[\widehat{U[p]X}(\omega)\hat{\phi}_J(\omega)] . \tag{4.2}$$

Since (4.2) holds true for any ω, we can take $\omega = 0$, obtaining

$$\mathbb{E}[\widehat{U[p]X}(0)] = \mathbb{E}[U[p]X \star \delta_0] = \mathbb{E}[U[p]X] = \bar{S}X(p) ,$$

which proves the thesis. □

The interesting fact is that, when dealing with stationary processes, the same results obtained for deterministic signals, such as non-expansion property, isometry, ecc., can be expressed in terms of mean square norm $\mathbb{E}(|X|^2)^{1/2}$, instead of euclidean L^2 norm. For instance when the Littlewood-Pailey condition is met by the wavelet ψ, operators $\bar{S}$ and $S_J[\mathcal{P}_J]$ are non expansive, as stated by the upcoming proposition.

Proposition 4.2.3. *Let ψ be a wavelet satisfying the Littlewood-Pailey condition (2.4). If X and Y are two random processes with finite second order momenta and stationary increments, then*

$$\mathbb{E}[\|S_J[\mathcal{P}_J]X - S_J[\mathcal{P}_J]Y\|^2] \leq \mathbb{E}(|X - Y|^2) \tag{4.3}$$

and

$$\|\bar{S}X - \bar{S}Y\|^2 \leq \mathbb{E}(|X - Y|^2)\,. \tag{4.4}$$

We observe that if Y is a constantly null process, therefore $\|\bar{S}X\|^2 \leq \mathbb{E}(|X|^2)$. However under more restrictive hypothesis on the wavelets, such result can be refined in terms of energy conservation.

Theorem 4.2.1. *If a wavelet ψ is admissible, and if X is a stationary random process, then*

$$\mathbb{E}[\|S_J[\mathcal{P}_J]X\|^2] = \mathbb{E}(\|X\|^2)\,.$$

Also the convergence properties stated in the deterministic case apply in the random, therefore the process $U[p]X$ should be ergodic. In such case it results that for any path p, as $J \to \infty$

$$S_J[p]X \xrightarrow{\mathbb{P}} \bar{S}X\,.$$

Definition 4.2.2. *A random process X with stationary increments is said to have a **mean squared consistent scattering** if the total variance of $S_J[\mathcal{P}_J]X$ goes to zero as J increases, i.e.*

$$\lim_{J\to\infty} \mathbb{E}[\|S_J[\mathcal{P}_J X - \bar{S}X]\|^2] = \sum_{p\in\mathcal{P}_J} \mathbb{E}[|S_J[p]X - \bar{S}X(p)|^2]] = 0\,.$$

In other words, if X have a mean squared consistent scattering, with probability 1 $S_J[\mathcal{P}_J]X$ converges to $\bar{S}X$ as $J \to \infty$. Since this property is so powerful, it is reasonably natural to ask when a process has mean squared consistent scattering. Mallat conjectured that any Gaussian process with integrable autocorrelation has a mean squared consistent scattering [6]. A further consequence of Theorem 4.2.1 and square mean consistent scattering processes is the following corollary.

Corollary 4.2.1. *For admissible wavelets, the process $S_J[\mathcal{P}_J]X$ is mean squared consistent if and only if $\|\bar{S}X\|^2 = \mathbb{E}(|X|^2)$.*

Further theoretical results and properties of wavelet scattering transform and stochastic processes are discussed in [6].

This result ends the mathematical introduction of the wavelet scattering transform and its properties, establishing an understanding of its theoretical foundations and significance in signal representation. In this part we illustrated the definition

and the most important properties of the wavelet scattering transform and compared them through examples and remarks with common signal representation techniques, highlighting differences and similarities. This section focused on discussing invariance properties of the wavelet scattering operator which are fundamental in machine learning applications. In the aim of setting the base for gravitational wave analysis with this new technique, it was crucial to deeply investigate the wavelet scattering transform from a formal point of view. The acquisition of fundamental knowledge on signal representation techniques has been fundamental to conduct and construct the experiments portrayed in the following pages. We presented an analysis of the wavelet scattering transform, and compared it with common Fourier based signal representation techniques. We studied the fundamental properties, i.e. stability to additive noise, local translation invariance and Lipschitz continuity to C^2-diffeomorphisms of the wavelet scattering metric. The great potentiality of the wavelet scattering transform is furtherly assessed in the fact that just the first layers are needed to represent almost all the energy of the signal, as the energy of high order layers rapidly goes to zero. In this part we also discussed how the properties of the wavelet scattering transform can be extended to stochastic processes, in particular to stationary processes. In the following pages, we applied the wavelet scattering transform to auditory data obtained from the Free Spoken Digits dataset [9], a dataset containing human voice labeled recordings of digits from 0 to 9. In the third part of the thesis, we extended our exploration by attempting to represent gravitational wave signals with the wavelet scattering transform. Our objective was to investigate the application of the wavelet scattering transform to extract patterns and discern glitches that affect data collected by interferometers. We proposed a new approach to discriminate glitches, in order to enhance the analysis of gravitational wave data. Moreover, the acquisition of mathematical knowledge concerning state-of-the-art signal representation techniques has been crucial to develop and conduct our experiments. Relying upon this knowledge, we have laid the basis for the application of the wavelet scattering transform in novel domains, leading to promising results towards its utilization in the analysis of gravitational wave data.

Part II

Wavelet Scattering Transform. An Application to Free Spoken Digits

Chapter 5

Methods and Materials

This chapter introduces fundamental tools for mathematical processing of auditory data, from signal representations to the used statistical learning algorithms for classification. In the first part of the chapter we present the state-of-the art methods for extracting information from one dimensional time-series, i.e. the Short Time Fourier Transform. The treatment proceeds with an introductory description to classification algorithms that have been used throughout the work - Decision Tree, Random Forest, Logistic Regression, Feed Forward Neural Net and K-Nearest Neighbors [23], [33], [24]. The last part of the chapter delves into the preparation of auditory data from the Free-Spoken Digits dataset, e.g. standardization and alignment. The analysis was developed using *Numpy* [34] and *Sci-kit-learn* [35] Python libraries as far as the analysis and the classification task are concerned, while for the STFT computation we relied on *Sci-Py* implementation. All the computations on the wavelet scattering transform are optimized on *Kymatio* library [36].
We choose to work on Free Spoken Digits dataset [9] due to the many similarities that are found between human voice recordings and gravitational waves. Either families consist of one dimensional non stationary signals, therefore studying a real, but simple case like the Free Spoken Digits Dataset allow to determine if the way paved by the wavelet scattering transform approach is promising for gravitational wave data analysis.
It has been revealed from a careful literature analysis, that the detailed research we address in this section on the Free Spoken Digits dataset has remained relatively unexplored. Prior research has delved into the use of the wavelet scattering transform on different datasets for various tasks, e.g. music genre classification on GTZAN dataset [37], or emotion prediction from auditory speech sources [38]. The code is available at `https://github.com/alelicciardi99/wavelet_scattering_transform_thesis`.

5.1 Methods

5.1.1 Fourier Spectrogram

Spectrogram representation of one-dimensional signals is one of the most common technique used in signal representation theory [39]. It provides information about the energy spectrum with respect to the couple of time-frequency variables (t, ω). Spectrograms are based on the *Short Time Fourier Transform* (STFT).
We briefly introduce the construction of the Short-Time-Fourier-transform in the case of continuous time, therefore we suppose that the time variable t is a positive real number, i.e. $t \in \mathbb{R}^+$. Let us fix a function $w(t)$ called *window function*, most common choices are *Hann window* or *Gaussian window*. Hann window, with support length $T > 0$, has the following form

$$w(t) = a \cos^2\left(\frac{\pi t}{T}\right) \mathbf{1}_{\{|t| \leq T/2\}}(t)$$

while Gaussian window is a centered Gaussian function with amplitude a and spread σ, i.e.

$$a \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Definition 5.1.1. *For a given signal $x(t)$ and a fixed window function $w(t)$, the **Short Time Fourier Transform** is defined as*

$$\mathbf{STFT}\{x\}(t, \omega) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-i\omega\tau} \, d\tau.$$

Note that STFT is strictly related to the Fourier transform, due to the immediate relation

$$\mathbf{STFT}\{x\}(t, \omega) = \mathcal{F}\{x(\tau) w(\tau - t)\}(\omega)$$

i.e. the Fourier transform of the signal $x(\tau)$ multiplied by a moving window $w(\tau - t)$, for any $t > 0$. A trivial extension of the definition to the discrete time case is possible, by replacing the integral with an infinite sum.

Definition 5.1.2. *For any $t > 0$ and $\omega > 0$, the **spectrogram** of a signal x is defined as the power spectrum of $x(\tau)w(\tau - t)$, i.e.*

$$|X(t, \omega)|^2 = |\mathbf{STFT}\{x\}(t, \omega)|^2,$$

where $w(t)$ is the chosen window function for computing the STFT.

5.1.2 Spectrograms on the Free Spoken Digits Dataset

We present at the end of this chapter our proposed approach to clean and preparate auditory data from Free Spoken Digits dataset [9]. However pre-processed time-series themselves do not provide enough information. In order to find patterns that help classifiers to learn we need to transform the data into another domain. State-of-the-art method involve computing the spectrogram for each signal. Each spectrogram is a matrix of positive numbers of dimension 128×96. Dealing with large data points slows the training of the classifiers, therefore in Section 6.2 we present more complete results after the spectrogram dimension has been suitably modified with adjusted techniques. Note that cropping the spectrogram is not a good procedure, since the STFT does not produce a translation invariant representation. On the other hand, if we use wavelet scattering transform this procedure is theoretically justified by translation invariance as discussed in the first part of the present work.

5.1.3 Classification Algorithms

In this section we briefly introduce the classification algorithms that have been tested throughout this part of the thesis.
The **Decision Tree** is the simplest classification model with a tree structure [24],[33]. It consists of nodes that can be internal or leaf nodes, and a first node called the root. Based on the samples belonging to the region defined by each node, a splitting rule of the form $\mathbf{1}_{\{x_j \leq \xi\}}$ is chosen to maximize the partitioning of the samples into subsequent child nodes. This partition is evaluated by calculating the improvement in the degree of impurity achieved after the split using either the Gini criterion (5.1) or by evaluating the entropy (5.2), as indicated below.

$$\frac{1}{2}\left(1 - \sum_{z=0}^{c-1} (p_z^i)^2\right) \tag{5.1}$$

$$-p_z^i \sum_{z=0}^{c-1} \log_2 p_z^i \tag{5.2}$$

In particular, p_z^i is the proportion of elements in the region R_i of samples assigned to class z, where R_i represents the region of the space associated with node i. As a final result, a partition of the sample space is obtained, where it is possible to classify and predict the label of a given sample based on its corresponding region.
The **Random Forest** is an *ensemble* model commonly used for classification, built by bagging decision trees [24],[33]. In a random forest, each decision tree is trained on a subset of the training set to avoid the risk of over-fitting training data. The final prediction is obtained by applying a majority voting on the predictions of the

individual trees.
The **logistic regression** is a statistical learning model commonly used for binary classification, but easily extendable to multiclass tasks, such as the one analyzed in this part of the work [24],[33]. It models the relationship between a set of input variables and the probability of a binary outcome. The model applies *Boltzmann's Softmax function* to map the linear combination of the input variables to the range $[0, 1]$, representing the posterior probability of the input data of belonging to each class, namely

$$\sigma : \mathbb{R}^K \to [0,1]^K$$

and

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, \ldots, K\,. \tag{5.3}$$

The parameters of the logistic regression model are estimated using maximum likelihood estimation. However in machine learning tasks it is common to use gradient descent to estimate the optimal parameters. Accuracy of logistic models can be boosted up by adding a regularization, or penalty term, namely C. This additional term added to the loss function during model training process aims to control the complexity of the model and to prevent overfitting. By including the regularization term in the loss function, logistic regression seeks to find a trade-off between minimizing the error on the training data and minimizing the magnitude of the coefficients of the model. Thus it helps the model to avoid fitting too closely to the training data and making it more robust to noise and outliers. The penalty term discourages the model from learning overly complex patterns that may not generalize well to unseen data.
A **Feed Forward Neural Network** (FFNN), or **MultiLayer Perceptron** (MLP), consists of multiple layers of interconnected neurons. Each neuron in a layer receives inputs from the previous layer, performs a weighted sum of the inputs, applies an activation function to mimic neuron activation, and passes the result to the next layer [24],[33]. The neural network learns by adjusting the weights between the neurons during the training process. This is typically done using backpropagation algorithm, where the loss between the predicted output and the true output is propagated backward through the network, and the weights are updated based on the gradient of the error. This iterative process continues until the network converges to a set of weights that minimize the loss function. Training a neural network is way more costly than other simple algorithms, as a lot of parameters need to be learnt. In this setting either *Adam* and *Stochastic Gradient Descent* have been tested. Common activation functions include the sigmoid function, hyperbolic tangent function, and rectified linear unit (ReLU).
The **K-Nearest Neighbors** algorithm (KNN) operates by assigning a class label to a test instance based on the majority class labels of its k nearest neighbors in

the training data. In the KNN algorithm, the choice of k, the number of nearest neighbors to consider, is an important hyperparameter, which tuning was properly addressed in this analysis. A larger value of k tends to make the decision boundary smoother but may also introduce more bias, while a smaller value of k can lead to a more complex and potentially unstable decision boundary. In order to find the nearest neighbors, KNN uses different distance metrics. In this framework Minkowski, Euclidean, Manhattan and Cosine distances were tested to calculate the distances between training points and test instance. The class label is hence assigned to the test instance based on the majority vote of the k nearest neighbors [24],[33].

5.1.4 Semi-Supervised Learning and Data Exploration with PCA

In this section, we aim to propose a possible approach for data description using the technique of PCA (Principal Component Analysis) [23]. Specifically, we want to highlight similarities and dissimilarities among signal classes.
Let us consider the three order wavelet scattering transformed datasets, namely

$$\tau = \{S^0[x_i], S^1[x_i], S^2[x_i]; y_i\}_{i=0}^N$$

where $x_i \in \mathbb{R}^n$ represents the $i-$th signal with class label $y_i \in 0, \ldots, 9$. The main idea is to group in sets the wavelet scattering representations according to the class label, and then perform three PCAs, one for each order.
Let us simply recall how PCA works from a linear algebra point of view. For each class $j \in 0, \ldots, 9$ we define the j-th feature matrix as follows

$$X_j = \begin{bmatrix} s_{1,j}^T \\ \vdots \\ s_{K,j}^T \end{bmatrix} \in \mathbb{R}^{K \times n}$$

where the vectors $s_{l,j}$ is the $l - th$ array of class j. Let $\mathcal{S}_j$ be the estimation of the covariance matrix of the (centered) features. Recall that the matrix $\mathcal{S}_j$ is symmetric by construction

$$\mathcal{S}_j = \frac{1}{K-1} \bar{X_j}^T \bar{X_j} \in \mathbb{R}^{n \times n}$$

where $\bar{X_j}$ is the centered matrix X_j. Let P_j be the orthogonal matrix that diagonalizes $\mathcal{S}j$, which exists due to the *spectral theorem*. In particular, P_j satisfies

$$\mathcal{S}_j = P_j \Lambda_j P_j^T$$

where $\Lambda_j = \text{diag}(\lambda_{1,j}, \ldots, \lambda_{n,j})$ is the diagonal matrix of eigenvalues of $\mathcal{S}_j$. Note that since the trace of a matrix is invariant under similarity transformations and

the trace of $\mathcal{S}_j$ estimates the total variance in class j, we can express the variance explained by the l-th principal component of X_j as

$$\frac{\lambda_{l,j}}{\mathrm{tr}(\Lambda_j)}.$$

Define

$$m = m(\epsilon) := \inf\left\{k = 1, \ldots, n : \frac{\sum_{l=1}^{k} \lambda_{l,j}}{\mathrm{tr}(\Lambda_j)} \geq \epsilon\right\}$$

where $\epsilon \in (0,1]$ represents the threshold of the variance ratio explained. Once the value of m is determined, the first m principal components can be identified.
Since we performed one PCA for each order we find a representative wavelet scattering transform.
For each class j the zero-th order representative is given by the sum of the first m eigenvectors, while for orders one and two has been taken the absolute value of the sum of the principal components.
We address the PCA based approach to show the wavelet scattering transform representants and to see if distinctive patterns emerge differently from each class. Furthermore we apply this method after the KNN in order to visually assess and compare the classification results obtained for each class. This technique presented in this section becomes crucial in the analysis conducted on gravitational wave data in the final section of the thesis to evaluate the dispersion of *scattered-light glitch* class.

5.1.5 Kymatio Library

Computing and handling Wavelet Scattering Transform in codes has become quite straightforward due to the recent development of a powerfool and malleable toolkit by researchers of the *Ecóle Normale Superior* of Paris [10].
Kymatio is a Python library that offers a numerically efficient implementation of the wavelet scattering transform. It allows to utilize wavelet scattering transforms in deep learning and signal processing applications, since it has been properly designed to be compatible with the most common deep learning frameworks, e.g. Numpy, PyTorch, Keras.
Kymatio supports a wide range of signal types, including 1D signals, 2D images, and 3D volumes. In this framework we used only the 1D wavelet scattering transform, since either audio recordings and gravitational waves, are mono-dimensional signals. It offers flexibility in choosing different wavelet families and filter banks, allowing users to tailor the transform to their specific needs. The library takes advantage of the computational power of modern hardware by implementing parallel processing and GPU acceleration for efficient computation of scattering transforms. Figure 5.1 pictures the difference between ScatNet [10] which is the first published practical

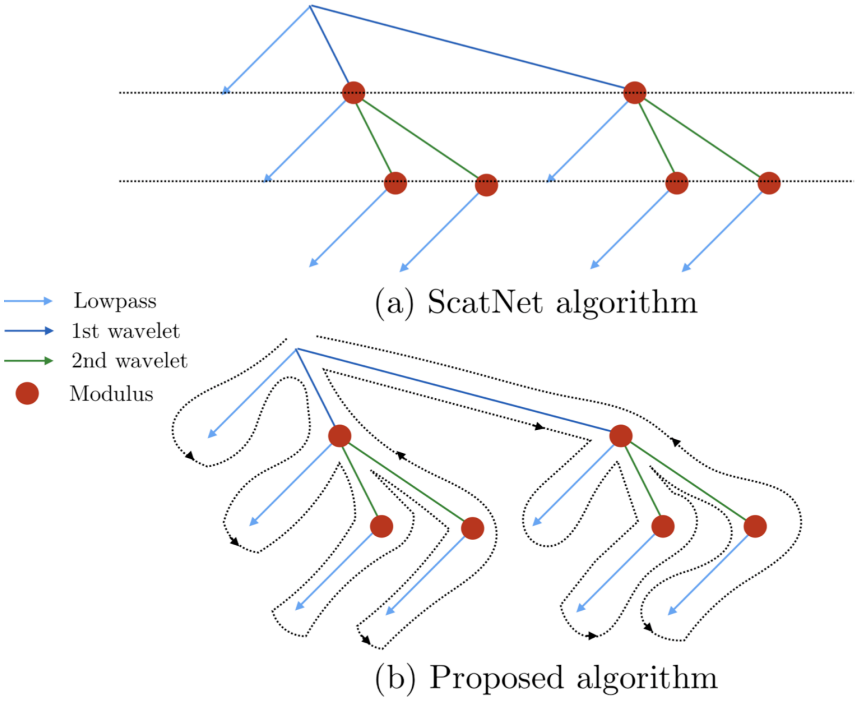


Figure 5.1: Comparison between ScatNet and Kymatio algorithms [40]

implementation (Matlab) of the wavelet scattering transform.
The pipeline developed for this work involved using Kymatio class ***Scattering1D***, from *kymatio.numpy*, which allows to obtain the wavelet scattering transform of a signal up to second order. The processed digit is stored in a *numpy* one-dimensional array. Scattering1D class relies on three parameters J, Q and T. The parameter T is simply the length of each input signal as numpy array structure. The two fundamental parameters are J and Q which, recalling Definition 2.2.3 are related to the maximum scale 2^J and the scaling factor, or octave, $a > 1$, according to the relation $a = 2^{1/Q}$. In the framework of Part II that involves audio processing the parameters chosen are $J = 4$ and $Q = 8$, according to the literature. While for Part III, that deals with gravitational waves data, a further tuning study for this parameters have been developed in the thesis.

5.2 Materials: Free Spoken Digits Dataset

Free Spoken Digits dataset [9] is a collection of auditory data of digits from 0 to 9 spoken by different voices and recorded in different conditions. The dataset is balanced and contains 1500 audio files (*wav* format), 150 for each class, with

sample rate of 8000 Hz. However, as pictured in Figure 5.2, the time durations of the signals are not homogeneous, the vast majority of the data last at least one second. A second observation on the data, highlighted by the histogram in Figure 5.3, is that the maximum amplitude is different for each sample.

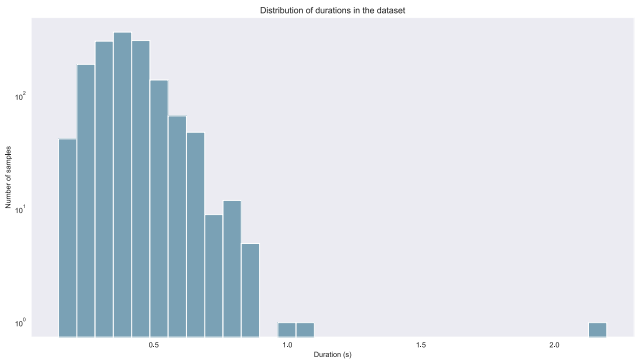


Figure 5.2: Histogram showing the distribution of duration on the dataset

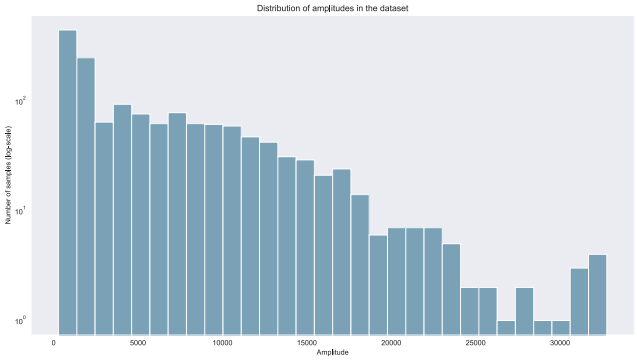


Figure 5.3: Histogram showing the distribution of maximum amplitudes on the dataset

Furthermore, by a simple visualization of four signals, it results quite obvious that the time instant in which the speaker starts to talk differ from sample to sample, as portrayed in Figure 5.4. Therefore in this setting it is fundamental to take into account all the aforementioned characteristics of dis-homogeneity of the dataset. Before applying any sort of representation we should pre-process the time series in order to *equate their length*,*standardize them* and *align them*. This practice would be common both for the spectrogram state-of-the-art analysis, both

for the wavelet scattering transform pipeline. In either cases the same classifiers are trained on different data, with the same parameters and finally, and finally compared, to see if the use of the wavelet scattering transform provides better results.

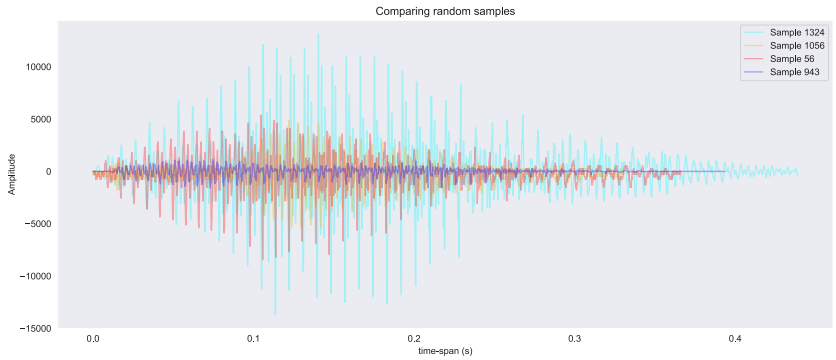


Figure 5.4: Four different signals randomly chosen

5.2.1 Audio Pre-processing

As previously introduced, before transforming the data and testing the difference between signal representations techniques, namely Fourier spectrogram against the wavelet scattering transform, it is important to manipulate the time series. Therefore we delve into the description of the following steps:

- standardize the signals
- equate the duration of the signals
- align the beginning of the signals

Due to different recording equipment, speakers and other conditions there is great variability in the data as far as maximum amplitudes are concerned. There are two possible ways to overcome this problem, *normalization* and *standardization* [24], [33]. Note that here we refer to normalization as mapping a feature vector into the range $[-1,1]$ simply applying a linear scaling. However this technique does not modify the variability internal to the time series. Instead standardization is a more powerful tool for signal analysis, as it provides also important information that can be used to align the recordings. In this case the signals are vectors of points, therefore the dataset can be written as $\mathcal{D} = \{x_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{m_i}$. Let

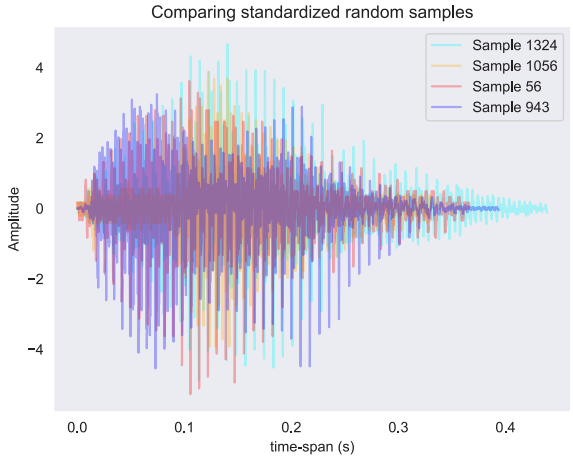


Figure 5.5: Four recordings shown after applying *standardization*

$\hat{\mu}_i$ be the sample mean of x_i and s_i the estimate of its standard deviation, namely

$$\hat{\mu}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_i(j)$$

and

$$s_i = \sqrt{\frac{1}{m_i - 1} \sum_{j=1}^{m_i} \left(x_i(j) - \hat{\mu}_i \right)^2} .$$

The standardized vector $\hat{x}_i$, defined component-wise

$$\hat{x}_i(j) = \frac{x_i(j) - \hat{\mu}_i}{s_i} , \quad j = 1, \dots, m .$$

Due to this definition it is trivial to note, as also recalled by the term *standardization*, that each standirdezed sample is centered, i.e. has a null sample mean, and has unitary sample standard deviation. Figure 5.5 shows that after standardization has been applied, with respect to the original signals displayed in Figure 5.4, the amplitudes of the signals are comparable and vary in similar ranges.
A central aspect of the analysis focused on addressing the challenge of signal alignment. Overcoming this problem was crucial to ensure accurate and reliable results in the study. We developed a method based on the heuristic assumption that in the neighborhood of the time instant in which the speaker pronounces the digits the variance suddenly increases. Standardization technique embeds a threshold for discriminating variance jumps. To construct the neighborhood each recording has been split in windows of 0.375×10^{-2} s, i.e. each window consisted of 30 scalar values.

Algorithm 1 Algorithm for signal alignment

Require: standardized signal $\hat{x}_i \in \mathbb{R}^{m_i}$
Ensure: index $\bar{k}$ of the window in which the voice begins
  split $\hat{x}_i = (z_1, \ldots, z_K)$, where $z_k \in \mathbb{R}^{30}, k = 1, \ldots, K$
  while k=1,...,K **do**
    compute the standard deviation σ_k for z_k
    if $\sigma_k > 1$ **then**
      $\bar{k} \leftarrow k$
    end if
  end while

After the time window index $\bar{k}$ is detected for each standardized time series, the previous windows, having index $k < \bar{k}$ are set to zero. After that it becomes trivial

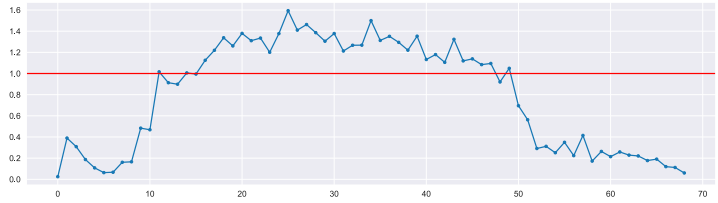


Figure 5.6: Variance plot window by window (threshold variance in red)

to align signals, as before the beginning of the digits are pronounced there is a sequence of zeros. Figure 5.6 pictures the idea behind the alignment algorithm. In

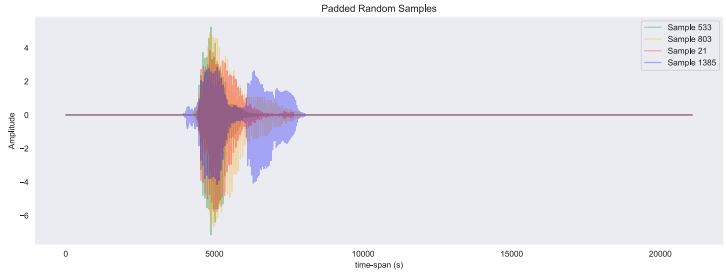


Figure 5.7: Four recordings after standardization, alignment and zero-padding

order to equate the length of the signals it is applied simply a *zero-padding*, hence each recording in the dataset is padded with null values up until it reaches the length of the longest one. This is a common technique in signal processing. Figure 5.7 shows four recordings after zero-padding was performed, as a result all the

samples in the dataset have equal duration. Let us observe that it is necessary to standardize and align the signals before *zero-padding*, since adding zeros affects the estimate of the variance and of the sample mean. Even though this implemented technique is based on heuristic assumptions, plots and results proved empirically the efficiency of this data preparation pipeline. In the following chapter we present the state-of-the-art signal representation (Fourier spectrogram) and, thus, the classification results obtained.

Chapter 6

Experimental results

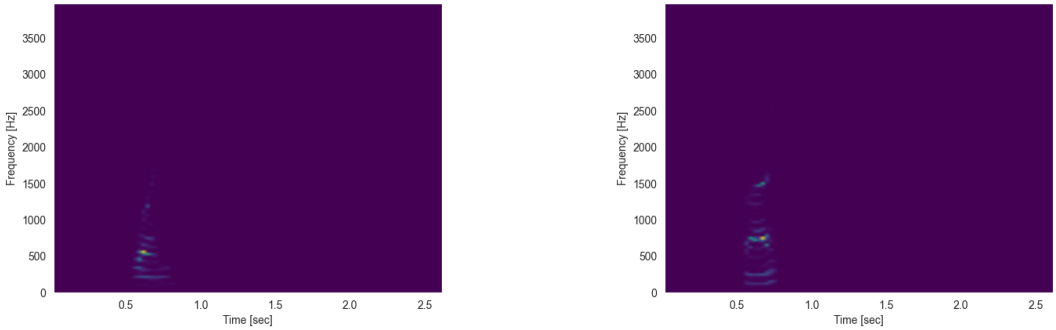
This chapter delves into a comprehensive analysis and comparison of the results obtained using state-of-the-art representation technique, the Short-Time-Fourier-Transform (STFT), against the ones obtained with a wavelet scattering transform based approach. Our investigation includes a meticulous examination of the performance achieved by four classification algorithms, namely Decision Tree, Random Forest, Logistic Regression, and a fully connected Neural Net, within each representation method. Furthermore, we explore the effectiveness of different metrics in the transformed space in terms of data separation. To assess this, we employ a distance-based simple algorithm, namely K-Nearest Neighbors (KNN). Drawing upon the mathematical theory behind signal representations, we anticipate that the KNN algorithm in the wavelet scattering domain will exhibit superior performance compared to its counterpart in the Fourier domain. Throughout the analysis, we provide visual aids such as plots and tables to facilitate a comprehensive understanding and visualization of the results.
The predictions have been evaluated using either **accuracy** and **F1-score** [33]. Accuracy measures the overall correctness of predictions by calculating the ratio of correctly classified instances to the total number of instances. On the other hand, the F1-score leverages either precision and recall, providing a balanced assessment of the performance, particularly in cases where class imbalances subsists.
For each class we define the *precision* as $P = TP/(TP + FP)$, while the *recall* is defined as $R = TP/(TP + FN)$, where TP, FP and FN stand for *true positive*, *false positive* and *false negative*, respectively. While the $F1-$score is obtained as

$$2 \times \frac{P \times R}{P + R},$$

where P and R are respectively precision and recall of the class. These metrics play a crucial role in quantifying the accuracy and effectiveness of the classification models discussed in this chapter.

6.1 State-of-the-art methods: STFT

In Chapter 5 we briefly introduced the definition of STFT, here we provide some visual examples of specograms computed on signals randomly sampled from the Free Spoken Digits dataset. Figure 6.1 compares the spectrograms of a *one* and a *five*. Each spectrogram is a matrix of positive numbers of dimension 128×96.



(a) Spectrogram of a *one*

(b) Spectrogram of a *five*

Figure 6.1: Examples of spectrograms of two recordings

Dealing with large data points slows the training of the classifiers, therefore afterwards we present a techinque, based on the *average pooling* that allows the dimension to remarkably decrease, and, moreover, to highlight the information on more representative features. Note that cropping the spectrogram is not a good procedure, since the STFT does not produce a translation invariant representation, while cropping wavelet scattering transform signals around a region of interest is theoretically supported by the invariance theory shown in the first part of the thesis.

6.1.1 Classification Performances

This paragraph reports results obtained with the most common classifiers, that are trainable considering the dimensions of the spectrogram. Note that in the following section after proposing a data reduction method other algorithms are tested.
In this setting we split the dataset into *training set* and *validation set*. For each classifier different combination of parameters are tested, and then we report the scores obtained on the validation set, in terms of average F1-score.
Among all the classifiers the *Random Forest* is the one that achieved performed better on the validation set, reaching a score of 0.96.
The classification report presented in Table 6.1 showcases the performance evaluation of the Decision Tree classifier with optimized hyperparameters in the STFT

space. Upon analysis, it becomes evident that the classifier encounters challenges in effectively discerning between different signal classes.

Class	**Precision**	**Recall**	**F1-score**
0	0.64	0.57	0.60
1	0.76	0.66	0.70
2	0.61	0.71	0.66
3	0.53	0.61	0.57
4	0.72	0.75	0.74
5	0.54	0.71	0.62
6	0.59	0.46	0.52
7	0.52	0.52	0.52
8	0.62	0.57	0.59
9	0.79	0.68	0.73
Overall			
Accuracy			**0.62**
Macro Avg	0.63	0.62	**0.62**
Weighted Avg	0.63	0.62	0.62

Table 6.1: *Decision Tree Classification report* with optimal parameters (STFT domain)

Examining the *precision* values, we observe varying levels of accuracy in correctly identifying instances for each class. For instance, classes 1 and 9 demonstrate relatively higher precision scores of 0.76 and 0.79, respectively, indicating that the classifier is more adept at correctly classifying signals belonging to these classes. However, classes 6 and 7 exhibit lower precision scores of 0.59 and 0.52, respectively, suggesting a higher likelihood of mis-classifications. The *recall* values provide insights into the classifier ability to correctly identify instances from each class. While classes 4 and 5 demonstrate recall scores of 0.75 and 0.71, respectively, indicating good performance in correctly capturing instances from these classes. Classes 0 and 6 show relatively lower recall scores of 0.57 and 0.46, respectively, indicating a higher rate of false negatives. The F1-scores offer a balanced evaluation of the classifier's overall performance for each class. The F1-scores range from0.52 to 0.74, with classes 4 and 9 achieving higher scores, suggesting better classification performance for these classes. However, classes 6 and 7 exhibit lower F1-scores, indicating a more challenging discrimination task for the classifier. The overall accuracy of 0.62 highlights the classifier ability to correctly predict instances across all classes. While this accuracy score indicates a moderately low level of performance, it also suggests the presence of mis-classifications and much room for improvement.

Based on the presented classification report, it becomes apparent that the Decision Tree classifier struggles to effectively discriminate signals in the STFT space. The variations in precision, recall, and F1-scores across different classes emphasize the challenges faced by the classifier in accurately identifying instances from each class. Considering the classification report presented in Table 6.1 and the observed limitations of the Decision Tree classifier, we found crucial to explore alternative algorithms to potentially improve classification performance. One possible approach is to utilize *ensemble* methods such as Random Forest, which can harness the collective decision-making of multiple decision trees. By aggregating the predictions of multiple trees, Random Forest can potentially mitigate the limitations of individual Decision Trees and enhance the overall discriminatory power in the STFT space. In the STFT domain Random Forest was the best performing algorithm. Figure 6.2 portraits how changing the number of trees in the forest affects the F1-score. This grid-search led to train a forest with 90 trees. The classification report presented in

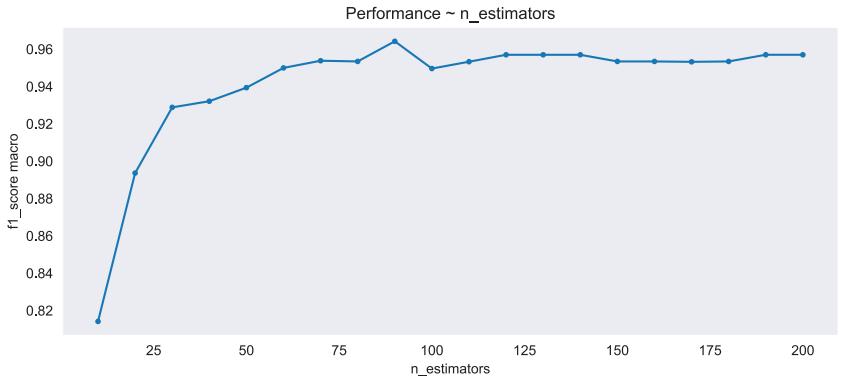


Figure 6.2: F1-score vs number of estimators in the Random Forest trained on the spectrograms

Table 6.2 highlights the performance of the Random Forest classifier with optimized hyperparameters in the STFT domain. Among the various algorithms tested within this experiment, the Random Forest classifier emerges as the top performer, delivering good results. The precision, recall, and F1-score values for each class demonstrate remarkable accuracy in correctly classifying instances. With precision scores ranging from 0.89 to 1.00 and recall scores ranging from 0.89 to 1.00, the classifier exhibits consistently high performance across different classes.

Class	Precision	Recall	F1-score
0	0.96	0.96	0.96
1	1.00	1.00	1.00
2	0.93	0.93	0.93
3	1.00	1.00	1.00
4	1.00	1.00	1.00
5	0.97	1.00	0.98
6	0.96	0.89	0.93
7	0.97	0.97	0.97
8	0.89	0.89	0.89
9	0.97	1.00	0.98
Overall			
Accuracy			**0.96**
Macro Avg	0.96	0.96	**0.96**
Weighted Avg	0.96	0.96	0.96

Table 6.2: *Random Forest Classification report* with optimal parameters (STFT domain)

Notably, the overall accuracy of 0.96 denotes the ability of the algorithm to accurately predict class labels for the entire dataset, solidifying its position as the best-performing algorithm within the STFT domain. The achieved accuracy score indicates the successful use of ensemble methods to leverage the collective decision-making of multiple trees and achieve superior classification performance. The exceptional performance of the Random Forest classifier, as demonstrated by the classification report in Table 6.2, is further substantiated by the corresponding confusion matrix presented in Figure 6.3. The visually depicted distribution of correctly classified instances across different classes provides a comprehensive and conclusive confirmation of the Random Forest effectiveness in accurately distinguishing signals within the STFT domain. Furthermore, the macro average and weighted average F1-scores of 0.96 indicate a well-balanced performance across all classes, considering both the class distribution and the classifier's ability to discriminate between classes effectively. These results reinforce the notion that the Random Forest classifier, without resorting to deep convolutional structures, can achieve highly satisfying results in signal classification tasks.
In order to evaluate the impact of applying the wavelet scattering transform instead of the STFT on the classification performances, we present the behaviour of other classifiers in the following tables. Specifically, we report the results for the Multiclass Logistic Regression and the Feed-Forward Neural Network.

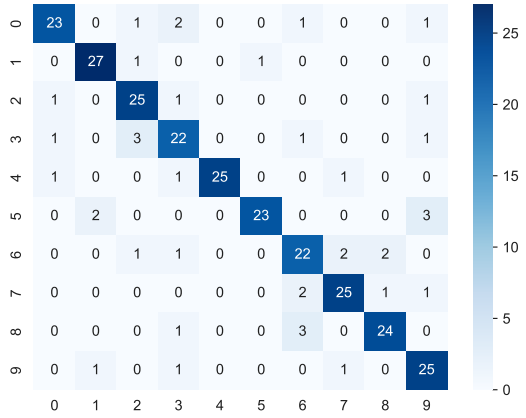


Figure 6.3: *Confusion Matrix* for the best performing model in the STFT domain (*Random Forest*)

Class	Precision	Recall	F1-score
0	0.79	0.82	0.81
1	0.82	0.93	0.87
2	0.86	0.89	0.88
3	0.76	0.79	0.77
4	1.00	0.89	0.94
5	0.96	0.86	0.91
6	0.72	0.75	0.74
7	0.89	0.83	0.86
8	0.85	0.79	0.81
9	0.80	0.86	0.83
Overall			
Accuracy			**0.84**
Macro Avg	0.85	0.84	**0.84**
Weighted Avg	0.85	0.84	0.84

Table 6.3: *Logistic Regression Classification report* with optimal parameters (STFT domain)

Tables 6.3 and 6.4 demonstrate that both classifiers exhibit similar performances. However, the Feed-Forward Neural Network performs slightly better the Logistic Regression. It is important to note that training logistic regression is computationally less demanding and therefore a more cost-effective option. We will compare the results displayed in this section, obtained training the classifiers in the Fourier

domain, with the ones obtained in the wavelet scattering domain.

Class	Precision	Recall	F1-score
0	0.88	0.82	0.85
1	0.90	0.93	0.92
2	0.81	0.89	0.85
3	0.76	0.79	0.77
4	1.00	0.89	0.94
5	0.96	0.82	0.88
6	0.76	0.79	0.77
7	0.86	0.86	0.86
8	0.89	0.86	0.87
9	0.78	0.89	0.83
Overall			
Accuracy			**0.85**
Macro Avg	0.86	0.85	**0.86**
Weighted Avg	0.86	0.85	0.86

Table 6.4: *FeedForward Neural Net Classification report* with optimal parameters (STFT domain)

6.1.2 Feature Reduction on the STFT and Classification Performances

As previously mentioned spectrogram representation obtained for this dataset provides for each signal large sparse matrices, and furthermore, due to the fact that STFT is non-translation invariant, it is not possible to reduce the data simply cropping the more informative areas. Inspired by average pooling operation in *ConvNets* architectures, to reduce the features space and gain information the following manipulation on the raw spectrograms was performed.
This particular average pooling allows to reduce the input data from a matrix of 128×96 to a matrix of size 32×48. The approach consists in dividing each raw spectrograms in a grid of sub-matrices of size 4×2, for each of them we compute the sample mean. Figure 6.4 displays the grid overlapped to a spectrogram. As Figure 6.5 suggests the information provided by the data is more concentrated. Due to the consistent reduction of the features space, the algorithm run faster, and, surprisingly, algorithms perform better as shown in Table 6.5. Predictions are more accurate because applying the average pooling reduced the number of null features, and led to a higher level of invariance with respect to the original spectrogram.

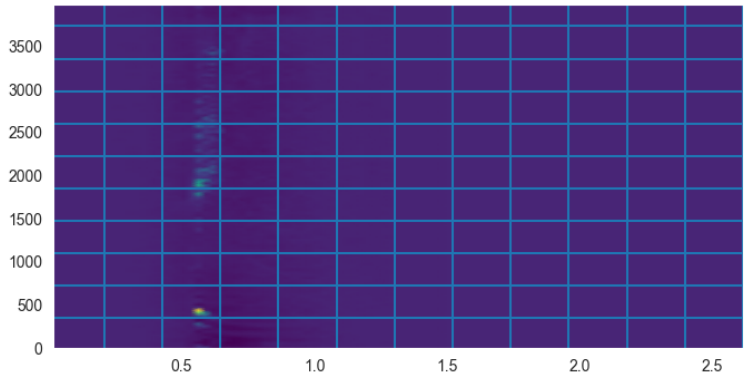


Figure 6.4: Grid for constructing average pooling on a spectrogram

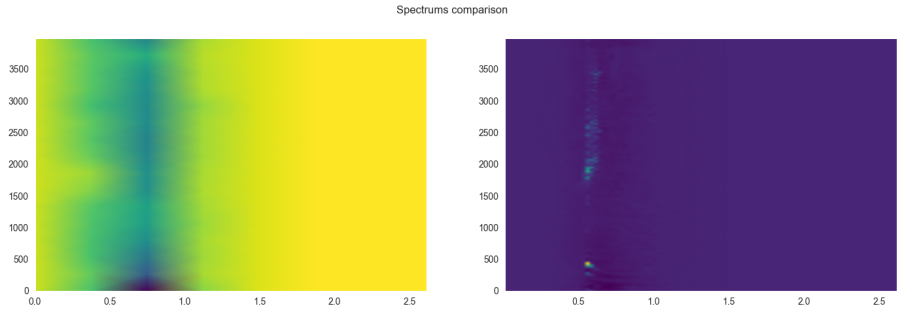


Figure 6.5: Average pooling (left) on a spectrogram (right)

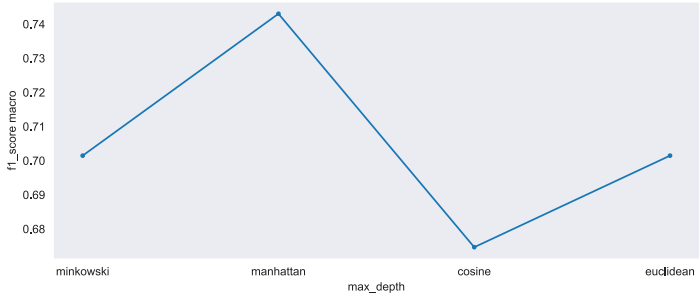
The reduction in feature space not only enhances the algorithm's runtime performance but also yields surprising improvements in performance metrics, as evidenced by the higher F1-scores reported in Table 6.5. The predictions become more accurate due to the reduction of null features and the higher level of invariance achieved compared to the original spectrogram. The comparison of classifiers using pooled spectrograms and raw spectrograms further highlights the effectiveness of the average pooling technique in enhancing classification results.

Classifier	F1-score	Accuracy	F1-score w/ pooling	Accuracy w/ pooling
DecisionTree	0.68	0.68	0.62	0.62
RandomForest	**0.98**	**0.98**	0.96	0.96
LogReg	0.90	0.90	0.84	0.84
FFNN	0.91	0.91	0.86	0.85

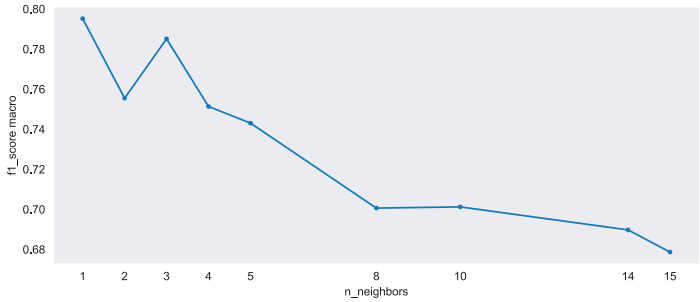
Table 6.5: *Comparison:* performances of the algorithm traind with pooled spectrograms and raw spectrograms

6.1.3 Distance Evaluation: K-Nearest Neighbors

The following analysis investigates the separability of the data in the Fourier domain. We applied the K-Nearest Neighbors (KNN) algorithm with the STFT representation; in a similar manner used with the wavelet scattering transform, we tuned the distance metric and the number of neighbors.



(a) *K-Nearest Neighbours* with STFT: F1-score with respect to the metrics, with $k = 5$



(b) *K-Nearest Neighbours* with STFT: F1-score with respect to the number of neighbors k

Figure 6.6: Tuning: *K-Nearest Neighbours* with STFT

Surprisingly, the best-performing metric was found to be the Manhattan distance, with $k = 1$. However, it should be noted that using a single neighbor does not provide a stable outcome, indicating a lack of invariance in the STFT spectrograms. This instability can be attributed to the fact that the STFT representation does not effectively cluster the data into distinct groups of the same class. Figure 6.6 visually demonstrates that as the number of neighbors increases, the prediction performance rapidly deteriorates. The results presented in Table 6.6 further support this observation, with relatively lower precision, recall, and F1-scores compared to other classifiers. The accuracy achieved with STFT-based KNN is 0.79, indicating that the data is not well separable in the STFT space.
Table 6.6 presents the classification report for the K-Nearest Neighbors (KNN) algorithm using the STFT representation. Upon analyzing the results, it is evident that the performance of KNN with STFT is unstable. The precision values range from 0.52 to 0.96, indicating variability in the accuracy of classifying different classes. The recall values range from 0.60 to 0.93, suggesting that the algorithm captures varying proportions of the true positive instances for each class.
The results highlight the challenges of achieving effective classification using the STFT representation. The relatively lower precision, recall, and F1-scores suggest that the data in the Fourier domain is not well separable, leading to difficulties in accurately distinguishing between different classes. These findings emphasize the limitations of using STFT spectrograms as a standalone feature representation for this classification task.

Class	**Precision**	**Recall**	**F1-score**
0	0.95	0.60	0.73
1	0.81	0.73	0.77
2	0.65	0.73	0.69
3	0.52	0.77	0.62
4	0.71	0.83	0.77
5	0.93	0.93	0.93
6	0.78	0.83	0.81
7	0.92	0.80	0.86
8	0.96	0.83	0.89
9	0.93	0.83	0.88
Overall			
Accuracy			**0.79**
Macro Avg	0.82	0.79	**0.80**
Weighted Avg	0.82	0.79	0.80

Table 6.6: *K-Nearest Neighbours Classification report* with optimal parameters with STFT

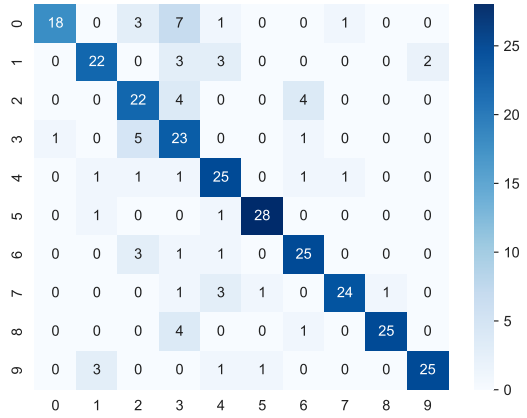


Figure 6.7: *Confusion Matrix* for the *K-Nearest Neighbours* with the optimal parameters in the STFT space

6.2 Wavelet Scattering Transform

This section discusses the classification results obtained with the wavelet scattering representation of the data.
In parallel to the analysis presented in the previous section, this paragraph exhibits the application wavelet scattering transform to some sampled signals from the dataset. As anticipated by the theoretical results discussed in Part I, the wavelet scattering representation effectively captures the most significant information in the same spatial region. This observation aligns with the translation invariance properties of the wavelet scattering transform, justifying our decision to crop the transforms around the region of interest.
Figure 6.8 displays the wavelet scattering transforms for sampled signals in classes 0, 1, and 2. It is clear that the wavelet scattering representation highlights relevant patterns and structures within the signals, enabling robust and discriminative feature extraction. To exploit the stability and invariance properties of the wavelet scattering representation, we concatenated the three layers after flattening each one of them. Unlike image classification tasks where such an approach could lead to poor performance, this concatenation strategy proves to be effective for the wavelet scattering transform. It bears resemblance to the architecture of convolutional neural networks. As discussed in Chapter 1, after the initial convolutional layers process an image, the resulting filtered data is aligned and flattened before being classified using a fully connected neural network. Similarly, we can consider the wavelet scattering transform as analogous to the convolutional layers of a CNN.

However, the wavelet scattering transform offers stability properties that were introduced and demonstrated in the first part of this thesis.

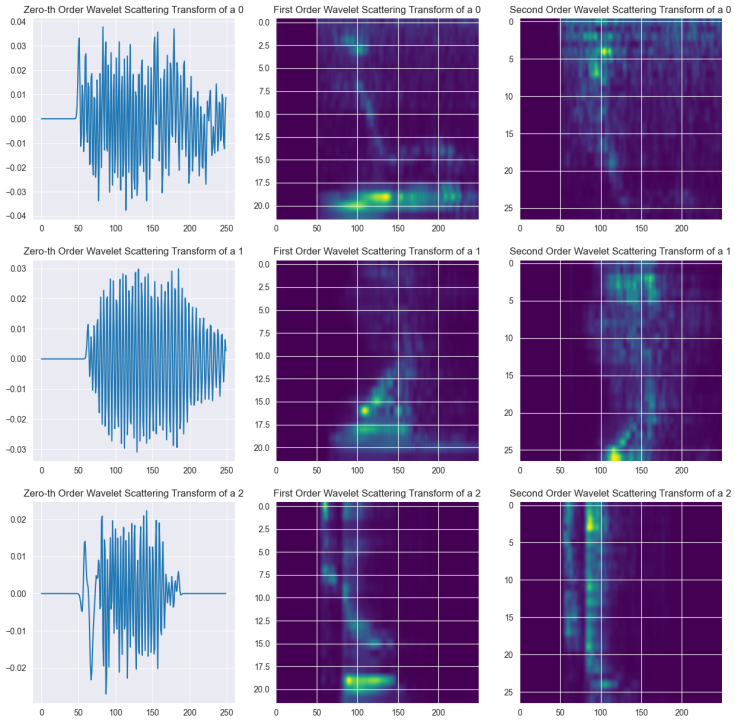


Figure 6.8: Wavelet scattering transform for sampled signals in class 0, 1 and 2

6.2.1 Classification Algorithms

In this section, we present the classification results obtained using the wavelet scattering transform representation. We focus on evaluating the performance of several algorithms, including Decision Tree, Random Forest, Logistic Regression, and Feed Forward Neural Net. By analyzing the impact of various hyper-parameters on the classification metric, specifically the F1-score, we provide insightful plots that depict the relationship between these parameters and the performance of the algorithm. Additionally, an exhaustive classification report summarizes the obtained results, highlighting the effectiveness of these algorithms when applied to the Wavelet Scattering Transform representation. Through this dedicated analysis, we aim to uncover valuable insights into the capabilities and performance of these

algorithms within the context of the wavelet scattering transform.
The first algorithm examined is the Decision Tree, applied to the wavelet Scattering Transform representation.

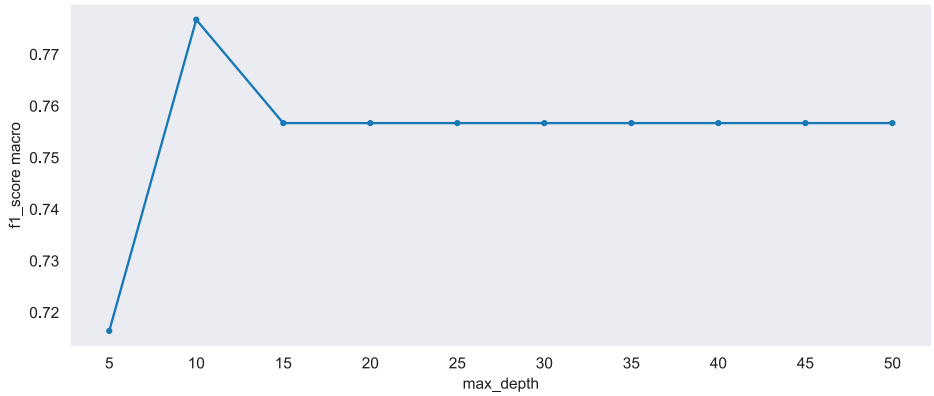


Figure 6.9: Decision Tree F1-score and maximum depth in the wavelet scattering domain

Figure 6.9 illustrates the impact of varying the maximum depth of the decision tree on its performance, measured in terms of F1-score. It is interesting to observe that as the depth increases, the F1-score initially rises and reaches its peak at a maximum depth of 10, after which it stabilizes for deeper trees. To further optimize the performance, a grid search was conducted to fine-tune the other hyper-parameters, yielding the results presented in Table 6.7. A noteworthy observation from the comparison with the results obtained using the STFT representation, as shown in Table 6.1, is that the Wavelet Scattering Transform space appears to offer better separability of the data. The improved performance suggests that the wavelet scattering representation captures more discriminative features, leading to enhanced classification accuracy. The classification report in Table 6.7 provides a detailed assessment of the Decision Tree performance. The overall accuracy achieved with the optimal parameters in the wavelet scattering transform domain is 0.81, indicating that the algorithm is able to correctly classify a significant portion of the validation data. The macro-average and weighted-average F1-scores are both 0.81, further validating the the stability across different classes. These results demonstrate the potential of the Decision Tree algorithm when applied to the Wavelet Scattering Transform representation, offering promising prospects for various classification tasks.

Class	Precision	Recall	F1-score
0	0.86	0.80	0.83
1	0.66	0.90	0.76
2	0.63	0.73	0.68
3	0.77	0.77	0.77
4	0.96	0.80	0.87
5	0.88	0.70	0.78
6	0.85	0.77	0.81
7	0.79	0.90	0.84
8	0.96	0.77	0.85
9	0.88	0.93	0.90
Overall			
Accuracy			**0.81**
Macro Avg	0.82	0.81	**0.81**
Weighted Avg	0.82	0.81	0.81

Table 6.7: *Decision Tree Classification report* with optimal parameters in the wavelet scattering domain

Building upon the promising performance of the Decision Tree algorithm in the wavelet scattering transform space, we can explore the potential improvements achievable by employing a Random Forest classifier, as presented for the state-of-the-art STFT.

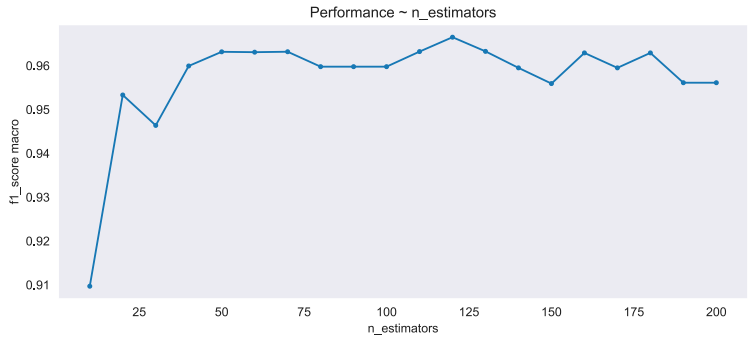


Figure 6.10: Random Forest F1-score and the number of estimators in the wavelet scattering domain

Figure 6.10 illustrates the impact of varying the number of estimators, i.e. trees, in the Random Forest classifier on the F1-score in the wavelet scattering transform domain. It is evident that increasing the number of trees initially leads to a significant improvement in the classification performance. However, beyond

approximately 125 trees, the F1-score decreases, indicating diminishing returns in terms of accuracy.

Class	**Precision**	**Recall**	**F1-score**
0	0.88	1.00	0.94
1	0.97	1.00	0.98
2	0.96	0.90	0.93
3	0.96	0.87	0.91
4	0.97	0.97	0.97
5	1.00	0.97	0.98
6	0.94	1.00	0.97
7	1.00	0.97	0.98
8	1.00	1.00	1.00
9	1.00	1.00	1.00
Overall			
Accuracy			**0.97**
Macro Avg	0.97	0.97	**0.97**
Weighted Avg	0.97	0.97	0.97

Table 6.8: *Random Forest Classification report* with optimal parameters in the wavelet scattering domain

The results presented in Table 6.8 provide a more detailed view of the classification performance with the optimal Random Forest parameters. Analyzing the classification report (Table 6.8), we observe that the Random Forest model achieves good results in classifying the different classes. Notably, classes 0, 4, 8, and 9 exhibit perfect precision, recall, and F1-scores of 1.00, indicating the high discriminative capability of the wavelet scattering transform representation in capturing the distinguishing features of these classes. Other classes, such as 1, 5, and 7, also achieve high scores, further confirming the effectiveness of the Random Forest approach in exploiting the informative characteristics of the transformed data.
The overall accuracy of 0.97 achieved by the Random Forest demonstrates its robustness and ability to generalize well to unseen samples. Comparing the results obtained with the Random Forest algorithm in the wavelet scattering domain to those of the Decision Tree discussed earlier, we observe substantial improvements in the classification metrics. As expected, the Random Forest model outperforms the Decision Tree. Such result can be attributed to the ensemble nature of Random Forest, which leverages multiple trees to capture a more comprehensive representation of the data and make more accurate predictions. We further explore the performance of two powerful classification algorithms: Logistic Regression and Feed

Forward Neural Net. These algorithms have already been examined in the previous section for the STFT representation. In the wavelet scattering domain, Logistic Regression was employed to classify the data, and its performance was assessed by analyzing the F1-score in relation to the regularization coefficient,namely C, as depicted in Figure 6.11. Notably, it was observed that increasing the penalty term led to improved predictions, indicating the importance of regularization in achieving optimal performance.

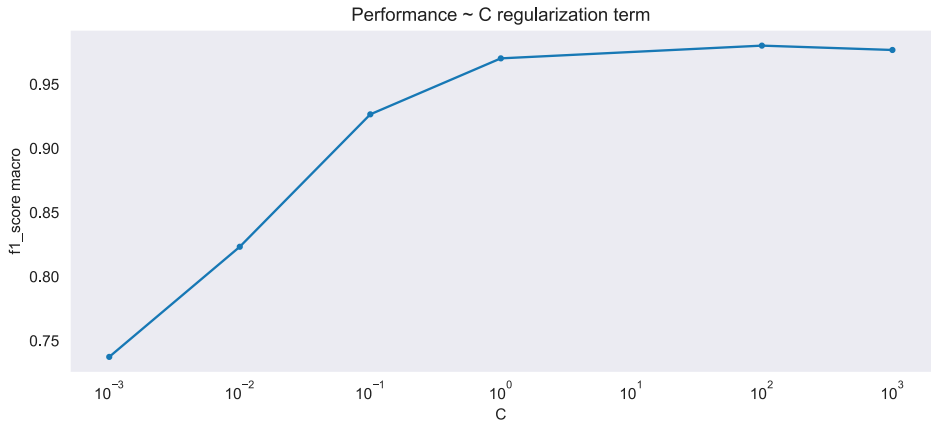


Figure 6.11: Logistic Regression F1-score with respect to the regularization coefficient in the wavelet scattering domain

A complete evaluation of the classification results is provided in Table 6.9. All the different classification metrics for each class demonstrate the effectiveness of Logistic Regression combined with the wavelet scattering transform. Notably, the F1-scores for all classes are high, ranging from 0.95 to 1.00, indicating accurate and reliable predictions. Comparing the performance of Logistic Regression with the wavelet scattering transform to that of the STFT representation (as presented in Table 6.3), it is evident that the wavelet scattering transform yields superior results.

Class	Precision	Recall	F1-score
0	1.00	0.97	0.98
1	1.00	0.97	0.98
2	0.94	0.97	0.95
3	1.00	0.97	0.98
4	0.97	1.00	0.98
5	1.00	1.00	1.00
6	0.97	1.00	0.98
7	1.00	0.97	0.98
8	1.00	0.97	0.98
9	0.94	1.00	0.97
Overall			
Accuracy			**0.98**
Macro Avg	0.98	0.98	**0.98**
Weighted Avg	0.98	0.98	0.98

Table 6.9: *Logistic Regression Classification report* with optimal parameters in the wavelet scattering domain

Following the evaluation of Logistic Regression in the wavelet scattering domain, the next algorithm tested is the Feed Forward Neural Network. The performance of the FFNN was assessed based on various hyperparameters, including the number of hidden layers, the number of neurons in each layer, the learning rate and the optimizer. The results obtained with the Feed Forward Neural Network (FFNN) in conjunction with the wavelet scattering transform are truly remarkable.

Figure 6.12: Feed Forward Neural Net F1-score with respect to the hidden layer size in the wavelet scattering domain

The FFNN achieved an exceptional accuracy of 0.99 on the validation data, demonstrating its ability to learn the underlying patterns in the data. The classification report in Table 6.10 provides detailed insights into the performance of the FFNN across different classes. It is evident that the FFNN excels in accurately predicting all classes, with high precision, recall, and F1-score values consistently above 0.98. These outstanding results highlight the potential of leveraging the wavelet scattering transform in machine learning tasks, and specifically, its effectiveness in enhancing the performance of the FFNN model.

Class	**Precision**	**Recall**	**F1-score**
0	1.00	1.00	1.00
1	1.00	1.00	1.00
2	1.00	0.97	0.98
3	1.00	1.00	1.00
4	0.97	1.00	0.98
5	1.00	1.00	1.00
6	0.97	1.00	0.98
7	1.00	0.97	0.98
8	1.00	0.97	0.98
9	0.97	1.00	0.98
Overall			
Accuracy			**0.99**
Macro Avg	0.99	0.99	**0.99**
Weighted Avg	0.99	0.99	0.99

Table 6.10: *Logistic Regression Classification report* with optimal parameters in the wavelet scattering domain

The classification report for the STFT-based FeedForward Neural Network (FFNN) model, as shown in Table 6.4, reveals that it achieved an overall accuracy of 0.85 on the validation data, and the macro-average F1-score is 0.86.On the other hand the FFNN model paired with the wavelet scattering transform demonstrated superior performance. Table 6.10 reports an accuracy and F1-score of 0.99 on the validation data. In contrast, the FFNN model with the wavelet scattering transform demonstrated superior performance. The classification report provided earlier for the FFNN with wavelet scattering, as shown in Table 6.10, reported an accuracy of 99In conclusion, the best pipeline for classifying digits in this study involved the combination of the wavelet scattering transform and a fully connected neural network. This architecture, resembling that of a ConvNet but with significantly fewer parameters, achieved exceptional performance. The confusion matrix depicted in Figure 6.13 provides a visual representation of the classification results for the

best pipeline. It reveals that out of the 300 signals, only 3 were misclassified, resulting in a remarkable accuracy and F1-score of 99

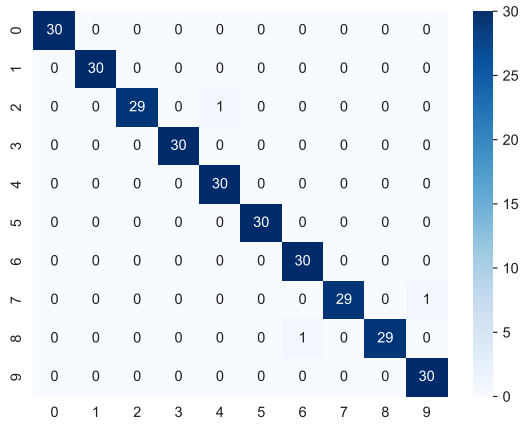
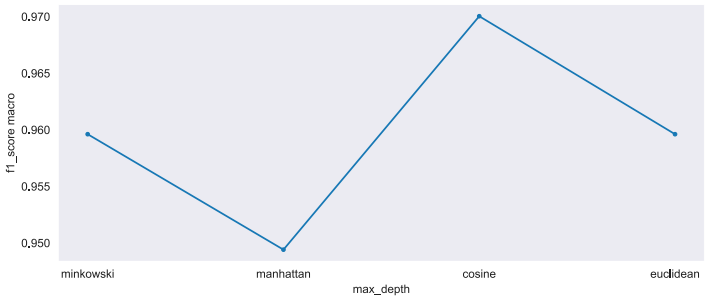


Figure 6.13: *Confusion Matrix* for the best performing model in the wavelet scattering domain (*Feed Forward Neural Network*)

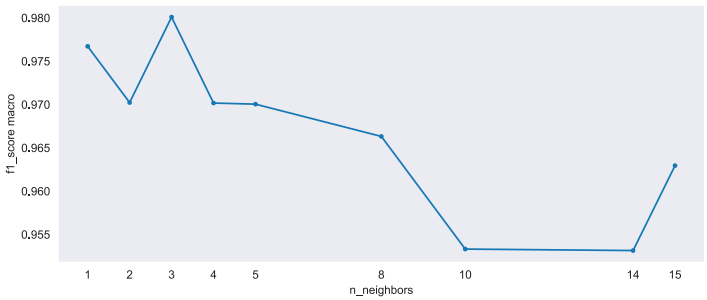
This outcome underscores the effectiveness of the wavelet scattering transform in extracting discriminative features from the auditory data and the remarkable performance of the resulting pipeline. By leveraging the wavelet scattering transform and a simplified architecture, the model achieved near-perfect accuracy in digit classification. These findings highlight the potential of the wavelet scattering transform and its ability to enhance the performance of classification models with reduced computational complexity. The combination of the wavelet scattering transform and a fully connected layer demonstrates the efficacy of leveraging advanced mathematical tools and thoughtful architectural choices in developing high-performing machine learning models.

6.2.2 Distance Evaluation: K-Nearest Neighbors and PCA

Since the wavelet scattering transform induces an embedded metric which enjoys stability and invariance properties, we aim to show in the following section from an empirical point of view that this can be easily verified. The idea is quite straightforward. If the wavelet scattering metric is stable simple algorithms based on distances should reach high accuracy. The following section addresses this discussion testing *K-Nearest Neighbour* (KNN) algorithm over wavelet scattering transforms.



(a) *K-Nearest Neighbours* with WST: F1-score with respect to the metrics, with $k = 5$



(b) *K-Nearest Neighbours* with WST: F1-score with respect to the number of neighbors k, with *cosine* similarity

Figure 6.14: Tuning: *K-Nearest Neighbours* with WST

Figure 6.14 reports F1-score with respect to distances and number of neighbors. The best combination has been proven to be with $k = 3$ neighbors and *cosine* similarity metric, reaching a 0.98 of accuracy. For a further comprehension of the results we present for KNN classifier the classification report (Table 6.11) and the confusion matrix (Figure 6.15). Table 6.11 presents the detailed classification report for the KNN classifier with the optimal parameters in the wavelet scattering domain. The precision, recall, and F1-score for each class demonstrate the effectiveness of the WST in distinguishing between different digits. The overall accuracy is reported as 0.98, with both the macro average and weighted average F1-scores also reaching 0.98. The confusion matrix shown in Figure 6.15 provides a visual representation of the classification results. It is clear from the matrix that the majority of the digits are correctly classified, with only a few mis-classification errors. This further supports the high accuracy achieved by the KNN algorithm with the wavelet scattering transform.

Class	**Precision**	**Recall**	**F1-score**
0	1.00	0.97	0.98
1	1.00	1.00	1.00
2	0.91	1.00	0.95
3	1.00	0.93	0.97
4	1.00	1.00	1.00
5	0.97	1.00	0.98
6	0.97	0.97	0.97
7	0.97	0.97	0.97
8	1.00	1.00	1.00
9	1.00	0.97	0.98
Overall			
Accuracy			**0.98**
Macro Avg	0.98	0.98	**0.98**
Weighted Avg	0.98	0.98	0.98

Table 6.11: *K-Nearest Neighbor Classification report* with optimal parameters in the wavelet scattering domain

Table 6.11 shows a notable difference between STFT and wavelet scattering representations, that can also be appreciated in Figure 6.15. While STFT does not allow simple algorithms to reach extremely accurate performances in classifying digits, the wavelet scattering transform does. The classification report for KNN using the WST representation (Table 6.11) depicts high performances values among all the classes. The overall accuracy achieved by KNN with the wavelet scattering transform representation is 0.98, indicating a strong clustering performance. In contrast, the performance of KNN with the STFT representation was comparatively lower. The classification report and confusion matrix for KNN using the STFT representation are not provided in the given information. However, previous discussions mentioned an accuracy of 0.79 for STFT-based KNN classification. The superior clustering ability of the wavelet representation can be attributed to its stability and invariance properties. The wavelet scattering transform captures both frequency and higher-order statistical information in a multi-resolution manner, providing a more comprehensive representation of the signals. This enables the KNN algorithm to effectively distinguish between different classes, resulting in higher accuracy and F1-scores.

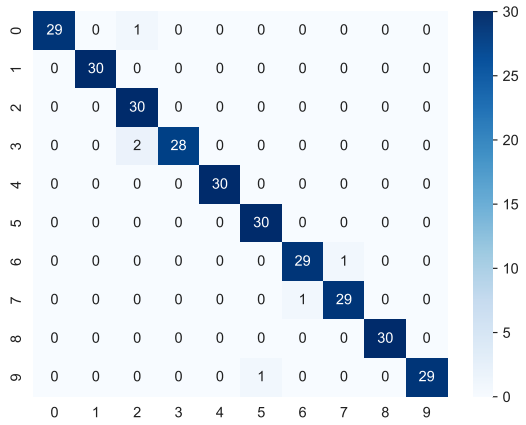


Figure 6.15: *Confusion Matrix* for the *K-Nearest Neighbours* with the optimal parameters in the WST space

The result obtained in this section are very good, and perfectly aligned with all the theory explained in Chapter 1, and perfectly captures the essence of the wavelet scattering transform. To achieve accurate results after applying the wavelet scattering transform it is not needed to run expensive algorithms like ConvNets: **data is well-separable in the wavelet scattering domain!**
To further assess, from a qualitative point of view, the separability capacity of the wavelet scattering we try to find representative spectrograms, in a way similar to what is done to find *eigen-faces*, using the Principal Component Analysis (PCA). We expect to find similarities with the results of KNN portrayed in Table 6.11. Figure 6.16 and Figure 6.17 show the representatives for each class. This result is in armony with the results provided by KNN in the previous section, in fact there seem not to be, from simple observations, any similar patterns among classes. Furthermore energy regions which explain more variance are different for each class, thus an empirical proof of the fact that wavelet scattering transform separates well the data. This approach allows us to capture the dominant patterns within each class and obtain representative samples that best represent the characteristics of that particular digit class. By examining these class representatives, we can gain insights into the distinctive features and variations present within each digit category.

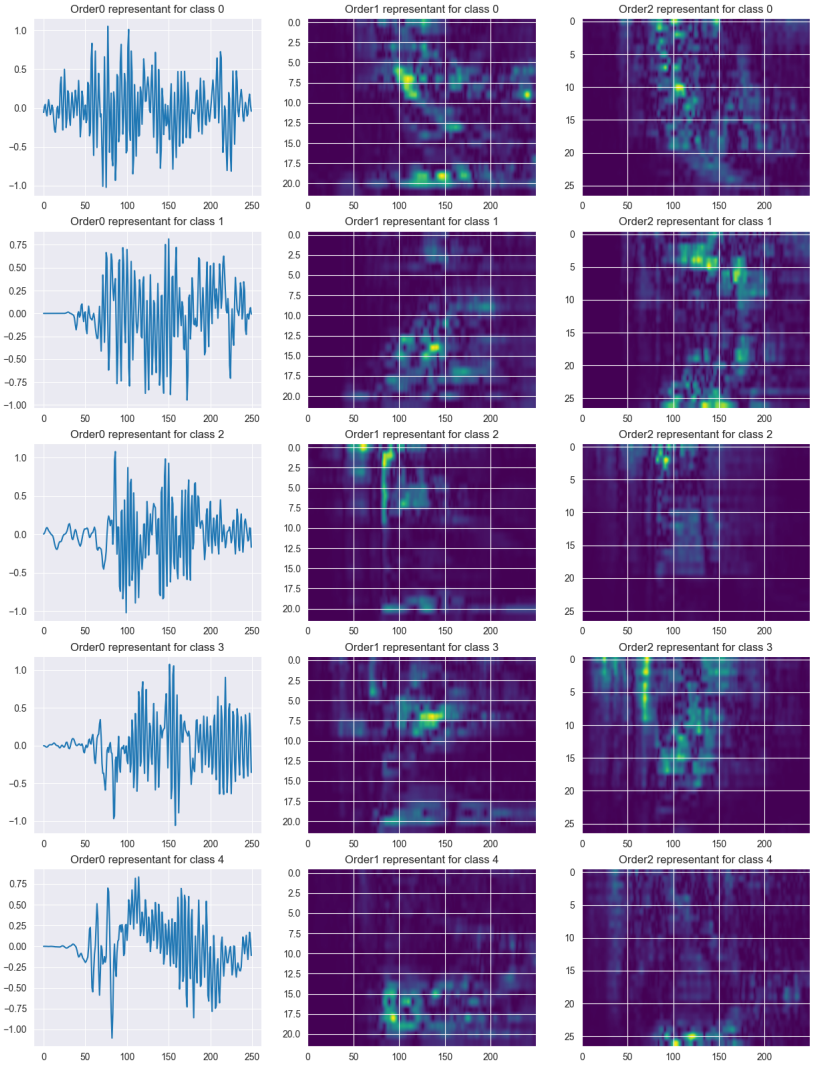


Figure 6.16: WST representants obtained via PCA, for classes 0 to 4

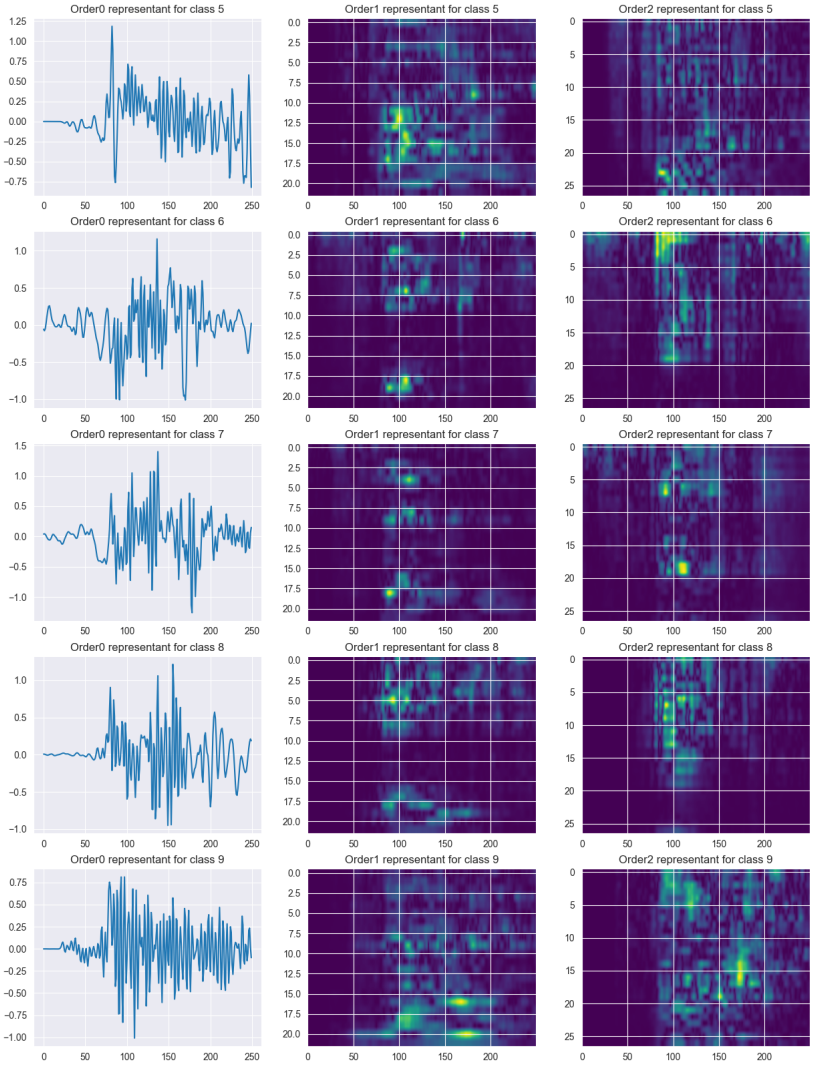


Figure 6.17: WST representants obtained via PCA, for classes 5 to 9

Comparing the class representatives across different classes can help us identify common patterns and differences, providing a deeper understanding of the discriminative power of the wavelet scattering transform and its ability to capture the inherent structure of the data. Analyzing the class representatives obtained through PCA can provide visual evidence of the effectiveness of the wavelet scattering transform in separating and clustering the digit classes, further supporting the high accuracy and F1-score achieved in the KNN-WST classification.
In addition to its application in digit classification, the approach of computing PCA and generating class representatives can also be utilized in other domains, such as the analysis of gravitational waves. In the study of gravitational wave data, one of the challenges is to identify and classify different types of noise or glitches that can contaminate the signals. By applying a similar PCA based technique to construct class representative, we can extract the dominant features from the data and create representative samples for each class of glitches. These representative samples can then be used to study the characteristics and patterns of different types of glitches, aiding in their identification and classification. The utilization of this approach shows its versatility and effectiveness not only in digit classification but also in various scientific domains where pattern recognition and clustering analysis play a central role.

Chapter 7

Conclusive Comments on the Analysis

This chapter aims to summarize the results obtained from the experiments conducted on the Free Spoken Digits Dataset. Table 7.1 provides valuable insights into the classification results obtained in Chapter 6. Comparing the results, we observe distinct patterns and variations in the classification performance across the STFT and wavelet scattering transform representations.

Classifier	**Representation**	**Avg F1-score**	**Accuracy**
DecisionTree	STFT pooled	0.68	0.68
RandomForest	STFT pooled	**0.98**	**0.98**
LogReg	STFT pooled	0.90	0.90
FFNN	STFT pooled	0.91	0.91
DecisionTree	STFT	0.62	0.62
RandomForest	STFT	0.96	0.96
LogReg	STFT	0.84	0.84
FFNN	STFT	0.86	0.85
KNN	STFT	0.80	0.79
DecisionTree	WST	**0.81**	**0.81**
RandomForest	WST	0.97	0.97
LogReg	WST	**0.98**	**0.98**
FFNN	WST	**0.99**	**0.99**
KNN	WST	**0.98**	**0.98**

Table 7.1: Performance summary of Free Spoken Digits analysis

As far as the results obtained with STFT representation and average pooling are

concerned, the Random Forest classifier stood out as the top performer, achieving an impressive average F1-score and accuracy of 0.98. This highlights the ability of Random Forest to effectively capture the discriminative patterns present in the pooled STFT spectrograms. Other classifiers such as Decision Tree, Logistic Regression, and FFNN also demonstrated moderate performances, ranging from F1-scores of 0.68 to 0.91. However, they fell slightly short compared to the Random Forest in terms of predictions accuracy.

Moving to the STFT spectrograms without pooling, the Random Forest classifier continued to exhibit strong performance with an average F1-score and accuracy of 0.96. This reinforces the fact that this algorithm is capable to handle the raw spectral information with effectiveness. On the other hand Logistic Regression, FFNN, and KNN classifiers provided decent results, achieving F1-scores ranging from 0.8 to 0.86, without achieving the results provided by the Random Forest.

As far as the wavelet scattering transform representation is concerned, Table 7.1 shows that this mathematical operator outperformed the STFT representations. The FFNN classifier excelled in this domain, achieving a F1-score and accuracy of 0.99. Logistic Regression closely followed with an F1-score and accuracy of 0.98. These results clearly highlight the superior discriminative power of the wavelet scattering metric, enabling the classifiers to effectively differentiate between different digit classes. The Random Forest and KNN also demonstrated strong performance, achieving F1-scores and accuracies of 0.97 and 0.98, respectively. The Decision Tree achieved a respectable F1-score and accuracy of 0.81.

Moreover, classification pipelines based on the wavelet scattering transform hold promise beyond digit classification. The stability and invariance properties of the wavelet scattering transform make it suitable for analyzing and classifying other types of data, including gravitational wave glitches, as portrayed in Part III. By adapting the techniques discussed in this study, it is possible to build a pipeline to classify and analyze glitches in gravitational wave data.

In conclusion, the results disserted in this chapter demonstrate the promising perspectives for the wavelet scattering transform in capturing discriminative features for digit classification. The high performances achieved by simple algorithms such as K-Nearest Neighbors highlights the effectiveness of the wavelet scattering representation in separating the data. The comparison between STFT and WST representations further underscores the superiority of the wavelet scattering transform, emphasizing its stability and invariance properties.

Part III

Wavelet Scattering Transform. An Application to VIRGO Gravitational Waves Data

Chapter 8

Gravitational Waves and VIRGO Interferometer

Gravitational waves have gained more interest in the scientific community in the last decade, since the detection of the first gravitational wave in 2015. The remarkable event, marked as *GW150914*, was detected in Laser Interferometer Gravitational-Wave Observatory (LIGO) in the United States. The detection of GW150914 provided strong confirmation of Einstein's general theory of relativity [8]. According to general relativity, massive objects, such as black holes, create ripples in the fabric of space-time, known as gravitational waves, when they accelerate or undergo violent events like merging. In the case of GW150914, the observation of gravitational waves matched the predictions made by general relativity. The detected signal exhibited specific characteristics that aligned with what Einstein's theory described. These characteristics included the waveform shape, the frequency evolution over time, and the observed amplitude of the gravitational waves. The confirmation of general relativity in this context was significant because it provided direct evidence for the existence of gravitational waves, which had long been predicted but not directly detected until then. This discovery opened up a new avenue of exploration in astrophysics, enabling scientists to study the universe through gravitational waves and deepen our understanding of cosmic phenomena, such as black holes, neutron stars, and the nature of gravity itself.
This chapter, therefore, aims to provide a brief simple introduction to gravitational waves and the main historical steps that led to the detection of GW150914 that confirmed of Einstein's general relativity theory [41]. Furthermore we describe how gravitational wave detectors work, with particular emphasis to VIRGO interferometer. VIRGO is a gravitational wave observatory located in Pisa, operated by a European collaboration between the Italian *Istituto Nazionale di Fisica Nucleare* (INFN) and the French *Centre National de la Recherche Scientifique* (CNRS), which

contributes to the global effort in detecting and studying gravitational waves.

8.1 Introduction to Gravitational Waves: a General Framework

In a similar manner to Maxwell's groundbreaking work in the 19-th century on electromagnetic waves, Einstein's purely theoretical predictions in 1916 anticipated the existence of gravitational waves. Gravitational waves possess a nature entirely distinct from electromagnetic waves, yet they emerge from a mechanism that bears striking similarities. While electromagnetic waves originate from the fluctuation of electric charges, the generation of gravitational waves only requires the substitution of these charges with masses. It can be intuitively stated that if a mass undergoes accelerated motion, it induces oscillations in the gravitational field, giving rise to propagating gravitational waves at the speed of light. As a result, masses subjected to gravitational waves are compelled to move in accordance with this dynamic gravitational field.
However, there are significant distinctions between electromagnetism and gravitation. For instance, electric charges can possess both positive and negative values, resulting in either attraction or repulsion, while masses are uniformly attractive and share the same sign. This distinction is tied to the fact that gravitational waves are generated by accelerated masses only when the motion lacks spherical symmetry (for example, two colliding stars produce gravitational waves, whereas a collapsing star maintaining perfect sphericity does not). However, gravity stands apart from any other phenomenon, necessitating the involvement of *special relativity* and *general relativity* with astonishing effects: the equivalence of mass and energy, the presence of mass warps the fabric of space-time, the fact that freely moving objects follow curved trajectories and finally time is not invariant across different frames of reference. In a concise statement: *masses determine the deformation of space-time, while space-time governs the motion of masses.*
All the elements are set to comprehend the effects of a passing gravitational wave. From a Newtonian perspective, we would describe it as fluctuations in the gravitational field. Instead, according to Einstein's perspective, we assert that the fabric of spacetime undergoes variable deformation. In greater detail, this process is illustrated in the Figure 8.1. In the presence of gravitational waves descending from above (depicted as undulating red arrows), the traversed space exhibits the behavior of a substantial block of gelatin: the undisturbed space, represented in black, periodically transforms into red and blue parallelepipedes. Consequently, the mutual distances between freely moving objects, masses, within this region of space undergo specific variations. These distances first increase along one direction while simultaneously decreasing along the orthogonal direction, and then the pattern is

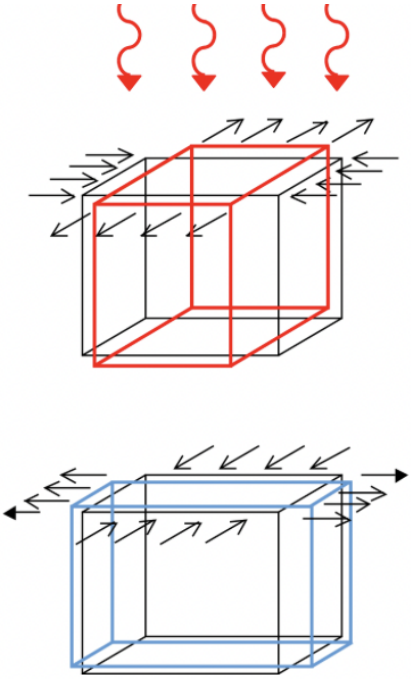


Figure 8.1: Effect on masses of gravitational waves [42]

reversed, repeating this alternating sequence multiple times.
Gravitational waves are of primary importance as they offer a new perspective on our understanding of the Universe. Until now, our exploration has largely relied on capturing diverse electromagnetic radiations (visible light, infrared and ultraviolet radiation, microwaves, X-rays, gamma rays, etc.) originating from the electric charges present in celestial bodies. However, gravitational waves, as we have seen, are a result of mass, a fundamentally different and charge-independent property. Consequently, we can anticipate that they will provide us with entirely novel and astonishing information. Significantly, the study of gravitational waves generated during catastrophic astrophysical events, particularly in the presence of intense gravitational fields, holds paramount importance. Presently, our ability to solve Einstein's equations is limited to scenarios with weak gravitational fields, necessitating further experimentation to comprehend their behavior in situations characterized by strong fields. By continuously enhancing the sensitivity of our instruments, we aspire to unveil the cosmological background of gravitational waves—an equivalent to the cosmic microwave background—offering invaluable insights into the primordial moments that followed the Big Bang.
According to Einstein's general relativity, the generation of gravitational waves

occurs whenever a physical object accelerates- humans, cars, celestial bodies. However, the masses and accelerations of objects on Earth are too small to produce gravitational waves of sufficient magnitude that can be detected by our instruments. To identify substantial gravitational waves, we must extend our observations far beyond the confines of our own solar system.
The vast expanse of the Universe is populated with immensely massive objects that undergo rapid accelerations, giving rise to gravitational waves that we are now capable of detecting. Notable sources of gravitational waves include pairs of orbiting black holes or neutron stars, as well as **binary systems** comprising a neutron star and a black hole. Additionally, gigantic stars reaching the end of their lives and undergoing explosive supernova events also generate gravitational waves. Astronomers have classified gravitational waves into four categories based on the objects or systems that generate them: **continuous**, **compact binary inspiral**, **stochastic**, and **burst**.

8.1.1 Continuous Gravitational Waves

Continuous gravitational waves play a significant role in the study of spinning massive objects such as neutron stars, aligning with the research focus of this thesis. When a neutron star possesses irregularities or deformities on its spherical surface, they give rise to gravitational waves as the star spins. As long as the spin rate of the neutron star remains constant, the emitted gravitational waves will also maintain a consistent frequency and amplitude. This behavior is similar to a musical instrument holding a single note, therefore the term continuous gravitational waves has been applied to describe this phenomenon. To better understand and appreciate the nature of continuous gravitational waves, researchers have conducted simulations to convert the signals detected by LIGO into audible sounds. These simulations allow us to perceive the gravitational waves produced by a spinning neutron star as if they were transformed into sound.

8.1.2 Compact Binary Inspiral Gravitational Waves

The only class of gravitational waves that interferometers have detected by now belong to the class of compact binary inspiral gravitational waves. These signals are produced by orbiting pairs of massive and dense objects, such as white dwarf stars, black holes, and neutron stars. They divide into three subclasses

- Binary Neutron Star (BNS), i.e. two neutron stars orbiting each other
- Binary Black Hole (BBH), i.e. two black holes orbiting each other
- Neutron Star-Black Hole Binary (NSBH), i.e. a neutron star and a black hole orbiting each other.

The phenomenon of inspiral is of great importance in the study of compact objects and their gravitational interactions. Over the course of millions of years, pairs of dense and compact objects, such as neutron stars or black holes, engage in an intricate orbital dance, Figure 8.2. As they revolve around each other, gravitational waves are emitted, carrying away a portion of their orbital energy. This continuous radiation of gravitational waves gradually brings the objects closer together. As the distance between them decreases, their orbital velocity increases, intensifying the emission of even stronger gravitational waves. Consequently, the objects lose more orbital energy, leading to a perpetual cycle of drawing closer, orbiting faster, losing energy, and moving even closer still. They are entrapped in an inescapable spiral of acceleration—a fate sealed by the relentless force of gravity. To illustrate

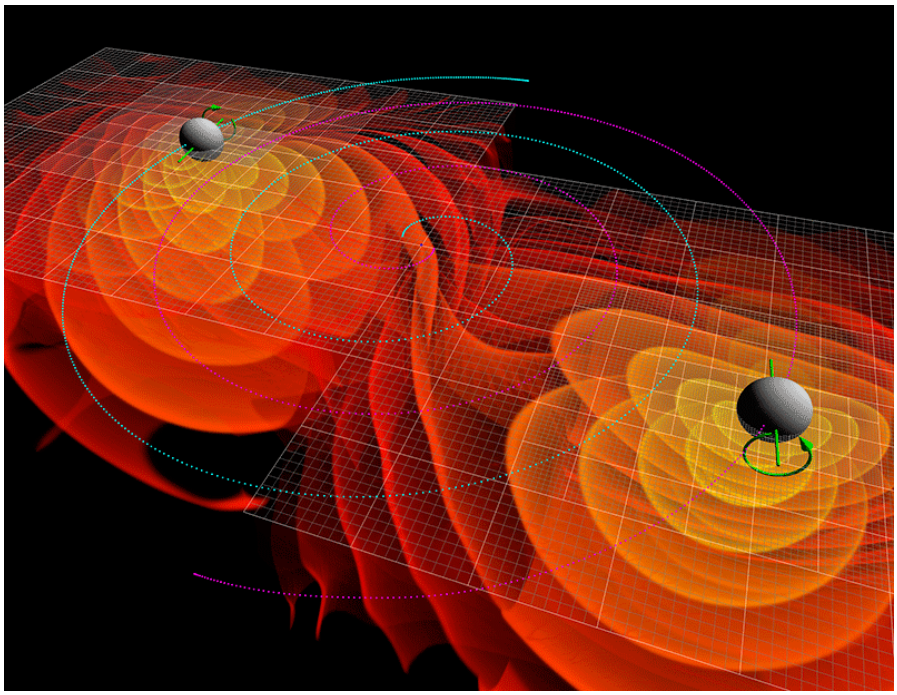


Figure 8.2: Two black holes spinning and generating a BBH gravitational wave [43]

this process, envision a figure skater performing a spinning routine. Imagine their outstretched fists representing the neutron stars or black holes, while their body symbolizes the gravitational force binding these objects together. Similar to how a skater spins faster by bringing their fists closer to their body, the pairs of compact objects also accelerate their orbital motion as they orbit closer to one another. This is an embodiment of the conservation of angular momentum in action. Unlike the skater, however, the pairs of compact objects cannot halt their rotation. The emission of gravitational waves and the ever-decreasing orbit initiate an unstoppable

sequence of events that can only culminate in a cataclysmic collision between the objects. This description of the inspiral process provides valuable insight into the dynamics and fate of compact object pairs. The thesis aims to delve deeper into this phenomenon, exploring its implications, and expanding our understanding of the intricate interplay between gravity, orbital motion, and the emission of gravitational waves.

8.1.3 Stochastic Gravitational Waves

In the realm of gravitational wave detection, astronomers have anticipated that the occurrence of significant continuous or binary inspiral gravitational wave sources in the Universe is exceedingly rare. Consequently, research institutes, such as VIRGO or LIGO, do not concern themselves with the possibility of multiple such sources passing by Earth simultaneously, which could potentially generate confusing signals in the detectors. However, it is presumed that numerous small gravitational waves permeate the Universe, constantly traversing through space from various directions. These waves, originating from diverse sources, become intermingled in a random fashion, giving rise to a stochastic signal. The term stochastic denotes a pattern characterized by randomness, which can be analyzed statistically but cannot be precisely predicted. This Stochastic Signal comprises the aggregation of these small gravitational waves arriving from all corners of the cosmos. Detecting and discerning this signal poses a formidable challenge as it represents the faintest and most elusive gravitational waves to detect. Nonetheless, there is a possibility that at least a portion of this stochastic signal stems from the Big Bang. The detection of relic gravitational waves originating from the Big Bang holds immense significance. It promises to unveil a window into the earliest moments of the Universe, granting us the ability to peer further back into cosmic history than ever before. By capturing and analyzing these primordial waves, astronomers can gain unprecedented insights into the dynamics, evolution, and fundamental properties of the Universe in its infancy. By investigating the statistical properties of these signals, researchers aim to extract valuable information about cosmic events and phenomena that have shaped the cosmos throughout its vast and enigmatic history.

8.1.4 Burst Gravitational Waves

In the quest for *burst gravitational waves*, scientists find themselves embarking on a captivating journey into the realm of the unknown. The nature of these waves remains elusive, as the main detectors have yet to detect them, and our understanding of their properties is shrouded in uncertainty. One of the challenges lies in the fact that our current understanding of the underlying physics may be inadequate to precisely anticipate the appearance and characteristics of gravitational

waves emanating from such sources. The complexity of these systems may render them enigmatic, defying our attempts to predict their gravitational wave signatures with certainty. This knowledge gap necessitates a humble approach, recognizing that we may encounter systems and phenomena that eluded our awareness until now. The search for burst gravitational waves demands a willingness to recognize patterns even in the absence of prior modeling or predictions. Scientists must remain receptive to unexpected signals, being prepared to discern and interpret these patterns without preconceived biases. While the pursuit of burst gravitational waves poses a formidable challenge, it also holds immense promise. Uncovering and detecting these elusive waves have the potential to unlock revolutionary insights about the workings of the Universe.

8.2 Gravitational Waves Detection

8.2.1 Importance of Studying Gravitational Waves

Historically, scientists have primarily relied on electromagnetic radiation, such as visible light, X-rays, radio waves, and microwaves, to study the Universe. Some have also explored the use of subatomic particles, e.g neutrinos. Each of these information carriers provide scientists with a unique perspective on the Universe, complementing one another. However, gravitational waves are fundamentally different from EM radiation. They are as distinct from light as hearing is from vision. Imagine if humans were a species that solely relied on their eyes to observe and comprehend the Universe. Through the study of light emitted by celestial objects, astronomers have made significant discoveries in the past century. But then, imagine a breakthrough invention: an *ear* that can sense distant vibrations in air or water, vibrations that have always been present but were previously undetectable to the eyes. This new sense would enable humans to gain insights about the Universe that were previously unattainable through light alone. This analogy helps us grasp how LIGO has opened a new "window" to the Universe. LIGO and VIRGO act as a gravitational wave antenna, capable of detecting vibrations in the very fabric of space-time, emanating from the farthest corners of the cosmos. Events like colliding black holes, which are completely invisible to EM astronomers, become beacons in the vast cosmic sea for LIGO and Virgo.
Furthermore, gravitational waves interact with matter very weakly, unlike EM radiation, which can be absorbed, reflected, refracted, or bent by gravity itself. As a result, gravitational waves traverse the Universe with minimal interference, carrying pristine information about their origins.
Gravitational waves are generated by some of the most cataclysmic events in the Universe, including colliding black holes, merging neutron stars, exploding stars, and possibly even the birth of the Universe itself. Detecting and analyzing the

information carried by gravitational waves has provided us with an unprecedented opportunity to observe the Universe in ways that were previously impossible. It offers astronomers and other scientists their first glimpses of phenomena that were once unobservable. These incredibly advanced detectors, like VIRGO, have lifted a veil of mystery from the Universe and has sparked exciting new research in the fields of physics, astronomy, and astrophysics.

8.2.2 Early Detections of Gravitational Waves

On September 14, 2015, these gravitational waves were first detected by the two LIGO detectors, one in Louisiana and the other in the state of Washington. This gravitational wave is known in literature with the name GW150914 and it has been the first to ever been detected. The map in Figure 8.3 shows where the locations of the gravitational waves intereformeters. The Franco-Italian Virgo detector in Cascina, near Pisa, could not capture the signal as it was undergoing major component upgrades to enhance its sensitivity by a factor of ten. Subsequently, Virgo resumed its experimental activity alongside the American detectors, contributing to the detection of several other signals. The announcement of this significant discovery was made through simultaneous press conferences in Washington and Cascina on February 11, 2016, showcasing Virgo's important contribution. In fact, the perfectly reflective surfaces of the LIGO mirrors were realized in the Virgo laboratory in Lyon. The detected signal (referred to as GW150914) was generated by the merger of two black holes, each with a mass of approximately 30 solar masses, which occurred 1.3 billion years ago at a distance of 1.3 billion light-years from us. All of this information was deduced through a detailed analysis of the signals provided simultaneously by the two LIGO instruments. Subsequently, during the observation period that extended until January 12, 2016, additional candidate events were detected, two of which passed the rigorous analysis and were published on June 15, 2016, known as GW151226 and LVT151012, respectively. The delay in publication was due to the complexity of the required analysis. This result greatly reinforces our previous belief in the existence of gravitational waves. It was based, in perfect agreement with general relativity, on the study of binary systems of compact stars. All known systems of this type, around ten in total, consisting of pairs of neutron stars or black holes orbiting each other rapidly, exhibit a gradually decreasing orbital period over time, exactly as expected since the system loses energy due to the emission of gravitational waves.
Now we proceed explaining how the existence of gravitational waves was confirmed in an intuitive way. Let's start with a classic example: two people pulling the ends of a taut rope. If one of them shakes their end transversely, we observe a deformation of the taut rope (a wave) that propagates along the rope and reaches the opposite end after a certain time. Now suppose a large asteroid collides with

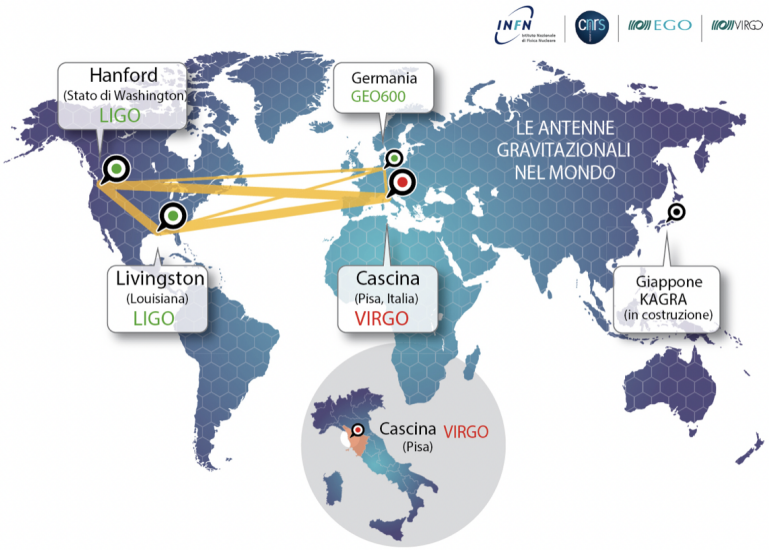


Figure 8.3: Map of the principal GW observatories [44]

the Earth, causing it to move suddenly. The Moon will not instantly notice this change; it is not reasonable to think that the Earth's gravitational field rigidly extends to any distance. Even Newton was troubled by the fact that his theory of gravitation predicted this. On the other hand, special relativity states that nothing can propagate or move at a speed faster than that of light. It is more reasonable to assume that information about the Earth's displacement reaches the Moon (approximately 400,000 km away) in a short but non-zero time. It is a deformation of the gravitational field that propagates and reaches the Moon – it is a gravitational wave – precisely as derived from Einstein's equations. Like all massless entities, gravitational waves travel through vacuum at the speed of light. After 1.28 seconds, the Moon *senses* the displacement of the Earth and moves accordingly. Figure 8.4 reconstructs the spatial diffusion in the Milky Way the most famous detected gravitational.

8.2.3 From Resonant Mass Detectors to the Interferometer

Before interferometers, the primary instruments used to detect gravitational waves were known as resonant mass detectors or bar detectors. These detectors consisted of large metal bars or spheres that were extremely sensitive to vibrations. When a gravitational wave passed through, it would cause the bar or sphere to vibrate, and these vibrations could be detected and analyzed. Resonant mass detectors operated on the principle of resonant frequencies. The detector was designed to

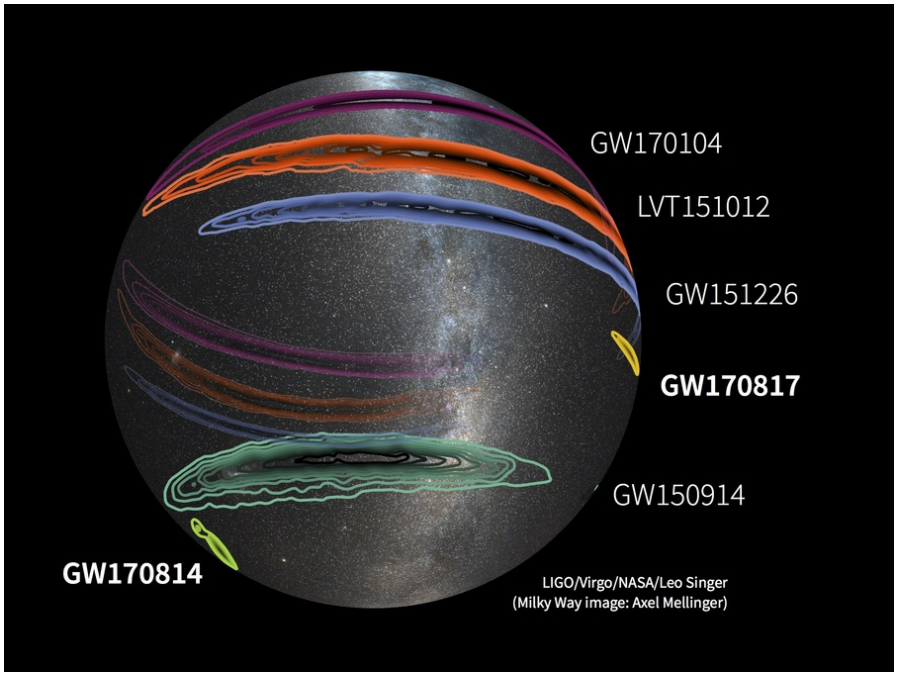


Figure 8.4: Detected gravitational waves and their spatial localization in the Milky Way [45]

vibrate at a specific frequency, and when a gravitational wave of the corresponding frequency interacted with the detector, it would induce resonance and amplify the vibrations, making them easier to detect. Some examples of early gravitational wave detectors include the Weber bar detector, developed by Joseph Weber in the 1960s, and the cryogenic resonant bar detectors such as Nautilus, Explorer, and Auriga, [46] and [47]. These detectors laid the groundwork for the development of more advanced interferometric detectors like LIGO and VIRGO, which have greatly improved the sensitivity and precision in detecting gravitational waves. A Michelson interferometer, Figure 8.5, such as those used in Virgo and LIGO, is the ideal instrument for detecting spatial deformations of this kind. In an interferometer, two laser beams are sent in two perpendicular directions, and the interference pattern between the two beams reflected back from the end mirrors of the long arms is continuously monitored using a photodetector, which is an array of light-sensitive diodes. Changes in the interference pattern are caused by variations in the length of the arms, which in turn are due to the deformations of space generated by the passage of gravitational waves. The longer the arms, the greater their length variations. To further enhance the effect of length variations, the light beams are made to travel back and forth through the arms hundreds of times before recombining. These gigantic instruments are sensitive to length

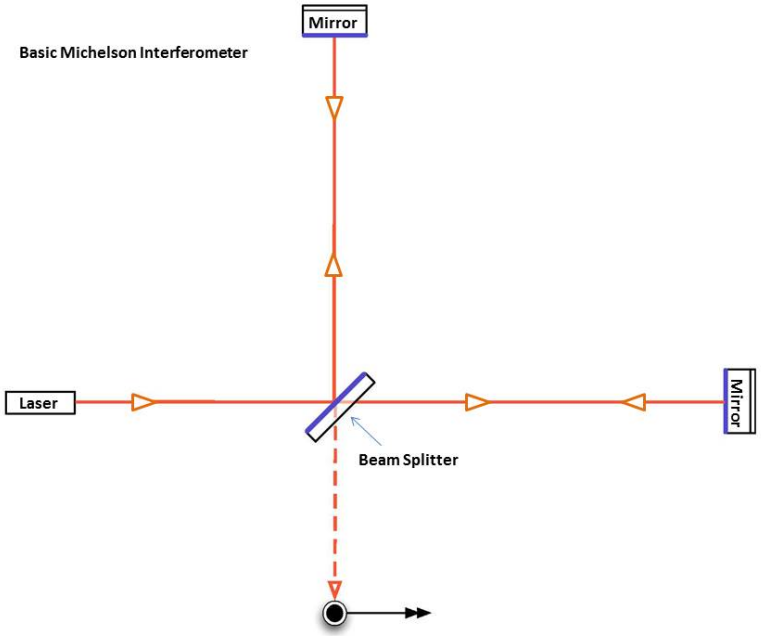


Figure 8.5: Michelson Interferometer [48]

variations much smaller than the size of an atomic nucleus; that's how tiny the effects produced by the most intense gravitational waves that can reach Earth are. Due to the extremely small magnitude of the effects of a gravitational wave, it is necessary to try to detect the most intense waves available, which are estimated to be produced in catastrophic astrophysical events involving enormous masses experiencing intense accelerations. It would be impossible to produce waves with perceptible effects in the laboratory. In the document where Einstein predicted the existence of gravitational radiation, he wrote that no one would ever be able to experimentally confirm this phenomenon.
Today, we can contradict Einstein thanks to technologies that he could not even imagine. Such small length variations (or displacements of the mirrors that define the interferometer arms) can be caused by gravitational waves but also by noise. This term refers to all spurious phenomena that can simulate or conceal the genuine effects of gravitational waves. The experimental apparatus must be designed in a way that eliminates or reduces the effect of noise. Then, distinguishing the origin of the detected signals and selecting the good ones while eliminating those due to noise is the task of data analysis and machine learning.

8.3 VIRGO Interferometer

VIRGO, located in Italy, is an interferometer that follows the design of the Michelson interferometer, with arms stretching over a distance of 3 km, see Figure 8.6. The main components of VIRGO are the mirrors that define the length of the arms. These mirrors are not simply fixed in place but are suspended using chains of pendulums within massive vacuum chambers. These vacuum chambers, measuring 2 meters in diameter and 11 meters in height, serve the important purpose of isolating the mirrors from the constant seismic vibrations that occur on the ground as well as the sound vibrations transmitted through the air. By suspending the mirrors with pendulum chains, VIRGO effectively minimizes any disturbances that could affect the precision of its measurements. To ensure the utmost sensitivity and accuracy, the laser beams used in VIRGO propagate within ultra-high vacuum tubes. These tubes have a diameter of 1.2 meters and stretch over the entire length of the arms, covering a distance of 3 km. The purpose of maintaining such a high vacuum is to eliminate any interference caused by convective motions of the air. By reducing air particles and their movements, VIRGO ensures that the laser beams remain undisturbed and can travel through the arms unimpeded, allowing for precise interference patterns to be detected. The mirrors themselves are meticulously crafted with exceptional smoothness, reaching a level of precision within a few nanometers. Such high-quality mirror surfaces are vital to minimize

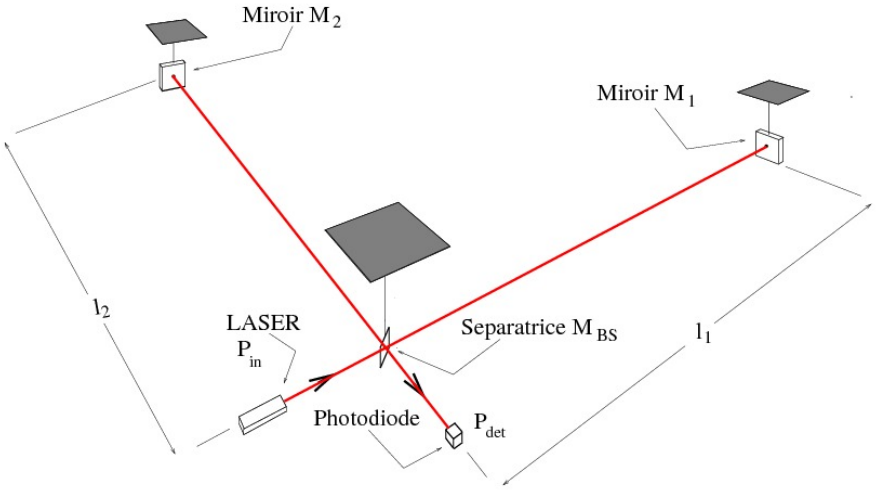


Figure 8.6: Virgo Interferometer [49]

any scattering or diffusion of the laser light, ensuring that the interferometer can accurately capture the interference patterns between the laser beams. The laser used in VIRGO is not only finely tuned in terms of its wavelength but also stabilized

with exceptional precision in terms of intensity. This ensures that the laser beams maintain their desired characteristics throughout the interferometer, allowing for reliable and consistent measurements. American interferometers, LIGO , share structural similarities with VIRGO. LIGO also follows the Michelson interferometer design and operates based on the same fundamental principles. However, LIGO's arms extend over a greater distance, reaching 4 km in length. Despite these differences in arm length, both VIRGO and LIGO contribute to the global network of gravitational wave detectors, collaborating to capture and analyze the faint signals of gravitational waves that traverse the cosmos.

Chapter 9

Machine Learning and AI. A New Frontier in Gravitational Waves Research

This chapter provides an in-depth exploration, building upon the framework presented in [50], of the significant difficulties that arise in the study of gravitational waves. Gravitational wave detection and analysis pose interesting challenges due to various factors, such as the inherently weak nature of the signals, the presence of noise and disturbances, and the complex and diverse astrophysical sources. To tackle these challenges, the integration of artificial intelligence and advanced machine learning techniques, particularly deep learning algorithms, has emerged as a powerful approach. AI offers the potential to revolutionize gravitational wave analysis by enabling automated and intelligent data processing, interpretation, and classification. By leveraging the capabilities of AI, researchers can develop sophisticated algorithms that excel at identifying subtle gravitational wave signatures buried within noisy data. These algorithms can learn from vast amounts of labeled data, allowing them to discern patterns, extract relevant features, and make accurate predictions. Deep learning, in particular, due to its ability to model complex relationships and hierarchies in data, proves to be well-suited for analyzing the nature of gravitational wave signals. The application of AI and deep learning in the field of gravitational wave astronomy not only aids in the detection and characterization of signals but also contributes to a deeper understanding of the physical processes underlying these phenomena. By analyzing the data with AI-driven algorithms, researchers can extract valuable insights into the astrophysical

sources of gravitational waves, such as black hole mergers, neutron star collisions, and cosmic explosions, introduced in Chapter 8. These insights can shed light on fundamental questions regarding the nature of gravity, the structure of spacetime, and the origins and evolution of the Universe itself.

9.1 Challenges in Gravitational Wave Detection

The future of gravitational wave astronomy faces various challenges, including the need to process and analyze a higher rate of detections and to refine astrophysical investigations. Accurate reconstructions of gravitational waves signals and the identification of errors are crucial. Additionally, addressing instrumental and environmental data artifacts requires efficient methods for characterization. Machine Learning algorithms offer promising solutions to these challenges. The LIGO and Virgo Scientific Collaborations employ different techniques for gravitational waves searches, and machine learning algorithms can enhance their sensitivity and robustness. Machine learning techniques have been successfully applied in automated data analysis, noise classification, searches for compact binary coalescences, parameter estimation and noise removal.
Gravitational wave interferometers capture the strain of spacetime, i.e. a time-series that will be referred to as $h(t)$. The sensitivity of an ideal detector is determined by its intrinsic design physics and is limited by fundamental sources of noise, such as quantum and thermal noise. However, real-world detectors also face additional challenges from technical noise sources and non-stationary disturbances. These disturbances can arise from feedback control systems, instrumental and environmental factors, and non-linear coupling with the detector strain.
The presence of non-stationary and non-Gaussian noise, as well as noise artifacts, can impact data quality, interferometr performance, and increase the false alarm rate of gravitational wave searches. In particular, short-lived noise disturbances known, as **detector glitches**, have been observed to affect the detection and parameter estimation of astrophysical transient signals. The detection of the GW170817 binary neutron star merger event highlighted the importance of addressing these issues.
The performance of gravitational waves searches is further affected by non-stationary and non-linear transients, as well as continuous noise signals in the form of spectral lines. These factors necessitate the development of techniques to characterize and reduce detector noise in order to improve the sensitivity. Previous studies conducted by LIGO-Virgo researchers described various methods for identification and mitigation of gravitational waves data quality issues. The following pages portrait the most challenging open points that scientific community is addressing. Within this framework, our analysis positions itself as a study that introduces the

wavelet scattering transform as a powerful mathematical tool for the analysis of gravitational signals and the detection of glitches. In the context of gravitational wave analysis, the wavelet scattering transform holds significant potential. Its ability to capture both temporal and spectral information makes it well-suited for the study of gravitational wave signals, which are characterized by complex frequency components and time-varying characteristics. By decomposing the signals into a hierarchical structure, the wavelet scattering transform facilitates the extraction of relevant features and the identification of subtle variations in the presence of noise and other interference.

9.2 Machine Learning for Glitch Classification

Characterizing the detection of transient noise involves the crucial step of distinguishing glitches from potential astrophysical signals and classifying them into different families. This task can be challenging due to the complex temporal and frequency evolution exhibited by glitches, making their characterization with a fixed number of features difficult. Furthermore, as the sensitivity of detectors increases, the number of glitch morphologies also grows, further complicating the analysis. Machine learning techniques offer a solution to the problem of glitch classification. Previous research has demonstrated the effectiveness of various unsupervised methods [51] and neural networks [52] in this perspective .
One particularly promising approach for glitch classification is the use of Deep Convolutional Neural Networks that excel at extracting features from time-frequency representations. CNNs are designed to analyze 2D matrices, such as images, and utilize the extracted features for classification purposes. By feeding time-frequency transforms, such STFT Q-transforms, to a CNN-based deep network, glitch classification can be achieved effectively. A study implemented an image-based detection and classification pipeline using 2D CNN layers, achieving around 99% accuracy in the classification and differentiation of glitches from chirp-like signals. ConvNets have shown superior performance in distinguishing glitches with similar morphology compared to other modern approaches. Time-frequency maps, such as spectrograms, can be constructed to represent glitch appearance, and ConvNets possess the ability to automatically extract significant features from such images, enabling differentiation between different glitch types. However training these structure becomes extremely costly as the number of glitches increases. Therefore introducing a new representation, like the wavelet scattering transform, may significantly reduce the number of parameters on which the model depends. Ideally using a wavelet scattering transform representation for the glitches should allow to reach accurate results, but reducing the computational effort, hence making the pipeline more efficient and faster.

9.3 Glitch Characterization from Data Collected by Auxiliary Channels

The LIGO and Virgo detectors collect data from various subsystems that control different aspects of the instruments and monitor their state. These subsystems include instrumental and environmental sensors like photodetectors and seismometers, which can detect noise sources affecting the interferometers. Analyzing the data from these auxiliary channels is challenging due to the large number of sensors involved. Machine learning techniques have proven valuable in handling the vast amounts of auxiliary channel data. Researchers have extensively studied methods for identifying glitches in the data using auxiliary channels within the LIGO and Virgo collaborations. These approaches use the gravitational wave channel to determine labels for training samples, while the glitch identification process relies only on information from the auxiliary channels. Statistical learning algorithms, e.g. Random Forests, neural networks, and support vector machines have been explored for this purpose. One example is the *iDQ glitch detection pipeline* [53], which operates in real-time and utilizes features from auxiliary channels to train machine learning algorithms for identifying glitches in the target channel. iDQ played a crucial role in the rapid release of the GW170817 binary neutron star event by identifying the glitch coincident with the gravitational wave trigger and confirming the presence of an astrophysical signal.
LIGO Scientific Collaboration has also developed additional approaches for glitch identification using auxiliary channels, particularly in high-latency settings. Two fast algorithms based on Random Forest and Genetic Search aim to track the causes of glitches with minimal tuning. These algorithms use features derived from real-time data quality pipelines and can be quickly trained and run on computing clusters. Another machine learning tool called *Elastic-net based machine learning for Understanding*, EMU, [54], utilizes data from all auxiliary channels per detector site to predict the probability of a glitch. EMU provides insights into the significance of auxiliary channels in predicting glitches and helps uncover noise couplings to the gravitational waves data stream.

9.3.1 Further Works Relying on Auxiliary Channels Collected Data

Virgo and LIGO researchers are putting much effort in using real-time data collected by auxiliary channels. Applications other than glitch classification have been proposed. For example, regression and clustering methods have been used to infer ground velocities from Earthquake Early Warning alerts, allowing the detector

control configuration to be adjusted during periods of excessive ground motion. Additionally, *Hey LIGO* [55] is an AI-based information retrieval tool that supports the commissioning and characterization efforts of gravitational waves observatories. It utilizes natural language processing (NLP) to search the detector logbook data and provide information on detector operation, maintenance, and characterization tasks.

9.4 Machine Learning Based Denoising Techniques

Among the wide range of open challenges in gravitational waves research, machine learning may also help to construct more accurate techniques to denoise astrophysical signals.
While linear noise can be filtered out using Wiener filtering [56], there remains a significant amount of non-linear couplings that persist in the output signal. Machine learning algorithms, particularly neural networks, can be employed to model and subtract these non-linear couplings by utilizing environmental and control data streams as input. Methods like *NonSENS* [57] and *DeepClean* [58] have been developed to implement this approach. They leverage the ability of machine learning algorithms to infer non-linear functions and find transfer functions that describe the systems producing non-linear noise in the detector output. Once trained, the neural network can effectively subtract these non-linear couplings in real-time, resulting in a lower overall noise floor. Compared to traditional filtering techniques, AI algorithms are faster and more efficient in this context.
An interesting scenario arises when the noise source can be described as linear on short time scales but exhibits varying transfer functions on longer time scales. In such cases, the coupling can be efficiently tracked using interferometer angular control signals, allowing for the development of a stable and parametric model of the time-varying noise. This model enables a time-domain subtraction approach that outperforms linear and stationary schemes. The successful application of this non-stationary noise subtraction scheme to LIGO data during the third observing run allowed for the removal of non-stationary power supply line coupling and improved detector sensitivity.
Deep learning techniques can also be employed to uncover underlying signals by applying various denoising algorithms to the data. These algorithms include the total-variation method, dictionary learning, deep learning with WaveNet [59] implementation, and deep recurrent neural networks in denoising auto-encoders architecture.

9.5 Gravitational Waves Search

This paragraph discusses how AI and machine learning can address the problem of seeking the family to which the detected gravitational wave belongs. Let us recall that there are four classes of gravitational waves: compact binary cohelescence (CBC), burst, continuous and stochastic.
The old-fashioned technique for CBC search is matched filtering, which involves cross-correlating the gravitational wave data with a bank of template waveforms. Recent developments have explored the use of Random Forest and ConvNets as alternatives to matched filtering, showing potential for improved detection efficiencies.
Gravitational wave bursts are transient signals with unknown or partially modeled waveforms. These bursts can be produced by various astrophysical phenomena such as core-collapse supernovae (CCSNe). The *coherent Wave Burst (cWB) algorithm* [60] is commonly used for burst searches, which measures excess power in the time-frequency domain. ConvNets based approaches have been applied to improve the classification of burst signals and reduce the impact of noise artifacts.
Continuous gravitational wave signals are emitted by rotating neutron stars. These signals have not yet been observed but are challenging to detect due to their small amplitude. AI offers a promising alternative to traditional grid search methods, as they can be faster once trained.

9.6 Parameter Estimation

In the context of gravitational waves, *parameter estimation* refers to the process of determining the physical properties of the astrophysical sources that produce these signals. Parameter Estimation is fundamental in gravitational waves research as it provides significant information about cosmic events that generated the signal, thus leading to a further comprehension of the Universe.
When interfermoters detect a gravitational signal, it is crucial to extract information about the parameters of its source, e.g. masses and spins of the compact objects involved in a binary merger or the properties of a continuous wave emitted by a rotating neutron star. Traditional methods for parameter estimation often rely on computationally expensive Bayesian inference techniques, such as Markov Chain Monte Carlo algorithms. However AI offers an alternative approach to parameter estimation, relying on the capability of models to learn complex relationships between signals and their astrophysical parameters, enabling more efficient and accurate estimation processes. Possible approaches for parameter estimation may be developed from generative models, such as Variational Autoencoders (VAEs)

[61] or Generative Adversarial Networks (GANs) [62]. These models can learn the underlying distribution of the gravitational signals and generate synthetic waveforms with varying parameter values. By comparing the synthetic waveforms to the observed data, machine learning algorithms can infer the most likely parameter values that produced the detected signal.
Note that the current studies on parameter estimation in gravitational waves are still in the proof-of-principle stage. However, they demonstrate the potential of AI as a promising tool for future parameter estimation. As the sensitivity of detectors continues to improve, the number of detections is expected to significantly rise. One notable advantage of employing machine learning techniques is their ability to rapidly measure the astrophysical parameters of gravitational waves sources compared to traditional methods. This advantage becomes especially valuable when dealing with a large volume of waves alerts, allowing to efficiently process them.

Chapter 10

Wavelet Scattering Transform for Gravitational Waves Data. A Collaboration with INFN and VIRGO

In this chapter, we delve into the aim and scope of our collaboration with the respected Istituto Nazionale di Fisica Nucleare (INFN) and the VIRGO observatory. Our collaborative efforts were part of a larger project, ***interTwin*** with a primary objective to discriminate glitches from genuine gravitational wave signals through real-time analysis of auxiliary channels data. Specifically, our focus was on illustrating the promising applications that the wavelet scattering transform offers, as an alternative to the state-of-the-art method, the Q-transform.
The first part of this chapter provides a detailed overview of the challenges we encountered and the goals we set out to achieve through this collaboration. We highlight the underlying reasons that make the wavelet scattering transform a suitable ally in handling gravitational wave data. By exploring the unique characteristics and capabilities of this transform, we aim to shed light on its potential in the field of gravitational wave research.
The second part of this chapter delves into the current state-of-the-art methods employed for representing the signals acquired by interferometers. Furthermore, we provide a concise description of the Python library, GWpy [11], which played a crucial role in manipulating and processing gravitational data throughout our collaboration.

Through this chapter, we aim to lay the foundation for our collaborative work and set the stage for the subsequent discussions and findings that will unfold in the following sections. In this work in collaboration with INFN and VIRGO we strive to make a contributions to the field of glitch discrimination and real-time analysis of gravitational wave data, exploiting the properties of wavelet scattering transform and statistical learning algorithms.

10.1 Project Framework: *interTwin*

This paragraph focuses on the *interTwin* project led by the Istituto Nazionale di Fisica Nucleare (INFN), which aims to develop a digital twin based on generative architectures, specifically Generative Adversarial Networks (GANs). The primary objective of this project is to create a digital representation that can generate expected spectrograms for astrophysical data collected by interferometers. To achieve this, the project utilizes auxiliary channel data obtained from various instrumental and environmental sensors, including photodetectors and seismometers.
The project involves comparing the generated spectrograms with real spectrograms to identify differences and enhance the detection and discrimination of gravitational wave events and glitches. This process not only aids in identifying astrophysical signals but also helps to determine the non-astrophysical nature of triggers.
Glitches, which are short-lived noise disturbances, can originate from anthropogenic sources, weather conditions, equipment malfunctions, or unknown sources. These glitches mimic the true astrophysical transient signal and exhibit non-stationary and non-Gaussian behavior, thus it is not trivial to characterize them. It is important to note that glitches do not necessarily couple linearly into the interferometer, making their identification and discrimination challenging.
One key aspect of the interTwin project is the utilization of auxiliary data streams to uncover the transfer function of the system responsible for producing non-linear noise in the detector output. Machine learning techniques are employed to analyze the auxiliary data and identify the underlying transfer function. This machine learning based approach enables a deeper understanding of the non-linear noise sources, facilitating the development of improved detection and discrimination algorithms.
In this framework this thesis focuses on investigating the impact of using the wavelet scattering transform as a representation method to discriminate signals, instead of the traditional Q-transform. As shown by the theoretical results presented in Part I, wavelet scattering transform offers promising advantages, such as improved discrimination capabilities and the potential for lighter ML algorithms and architectures that are faster to train. The comparative analysis of the two representation methods and their effectiveness in discriminating signals will provide

valuable insights for future advancements in glitch detection and characterization.

10.2 Methods and Materials

10.2.1 State-of-the-art Representation: Q-Transform

This section introduces the state-of-the-art representation technique for processing gravitational wave data, the Q-transform [63].
The Q-transform, also known as the Constant Q-transform (CQT), is a time-frequency analysis technique that provides a representation of a signal in the joint time-frequency domain with a constant quality factor, denoted as Q. Let $h(t)$ be a continuous-time signal, and consider a family of window functions $g_Q(t)$, e.g. Hann window functions, parameterized by the quality factor Q. The window functions have a constant bandwidth, centered at logarithmically spaced frequencies ω_k. The Q-transform of the signal $h(t)$ at a specific frequency ω_k and factor Q is obtained by the convolution:

$$H_k(t) = \int_{-\infty}^{\infty} h(\tau)\, g_Q(t-\tau)\, e^{-2\pi i \omega_k \tau}\, d\tau \tag{10.1}$$

where $H_k(t)$ represents the Q-transform coefficient at time t and frequency f_k.
To obtain the full Q-transform of the signal, we compute the Q-transform coefficients for a range of frequencies ω_k in the chosen frequency grid. The resulting Q-transform provides a time-frequency representation of the signal, where the time axis corresponds to the original signal's time domain, and the frequency axis represents the logarithmically spaced frequencies. The Q-transform is particularly useful for analyzing signals with varying frequency content, as it provides a higher resolution at lower frequencies and a lower resolution at higher frequencies. This logarithmic frequency scaling ensures that the Q-transform captures the details of signals with complex harmonic structures and varying frequency components. It enables the analysis of signals with complex frequency content and is widely applied in various domains, including audio processing, music analysis, and signal processing tasks.
From equation (10.1) it is possible to deline similarities and dissimilarities with the STFT. On one hand the STFT divides a signal into short, overlapping segments and applies the Fourier Transform to each segment. This results in a time-frequency representation where the time axis is preserved, but the frequency resolution is constant across all frequencies. The STFT provides good frequency localization at the expense of time resolution.
On the other hand, the Q-transform uses a family of window functions with constant bandwidths and logarithmically spaced frequencies, allowing for a higher frequency resolution at lower frequencies and a lower resolution at higher frequencies. The

Q-transform provides a time-frequency representation with a constant quality factor, denoted as Q, which captures the details of signals with complex harmonic structures and varying frequency components. Therefore the main difference between the STFT and the Q-transform lies in their frequency resolution characteristics. The STFT offers a constant frequency resolution but sacrifices time resolution, while the Q-transform provides variable frequency resolution with a constant quality factor, allowing for better representation of signals with varying frequency content.

10.2.2 GWpy: A Python Package for Gravitational Wave Data

This section briefly introduces GWpy [11], the package that allows to handle gravitational waves data on Python. GWpy offers a range of classes and utilities designed for analyzing data from gravitational-wave detectors, catering to both astrophysical and instrumental purposes. The core infrastructure of GWpy builds upon and extends the functionality of the Astropy package, a highly regarded suite of tools for astrophysical analysis primarily focused on FITS images. Furthermore, the methodology employed in GWpy draws inspiration from and expands upon the LIGO Algorithm Library Suite (LALSuite) [64], which comprises a comprehensive collection of *C*99 routines for manipulating and analyzing data from gravitational-wave detectors. By utilizing the SWIG program to generate Python interfaces for all C modules, GWpy harnesses the full capabilities and efficiency of these libraries. GWpy brings together various code snippets from different sources, amalgamating them into a cohesive package that simplifies their utilization. By leveraging existing resources and providing a user-friendly interface, GWpy empowers researchers to perform comprehensive data analysis in the field of gravitational-wave research.
One of the key features of GWpy is its ability to handle timeseries data using the **TimeSeries** class. This class provides a convenient and intuitive interface for loading, manipulating, and analyzing timeseries data. Users can create instances of TimeSeries objects by providing the data and relevant metadata, such as the start time, sampling frequency, and units. GWpy supports a variety of data formats commonly used in gravitational-wave research, including HDF5 (the one used in this work), ASCII, and GWF (Gravitational Wave Frame) files. Once the timeseries data is loaded into a TimeSeries object, GWpy offers a rich set of built-in methods and functions for data manipulation and analysis. Users can easily perform common operations such as time shifting, resampling, and filtering to preprocess the data according to their specific requirements. GWpy also provides convenient tools for computing statistical measures, such as mean, standard deviation, and median, as well as more advanced operations like Fourier transforms and spectrograms. Furthermore, GWpy integrates with other scientific Python packages, such as NumPy, SciPy, and Matplotlib, allowing users to leverage their functionalities

for advanced data analysis and visualization. This interoperability with popular scientific libraries makes GWpy a versatile tool to work with gravitational-wave timeseries data.

Chapter 11

Wavelet Scattering Transform. A Preliminar Analysis

This first chapter describes the first part of the work we conducted and presented to VIRGO as a preliminar analysis. We developed a qualitative comparative analysis to portrait differences among Q-transform and wavelet scattering transform representations.
The first part of the study aims to understand the standard pre-processing pipeline that researchers of LIGO and VIRGO built to enhance significant features of the time series.
We worked with four different signals captured by LIGO Hanford observatory, more precisely two gravitational waves, **GW150914** (Figure 11.1) and GW170817 (Figure 11.2), and two glitches belonging to two different classes, **Injection Noise glitch** (Figure 11.3) and **Koi Fish glitch** (Figure 11.4). The analysis is presented by the metric perspective, therefore we carefully develop the discussion with the help of distance matrices and plots. Another central part of the discussion aims to seek for the parameters J and Q of the wavelet scattering transform, since their tuning depends on the frequencies that we want to highlight. Since this is one of the first attempts of applying the wavelet scattering transform to gravitational wave data, in this thesis we try to set the basis for a method that allows to make considerations over the parameter tuning. The method should account for two main points. Firstly visual differences and patterns in the wavelet scattering transform should be present, and, furthermore, the choice of J and Q should be such that the distance between glitches and gravitational waves is large, while keeping a small distance between the gravitational waves.
In this chapter we also perform a comparative study for the choice of the metric,

in particular the *euclidean distance* and the *mean squared error* (MSE), and we try to see how considering deeper layers of the scattering transform provides different amount of information.

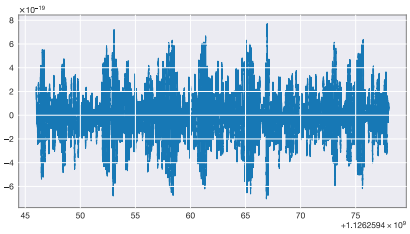


Figure 11.1: Gravitational wave GW150914

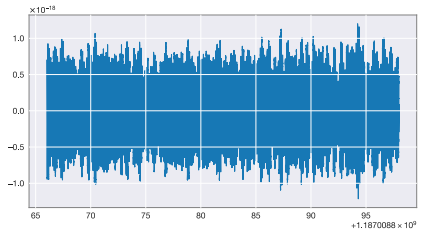


Figure 11.2: Gravitational wave GW170817

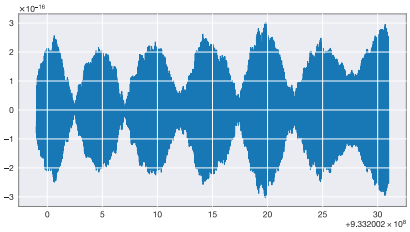


Figure 11.3: Injection Noise glitch

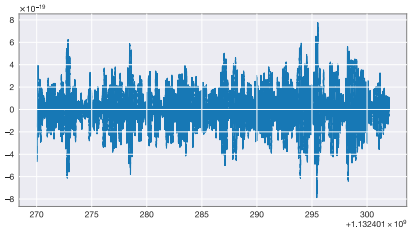


Figure 11.4: Koi Fish glitch

11.1 Data Preparation

This section aims to provide an overview of the initial steps involved in preparing the time series data collected by the interferometers, which consists of two primary phases: the *cropping* phase and the *whitening* phase. These steps are crucial for the subsequent analysis of astrophysical signals.
The nature of the collected data is such that, due to the exceptionally high sensitivity of the measurement instrument, most signals in the time domain appear to be akin to random noise, with the exception of the injection noise which exhibits a more discernible pattern. As a result, it becomes imperative to identify and employ an appropriate transform operator to effectively analyze astrophysical signals.
The data is initially stored in the *'hdf5'* format and is subsequently converted into TimeSeries objects using dedicated methods provided by GWpy. Each TimeSeries object retains the information about the time at which the event occurs, along with the associated sampling rate. In this framework, we adopt a cropping strategy where the signals are segmented using a window of 32 seconds, centered around the event time t_0. This allows for a focused analysis of the specific time intervals

relevant to the events of interest.

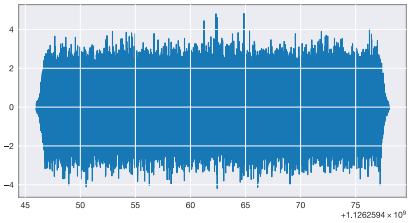


Figure 11.5: Whitened GW150914

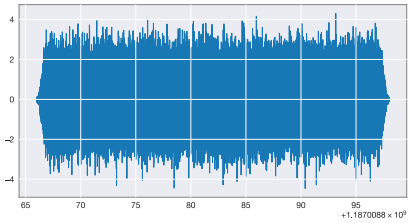


Figure 11.6: Whitened GW170817

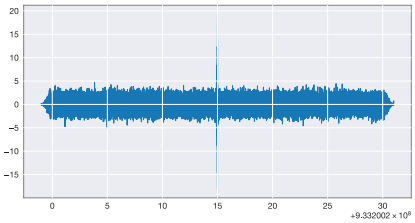


Figure 11.7: Whitened Injection Noise glitch

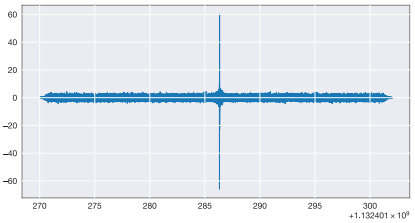


Figure 11.8: Whitened Koi Fish glitch

Figures 11.1, 11.2, 11.3, and 11.4 visually demonstrate the characteristics of the signals, highlighting the prevalence of random noise-like behavior in the time domain. These figures serve to underscore the significance of employing an appropriate transform operator to enhance the analysis of astrophysical signals within this data.

11.1.1 Whitening the Signals

Most data recorded from a gravitational-wave interferometer carry information across a wide band of frequencies, typically up to a few kiloHertz. However, it is often observed that the low-frequency amplitude dominates the high-frequency content, making it challenging to distinguish high-frequency features in the data. To address this issue, we employ the whitening technique, already implemented within the GWpy package.

The purpose of whitening is to normalize the power at all frequencies, ensuring that excess power at any particular frequency becomes more evident. By applying a whitening filter to the data, it becomes possible to equalize the power spectrum effectively, reducing the dominance of low-frequency components and highlighting the visibility of high-frequency features.

The whitening technique implemented in GWpy involves estimating the power

spectral density of the noise and then dividing the data by the square root of the power spectral density. This normalization effectively flattens the noise contribution across different frequencies, making it easier to identify and analyze specific features, such as the glitches in our example. By normalizing the power at all frequencies, the whitening process enhances the visibility of subtle features and enables a more comprehensive analysis of the data. This contributes to a deeper understanding of the underlying astrophysical phenomena that generate gravitational waves and facilitates the accurate characterization of these signals. Figures 11.5, 11.6, 11.7 and 11.8 represent, respectively, the result of whitening GW150914, GW170817, Injection Noise glitch, and Koi Fish glitch. It is interesting to note that the whitening technique provides appreciable differences between astrophysical signals and glitches. Indeed it is immediate to see that the spikes in whitened gravitational waves are not identifiable by qualitative observations, while as far as glitches are concerned the spike is quite obvious to spot in the time domain.

11.2 Wavelet Scattering Transform Analysis

This section provides a detailed analysis of the preliminary research conducted, focusing on the utilization of the wavelet scattering transform as a representation technique. The section is structured into three main phases, each addressing different aspects of the analysis.
In the first phase, the advantages of employing the wavelet scattering transform on the whitened signals are thoroughly examined and compared to the use of raw gravitational data. The second phase delves into a fundamental aspect of working with the wavelet scattering transform, which is the challenge of tuning the scattering scale parameter J and the octave parameter Q. We explore strategies to effectively address this challenge. By refining the parameter selection process, the performance and ability of finding hidden discriminative patterns in the data of the wavelet scattering transform can be optimized. The subsequent phase focuses on the analysis of different metrics and their implications in evaluating the transformed data. Through a series of insightful plots and considerations, the effectiveness of various metrics in discerning important features and patterns is examined. This analysis serves to enhance the understanding of the strengths and limitations of the wavelet scattering transform and provides valuable insights for further refinement and improvement. Finally, the section concludes by testing the application o median normalization to the transforms. This technique aims to improve the performance and consistency of the wavelet scattering transform by mitigating potential biases and variations. The benefits of this normalization approach are discussed, highlighting its potential impact on the overall analysis process.

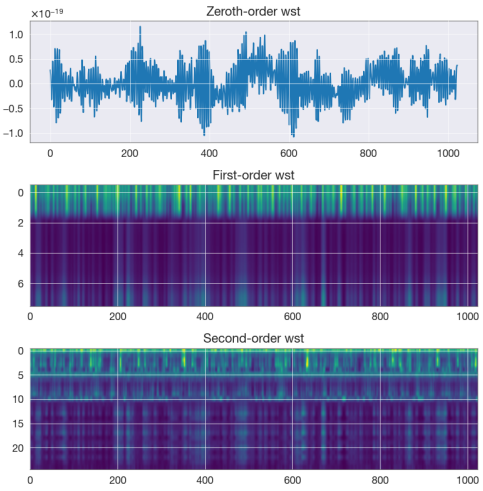


Figure 11.9: WST of raw non-whitened GW150914

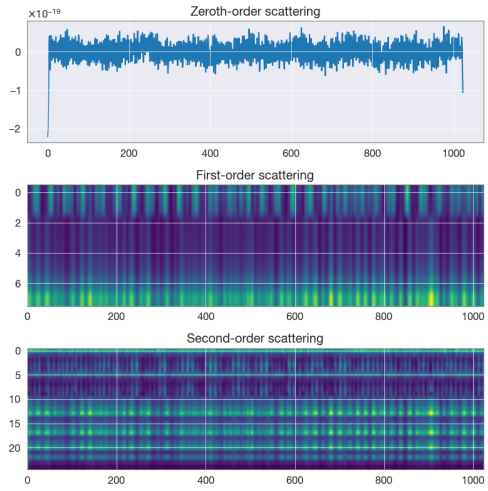


Figure 11.10: WST of raw GW170817

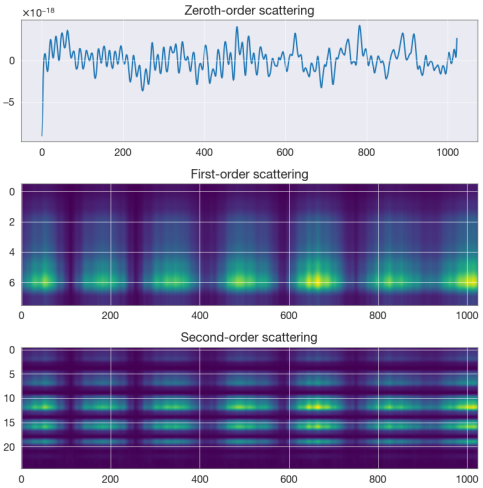


Figure 11.11: WST of raw Injection Noise Glitch

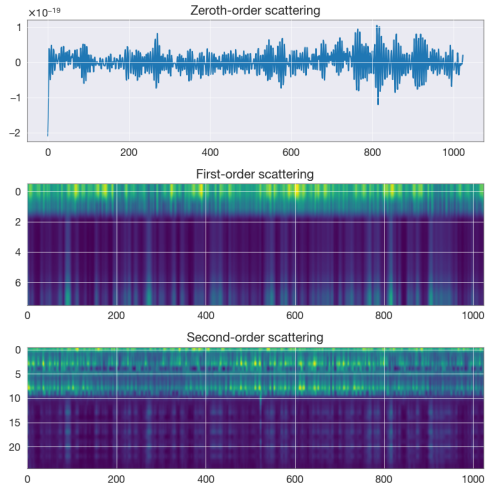


Figure 11.12: WST of raw Koi Fish Glitch

Figures 11.9, 11.10, 11.11, and 11.12 provide visual representations of the wavelet scattering transform computed up to the second order for both the gravitational signals and the glitches. Notably, these figures demonstrate that the wavelet scattering transform captures intricate details and structures in the data, even in the absence of readily discernible patterns. This reinforces the necessity of working with whitened signals, as it allows for a more reliable and comprehensive analysis,

akin to the approach employed with the Q-transform method.

11.2.1 Wavelet Scattering Transform on Whitened Signals

This paragraph discusses the possible effectiveness of whitening the signals in the wavelet scattering transform framework.

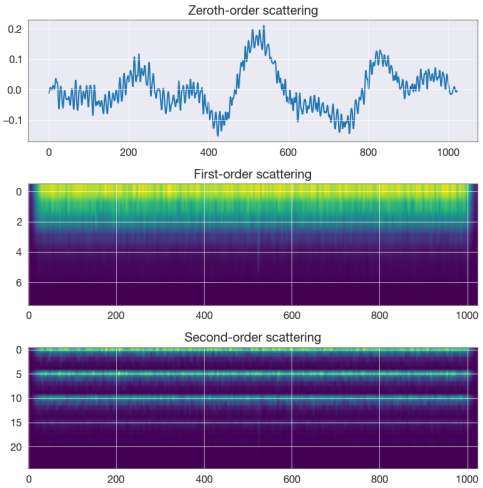


Figure 11.13: WST of raw non-whitened GW150914

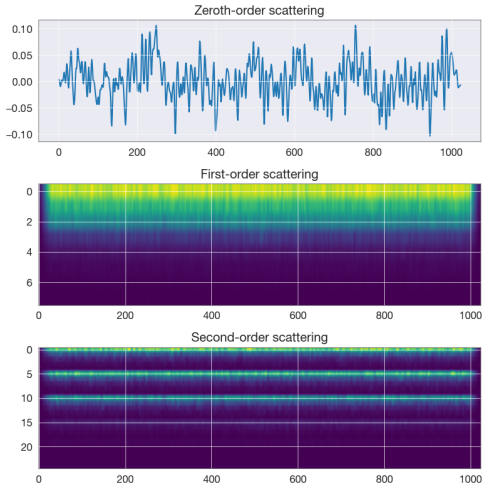


Figure 11.14: WST of raw GW170817

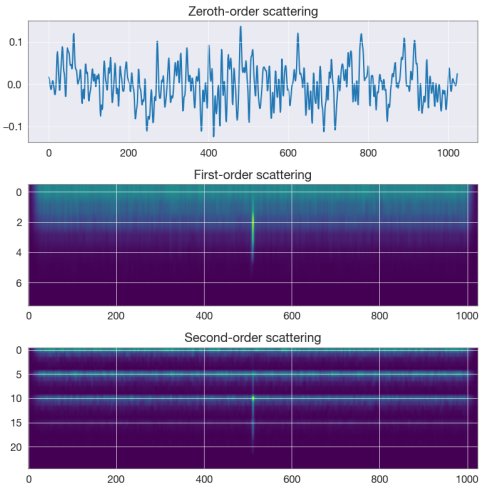


Figure 11.15: WST of raw Injection Noise Glitch

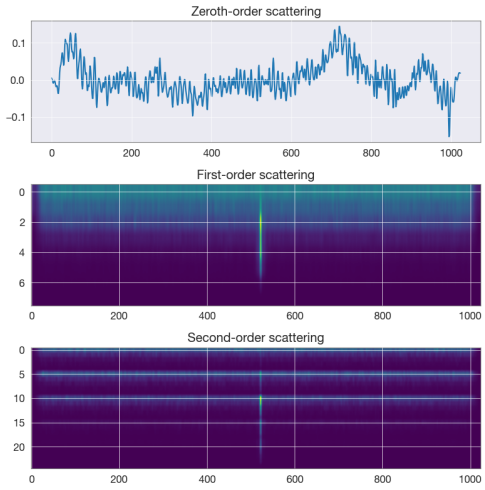


Figure 11.16: CWST of whitened Koi Fish Glitch

As previously mentioned the wavelet scattering transform computed over raw signals does not seem to provide, from the simple and empirical point of view of visualization, any kind of further information. If we compare the transform of the whitened signals, displayed in Figures 11.13, 11.14, 11.15 and 11.16, with the transforms of the raw signals, the difference becomes clear. Since whitening technique cuts off the noisy frequencies the representation of the glitches is closer, with an evident spike corresponding to the event itself. As far as the wavelet scattering transform of the two waves is concerned we can observe that the two representations in Figures 11.13 and 11.14 are closer as they are depicted in Figures 11.9 and 11.10. A further comment can be done by looking at the frequencies recovered by applying the low-pass filter ϕ_J, i.e. the zero-th order wavelet scattering transform. If we simply observe the case in which whitening has been performed, on average, the zero-th order wavelet scattering transform is more regular and less noisy.
This brief analysis, coupled to the accomplished data preparation pipeline that Virgo and LIGO implemented, gave us confirmation that also with wavelet scattering transform whitening the signals becomes fundamental to extract discriminative patterns from the time series. In order to support this thesis with quantitative results, in the following paragraph we report the distance matrices either for the non-whitened wavelet scattering transform, either for the whitened.

11.2.2 Hyper-parameters Tuning

Here we provide the details and the results of the distance based experiments that have been performed in order to estimate the most suitable values of the scale parameter J and the octave parameter Q for the scattering representation of gravitational waves data.
A parallel analysis was performed to evaluate how much discriminative information the second order added to the amount of the first order. To assess the suitability of different parameter values, we considered two similarity metrics: the Mean Squared Error (MSE) and the Euclidean distance (L^2). These metrics are well-suited for comparing the similarity between two matrices, as the rappresentations account for displacements and translations. Let us recall that for two matrices X and Y, of size $n \times m$, the MSE is a measure of similarity, and is defined as:

$$\mathrm{MSE}(X - Y) = \frac{1}{n \times m} \sum_{i,j} (X_{ij} - Y_{ij})^2 \,, \tag{11.1}$$

while the *Euclidean* distance is simply defined as

$$d_2(X, Y) = \sqrt{\sum_{i,j} (X_{ij} - Y_{ij})^2} \,. \tag{11.2}$$

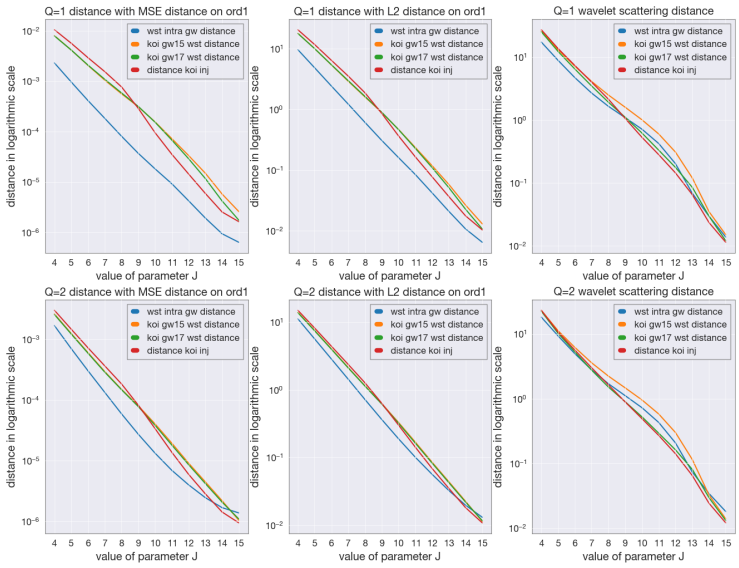


Figure 11.17: Distances Variations with respect to Q and J on first order wst

In our analysis, we also employed the embedded metric defined in the proof of Proposition 2.3.1 specifically for the wavelet scattering transform.
Figure 11.17 reports for different values of the octave $Q \in \{1,2\}$ and for different values of $J \in \{4,5,\ldots,15\}$, the pair-wise distances between the four signals. Figure 11.17 depicts the variations of distances with respect to different values of the octave parameter $Q \in \{1,2\}$ and the scale parameter $J \in 4,5,\ldots,15$ for the first-order wavelet scattering transform. The figure illustrates the pair-wise distances between the four signals. It is important to note that Figure 11.17 focuses on the distances computed using MSE and L^2 metrics for the first-order wavelet scattering transform, while it considers orders 0, 1, and 2 for the wavelet scattering distance. The objective is to identify values of J and Q that yield a desirable separation of data in the wavelet scattering domain, aligning with the goals of the interTwin project. Specifically, suitable hyperparameter values should exhibit small distances between the gravitational signals GW150914 and GW170817, while exhibiting larger distances between the two glitches. The best combination of values for gravitational wave found with this grid search is provided by $Q = 1$ and $J = 5$. This choice was adopted for the rest of the analysis. For this small size experiment the metric that resulted preferable was the MSE, since it provides fair comparisons with the Q-transform representation, that, due to the non-invariance properties of Fourier transform, is not translation invariant and stable to the action of diffeomorphisms. After tuning of the hyper-parameters J and Q, the study of the distances proceeds

with reporting the distance matrices. Moreover the analysis focuses on comparing three main aspects: MSE and L^2 metrics, whitening technique against raw signals and, finally, the number of discriminative information provided by order 2. We expect from the theoretical results that second order transform enhances finer grained information, while first order highlights more evident features.

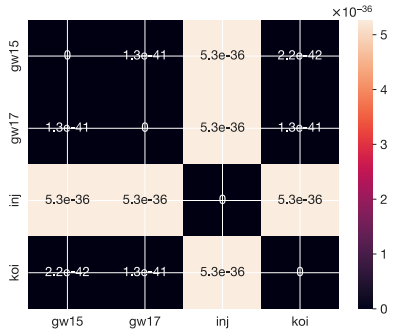


Figure 11.18: Distance matrix for first order WST with MSE

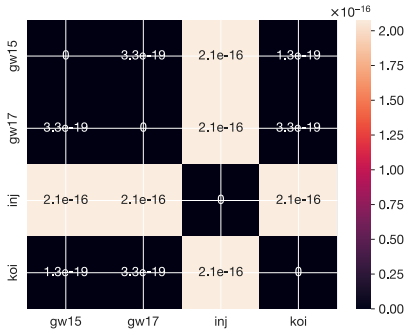


Figure 11.19: Distance matrix for first order WST with L^2

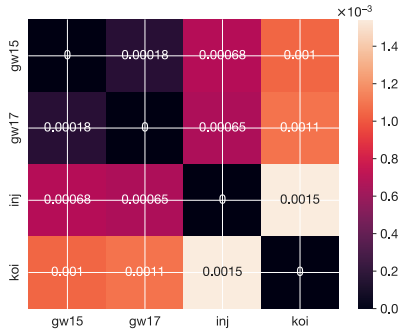


Figure 11.20: Distance matrix for whitened first order WST with MSE

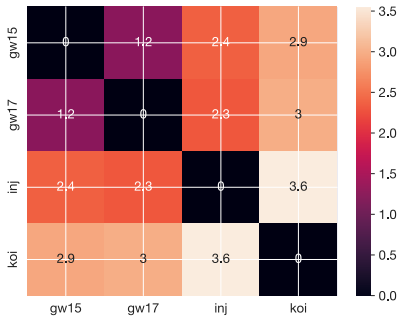


Figure 11.21: Distance matrix for whitened first order WST with L^2

If we compare Figures 11.18 and 11.19 reporting the distance matrices for the non-whitened first order wavelet scattering transform, we can see that, apart from the order of magnitude, Injection Noise glitch has the representation which is more distant to each one of the other signals. However Figures 11.20 and 11.21 clearly portrait from a strictly quantitative point of view the effectiveness of whitening the interferometer signals before applying the wavelet scattering transform. For both metrics it results that the GW150914 and GW170817 have the closest representations, while glitches are far from each other and from the

two gravitational waves. Moreover the heatmap of the distance matrices show that whitening the signals allow to obtain nice distances which appear in the same order of magnitude. Analyzing second order wavelet scattering transform leads to the following distance matrices.

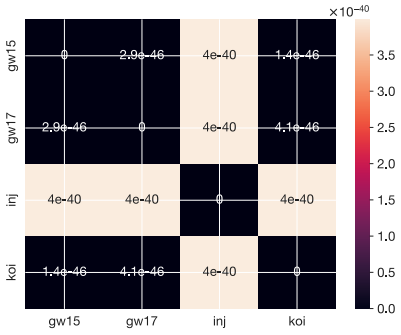


Figure 11.22: Distance matrix for second order WST with MSE

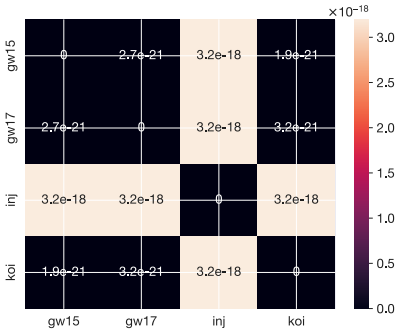


Figure 11.23: Distance matrix for second order WST with L^2

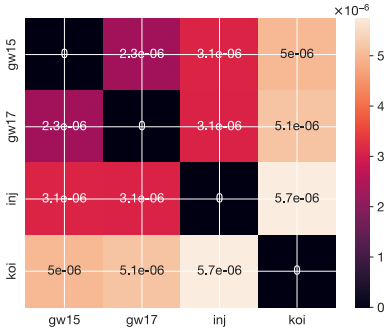


Figure 11.24: Distance matrix for whitened second order WST with MSE

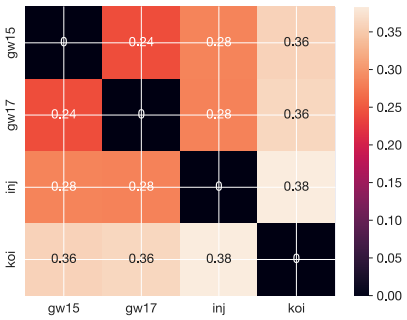


Figure 11.25: Distance matrix for second order whitened WST with L^2

Figures 11.22 and 11.23 show similarities with the Figures 11.18 and 11.19 in terms of order of magnitude. Injection Noise glitch is further from the other signals and Koi Fish glitch is closer to the waves, than the waves themselves. As in the previous case transforming the whitened signals outperforms transforming the raw signals. Figures 11.24 and 11.24 display the distance for MSE and L^2 metric in the case whitening has been applied before transforming the signals. It is interesting that also considering only the second order transform keeps the same clusters. In the following paragraph we try to establish a definitive pipeline to compute the

wavelet scattering transform over gravitational waves data. We prove from an empirical point of view that whitening the signal is fundamental, in the next pages we address the problem of normalizing the wavelet scattering transforms. Does median normalization lead to a better separation of the data?

11.2.3 Median Normalization on the Wavelet Scattering Transforms

In order to provide a full comparison with the methodologies that are applied to the Q-transform in the analysis of gravitational waves data, we address the study of the so-called *median normalization*.
Let us consider a data point X of size $n \times m$. In order to estimate the median X is necessary to rank the values in increasing order, then the median is obtained by

$$m = \mathbf{median}(X) = \begin{cases} X_{(mn+1)/2} & \text{if } mn \text{ is odd} \\ \dfrac{X_{(mn)/2} + X_{(mn+1)/2}}{2} & \text{if } mn \text{ is even}\,. \end{cases} \tag{11.3}$$

If we call $\hat{\sigma}$ the sample estimate of the standard deviation of X, the median normalized version of X is defined as

$$\bar{X} = \frac{X - m}{\hat{\sigma}}\,, \tag{11.4}$$

where m is the median of X.
To evaluate the median normalized wavelet scattering transforms we take into account the MSE and L^2 distance metrics.

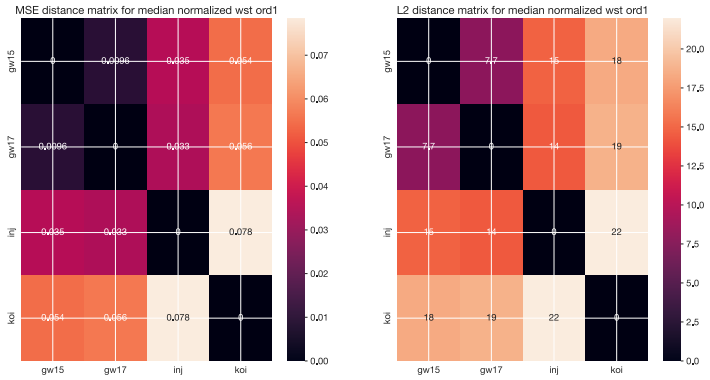


Figure 11.26: Distance Matrix for order 1 median normalized WST

Figure 11.26 shows the distance matrices computed on the first order wavelet scattering transformed after the median normalization have been applied. A first sight it immediatly appears that the order of magnitude of the pairwise distances is the same, which is a good result, as it means that the distances are comparable. Moreover, for both matrices, we obtain that the distance between the two gravitational waves is the smallest, as well as the distance between the two glitches is the largest. Moreover, as done in the previous section we aim to understand the behaviour of the second order scattering transform. Comparing Figure 11.27 and Figure 11.26 it is possible to observe that the desired order relationship is mantained, i.e. the two glitches are more distant and the two gravitational wave signals are closer.

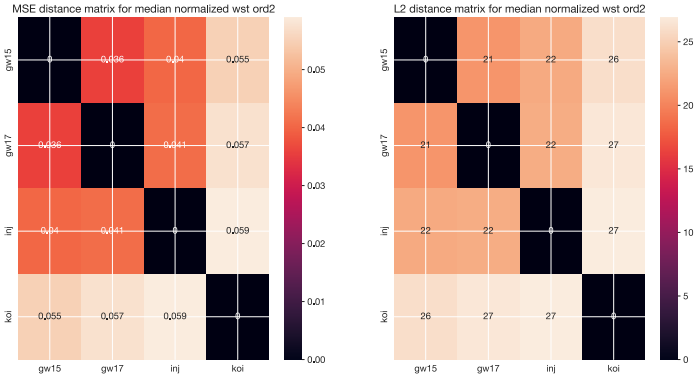


Figure 11.27: Distance Matrix for order 2 median normalized WST

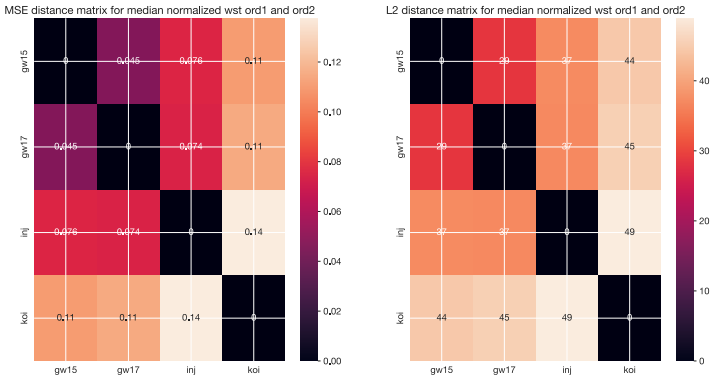


Figure 11.28: Distance Matrix for median normalized WST up to order 2

However, as perfectly highlighted by the colours of the heat map, compared to the first order, the distances between the second order wavelet scattering transform of the gravitational waves grows larger.
Results portrayed in Figure 11.28 consider the sum of the distancers of the first two layers of the wavelet scattering transform, after being normalized to the median. Since the order relationships are kept, this is a promising results. Even if this analysis have been conducted with very few signals, it suggests that, as for the auditory data, the wavelet scattering transform seem to separate well the data. This analysis led us to consider that the best pre-processing pipeline for applying the wavelet scattering transform to Virgo and LIGO data is to follow two fundamental step. The first consist in whitening the time series, while the second consist in independently normalize to the median each layer of the wavelet scattering transform. This approach allows to exctract more patterns and discriminative information from the signals with the wavelet scattering operator.

11.3 Comparison with the Q-transform Results

In this section we perform the same analysis that was previously performed with the wavelet scattering transform, but this time with the state-of-the-art representation technique, the Q-transform. In the first part of the section we show the spectrograms obtained with the Q-transform of the four signals, and in the second part we evaluate the distance matrix obtained with MSE and Euclidean distance in the Q-transform space.

11.3.1 Q-transform of the signals

As previously mentioned, the Q-transform is a widely used tool for analyzing gravitational wave data in the field of gravitational wave astronomy. In this analysis we rely on the implementation of the Q-transform offered by the GWpy package.
The q-transform in GWpy produces a spectrogram-like representation of the signal, where the intensity of each point in the time-frequency plane corresponds to the magnitude of the signal at that specific time and frequency. This representation allows us to visualize the frequency content of the signal over time and identify any transient or persistent features, such as gravitational wave signals or noise artifacts. The following images show the Q-transform of the four signals. Figure 11.29 portraits the Q-transform representations of the well-known gravitational wave events GW150914 and GW170817, spanning a time window of 32 seconds. In contrast, Figure 11.30 presents the Q-transform representations of the two glitches observed within the same time window.

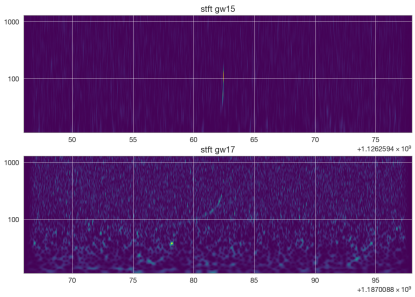


Figure 11.29: Gravitational Waves Q-transforms over full time range

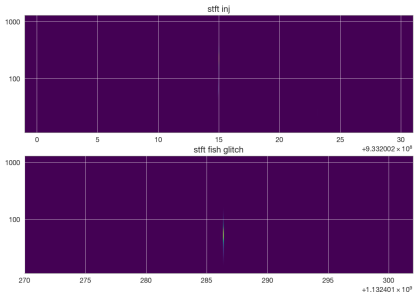


Figure 11.30: Glitches Q-transforms over full time range

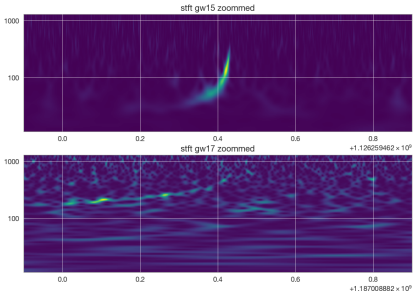


Figure 11.31: Gravitational waves Q-transforms around detection time

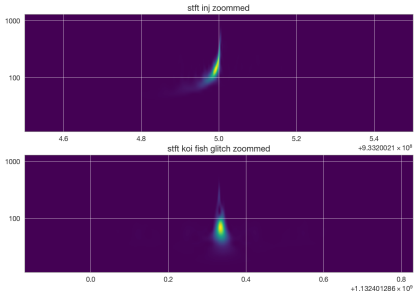


Figure 11.32: Glitches Q-transforms around detection time

Upon closer examination of these images, particularly when zoomed around the detection time (Figures 11.31 and 11.32), distinct patterns emerge. However, it becomes evident that the Q-transform operator lacks invariance and stability. Moreover, when compared to the wavelet scattering transforms of the gravitational waves, the spectrograms of GW150914 and GW170817 exhibit noticeable differences.

11.3.2 Considerations on Distances Using Q-transform

Building upon the methodology described in the previous section with the wavelet scattering transform, we now present the results obtained using the Q-transform method. It is important to note that the GWpy library performs the Q-transform on whitened signals and applies a median normalization. Figure 11.33 displays the distance matrix for the Q-transform representations.

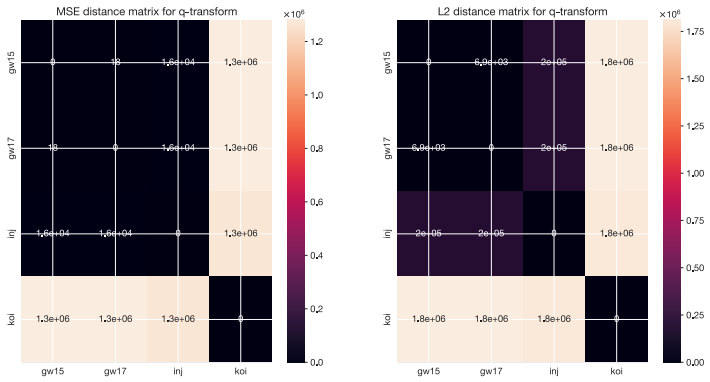


Figure 11.33: Distance Matrix for Q-transforms

It is important to compare these results with the distance matrix shown in Figure 11.28 obtained using the wavelet scattering transform. One immediately notices that the color ranges in the two matrices are very different. While the distances obtained using the wavelet scattering transform are comparable, the Q-transform distances exhibit much larger magnitudes. For instance, in the case of the L^2 metric, the distance between GW150914 and GW170817 reaches the order of 10^3. This significant increase in distances highlights the need for larger architectures, such as those based on convolutional networks, when working with the Q-transform. This observation provides a straightforward explanation as to why Q-transform-based analysis often requires complex and deep neural network architectures. In contrast, the wavelet scattering operator's properties allow for the use of lighter architectures. Exploiting the structure and the features extracted by the wavelet scattering transform, we can potentially achieve good results with simpler and more efficient neural network architectures. This opens up opportunities for efficient and effective analysis of gravitational wave data using the wavelet scattering transform.

Chapter 12

Wavelet Scattering Transform for Glitch Detection. An Analysis on Scattered-light Dataset

This chapter analyzes the results obtained conducting our study that we developed on the single class dataset of *scattered-light* glitches. According to VIRGO expertise the scattered-light glitches are the one on which the Q-transform fails the most in extracting discriminative features. Therefore this analysis tries to use the wavelet scattering transform to overcome the difficulties encountered by the Q-transform representation.
In order to conduct the study we find a mean representative element with either representation methods and try to quantify the dispersion, in terms of *coefficient of variation*. The coefficient of variation (CV) measures the relative variability or dispersion of a set of measurements in relation to its mean. We present to different methods to compute the representative element. The first follows the PCA based technique showcased in Chapter 7, while the second considers a component-wise average, leveraging an additive terms that accounts for the pixel-wise standard deviation.
In this chapter we follow the results obtained in Chapter 11, meaning that we assume as a standard technique to compute the wavelet scattering transform over the whitened signal, and then we normalize to the median each layer.

12.1 Scattered-light Dataset Overview

Scattered-light glitches occur within the frequency range of 10-120Hz, which coincides with the frequency band where we observe the inspiral and merger signatures of compact binary coalescences [65]. Scattered-light glitches are a result of laser light within the interferometer being scattered from its intended optical path by components within the detector. The movement of these components is influenced by seismic motion, causing a phase shift in the scattered light as the surfaces oscillate. This scattered light then combines with the primary laser, resulting in scattered-light glitches within the data. These glitches typically originate from objects placed on optical benches, such as lenses, mirrors, and photo-detectors.
To further investigate these scattered-light glitches, Virgo provided us a dataset comprising 855 samples. In order to compare and analyze the data in this section, we will consider the Q-transform and wavelet scattering transform from a visual standpoint.

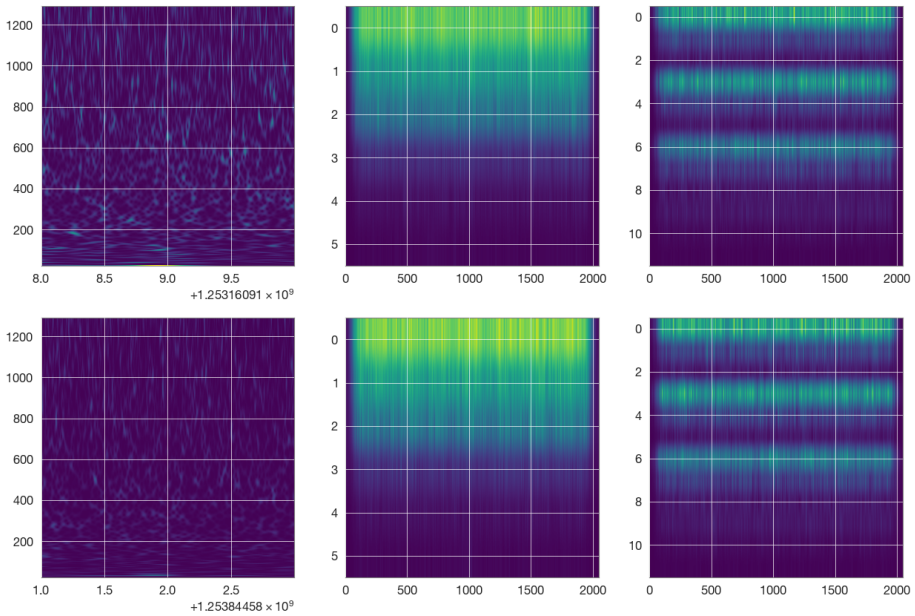


Figure 12.1: Comparing two signals: Q-transform (left), first order WST(center) and second order WST (left)

Figure 12.1 provides a visual comparison between the two transforms, showcasing two selected glitch samples from the dataset. From a cursory examination,

it is apparent that the wavelet scattering transforms exhibit a greater degree of similarity compared to the Q-transforms.
Expanding our analysis to a larger set of samples, the Q-transform representations appear more dispersed and lacking any discernible pattern in the time-frequency domain. Energy spikes are scattered sporadically throughout the representation, displaying a quasi-random distribution.
On the other hand, the wavelet scattering transforms reveal a consistent pattern across both first and second orders. The first order exhibits higher energy concentrations at lower frequencies, while the second order displays similarities across three distinct regions of the spectrogram. This consistency indicates the potential presence of underlying structures or features within the scattered-light glitches.
It seems that the wavelet scattering transform shows promise in capturing and highlighting the common patterns inherent in scattered-light glitches, suggesting its potential utility in further analysis and characterization of these phenomena.

12.2 Principal Component Analysis and Scattered-Light Glitches

In this section, we employ Principal Component Analysis (PCA) to extract meaningful features from the various representation techniques used. These features are then combined to create a representative average spectrogram for the scattered-light glitches. The methodology follows the approach outlined in Chapters 6 and 7 of the thesis.
First, we present an analysis and report the results obtained using the state-of-the-art method. This allows us to establish a baseline for comparison. Subsequently, we delve into the experiments conducted using the wavelet scattering transform, showcasing the advancements and insights gained through its application.

12.2.1 Results for the Q-transform

As mentioned earlier, scattered-light glitches are notorious among INFN researchers for their challenging characteristics that are not easily captured by the Q-transform representation. In this section, we aim to quantify the dispersion within the Q-transform using a representative element computed through Principal Component Analysis.
Since the standard processing of the Q-transform involves median normalization, we applied median normalization to the representative element obtained from PCA.
To begin, we performed PCA with the objective of achieving more than 95% explained variance.

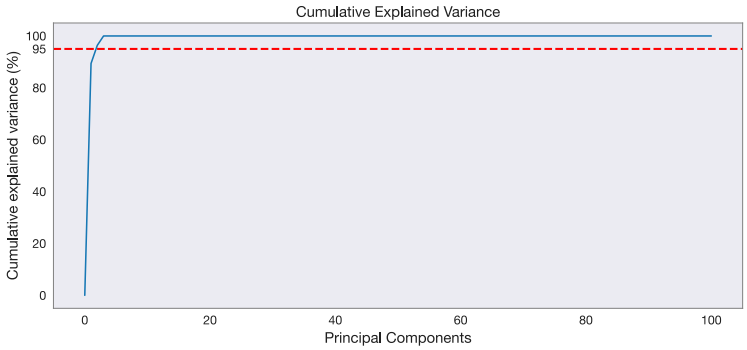


Figure 12.2: Cumulative explained variance in the Q-transform

Figure 12.2 illustrates the cumulative explained variance in the Q-transform, demonstrating that only a few principal components are necessary to explain the majority of the variance. This result indicates that PCA struggles to identify clear patterns across the Q-transforms. Therefore, we choose to utilize three principal components to compute the representative element. Each component is summed and then normalized to the median for the purpose of comparison.

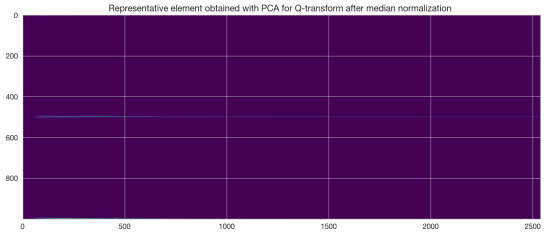


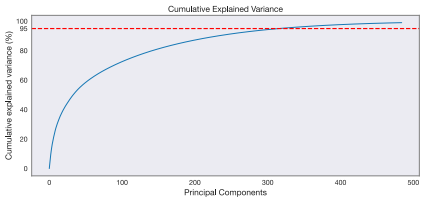
Figure 12.3: Median normalized PCA representant with Q-transform

The resulting representation is shown in Figure 12.3. It is evident that the representation lacks informative features. In terms of assessing average dispersion, we employ mean squared error (MSE) and Euclidean distance as metrics. In order to quantify the dispersion within the Q-transform, we introduce the coefficient of variation (CV). The CV is an adimensional index that allows for the quantification of dispersion within a dataset. It is defined as $CV = \hat{\sigma}/\hat{\mu}$, where $\hat{\sigma}$ is the sample standard deviation and $\hat{\mu}$ is the sample mean. A lower value of the coefficient of variation indicates smaller dispersion.
For the Q-transform, the calculated CV values are 26.22 for MSE and 20.23 for the Euclidean metric, indicating relatively high dispersion within the dataset.
These results emphasize the need for alternative approaches to accurately capture and analyze scattered-light glitches. In the following experiments, we turn our

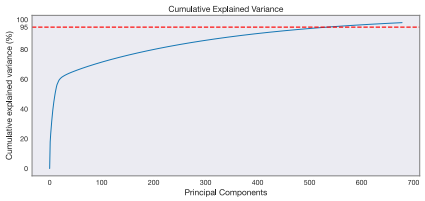
attention to the wavelet scattering transform, exploring its potential to overcome the challenges posed by scattered-light glitches.

12.2.2 Results for the Wavelet Scattering Transform

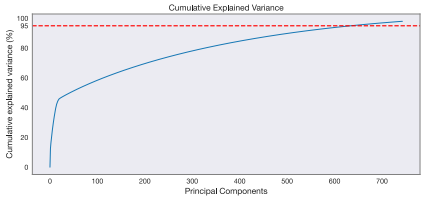
In this paragraph, we shift our focus to compare the dispersion of scattered-light glitches in the wavelet scattering domain. The objective is to assess whether the wavelet scattering offers improved performance and provides more informative representations compared to the previous Q-transform analysis.
We consider three metrics for evaluating the dispersion: mean squared error (MSE), Euclidean distance, and the wavelet scattering metric. The analysis is conducted up to the second layer of the wavelet scattering transform.
The approach used to compute the representative element in the wavelet scattering space is the same as that employed for the Q-transform. We determine a suitable number of principal components that reach a threshold of explained variance, and then sum these principal components before applying median normalization.



(a) Cumulative explained variance over zero-th order WST



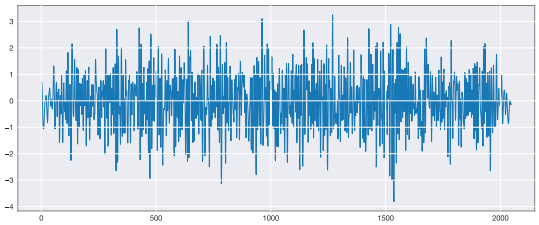
(b) Cumulative explained variance first over order WST



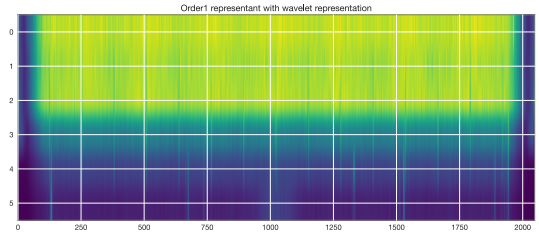
(c) Cumulative explained variance over second order WST

Figure 12.4: Cumulative explained variance in the WST space

Figure 12.4 displays the cumulative explained variance in the WST space. It is notable that a larger number of principal components is required to achieve 95% explained variance in the wavelet scattering domain compared to the Q-transform. This suggests that more diverse principal components are needed to describe and reproduce the discriminative patterns observed in the spectrograms.



(a) PCA representant for order 0 WST



(b) PCA representant for order 1 WST

(c) PCA representant for order 2 WST

Figure 12.5: PCA representant for WST

The resulting PCA representatives for each order of the WST are presented in Figure 12.5. Remarkably, the similarities between the representative elements and the samples depicted in Figure 12.1 are evident. This indicates that the wavelet scattering transform successfully captures and preserves the distinctive patterns present in the scattered-light glitches.
To quantify the dispersion and compare the performance of the metrics, we leveraged average distances along with their standard errors, accounting for the coefficient of variation. The MSE yields a CV of 0.72, the Euclidean distance 0.26, and the wavelet scattering metric results in 0.20. These results clearly demonstrate that in the wavelet scattering space, the representations exhibit less dispersion, with the wavelet scattering metric performing the best in terms of CV. This suggests that the wavelet scattering transform captures the underlying structure of scattered-light

glitches more effectively.

12.3 Average Representative Spectrogram

In this section, we present the results of an additional experiment conducted to compare the dispersion of scattered-light glitches in the wavelet scattering space and in the Q-transform domain. Instead of using PCA-derived representatives, we focus on the dispersion from the average spectrogram.
To compute the average spectrogram, we employ a pixel-wise sample mean, combined with half of the pixel-wise standard deviation: this approach takes into account the variability present in the data. Then the spectrogram is normalized to the median.
Empirical evidence demonstrates that in the wavelet scattering domain, the data exhibits less dispersion. This finding further reinforces the notion that the wavelet scattering transform captures the essential characteristics and patterns of scattered-light glitches, allowing more informative representations.

12.3.1 Q-transform

In this paragraph, we present the results obtained using the state-of-the-art technique of the Q-transform. The average Q-transform spectrogram for the scattered-light glitch is shown in Figure 12.6. It can be observed that the average representative computed using the Q-transform exhibits a very similar pattern to the one obtained via PCA (Figure 12.3). Consequently, we anticipate similar results in terms of average distance and standard error. MSE metric had a CV of 26.221, indicating a large dispersion in the distances from the representative element. Similarly, the Euclidean metric scored 20.222 in terms of CV, implying considerable variability. These values are comparable to the results obtained through PCA, highlighting the challenging nature of capturing distinct patterns and reducing dispersion using the Q-transform representation alone.

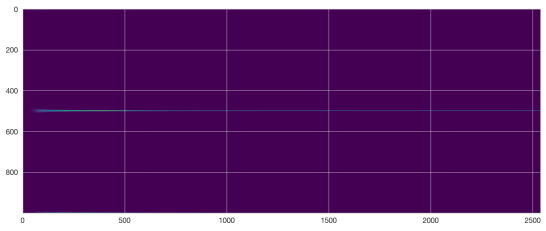
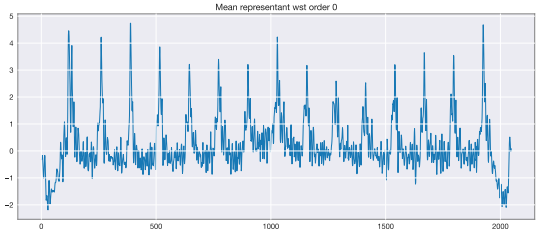


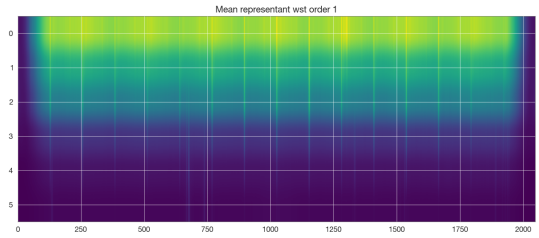
Figure 12.6: Median normalized average representant with Q-transform

12.3.2 Wavelet Scattering Transform

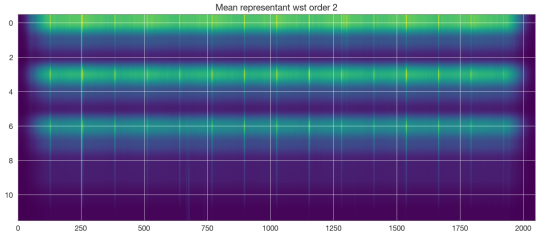
In this paragraph, we present the experiments and results obtained with the wavelet scattering transform. The average representants for each order of the wavelet scattering transform are shown in Figure 12.7, and they exhibit similar patterns to the PCA representative obtained in the previous methodology. We evaluate the dispersion of the scattered-light glitches using three metrics: MSE, Euclidean distance, and the wavelet scattering-induced metric. In terms of CV, MSE metric yielded a result of 1.650, the Euclidean distance resulted in 0.474, and the wavelet scattering metric showed a value of 0.278. Notably, the wavelet scattering-induced metric displayed less variability, indicating a more consistent representation of the data.



(a) Average representant for zero-th order WST



(b) Average representant for first order WST



(c) Average representant for second order WST

Figure 12.7: Average representant for WST

12.4 Summarizing the Results

Based on the experiments conducted on the Scattered-light Dataset, we have summarized the results in Table 12.1, which provides the Coefficient of Variation for each representation and technique in terms of different metrics.

REPRESENTATION	**TECHNIQUE**	**Metric**	**CV**
Q_transform	PCA	MSE	26.221
Wavelet Scattering Transform	PCA	MSE	0.725
Q_transform	PCA	L2	20.222
Wavelet Scattering Transform	PCA	L2	0.264
Wavelet Scattering Transform	PCA	WST	0.204
Q_transform	mean	MSE	26.221
Wavelet Scattering Transform	mean	MSE	1.650
Q_transform	mean	L2	20.222
Wavelet Scattering Transform	mean	L2	0.474
Wavelet Scattering Transform	mean	WST	0.278

Table 12.1: Summary of the results obtained with *Scattered-light Dataset* in terms of Coefficient of Variation

The results show a clear advantage of the Wavelet Scattering Transform over the Q-transform in terms of capturing and representing the scattered-light glitches. When comparing the PCA-based representations, the wavelet scattering transform exhibits significantly lower CV values across all metrics. For example, in terms of the MSE metric, the CV for the Q-transform PCA representation is 26.221, while for the wavelet scattering transform PCA representation, it is only 0.725. Similarly, in terms of the Euclidean metric, the CV for the Q-transform PCA representation is 20.222, while for the wavelet scattering transform PCA representation, it is only 0.264. Furthermore, when considering the average representations obtained using the mean technique, the superiority of the WST over the Q-transform is further emphasized. The CV values for the wavelet scattering transform average representation are consistently lower than those for the Q-transform average representation, regardless of the metric used. For example, in terms of the MSE metric, the CV for the Q-transform mean representation is 26.221, while for the wavelet scattering transform mean representation, it is 1.650. These results highlight the improved performance and more informative representations provided by the Wavelet Scattering Transform compared to the Q-transform. The lower CV values indicate a more concentrated and less dispersed distribution of the data in the wavelet scattering transform representations, leading to a more reliable and consistent

characterization of the scattered-light glitches.
The wavelet scattering transform outperforms the Q-transform in terms of minimizing the dispersion and variability of the representation of the scattered-light glitches. This suggests that the wavelet scattering transform is more effective in extracting and preserving the discriminative patterns present in this class. The wavlet scattering representations exhibit lower variability and dispersion, as indicated by the significantly lower Coefficient of Variation (CV) values across all metrics compared to the Q-transform representations.
These results show in a preliminar setting the advantages of the wavelet scattering transform to provide more informative and less dispersed representations of scattered-light glitches compared to the state-of-the-art technique of the Q-transform.

Conclusions

The wavelet scattering transform provides a beneficial trade-off between the complexity of classification architectures and the computational cost of the representation, due to translation invariance and stability properties. Since the transformation is not learnt, unlike for convolutional neural networks, the wavelet scattering transform does not require a large amount of data to achieve a stable representation, relying on cheaper computations. These characteristics well fit for the analysis and discrimination of one dimensional non-stationary processes, such as human voice recordings and gravitational waves.
The results obtained on the Free Spoken Digits dataset are proof of the potential of the wavelet scattering transform in the enhancement of classification algorithms. The wavelet scattering transform unveiled the underlying hidden patterns correlating spoken digits classes and allowed to map the data into a feature space where the points appeared properly clustered. This property was confirmed by the 98% accuracy achieved by K-Nearest Neighbors in the wavelet scattering space, compared to the 79% accuracy in the Fourier domain.
In this thesis we introduced a novel methodology to apply the wavelet scattering transform to gravitational waves data. The metric evaluation performed led to argue that signals captured by interferometers should be whitened firstly, and then the obtained wavelet scattering transform should be normalized to the median. This data processing procedure provided a better separation of glitches and gravitational waves in the wavelet scattering domain. Furthermore we developed a dispersion analysis to compare the proposed approach with the state-of-the-art method based on the Q-transform. The results obtained with the wavelet scattering transform outperformed the state-of-the-art by achieving lower coefficient of variations with different metrics (euclidean, MSE and wavelet scattering metric). In particular the lowest coefficient of variation obtained for the wavelet scattering transform was 0.204, with respect to 20.222. of the Q-transform.
This thesis shows that promising results in gravitational waves research can be achieved by applying the wavelet scattering transform to the data collected by interferometers, properly clustering the glitches classes.
This study represents one of the first steps towards the design of wavelet scattering

networks for glitch classification. To further establish the advantages of applying the wavelet scattering transform to gravitational waves a larger multi-class dataset is needed. This would allow to develop experiments for comparing the performances of architectures based on the wavelet scattering transform with the ones based on the Q-transform, as presented with auditory data.
Moreover dealing with larger amount of labelled data from different classes could help to assess more accurately this methodology, providing further understandings and considerations about the advantages of using the wavelet scattering transform for gravitational waves, that in this thesis are limited to data provided by INFN.
In this perspective it might be possible to conduct detailed analysis to carefully determine computational and energy consumption of architectures based on the wavelet scattering transform.
Further developments in the interTwin project framework should address the co-operative research where the wavelet scattering transform is used to represent environmental noises collected by auxiliary channels, and to build an automated machine that distinguishes gravitational signals from glitches. In this scenario the wavelet scattering transform is the founding element to design a lighter generative architecture, improving performances and computational costs to train the model, with a significant saving of energy.
Moreover studying the properties of the wavelet scattering operator from a theoretical point of view, and developing significant applications in speech recognition and gravitational wave analysis, extended the horizon of potential applications of this innovative technology in other research and industrial fields, such as domain adaptation and texture discrimination in computer vision, bio-electric signals characterization in health-care, time-series analysis in finance.

Bibliography

[1] Jean Baptiste Joseph Fourier. *Théorie analytique de la chaleur*. Gauthier-Villars et fils, 1822 (cit. on p. 1).

[2] Dennis Gabor. «Theory of communication». In: *Journal of the Institution of Electrical Engineers - Part I: General* 94 (1946), pp. 58–58 (cit. on p. 1).

[3] Judith C Brown and Miller S Puckette. «An efficient algorithm for the calculation of a constant Q transform». In: *The Journal of the Acoustical Society of America* 92.5 (1992), pp. 2698–2701 (cit. on p. 1).

[4] Jean Morlet, Georges Arens, Eliane Fourgeau, and Dominique Glard. «Wave propagation and sampling theory—Part I: Complex signal and scattering in multilayered media». In: *Geophysics* 47.2 (1982), pp. 203–221 (cit. on p. 1).

[5] Stéphane Mallat. «Understanding deep convolutional networks». In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150203 (cit. on pp. 2, 7, 10).

[6] Stéphane Mallat. «Group invariant scattering». In: *Communications on Pure and Applied Mathematics* 65.10 (2012), pp. 1331–1398 (cit. on pp. 2, 6, 15, 27, 28, 31).

[7] Albert Einstein. «Die feldgleichungen der gravitation». In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin* (1915), pp. 844–847 (cit. on p. 2).

[8] Benjamin P Abbott et al. «Observation of gravitational waves from a binary black hole merger». In: *Physical review letters* 116.6 (2016), p. 061102 (cit. on pp. 2, 76).

[9] Zohar Jackson, Csar Souza, Jason Flaks, Yuxin Pan, Hereman Nicolas, and Adhish Thite. *Jakobovski/free-spoken-digit-dataset: v1.0.8*. `https://github.com/Jakobovski/free-spoken-digit-dataset`. Aug. 2018 (cit. on pp. 2, 32, 35, 37, 41).

[10] Joakim Andén, Laurent Sifre, Stéphane Mallat, Vincent Lostanlen, and Edouard Oyallon. *ScatNet*. `http://www.di.ens.fr/data/software/scatnet`. 2014 (cit. on pp. 3, 40).

[11] Duncan M Macleod, Joseph S Areeda, Scott B Coughlin, Thomas J Massinger, and Alexander L Urban. «GWpy: A Python package for gravitational-wave astrophysics». In: *SoftwareX* 13 (2021), p. 100657. ISSN: 2352-7110. DOI: 10.1016/j.softx.2021.100657. URL: https://www.sciencedirect.com/science/article/pii/S2352711021000029 (cit. on pp. 3, 96, 99).

[12] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. «Communication-efficient learning of deep networks from decentralized data». In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282 (cit. on p. 3).

[13] Pete Warden and Daniel Situnayake. *Tinyml: Machine learning with tensorflow lite on arduino and ultra-low-power microcontrollers*. O'Reilly Media, 2019 (cit. on p. 3).

[14] Karen Hao. «Training a single AI model can emit as much carbon as five cars in their lifetimes». In: *MIT technology Review* 75 (2019), p. 103 (cit. on p. 3).

[15] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. «Towards the systematic reporting of the energy and carbon footprints of machine learning». In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 10039–10081 (cit. on p. 3).

[16] Joan Bruna. «Scattering Representations for Recognition». Déposée Novembre 2012. Theses. Ecole Polytechnique X, Feb. 2013. URL: https://pastel.archives-ouvertes.fr/pastel-00905109 (cit. on pp. 6, 10, 15, 17, 21, 28).

[17] Bryan Rynne and Martin A Youngson. *Linear functional analysis*. Springer Science & Business Media, 2007 (cit. on pp. 9, 10).

[18] Gerald B Folland. *Fourier analysis and its applications*. Vol. 4. American Mathematical Soc., 2009 (cit. on pp. 9, 10).

[19] Elias M Stein and Rami Shakarchi. *Fourier analysis: an introduction*. Vol. 1. Princeton University Press, 2011 (cit. on pp. 9, 10).

[20] Matthew D Zeiler and Rob Fergus. «Visualizing and understanding convolutional networks». In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pp. 818–833 (cit. on p. 10).

[21] Sam Roweis and Zoubin Ghahramani. «A unifying review of linear Gaussian models». In: *Neural computation* 11.2 (1999), pp. 305–345 (cit. on p. 10).

[22] Andrew H Jazwinski. *Stochastic processes and filtering theory*. Courier Corporation, 2007 (cit. on p. 12).

[23] Gilbert Strang et al. *Linear algebra and learning from data*. Vol. 4. Wellesley-Cambridge Press Cambridge, 2019 (cit. on pp. 13, 35, 39).

[24] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012 (cit. on pp. 13, 35, 37–39, 43).

[25] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. «Activation functions in neural networks». In: *Towards Data Sci* 6.12 (2017), pp. 310–316 (cit. on p. 13).

[26] Herbert E. Robbins. «A Stochastic Approximation Method». In: *Annals of Mathematical Statistics* 22 (1951), pp. 400–407 (cit. on p. 14).

[27] Diederik P Kingma and Jimmy Ba. «Adam: A method for stochastic optimization». In: (2014) (cit. on p. 14).

[28] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. «Striving for simplicity: The all convolutional net». In: (2014) (cit. on p. 14).

[29] Mingxing Tan and Quoc Le. «EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks». In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html (cit. on p. 14).

[30] John E Littlewood and Raymond EAC Paley. «Theorems on Fourier series and power series (II)». In: *Proceedings of the London Mathematical Society* 2.1 (1937), pp. 52–89 (cit. on pp. 15, 17).

[31] Joan Bruna and Stéphane Mallat. «Invariant scattering convolution networks». In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1872–1886 (cit. on pp. 15, 23, 27).

[32] Robert G Gallager. *Stochastic processes: theory for applications*. Cambridge University Press, 2013 (cit. on p. 28).

[33] Dirk P Kroese, Zdravko Botev, Thomas Taimre, and Radislav Vaisman. *Data science and machine learning: mathematical and statistical methods*. CRC Press, 2019 (cit. on pp. 35, 37–39, 43, 47).

[34] Charles R. Harris et al. «Array programming with NumPy». In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2 (cit. on p. 35).

[35] Fabian Pedregosa et al. «Scikit-learn: Machine Learning in Python». In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 35).

[36] Mathieu Andreux et al. «Kymatio: Scattering Transforms in Python». In: *Journal of Machine Learning Research* 21.60 (2020), pp. 1–6. URL: http://jmlr.org/papers/v21/19-047.html (cit. on p. 35).

[37] Wai Ho Chak, Naoki Saito, and David Weber. «The Scattering Transform Network with Generalized Morse Wavelets and its Application to Music Genre Classification». In: *2022 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*. 2022, pp. 25–30. DOI: 10.1109/ICWAPR56446.2022.9947091 (cit. on p. 35).

[38] Premjeet Singh, Goutam Saha, and Md Sahidullah. «Deep scattering network for speech emotion recognition». In: *2021 29th European Signal Processing Conference (EUSIPCO)*. 2021, pp. 131–135. DOI: 10.23919/EUSIPCO54536.2021.9615958 (cit. on p. 35).

[39] Richard A Roberts and Clifford T Mullis. *Digital signal processing*. Addison-Wesley Longman Publishing Co., Inc., 1987 (cit. on p. 36).

[40] URL: https://www.kymat.io/_images/algorithm.png (cit. on p. 41).

[41] URL: https://www.ligo.caltech.edu/page/what-are-gw (cit. on p. 76).

[42] URL: https://scienzapertutti.infn.it/images/stories/percorsi/onde/schema_onde.jpg (cit. on p. 78).

[43] URL: https://physics.aps.org/assets/a2a363a3-e840-47b7-8e12-de770d140a1a/e17_2.png (cit. on p. 80).

[44] URL: https://scienzapertutti.infn.it/images/stories/percorsi/onde/illustrazione_antenne_gravitazionali_mondo_INFN_Centimetri3.jpg (cit. on p. 84).

[45] URL: https://www.ligo.caltech.edu/system/avm_image_sqls/binaries/92/page/SkyMap__CREDIT__LIGO_Virgo_NASA_Leo_Singer__Axel_Mellinger.jpg?1508029988 (cit. on p. 85).

[46] Michele Bonaldi et al. «Nonequilibrium Steady-State Fluctuations in Actively Cooled Resonators». In: *Physical Review Letters* 103.16 (2009), p. 160602 (cit. on p. 85).

[47] Livia Conti, Paolo De Gregorio, Gagik Karapetyan, Claudia Lazzaro, Matteo Pegoraro, Michele Bonaldi, and Lamberto Rondoni. «Effects of breaking vibrational energy equipartition on measurements of temperature in macroscopic oscillators subject to heat flux». In: *Journal of Statistical Mechanics: Theory and Experiment* (2013) (cit. on p. 85).

[48] URL: https://www.ligo.caltech.edu/system/media_files/binaries/237/original/Basic_michelson_labeled.jpg?1435862648 (cit. on p. 86).

[49] URL: https://upload.wikimedia.org/wikipedia/commons/5/5e/ITFMichelsonSuspendu.jpg (cit. on p. 87).

[50] Elena Cuoco et al. «Enhancing gravitational-wave science with machine learning». In: *Machine Learning: Science and Technology* 2.1 (2021), p. 011002. DOI: 10.1088/2632-2153/abb93a. URL: https://doi.org/10.1088/2632-2153/abb93a (cit. on p. 89).

[51] Daniel George, Hongyu Shen, and EA Huerta. «Classification and unsupervised clustering of LIGO data with Deep Transfer Learning». In: *Physical Review D* 97.10 (2018), p. 101501 (cit. on p. 91).

[52] Sara Bahaadini, Vahid Noroozi, Neda Rohani, Scott Coughlin, Michael Zevin, Joshua R Smith, Vicky Kalogera, and Aggelos Katsaggelos. «Machine learning for Gravity Spy: Glitch classification and dataset». In: *Information Sciences* 444 (2018), pp. 172–186 (cit. on p. 91).

[53] Reed Clasey Essick. «Detectability of dynamical tidal effects and the detection of gravitational-wave transients with LIGO». PhD thesis. Massachusetts Institute of Technology, 2017 (cit. on p. 92).

[54] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006 (cit. on p. 92).

[55] Nikhil Mukund, Saurabh Thakur, Sheelu Abraham, AK Aniyan, Sanjit Mitra, Ninan Sajeeth Philip, Kaustubh Vaghmare, and DP Acharjya. «An information retrieval and recommendation system for astronomical observatories». In: *The Astrophysical Journal Supplement Series* 235.1 (2018), p. 22 (cit. on p. 93).

[56] Norbert Wiener and Cyberneticist Mathematician. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. Vol. 113. 21. MIT press Cambridge, MA, 1949 (cit. on p. 93).

[57] Gabriele Vajente, Yiwen Huang, Maximiliano Isi, Jenne C Driggers, Jeffrey S Kissel, MJ Szczepańczyk, and Salvatore Vitale. «Machine-learning nonstationary noise out of gravitational-wave detectors». In: *Phys. Rev. D* 101 (4 Feb. 2020), p. 042003. DOI: 10.1103/PhysRevD.101.042003. URL: https://link.aps.org/doi/10.1103/PhysRevD.101.042003 (cit. on p. 93).

[58] Tom Edinburgh, Peter Smielewski, Marek Czosnyka, Stephen J. Eglen, and Ari Ercole. *DeepClean – self-supervised artefact rejection for intensive care waveform data using deep generative learning*. 2020. eprint: 1908.03129 (stat.ML) (cit. on p. 93).

[59] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. *WaveNet: A Generative Model for Raw Audio*. 2016. eprint: 1609.03499 (cs.SD) (cit. on p. 93).

[60] Sergey Klimenko et al. «Method for detection and reconstruction of gravitational wave transients with networks of advanced detectors». In: *Physical Review D* 93.4 (2016), p. 042004 (cit. on p. 94).

[61] Danilo Rezende and Shakir Mohamed. «Variational inference with normalizing flows». In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538 (cit. on p. 95).

[62] Aaron Courville and Yoshua Bengio. «Generative adversarial nets». In: *Advances in Neural Information Processing Systems (NIPS)* (2014) (cit. on p. 95).

[63] Warren G Anderson and R Balasubramanian. «Time-frequency detection of gravitational waves». In: *Physical Review D* 60.10 (1999), p. 102001 (cit. on p. 98).

[64] LIGO Scientific Collaboration et al. «LALSuite: LIGO Scientific Collaboration Algorithm Library Suite». In: *Astrophysics Source Code Library* (2020), ascl–2012 (cit. on p. 99).

[65] Arthur E. Tolley, Gareth S. Cabourn Davies, Ian W. Harry, and Andrew P. Lundgren. *ArchEnemy: Removing scattered-light glitches from gravitational wave data*. 2023. eprint: 2301.10491 (gr-qc) (cit. on p. 117).