POLITECNICO DI TORINO

Master's Degree in
Mathematical Engineering

Master's Degree Thesis

Machine Learning for Flood Prediction: The Contribution of SAR and Multispectral Indexes

Supervisor
prof. Stefano Berrone

Co-Supervisor
Simonetta Bodojra

Candidate
Samuele Maria Garofalo

Academic Year 2024-2025

Abstract

Flood forecasting is crucial for mitigating the impact of extreme hydrological events, especially in urban environments where sudden flooding can cause severe damage to infrastructure and human lives. This thesis explores a machine learning approach to predicting river flood events based on historical hydrological and meteorological data and leveraging information provided by two types of satellite images: SAR images and Multispectral images. The study focuses on a binary classification problem and aims to investigate whether SAR and multispectral indices can provide valuable information when combined with meteorological and topographical features in order to forecast a flood event.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Floods are among the most frequent and devastating natural disasters globally, causing huge human, economic and environmental losses each year. According to the World Meteorological Organization (WMO), the number and severity of floods have increased significantly in recent decades, in part due to climate change and increasing urbanization. In particular, WMO stated that between 1994 and 2013 floods affected nearly 2.5 billion worldwide, causing more than $40 billion in damage each year around the world. In addition, it is estimated that nearly a quarter of the world's population is exposed to significant flood and flood risks. This is according to the latest Nature Portfolio report on flood exposure. According to that report, about 1.81 billion people (or 23 percent of the people on the planet) are directly exposed to flooding of more than 15 centimeters in a one-in-100-year flood. These events, characterized by the sudden or persistent inundation of normally dry areas, can have a variety of origins, such as intense rainfall, rapid snowmelt, or the breaching of levees and dams. The consequences, however, are always dramatic: destruction of infrastructure, loss of life, and damage to ecosystems and agricultural production.

An emblematic example is the floods that regularly affect regions such as South Asia, Europe and the United States, causing billions of dollars of damage each year. In addition, so-called **flash floods** represent one of the most dangerous forms of flooding. These events occur rapidly, often in less than six hours after heavy rains or other triggers, and leave very little time to respond. Flash floods are particularly deadly, causing about 60 percent of deaths associated with floods globally, and their frequency is increasing due to phenomena such as intensification of rainfall due to climate change.

The effects of climate change are evident in the increased frequency and intensity of extreme precipitation. According to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), each degree Celsius increase in global temperature results in a 7 percent increase in the atmosphere's ability to hold moisture, making extreme precipitation events more likely. In this context, studying flash floods becomes a priority not only to save lives but also to reduce economic and environmental impacts. Their unpredictability and rapidity of development require advanced forecasting and monitoring tools.

In Italy, events such as the 1966 Florence flood or more recent ones in Emilia-Romagna (both in 2023 and 2024) have highlighted the urgency of implementing more effective monitoring and forecasting systems to mitigate the risks associated with these phenomena. So, floods are also a significant risk in Italy, a country characterized by a geographical conformation that makes it particularly vulnerable to extreme events such as intense and rapid rainfall. According to data from Legambiente, more than 500 floods (and flooding from heavy rains that caused damage) have been recorded in the Peninsula since 2010. To address this problem, Italy has adopted the so-called Floods Decree (Directive 2007/60/EC), which provides for integrated flood risk management. This decree requires, among other things, the creation and updating of **hazard maps** and **risk maps** for areas exposed to flooding, with the aim of improving spatial planning and prevention measures.

Specifically, risk maps consider the socioeconomic impact of floods, assessing potential damage to infrastructure, population, and cultural property. Hazard maps, on the other hand, focus on the hydraulic characteristics of the event, highlighting flood-prone areas based on scenarios of different probability of occurrence, expressed in terms of return periods (e.g., 20, 100 or 500 years). Return periods represent a statistical indicator of the probability of an event occurring in a given time interval: a return period of 100 years, for example, indicates an event with a 1 percent chance of occurring in a specific year. An example of a hazard map can be found in Figure (1.2), where one can see for the entire Po River basin in blue the floodable areas in case of a 500-year return periods flood, in light blue the floodable areas in case of a 100-year return period flood, and in very light blue the floodable areas in case of a flood with a return period of up to 50 years. The following information can be seen from an ISPRA report to understand how much flood risk is present in Italy:

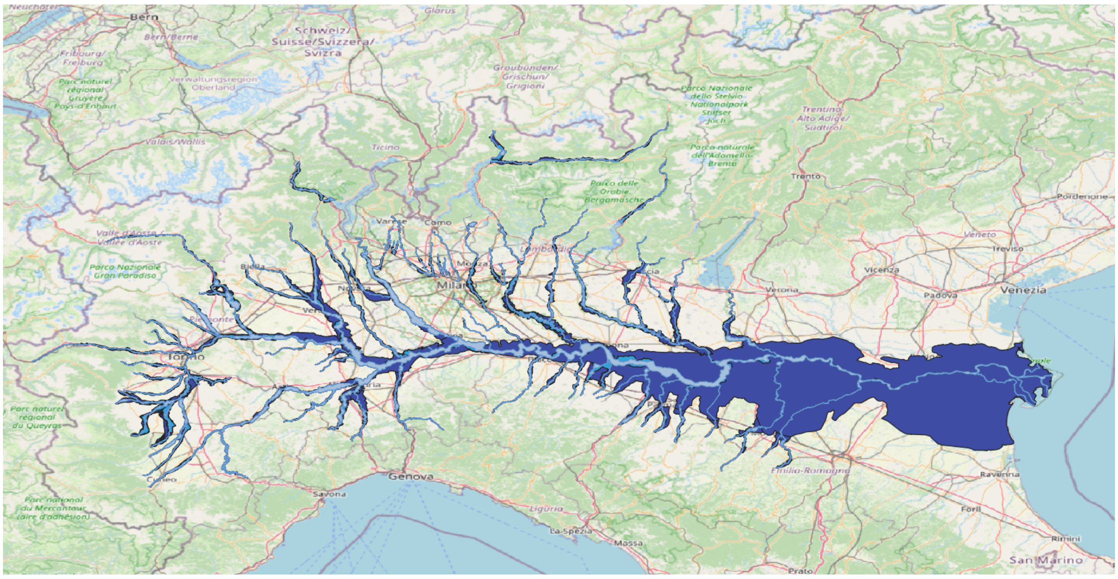
Neuchâtel
Bern
Bern/Berne
Schweiz/Suisse/Svizzera/Svizra
Glarus
Fribourg/Freiburg
Parc naturel régional Gruyère Pays-d'Enhaut
Verwaltungsregion Oberland
Graubünden/Grischun/Grigioni
Parco Nazionale dello Stelvio Nationalpark Stilfser Joch
Trentino Alto Adige/Südtirol
Valais/Wallis
Ticino
Parco naturale dell'Adamello-Brenta
Parco delle Orobie Bergamasche
Trento
Pordenone
Varese
Como
Valle d'Aosta/Vallée d'Aoste
Lombardia
Veneto
Treviso
Parco Nazionale Gran Paradiso
Biella
Monza
Milano
Vicenza
Verona
Padova
Venezia
Torino
Parc naturel régional du Queyras
Cuneo
Genova
Savona
Emilia-Romagna
Ravenna
Forlì
Liguria
La Spezia
Massa
Rimini
San Marino
Prato
Parc national du Mercantour (aire d'adhésion)

Figure 1.1. Hazard map for the Po river basin

- in Figure (1.2) can be seen the percentages of floodable territories divided by each region. In particular, the blue bar is referred to floods with 500-year return period, the light blue one is referred to floods with 100-year return period, while the very light blue one is referred to floods with return period up to 50 years;

- in Table (1.1) can be seen the total number of inhabitants exposed to flood risk.

ID	Region	Population	HPH		MPH		LPH	
			inhabitants	(%)	**inhabitants**	(%)	**inhabitants**	(%)
1	Piemonte	4.363.916	64.503	1,5	213.655	4,9	699.621	16,0
2	Valle d'Aosta	126.806	4.587	3,6	11.508	9,1	51.373	40,5
3	Lombardia	9.704.151	203.751	2,1	430.196	4,4	1.398.322	14,4
4	Trentino-Alto Adige	1.029.475	10	0,0	185.610	18,0	185.610	18,0
5	Veneto	4.855.904	422.659	8,7	568.131	11,7	1.557.994	32,1
6	Friuli Venezia Giulia	1.220.291	62.409	5,1	121.318	9,9	242.850	19,9
7	Liguria	1.570.694	164.897	10,5	273.583	17,4	365.762	23,3
8	Emilia-Romagna	4.342.135	428.568	9,9	2.714.773	62,5	3.014.805	69,4
9	Toscana	3.672.202	271.208	7,4	938.199	25,6	2.359.397	64,3
10	Umbria	884.268	33.992	3,8	63.947	7,2	103.416	11,7
11	Marche	1.541.319	2.664	0,2	79.717	5,2	186.471	12,1
12	Lazio	5.502.886	93.982	1,7	175.851	3,2	583.507	10,6
13	Abruzzo	1.307.309	39.814	3,0	94.563	7,2	259.237	19,8
14	Molise	313.660	0	0,0	7.152	2,3	7.152	2,3
15	Campania	5.766.810	115.490	2,0	293.525	5,1	346.535	6,0
16	Puglia	4.052.566	76.114	1,9	135.932	3,4	198.021	4,9
17	Basilicata	578.036	3.995	0,7	6.172	1,1	7.169	1,2
18	Calabria	1.959.050	236.707	12,1	250.035	12,8	282.577	14,4
19	Sicilia	5.002.904	126.751	2,5	215.545	4,3	246.130	4,9
20	Sardegna	1.639.362	78.485	4,8	128.963	7,9	268.893	16,4
	ITALY	**59.433.744**	**2.431.847**	**4,1**	**6.818.375**	**11,5**	**12.257.427**	**20,6**

Table 1.1. Regional data table for Italy.

In the city of Turin, flood risk is relevant because of the presence of four main waterways, the Po River, the Dora Riparia, the Sangone and Stura, which flow through densely populated and industrialized areas. Historic flood events, such as those in 2000 and 2016, have caused severe damage in the city and neighboring municipalities, highlighting the need for mitigation actions.

In last years, increasing urbanization and intensification of extreme weather events necessitate a constant review of prevention and response strategies.

1.1 The Importance of Flood Forecasting

Accurately predicting floods is a crucial challenge to save lives and minimize economic and environmental damage. The use of mathematical models and machine learning systems is emerging as a promising tool in this area due to their ability to analyze large amounts of data and identify complex patterns. These models can integrate historical data, weather forecasts, topographic information, and other variables to provide accurate estimates of flood risk in a given area.

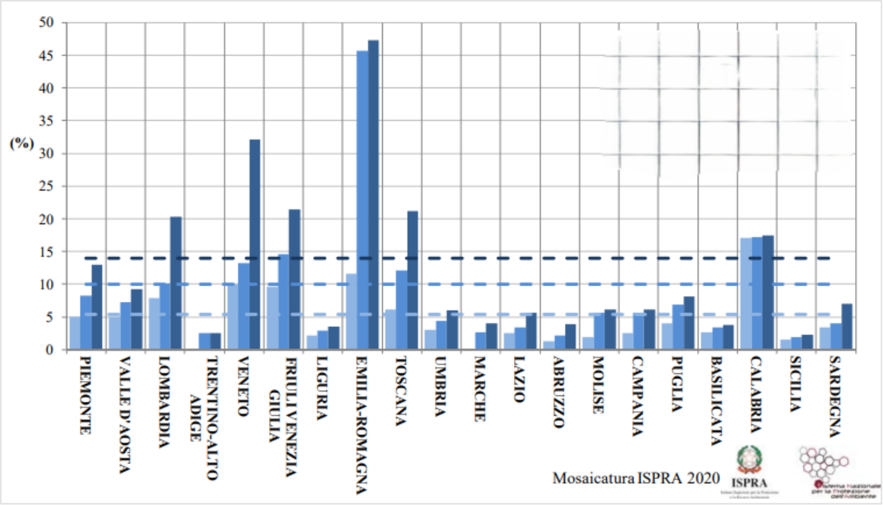

Figure 1.2. Percentage of Italian territory exposed to flood risk, divided by region

When analyzing state-of-the-art models such as LISFLOOD (developed by EFAS, the European Flood Awareness System), selecting the most relevant features is essential to train a machine learning model capable of accurately predicting floods. The key features include:

- Precipitation: Precipitation is the primary cause of flooding and one of the most critical variables to consider. In particular, runoff—the portion of precipitation that flows into rivers, increasing their discharge—is crucial for accurate flood modeling.

- Temperature, evapotranspiration, and climatic factors: These variables influence the volume of water flowing over land surfaces and are important for assessing flood risk under varying climatic conditions.

- River characteristics: Factors such as flow rate and bank capacity are essential for understanding a river system's response to extreme precipitation events.

- Vegetation and soil characteristics: Vegetation type and soil properties significantly affect water absorption and surface runoff potential.

- Land urbanization: Urbanized areas are at higher flood risk due to soil impermeability, which reduces water absorption and increases surface runoff.

- Topographical features: Elevation data is critical for understanding water runoff patterns and identifying areas more prone to flooding.

1.2 Thesis Outline

The thesis work is divided into 5 chapters, following the introduction chapter. Specifically, Chapter 2 will introduce SAR and multispectral images, which are captured by satellite and are very important in the study of flooding. Next, Chapter 3 will introduce the classification problem and the dataset constructed to address the problem, describing the methods adopted for filling it. Chapter 4 will carry out a theoretical discussion of the supervised machine learning models used to solve the classification problem, while Chapter 5 will show the experimental results of these models applied to the dataset. Finally, Chapter 6 will cover the conclusions of the work, including any limitations and suggestions for future improvements.

Chapter 2

Remote sensing and Earth observation

Remote sensing is the discipline based on the measurement of a property related to an object without making physical contact with the object itself. Sensors on Earth-orbiting satellites provide information about many natural aspects, such as dynamics of clouds, surface vegetation cover, surface morphologic structures, ocean surface temperature, and near-surface wind, as well as numerous other civilian and military information. The rapid wide coverage capability of satellite platforms allows monitoring of rapidly changing phenomena, particularly in the atmosphere. Moreover, the long duration and repetitive capability allows the study of seasonal, annual, and longer-term changes such as polar ice cover, desert expansion, and tropical deforestation.

2.1 Historical overview

Historically, the first remote sensing technique was photography. By 1858, balloons were being used to make photographs of large areas. This was followed by the use of kites in the 1880s and pigeons in the early 1900s to carry cameras to many hundred meters of altitude. The advent of the airplane made aerial photography a very useful tool because acquisition of data over specific areas and under controlled conditions became possible. The first recorded aerial photographs were taken from an airplane piloted by Wilbur Wright in 1909 over Centocelle, Italy (Figure (2.1)). These instruments were later widely used during World War I in order to monitor enemy positions, movements and defenses. For this reason, governments funding were made to further improve this technology. During the period of World War II, the main idea had become to broaden the spectrum that could be acquired by the camera. For this reason, it was started to acquire images not only using visible frequencies of the spectrum, but also using microwaves, ranging from wavelengths of 1 *mm* to about 30 *cm* or a portion of the spectrum known as near infrared (NIR), ranging from wavelengths of 0.70 μm to 1.0 μm. This change was made possible by improvements of radar (radio detection and ranging), thermal infra-red detection, and sonar (sound navigation ranging) systems. With the "space race" between 1950 and 1860,

Figure 2.1. One of the first aerial photographs taken in Centocelle, Italy

it began the development of satellite based remote sensing. In particular:

- In 1957 the Soviet Union launched the world's first artificial satellite, Sputnik 1;
- In 1960 the United States successfully lanched their first artificial satellite, Explorer 1

Due to the cold war between URSS and USA, the next decades brought about rapid developments in satellites and imaging technology. The first successful meteorological satellite (TIROS-1) was launched in 1960. In 1972 Landsat 1, the first earth resource satellite was launched by the US, in order to collect data from the Earth. The Landsat program has continued for over 50 years with Landsat 9 launched in 2021 (the entire timeline in Figure (2.2)). Since the launch of Sputnik in 1957 thousands of satellites have been launched. There are currently over 3,600 satellite orbiting the Earth, but only approximately 1400 are operational. Of these satellite, well over 100 are Earth observing satellites that carry a variety of different sensors to measure and capture data.

In Europe the main data provider is the European Space Agency (ESA) with the Copernicus Programme. It is the European Union's Earth observation program dedicated to monitoring the planet and its environment. The program is managed by the European Commission and is implemented in cooperation with the Member States, the European Space Agency, the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT), the European Centre for Medium-Range Weather Forecasts (CEPMMT), EU agencies and Mercator Océan.

The continuation of this chapter will present the two types of satellite images from the Copernicus program used in the thesis project, namely Synthetic Aperture Radar (SAR)

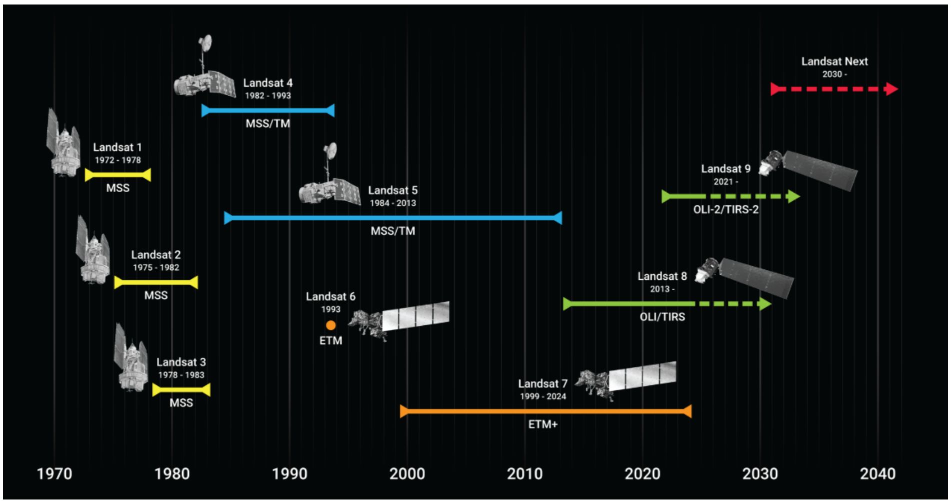

Figure 2.2. Timeline of the Landsat program

images and multispectral images.

SAR images are derived from the use of synthetic aperture rarar, which allows such images to be collected under any weather and lighting conditions. In fact, the use of such technology does not require sunlight and can also pass through clouds. This is a feature that has found its interest in surveillance of areas subject to natural events, such as flooding.

Multispectral imagery, on the other hand, captures data in different bands of the electromagnetic spectrum, thus providing much more detailed information about the condition of the Earth's surface. By integrating data from different spectral bands, an in-depth and accurate perception of the features present in a specific area can be obtained, thanks to the extraction of **spectral indices**, which are crucial for extracting some important features of the studied territories. Unlike SAR images, however, the collection of such images is hampered by clouds and absence of illumination, effectively rendering some of the collected images unusable.

However, before explaining in detail the operating principles of SAR and multispectral technologies, how the images are acquired, and their specific applications in the study of flooding, it is necessary to understand what spectroscopy is and what its importance is in the study of these satellite images.

2.2 Spectroscopy in remote sensing

Spectroscopy is the study of the absorption and emission of light and other radiation by matter. It involves the splitting of light (or more precisely electromagnetic radiation) into its constituent wavelengths (the so-called **spectrum**). Spectroscopy is really important for remote sensing; in fact, the first requirement for remote sensing is the necessity of an energy source illuminating the target that has to be acquired by the sensor. This energy is in the form of electromagnetic radiation. **Electromagnetic radiation** (Figure (2.3)) consists of an electrical field $\vec{E}$ that varies in magnitude in a direction that is perpendicular to the direction in which the radiation itself is traveling, and a magnetic field $\vec{M}$ which is perpendicular with respect to the electrical field. Both $\vec{E}$ and $\vec{M}$ move at the speed of light $c = 3 * 10^8\ m/s$. Electromagnetic radiations have two important properties that are related each other: **wavelength** and **frequency**. In particular, the wavelength λ is the length of one wave cycle, or the distance between two successive wave crests, and so it is measured in meters. On the other hand, frequency ν is the number of cycles of a wave passing a fixed point per unit of time. It is measured in hertz, that is equal to seconds to the power of -1. Wavelength and frequency are related by the following formula:

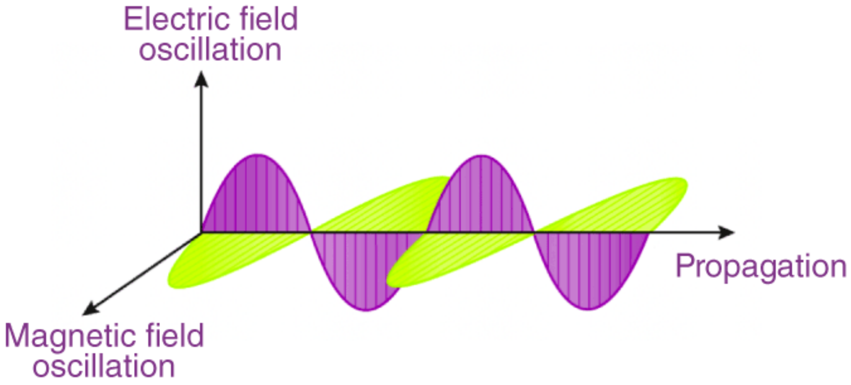

Figure 2.3. Electromagnetic radiation

$$c = \lambda\nu$$

and so are indirectly proportional. Another important characteristic of a periodic wave is the **phase**. In particular, it is an angle-like quantity representing the fraction of the cycle covered up to t, where t is time. The **electromagnetic spectrum** (Figure (2.4)) ranges from the shorter wavelengths to the longer wavelengths, and in remote sensing some portions of the spectrum are useful because the corresponding radiations can be transmitted and/or received by the sensors. In particular, the parts of the spectrum that will be considered by the satellite images used in this thesis are:

- the visible spectrum, including the frequency that human can see;
- infrared;
- microwaves.

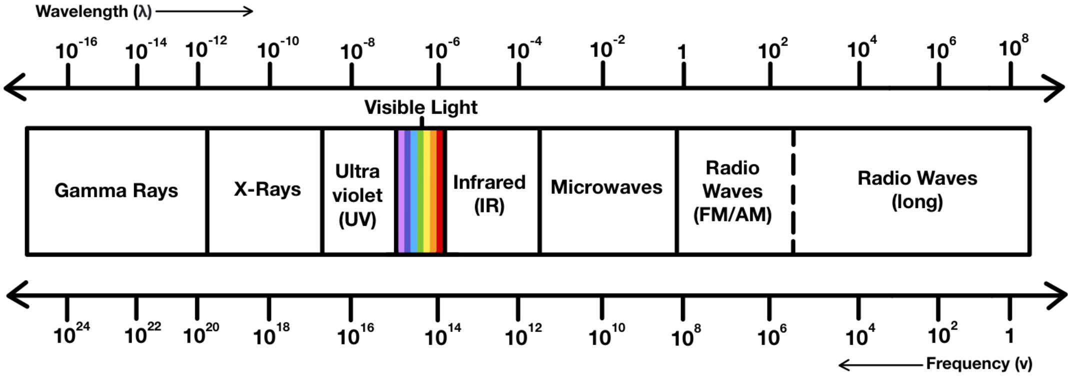

Figure 2.4. Electromagnetic spectrum

2.3 Basics of remote sensing

When discussing remote sensing using sensors and satellite images, the following distinction must first be made (Figure (2.5)):

- **active sensors**: sensors that provide their own source of energy.
- **passive sensors**: sensors that use natural energy from the Sun.

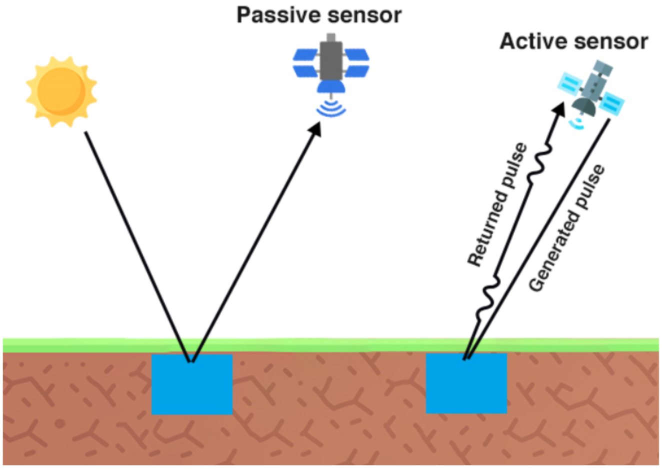

Figure 2.5. Difference between active and passive sensors

In both cases, an electromagnetic radiation marks the beginning of remote sensing. Such electromagnetic radiation, as it passes through the atmosphere, is subject to certain phenomena:

1. **Scattering**: it occurs when electromagnetic radiation, interacting with gases in the atmosphere, deviates from its original path. This effect depends on several factors, including the wavelength of the radiation, the abundance of particles or gases, and the distance traveled by the radiation through the atmosphere.

2. **Absorption**: through this phenomenon, molecules in the atmosphere absorb energy at various wavelengths. Because of this, there is a distinction about which parts of the spectrum are good candidates for remote sensing (as they have low absorption) and which are not. In addition to visible light, which clearly is not absorbed by the atmosphere, examples of waves that are not affected are microwaves, while waves such as gamma rays, X-rays, and UV rays have large percentages of absorption by the atmosphere.

At this point, the electromagnetic radiation finally hits the target object subject to measurement. The possible phenomena are as follows:

1. **absorption**: it occurs when the radiation is absorbed into the target;

2. **transmittance**: it occurs when the radiation passes through the target;

3. **reflection**: it occurs when radiation is reflected off a target and subsequently redirected. In the field of remote sensing, the primary focus is on measuring with some sensors the radiation that is reflected from these targets. In particular, when the surface of the target is smooth the reflection is **specular**, while when the surface of the target is rough the reflection is **diffuse** (Figure (2.6)).

Therefore, the key point in remote sensing is to use the right wavelength of electromagnetic radiation in order to extract some useful information. An example is provided in Figure (2.7): in fact, in this graph the response in terms of reflectance of vegetation and water using different wavelengths can be seen. In particular, while using wavelengths from 0.4 μm to 0.7 μm there is no big difference in the response, using wavelengths in the infrared part of the spectrum a high response can be noticed for vegetation and a very low response for water. For example, this can be exploited to extract very useful information about the presence of vegetation or water in specific areas of the image, as will be seen later.

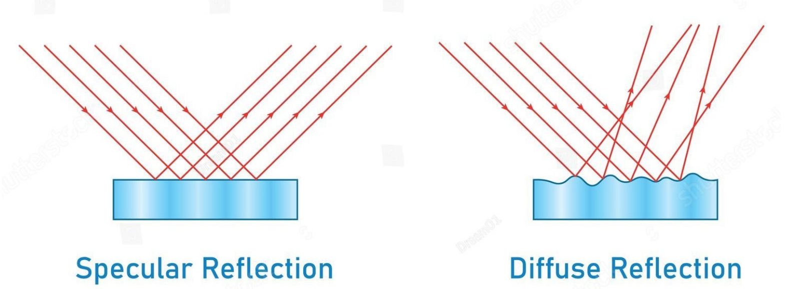

Figure 2.6. Specular and diffuse reflection

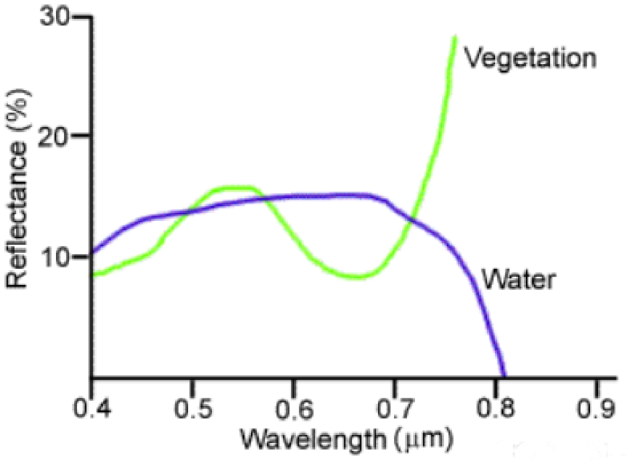

Figure 2.7. Response in terms of reflectance for vegetation and water, using different wavelengths

2.3.1 Resolutions of sensors

Before dealing specifically with the operation of SAR and multispectral imaging, some concepts regarding remote sensing using satellites will now be introduced. In particular, in order to understand the following sections more thoroughly, it is necessary to introduce the concepts of spatial resolution, spectral resolution, radiometric resolution, and temporal resolution.

- **Spatial Resolution**: it refers to the size of the smallest possible feature that can be detected. As shown in Figure (2.8), the spatial resolution of passive sensors (the specific case of active microwave sensors will be discussed later) is primarily determined by their Instantaneous Field of View (IFOV). The IFOV represents the sensor's angular cone of visibility (A) and defines the area on the Earth's surface that is "observed" from a specific altitude at a given moment (B). The extent of this observed area is calculated by multiplying the IFOV by the distance from the sensor to the ground (C). This ground area, known as the resolution cell, defines the maximum spatial resolution achievable by the sensor.

- **Spectral Resolution**: it indicates the sensor's ability to distinguish between bands of different wavelengths in the electromagnetic spectrum. It represents the width of the spectral bands that the sensor is able to detect and separate; the narrower these bands are, the higher the spectral resolution of the sensor. High spectral resolution enables the sensor to discriminate fine details in the spectral characteristics of materials on the Earth's surface. For example, hyperspectral sensors, which can have hundreds of very narrow bands, are able to identify specific spectral signatures of materials such as minerals, water, vegetation, etc., while multispectral sensors, with a few wider bands, provide a more synthetic and less detailed view.

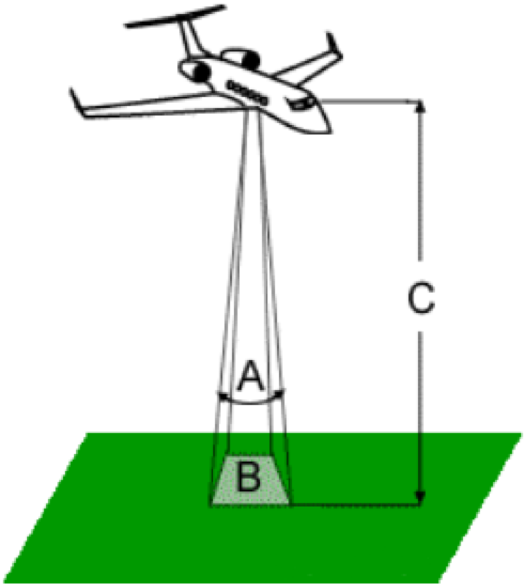

Figure 2.8. Spatial resolution for a passive sensor

- **Radiometric Resolution**: it describes the ability of the imaging system to discriminate very slight differences in energy. The finer the radiometric resolution of a sensor is, the more sensitive it is to detecting small differences in reflected or emitted energy. It depends on how many bits are used by the sensor to record the data. For example, if a sensor used 8 bits, there would be 2^8 digital values available, ranging from 0 to 255, to represent each pixel value in binary format.

- **Temporal Resolution**: it concerns the frequency with which a sensor can observe the same observation area. It is thus a measure of the time interval between two successive acquisitions of the same location by the sensor. Specifically, in the case of remote sensing by satellite, temporal resolution thus corresponds to the revisit time of a satellite in a given area, that is, the time interval it takes for the satellite to fly over the same area on the earth's surface again and collect new data.

2.4 Multispectral remote sensing

Most remote sensing satellite sensors offer multispectral images. Specifically, they are created through the use of multiple sensors mounted on the satellite. This makes it possible to receive information in different bands, both in the visible part of the spectrum and in other areas (such as NIR, Near-Infrared). This multispectral information is extremely valuable in identifying various objects in the image. For example, much information can be extracted regarding the vegetation in a given area. This happens because almost all plants need sunlight to survive, using chlorophyll to convert the sun's energy into organic energy. Chlorophyll itself has unique absorption characteristics, absorbing wavelengths around the visible red band (645 µm), while transparent at near-infrared wavelengths (700 µm). These chlorophyll characteristics are used to design indices to estimate the density of vegetation in a given area of a multispectral image, as will be seen later.

2.4.1 Applications of multispectral images

Multispectral imagery offers a wide range of applications because of its ability to capture and analyze detailed information across multiple spectral bands. In particular, there are applications in military, urban, and environmental disaster management. In fact, combining the different bands through specific formulas yields so-called **spectral indices**, which are useful for monitoring and detecting such things as the state of plant stress, soil type, or water quality.

2.4.2 Sentinel-2 mission

The multispectral images used in this thesis are taken from the Copernicus Programme. Specifically, they are part of the Sentinel-2 mission, which was launched on June 23, 2015 with the launch of the Sentinel-2A satellite and then continued with the launch of the Sentinel-2B satellite on March 7, 2017, and then with the launch of the Sentinel-2C satellite on September 5, 2024. Specifically, these satellites use the MultiSpectral Instrument (MSI), a sensor that detects 13 different spectral bands. These bands cover a range of wavelengths from visible (VIS) to shortwave infrared (SWIR). This sensor receives incoming light and uses a series of filters and detectors to separate them into the different spectral bands. The 13 different spectral bands are divided as follows:

- 4 bands with spatial resolution of 10 m;
- 6 bands with spatial resolution of 20 m;
- 3 bands with spatial resolution of 60 m.

More information about the bands is given in the table (2.1)

In addition, sentinel-2 mission images are available in 2 levels:

- L1C level: raw data (L0 level not available for download) are pre-processed with a series of algorithms to correct and georeference the images. However, no atmospheric correction is applied, i.e., the effects that the atmosphere has on an image are not removed.
- L2A level: an atmospheric correction is applied to L1C data.

2.5 Microwave remote sensing

Microwave Remote Sensing is an extremely useful technology that offers significant advantages. Although remote sensing with visible bands or infrared bands relies on sunlight

Band	**Num**	**Wavelength (nm)**	**Bandwidth (nm)**	**Spatial res.**
Coastal	1	442.7	21	60 m
Blue	2	492.4	66	10 m
Green	3	559.8	36	10 m
Red	4	664.6	31	10 m
Red Edge 1	5	704.1	15	20 m
Red Edge 2	6	740.5	15	20 m
Red Edge 3	7	782.8	20	20 m
NIR 1	8	832.8	106	10 m
NIR 2	8A	864.7	21	20 m
Water vapour	9	945.1	20	60 m
SWIR 1	10	1373.5	31	60 m
SWIR 2	11	1613.7	91	20 m
SWIR 3	12	2202.4	175	20 m

Table 2.1. Some information about the Sentinel-2 image bands

and can be hindered by adverse weather conditions, such as clouds or fog, microwave remote sensing has unique characteristics that make it especially suitable for monitoring areas that are difficult to reach or observe with other methods.

In fact, as can be seen in Figure (2.9), microwaves have the unique ability to penetrate clouds and other unfavorable atmospheric conditions, such as rain, fog or storms. Another crucial aspect is their ability to function at night, without depending on sunlight. In microwave remote sensing, the part of the spectrum containing wavelengths from about 1 mm to 30 cm is used. Specifically, this area of the spectrum is divided into the bands visible in Table (2.2). In contrast to remote sensing in the visible and infrared bands, where technologies based on passive sensors and camera-like lenses are used, in remote sensing by microwave, active sensors are used by exploiting radar technology. In particular, the sensor first functions as a transmitter, sending a microwave beam toward the area to be monitored. Subsequently, the same sensor acts as a receiver, measuring the intensity of the so-called "**backscatter signals**", which are the electromagnetic radiation reflected by

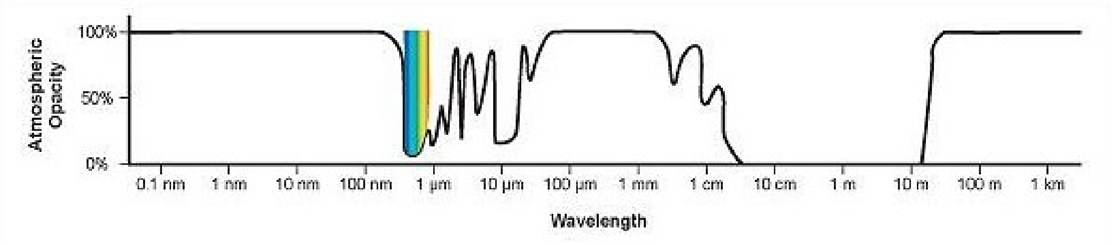

Figure 2.9. Percentage of light that is blocked by gases in the atmosphere, for different bands in the spectrum

Band	Wavelength Range (cm)	Frequency range (GHz)
UHF (P)	30-100	1-0.3
L	15-30	2-1
S	7.5-15	4-2
C	3.75-7.5	8-4
X	2.5-3.75	12-8
Ku	1.67-2.5	18-12
K	1.11-1.67	27-18
Ka	0.75-1.11	40-27
V	0.40-0.75	75-40
W	0.27-0.40	110-75
mm	0.10-0.27	300-110

Table 2.2. Standards microwaves bands

the targets that were intended to be measured. However, measuring only the intensity of these signals is not enough to build an image, as it would not provide information on the precise location of the objects illuminated by the initial beam.

At this point, the radar (radio detection and ranging) technology comes into play. In addition to saving the information on the intensity of the backscatter signal, the time it takes from when the electromagnetic radiation is transmitted by the sensor to when the reflected signal is received is recorded for each reflected signal. This time information allows the distance between the sensor and the target to be determined. Due to the use of radar technology just explained, however, it is necessary for remote sensing to be done by the antenna not illuminating the ground in a vertical orientation. In fact, if the antenna illuminated the ground vertically, features on the same distance from the antenna (as the two black rectangles in the right side of Figure (2.10)) would occupy the same pixel location in the final image.

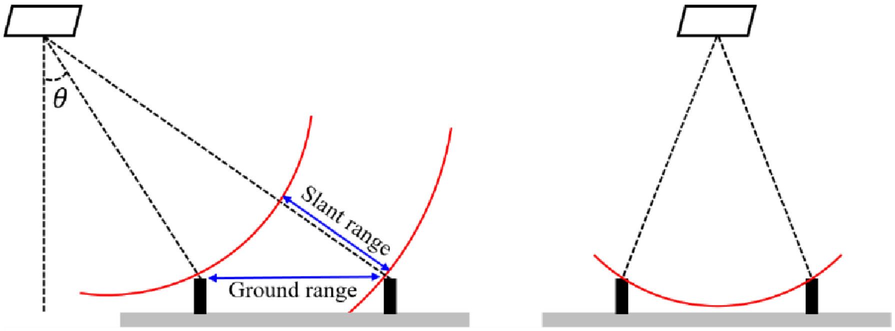

Figure 2.10. Side-looking and vertical illumination

Therefore, the imaging geometry of a radar system is different from the systems commonly employed for optical remote sensing. It is then necessary to introduce some terminology, looking at the left side of Figure (2.10) and Figure (2.11):

- **Nadir (B)**: it is the point directly beneath the platform (airborne or satellite);
- **Swath (C)**: it is the horizontal portion of the earth's surface that the sensor registers when it illuminates the ground;
- **Range (D)**: it refers to the across-track dimension perpendicular to the flight direction (A). In particular the portion of the image swath closest to the nadir track of the radar platform is called the **near range** while the portion of the swath farthest from the nadir is called the **far range**;
- **Azimuth (E)**: it refers to the along-track dimension parallel to the flight direction (A);
- **Slant range distance**: it is the radial line of sight distance between the radar and each target on the surface;
- **Ground range distance**: it is the true horizontal distance along the ground corresponding to each point measured in slant range;
- **Incidence angle**: it is the angle between the radar beam and the surface. It can be noted that this angle increases moving across the swath from near to far range;
- **Look angle**: it is defined as the angle between the vertical direction and the radar beam at the radar platform

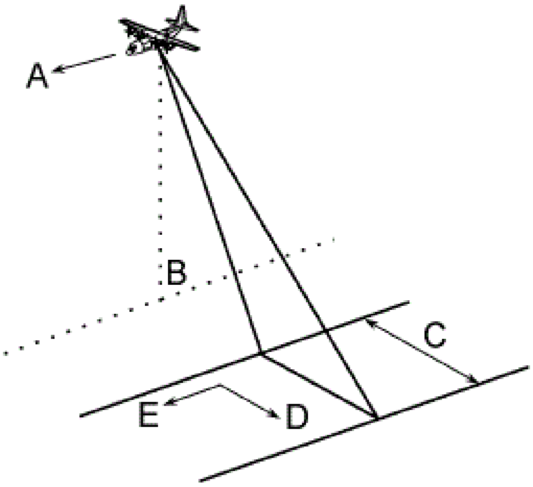

Figure 2.11. Geometry of a radar system

Resolution of radar system

Unlike optical and infrared remote sensing, where spatial resolution is typically a single measure, radar imagery requires spatial resolution to be considered in two distinct components: range resolution and azimuth resolution. In particular, the range resolution depends on the speed of light c (i.e., 3 x 10^8 m/s) and the duration of the electromagnetic pulse that is transmitted by the antenna, denoted by T; the specific formula is:

$$\text{Range Resolution} = \frac{cT}{2}. \tag{2.1}$$

On the other hand, azimuth resolution depends on the range R, on the wavelength of electromagnetic radiation λ and the diameter of the antenna D. Specifically, the formula is as follows:

$$\text{Azimuth Resolution} = R\,\frac{\lambda}{D} \tag{2.2}$$

The equations (2.1) and (2.2) present a significant problem. Indeed, consider an antenna of diameter $D = 6$ m (of considerable size) transmitting in the X-band ($\lambda = 3$ cm), positioned in space, with a range $R = 750$ km. Further, assume that it generates a pulse with a duration of $T = 1\ x\ 10^{-6}$ s. The results are then:

$$\text{Range Resolution} = 150\text{ m}$$

and

$$\text{Azimuth Resolution} = 3750\text{ m}$$

Clearly, an image with a resolution of 150 meters by 3750 meters per pixel is unusable. However, since neither range nor wavelength can be changed, it is possible to improve azimuth resolution by increasing the antenna diameter and range resolution by reducing the pulse duration. However, these solutions also present difficulties. Increasing the diameter of the antenna would involve complications in construction and transportation in space; an antenna of the order of kilometers in size would be required to achieve acceptable range resolution. Reducing the pulse duration, on the other hand, would require too much power to generate a pulse short enough and powerful enough for the antenna to receive its echo.

The solution to these problems lies in the use of the Synthetic Aperture Radar (SAR) technique.

2.6 Synthetic Aperture Radar (SAR)

Synthetic Aperture Radar (SAR) represents one of the most advanced and versatile technologies for remote sensing by microwave radar. This technology effectively meets the needs for high resolution in both range and azimuth directions, overcoming the limitations of conventional radars. Later in this section, it will be explored how SAR succeeds in synthesizing a long antenna to improve azimuth resolution and how the use of chirp signals enables improved range resolution.

Improvement of azimuth resolution Synthetic Aperture Radar improves the azimuth resolution of the image by simulating a much larger antenna through the movement of the radar itself. During flight, the radar does not transmit a single pulse toward a particular target, but as it moves, it transmits pulses in sequence and records signals reflected from the target (Figure (2.12)). These signals are then processed to construct a virtual aperture much larger than the physical aperture of the antenna. This process leads to an improvement in azimuth resolution. In fact, as seen from the formula (2.2), this resolution depends on the diameter of the antenna. Consequently, having synthesized this very large antenna, there will be an increase in azimuth resolution. In fact, denoted by θ the angle between the first moment of signal transmission toward a target and the last moment of transmission toward the same target, we have that the diameter L of the synthetic antenna can be approximated as the product of θ and range. This approximation is valid when the angle θ is very small, as in the case considered. In fact, in this situation the geometry is very different from that presented in Figure (2.12) (which serves only as an illustration for better understanding). In actual facts, given the size of the Earth, the distance traveled by the satellite is much shorter and therefore the angle θ can be approximated, as just explained. Therefore, substituting in the formula (2.2), the following is obtained:

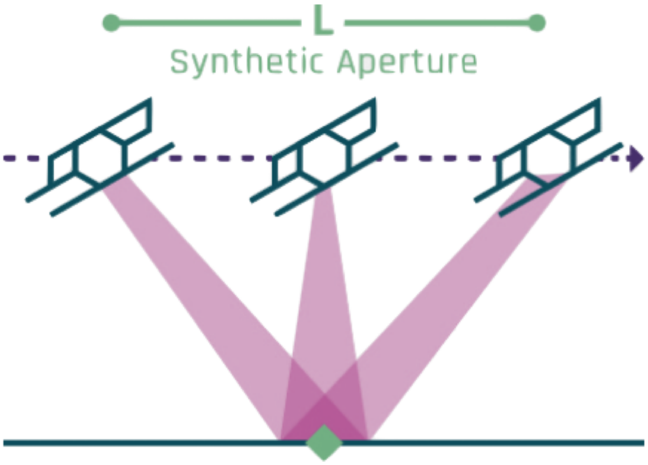

Figure 2.12. Synthetic Aperture

$$\text{Azimuth Resolution} = \frac{\lambda}{\theta}.$$

However, this equation only accounts for one-way ranges from some scatterer to the spread-out synthetic array elements. It assumes that the beams emanate from a single location, so the transmit distance to some object would be the same for all array locations. In such a case, the ranges of each reflection would naturally vary across the span of the array. In actual SAR imaging, the antenna transmits pulses sequentially from one array location to the next. In this case, both the transmit distance and the backscatter distance are different for each location. Thus, the phase differences are proportional to the round-trip distance, as opposed to the one-way distance above, and the phase changes are double in size. This improves the resolution by a factor of two. So, the final equation is:

$$\text{Azimuth Resolution} = \frac{\lambda}{2\theta}.$$

Improvement of range resolution An effective method to improve range resolution, used in SAR images, is the use of the chirp signal, also known as the Linear Frequency Modulation (LFM) signal (Figure (2.13)). A chirp signal is characterized by a frequency that varies linearly with time. When transmitted, the frequency of the signal increases (chirp up) or decreases (chirp down) steadily during the transmission interval. This frequency variation allows time (and therefore distance) information to be encoded in a frequency variation. Specifically, when stating the frequency or wavelength value of a SAR sensor, those values typically apply at the mid-way time of the pulse. This is known as the radar center frequency or wavelength. The chirp signal is exploited in the creation of SAR images in the following way. Suppose that a chirp signal that varies from high to low frequency is transmitted by the sensor. Then, the pulse reflects from a ground object and returns. The part that reflects first is the part that hit the object first, the high frequency portion, so the echo has a reversed form in comparison to the output pulse. In addition, the backscatter is extremely weak with a small amplitude. At this point, a reversed copy of the transmitted pulse is maintained so that it can be compared with the signal reflected from the target. A cross-correlation is then made between the received signal and the reversed copy of the transmitted one, and a strong signal is output exactly when the echo and the reference waves are aligned (i.e., in the situation visible in Figure (2.13)) This synthetic compressed pulse (Figure (2.15)) replaces the spread-out pulse and solves range resolution problems. In fact, in this way there is no need to decrease the duration of the transmitted signal, since the original signal is only used to synthesize the compressed signal, which has a much shorter duration. Specifically, defining bandwidth as the frequency change present in the chirp signal, it is known that a pulse with bandwidth B can be compressed into a pulse with a time duration of $1/B$. Following the formula (2.1), the range resolution then becomes:

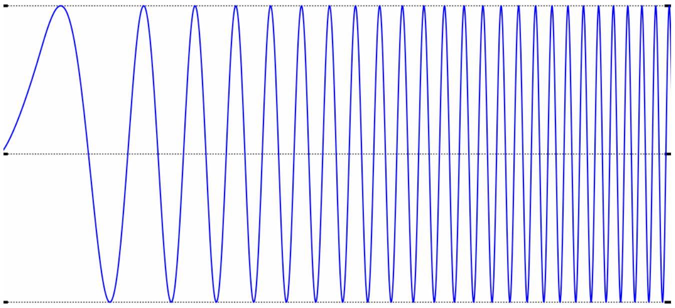

Figure 2.13. Chirp signal

$$\text{Range Resolution} = \frac{c}{2B}.$$

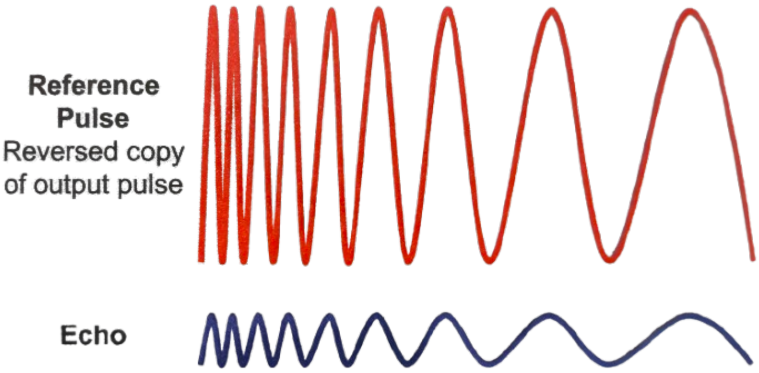

Figure 2.14. Reversed copy of transmitted signal and received signal aligned

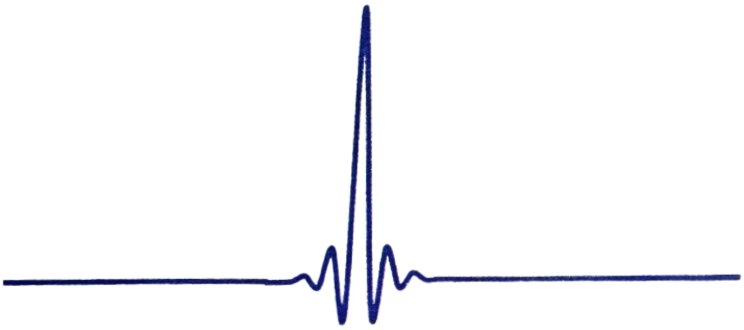

Figure 2.15. Compressed pulse

2.6.1 Characteristics and distortions of SAR images

Until now, the topic of SAR imaging has been treated by considering only the backscatter coefficient, that is, the intensity with which the pulse returns toward the antenna after hitting an object. However, it is important to note that such backscatter is not the only information saved in SAR images. In fact, by dealing with an active sensor, one has the ability to measure precisely the phase changes. This happens because, unlike passive sensors, the antenna transmits its own signal and thus the initial phase of the pulse is known. One can then compare that phase with the return signal to obtain the phase of the received signal. For this very reason, in many types of SAR images each pixel takes on a value expressed by a complex number, where:

- the real part corresponds to the intensity, backscatter coefficient;
- the complex part corresponds to the phase.

Another important feature of SAR imaging lies in the concept of **polarization**. Polarization refers to the orientation of the pulse's electric field relative to Earth's surface. Radar illumination is controlled and the sensors can emit pulses with vertically-oriented electric fields (V) or horizontal electric fields (H). The look of a SAR image is partly determined by the chosen polarization. Vertical objects reflect vertical waves strongly, such as buildings or waves in rough seas. Horizontal objects reflect horizontal waves strongly, such as a flat surface like water. A sensor can emit and receive the same polarization (HH, that is H-sent and H-receive or VV, that is V-sent and V-receive) or cross polarization (VH, that is V-sent and H-receive or HV, that is H-sent and V-receive). Because of their nature, SAR images also contain some distortions and noise that must be corrected in preprocessing:

- thermal noise: it is caused by microscopic motion of electrons of the radar receiver due to temperature;
- speckle noise: appears in the image in the form of granularity or spots, and is caused by the interference of reflected waves. Such interference occurs because the waves in question are **coherent** (that is, the phase difference between them remains constant over time). The interference just mentioned can be constructive or destructive, and causes random variations in the amplitude and phase of the received signal, which manifest themselves precisely as speckle noise in the image;
- radiometric distortion: incorrect backscatter values mainly because of the effect of atmosphere;
- geometric distortions: in the context of geometric distortions, three different effects (Figure (2.16)) due to relief are also to be considered: **foreshortening**, **layover** and **shadowing**. Foreshortening depends on the angle of incidence of the radar beam with respect to the slope of the object to be measured. When the radar pulse arrives at the base of a relief before it arrives at the top, foreshortening can occur, whereby the actual slope appears compressed and the magnitude of the distance from the base to the top incorrect. Layover occurs when the pulse arrives earlier at the summit than at the base. the summit return signal will be received earlier than the base signal, as a result there will be an incorrect position of the summit. Finally, the shadowing effect occurs when some surfaces (building facades, mountains, etc.) do not contribute to the radar echo.

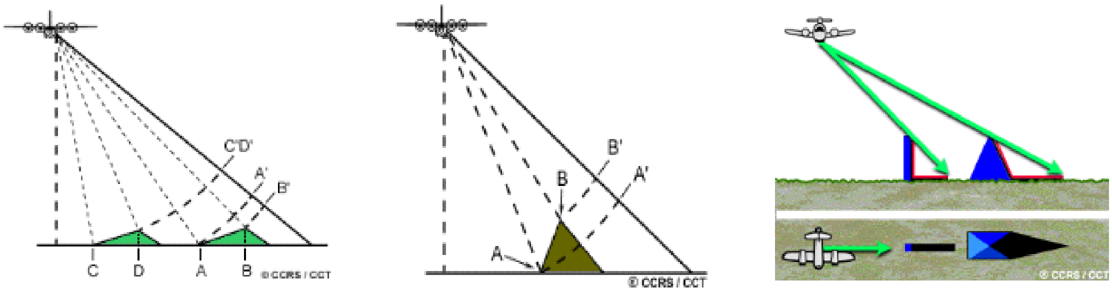

Figure 2.16. Foreshortening, layover and shadowing

2.6.2 Applications of SAR images

SAR images have multiple uses in natural disaster prevention, urban and security domains. The best known applications are in **change detection** and **interferometry**. Specifically, in change detection two images are compared, one before a particular event and the second after this same event. Thanks to change detection then, one can see the differences in the backscatter coefficients between the two images, and one can get a complete mapping of the areas affected by the event (e.g., flooded areas or areas that have suffered deforestation or fire). In the use of interferometry, on the other hand, the phase of two successive images is taken into account, and this is useful because by noting phase differences between the two different images it is possible to derive whether there has been subsidence or ground uplift, even if this movement should be on the order of millimeters. This finds wide use in the area of landslide or earthquake prevention and prediction. A final application of SAR imagery (the one that will be presented in this thesis) involves using backscatter coefficients through some mathematical formulas to take advantage of some indices aimed at monitoring flood risk.

2.6.3 Sentinel-1 mission

The SAR images used in this thesis are taken from the Copernicus Programme. Specifically, they are part of the Sentinel-1 mission, which was launched on April 3, 2014 with the launch of the Sentinel-1A satellite and then continued with the launch of the Sentinel-1B satellite on April 25, 2016, and then with the launch of the Sentinel-1C satellite on December 6, 2024. Specifically, these satellites transmit and receive in the C-band and are capable of operating with both cross-polarization and the same electric field polarization for transmitting and receiving. In addition, Sentinel-1 mission images are available in different acquisition modes. Specifically, the following information about spatial resolution pertains to the images used in this thesis. These images are Ground Range Detected (GRD) images, which contain only backscatter coefficients and do not include phase information, as the latter is not required for the thesis work that will be treated:

- Stripmap (SM): it acquires data with an 80 km swath at 9 m by 9 m spatial resolution. These are the best images to use in terms of spatial resolution, but unfortunately they often cover too narrow an area and are not available for all geographic areas;

- Interferometric Wide Swath: It acquires data with a 250 km swath at 20 m by 22 m spatial resolution. This mode is the main acquisition mode over land and it cover almost all the geographical areas;

- Extra Wide Swath: it acquires data over a wider area than for IW mode using five sub-swaths. EW mode acquires data over a 400 km swath at 50 m by 50 m spatial resolution;

- Wave (WV): this mode is specific for ocean or sea observation. In particular, it acquires data in 20 km by 20 km vignettes, at 5 m by 5 m spatial resolution.

Chapter 3

Dataset construction

This chapter outlines the process of constructing the dataset, which is designed for use in a binary classification problem. In particular, the positive class, labeled as “flooded,” will represent instances of flooding, while the negative class, labeled as “non-flooded,” will represent the absence of flooding. The goal is to predict flash floods in various urban areas.

As anticipated in the first chapter, the city of Turin was considered in this study, with the corresponding 4 rivers (Po, Dora Riparia, Sangone and Stura di Lanzo). In Figure (3.1) the geometry of the city and the waterways just mentioned can be seen. Since the city has an area of more than 130 km^2, it is now necessary to define a narrower study area. To do this, the flood hazard maps on the ADBPO (“Autorità di Bacino Distrettuale

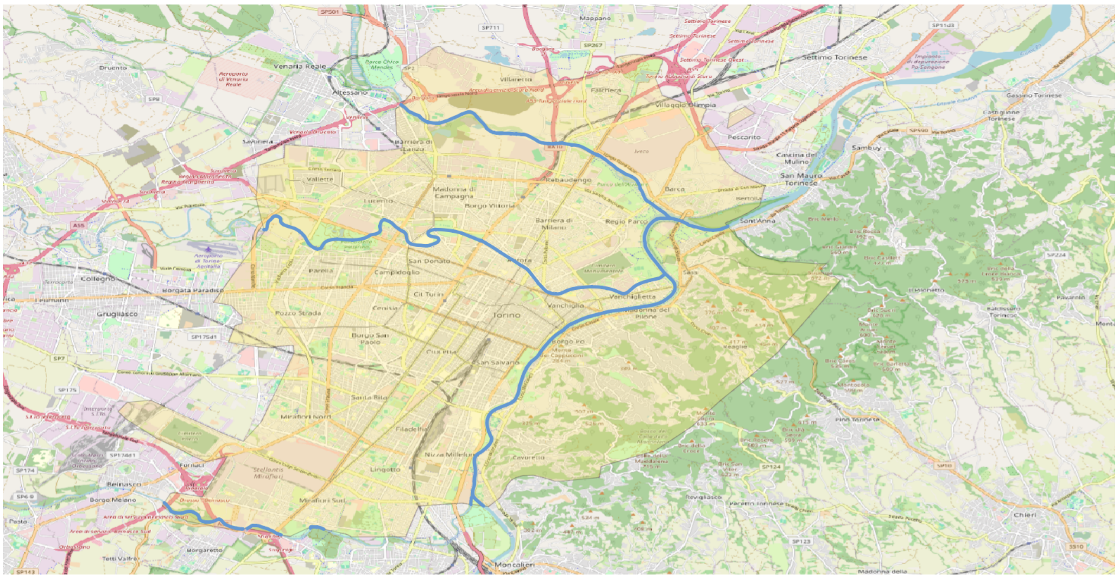

Figure 3.1. Geometry of Turin and rivers of the city

del Po") website were used. Following the Italian government's Directive 2007/60/EC, flood hazard maps were present for:

- return period up to 50 years;
- return period between 100 and 200 years;
- return period up to 500 years.

Among these three options, flood hazard maps with return periods up to 50 years were used (Figure (3.2)). This choice is justified by the fact that no past data regarding floods with return periods longer than 20 years exist in Turin, and consequently no predictive study would have been possible. In addition, in order to divide these study areas into smaller subareas that are easier to manage, so-called **Voronoi diagrams** were used. In particular, a Voronoi diagram is a partition of space into areas, or cells, that surround a set of geometric objects (usually points called **Voronoi centroids**). These cells, or polygons, must satisfy the criteria such that all locations within an area are closer to the object it surrounds than to any other object in the set.

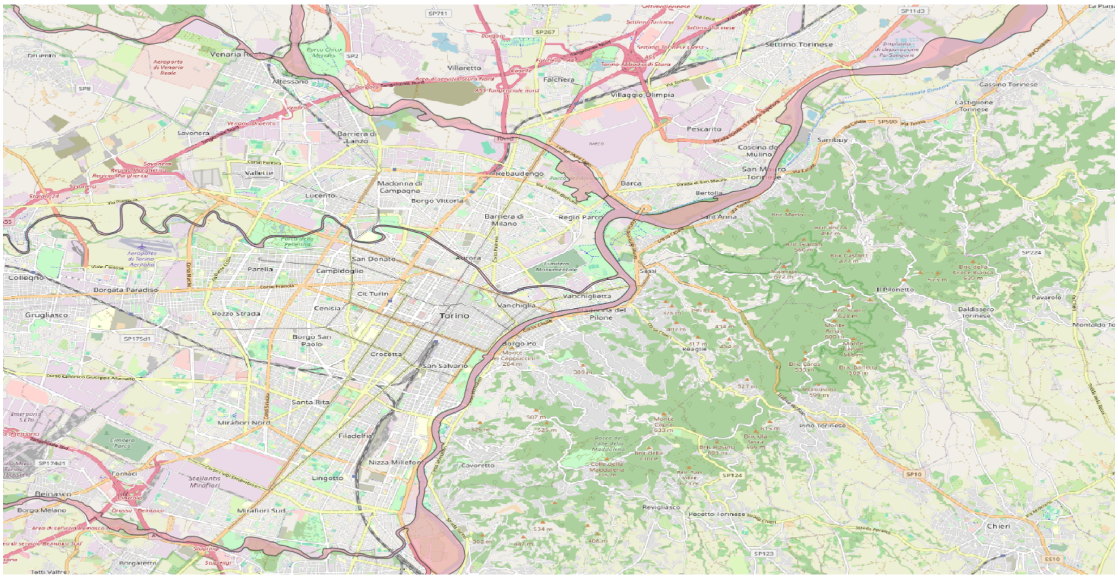

Figure 3.2. Flood hazard map with return periods up to 50 years

In this study, river points spaced 400 meters apart were used as centroids to construct Voronoi polygons, which were delineated based on these points and the flood hazard maps shown in Figure (3.2). The resulting partitioning of the Voronoi cells is illustrated in Figure (3.3), with the corresponding Voronoi centroids in yellow.

For the temporal definition of the dataset, all dates from January 1, 2016, to December 31, 2023, were included. Specifically, data was considered at four intervals each day, spaced six hours apart: midnight, 6 a.m., 12 p.m., and 6 p.m.. This choice was made with the idea of predicting flash floods using a sufficiently high temporal resolution of the data. In

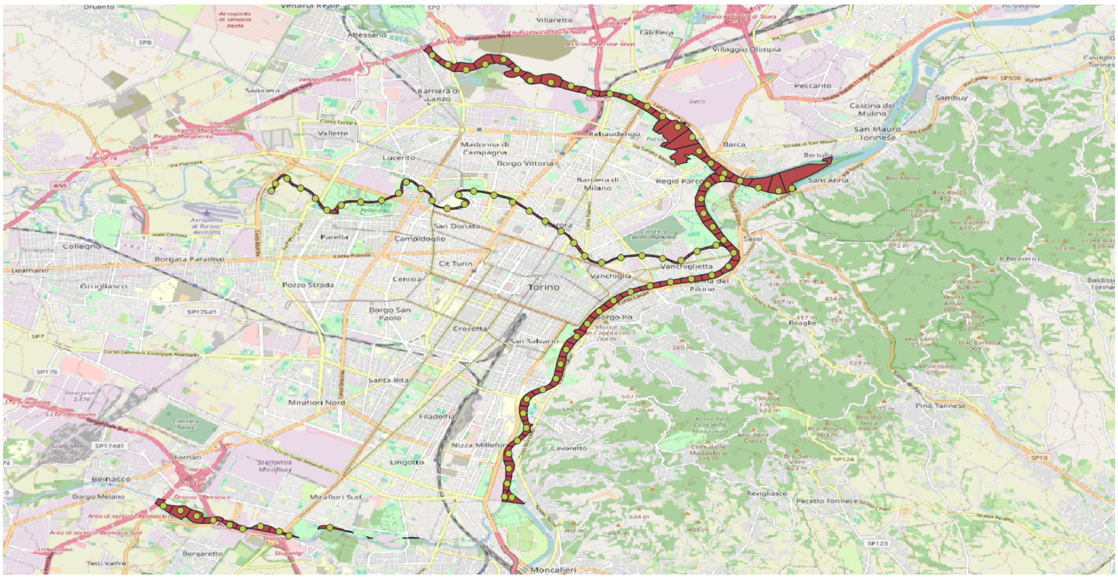

Figure 3.3. Voronoi cells and voronoi centroids

summary, each row of the dataset will contain the coordinates of the centroid of a specific Voronoi cell along with the corresponding date and time. The following sections will detail the features used to characterize these centroids across different dates and hours.

3.1 River discharge data

Since the ultimate goal is to create a binary classification model, initially the data used to compose the target variable of the problem will be processed. Specifically, thanks to EFAS (European Flood Awareness System) there is a dataset regarding stream flows every 6 hours. The data in question concerns the so-called “river discharge”, that is the volume of water flowing through a river channel at a given location and time and is expressed in m^3/s. In terms of spatial resolution, this dataset is not specific to streams but is presented through a regular grid of 0.05° x 0.05°. Therefore, it is necessary to filter the observations through an auxiliary file containing the upstream area of the considered river discharge; specifically, the upstream area of a river refers to the portion of the river and its watershed that is located above a specific point along the river’s course, typically moving toward the source or headwaters. Knowing this, the river discharge data can be filtered in the following way: if the upstream area is sufficiently high then the observation is indeed referring to a watercourse, otherwise it does not correspond to a river discharge value and therefore should be discarded as meaningless. Since the resolution of the dataset is lower than that of the centroids of the Voronoi polygons, it is necessary to adopt a method to populate all rows of the dataset. To achieve this, focusing on river stretches without inflows or outflows (where river discharge remains constant because of the principle of conservation of mass and equation of continuity in fluids), the following steps were implemented:

- initially, each river discharge value was assigned to the nearest centroid;
- for centroids located between two centroids with already assigned river discharge values, a linear interpolation was performed between the nearest discharge values to the east and west;
- for centroids not located between two centroids with assigned discharge values, the river discharge of the nearest centroid was attributed, leveraging the principle of discharge conservation in stretches of the river without inflows or outflows.

In doing so, each river stretch had its assigned river discharge for each time and date in the dataset. In order to obtain a classification problem, an EFAS auxiliary file containing river discharge thresholds not to be exceeded in order not to have flooding was then considered. Specifically, thresholds not to be exceeded for the occurrence of floods with a return period of 1.5 years were considered, consistent with the choice of flood hazard maps. Then next, for each centroid and for each date and time, the target variable was created as follows:

- if the river discharge is greater than or equal to the relative threshold, then the dataset row was marked as being part of the "flooded" class (positive class);
- if the river discharge is less than the relative threshold, then the dataset row has been marked as being part of the "not flooded" class (negative class).

3.2 Spectral indexes

As mentioned in the previous chapter, multispectral images can be used with the calculation of spectral indices, which are useful in monitoring certain environmental and land characteristics. All Sentinel-2 images of Turin captured between 2016 and 2023 were downloaded. Importantly, only images with less than 10% cloud cover were selected. In total, 198 images were retained, requiring approximately 156 GB of storage space. Looking at the bands in Table (2.1), the spectral indices used will now be described.

NDVI (Normalized Difference Vegetation Index)
It is used as a measure for abundant, healthy vegetation. It is sensitive to the effects of foliage chlorophyll concentration, canopy leaf area, foliage clumping and canopy architecture. This index's value lies between -1 and 1 and high values stand for dense vegetation.

$$\mathrm{NDVI} = \frac{\mathrm{NIR}_1 - \mathrm{Red}}{\mathrm{NIR}_1 + \mathrm{Red}} = \frac{\mathrm{B8} - \mathrm{B4}}{\mathrm{B8} + \mathrm{B4}}$$

MSAVI2 (Modified Soil Adjusted Vegetation Index 2)
It is used to estimate vegetation density in the same way as the NDVI index. However, the NDVI index is very sensitive to the presence of exposed soil (i.e., land areas that are not covered by vegetation or are only partially covered by vegetation), which can lead to underestimation of vegetation in sparsely vegetated areas. The MSAVI2 index solves this problem and is also more robust for urban settings where vegetation is sparse. In

particular, the MSAVI2 index is a variation of the MSAVI. In fact, for the MSAVI index the formula is:

$$\mathrm{MSAVI} = \frac{\mathrm{NIR_1} - \mathrm{Red}}{\mathrm{NIR_1} + \mathrm{Red} + \mathrm{L}} = \frac{\mathrm{B8} - \mathrm{B4}}{\mathrm{B8} + \mathrm{B4} + \mathrm{L}},$$

where L is a corrective term, chosen manually, which compensates for the effect of exposed soil. The formula of the second version of the index is the following:

$$\begin{aligned}\mathrm{MSAVI2} &= \frac{2 \cdot \mathrm{NIR_1} + 1 - \sqrt{(2 \cdot \mathrm{NIR_1} + 1)^2 - 8 \cdot (\mathrm{NIR_1} - \mathrm{Red})}}{2} \\ &= \frac{2 \cdot \mathrm{B8} + 1 - \sqrt{(2 \cdot \mathrm{B8} + 1)^2 - 8 \cdot (\mathrm{B8} - \mathrm{B4})}}{2}\end{aligned}$$

It can be noticed how here there is no need to manually choose L because it is automatically determined by the mathematical structure. The presence of the square root introduces a dynamic correction term that varies with the values of NIR and Red, making the index more adaptable to different land cover conditions.

ARI (Anthocyanin Reflectance Index)
The ARI index estimates the content of anthocyanins in vegetation. This is important because plants increase anthocyanin production under stressful conditions, such as intense light, drought, nutrient deficiency or pathogen attack. Therefore ARI is useful for monitoring plant health and identifying environmental stresses.

$$\mathrm{ARI} = \frac{1}{\mathrm{Green}} - \frac{1}{\mathrm{Red\ Edge_1}} = \frac{1}{\mathrm{B3}} - \frac{1}{B5}$$

NDWI (Normalized Difference Water Index)
It is sensitive to changes in water content of vegetative and it is able to detect subtle changes in water content of the water bodies. It ranges from -1 to 1.

$$\mathrm{NDWI} = \frac{\mathrm{Green} - \mathrm{NIR_1}}{\mathrm{Green} + \mathrm{NIR_1}} = \frac{\mathrm{B3} - \mathrm{B8}}{\mathrm{B3} + \mathrm{B8}}$$

3.2.1 Dataset filling for spectral indexes

After calculating the spectral indices for all the pixels in the Sentinel-2 images that were part of the flood hazard maps, the following steps were taken:

- for all pixels that were part of the same Voronoi cell was taken as the average, so that there was a single spectral index value for each Voronoi cell;

- having only 198 images many dates in the dataset would have been left with missing data. Therefore, it was opted to associate these dates with the latest available average spectral index values for each Voronoi cell.

3.3 Indexes based on SAR images

Some useful indices can also be calculated for SAR images to capture some of the city's environmental and urban features. However, before proceeding to explain these indices, it is necessary to correct the images from the distortions explained in the section about SAR images in Chapter 2. In particular, in the case of SAR images, one does not have to worry about cloud cover, and thus many images that would have been discarded due to the presence of too many clouds in this case are retained. Thus, 956 images were downloaded, totaling 1.61 TB of memory. All these images were preprocessed using SNAP, a software provided by the European Space Agency (ESA). In order to use such software on python and automate the process, it is necessary to download the software and create a virtual environment with the specific package "esa_snappy," found within the software itself. The preprocessing operations written in the order in which they were performed are:

1. Apply orbit file: the so called "orbit file" is applied to the image, This operation is crucial in order to georeference the image for the next operations. In particular, this file is already among the auxiliary files of the image;

2. Thermal noise removal: this operation is carried out through the use of an auxiliary file in the SAR product. In that auxiliary file, there is an estimate of thermal noise as range and azimuth change, for each pixel in the image;

3. Radiometric calibration: this is also done through the use of an auxiliary image file. In that auxiliary file, different information is saved for different types of radiometric correction. In order to take the backscatter coefficient the radiometric correction must be performed taking the σ^0 (**sigma nought**) coefficient;

4. Speckle noise filtration:Using a specific filter contained in the esa_snappy package, the speckle noise is removed;

5. Terrain correction: thanks to the inclusion of the orbit file, the image is georeferenced. This allows for the correction of geometric distortions such as foreshortening, layover and shadowing. Specifically, a Digital Elevation Model (DEM) can be utilized. In fact, each pixel within the DEM contains its absolute elevation as an attribute. The orbit file, therefore, provides information on which DEM to use, enabling the correction of geometric distortions that are directly related to elevation. Once the images were preprocessed and knowing that each of them has the backscatter coefficient σ^0 for both VV and VH polarization, the indices that will now be presented were calculated.

RVI (Radar Vegetation Index)

RVI is an index that measures the amount and density of vegetation in an area. In particular, it is between 0 and 1 and the higher it is, the more vegetation the area has.

$$\text{RVI} = \frac{4\sigma^0_{VH}}{\sigma^0_{VV} + \sigma^0_{VH}}$$

NDPI (Normalized Difference Polarization Index)
NDPI measures polarization differences between co-polarized radar signals and helps distinguish features such as surface roughness and orientation of structures. Specifically, it is between -1 and 1 and if it is high then it indicates that there are mostly horizontally oriented surfaces in the area in question, otherwise it indicates mostly vertically oriented surfaces.

$$\text{NDPI} = \frac{\sigma^0_{VV} - \sigma^0_{VH}}{\sigma^0_{VV} + \sigma^0_{VH}}$$

CPR (Cross Polarization Ratio)
CPR measures the ratio of cross-polarized to co-polarized backscatter intensity. In particular, high CPR indicates mostly complex or rough surfaces with strong multiple scattering (e.g., trees or rough terrain). In contrast, a low CPR indicates smooth or specular surfaces (e.g., still water or asphalt).

$$\text{CPR} = \frac{\sigma^0_{VH}}{\sigma^0_{VV}}$$

3.3.1 Dataset filling for SAR indexes

Similarly to multispectral images, after calculating the SAR indices for all the pixels in the Sentinel-1 images that were part of the flood hazard maps, the following steps were taken:

- for all pixels that were part of the same Voronoi cell was taken as the average, so that there was a single SAR index value for each Voronoi cell;
- Using the same approach of spectral indices, it was opted to associate the missing dates with the latest available average SAR index values for each Voronoi cell.

3.4 Weather data

Another important aspect for flood modelling is weather. In particular, it will be considered through the use of the ERA5 dataset. ERA5 is a reanalysis dataset developed by the European Centre for Medium-Range Weather Forecasts (ECMWF). It integrates a vast amount of observations from satellites, aircraft, land-based and maritime sensors, combining them with atmospheric model data into a global and consistent set, applying the laws of physics. This approach relies on numerical weather prediction models to estimate atmospheric parameters, which are then combined with in-situ or satellite observations in an optimal physical manner.

ERA5 provides hourly data on numerous atmospheric, land-surface, and sea-state parameters, arranged on regular grids with a spatial resolution of 0.25° × 0.25° (approximately 28 km). Within the ERA5 dataset family, there is also ERA5-Land, a version focused on land surface data with a higher spatial resolution (0.1° × 0.1°, approximately 9 km) and without additional data assimilation.

The weather parameters that were used taken by ERA5-Land are:

- Total precipitation (tp): it is the water that accumulates on Earth's surface in liquid or frozen form, such as rain or snow, and it originates from two main processes. There are large-scale precipitation, produced by widespread weather systems like troughs and cold fronts, and convective precipitation, which occurs when warmer, less dense air near the ground rises through cooler, denser air above, creating precipitation through convection. It is expressed in meters;
- Total evaporation (te): accumulated amount of water that has evaporated from the Earth's surface, including a simplified representation of transpiration (from vegetation), into vapour in the air above. It is expressed in meters of water equivalent;
- Skin reservoir content (src): amount of water in the vegetation canopy and/or in a thin layer on the soil. It represents the amount of rain intercepted by foliage, and water from dew. It is expressed in meters of water equivalent;
- Skin temperature (skt): temperature of the surface of the Earth. It is expressed in Kelvin.

Since the variables in the ERA5-Land dataset are available on an hourly basis, it was necessary to calculate an aggregate value every 6 hours to ensure consistency with the temporal granularity established in the construction of the dataset. To this end, averaging was chosen, thus obtaining average values for skin temperature, skin reservoir content, total precipitation, and total evaporation. Then, each centroid was assigned the value corresponding to the closest point, considering the different dates and times in the dataset.

3.5 Past flood data

To enhance the analysis of floods, a novel feature was created in the dataset leveraging the existing river discharge data and the established discharge thresholds auxiliary files (as discussed in previous sections). This feature, referred to as the **cumulative number of floods**, represents the total number of floods recorded for each geographical point (latitude and longitude) up to the day prior to the current timestamp. So, for each latitude-longitude pair and for each timestamp in the dataset, the total number of floods that occurred from January 1, 2016 to the previous day was calculated. This cumulative count aggregates floods over all prior time intervals, providing historical flood information up to four time steps earlier (since the dataset has a temporal resolution of 6 hours). This feature provides a backward-looking measure that encapsulates the historical frequency of flood events, effectively serving as a proxy for past flood activity.

3.6 Topographical data

Topographical data were provided by the so-called EU-DEM (Digital Elevation Model over Europe). In particular, it covers the 39 member countries of the European environment agency and other cooperating countries, as it can be seen in Figure (3.4). The dataset for EU-DEM was produced by the Copernicus programme, which is managed by the European

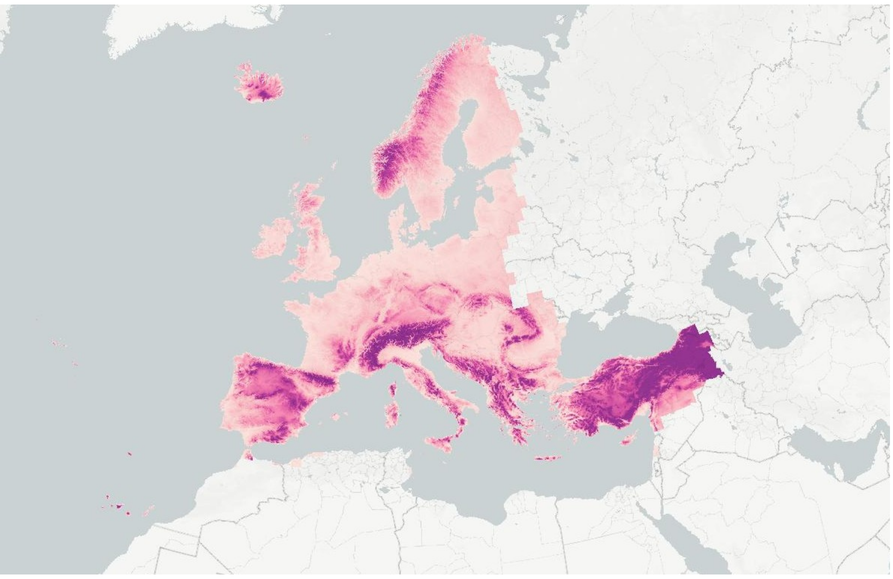

Figure 3.4. Render of EU-DEM elevation

Commission's Directorate-General for Enterprise and Industry. EU-DEM is a 3D raster dataset with elevations captured at 1 arc per second postings ($2.78 * 10^{-4}$ degrees), or about every 30 metres, while the unit of measure for the elevations is meter. In this study, elevation data provided by EU-DEM were aggregated to calculate an average elevation value associated with each Voronoi centroid coordinate. Since these data are static, the same elevation value was used for each centroid across all dates.

Chapter 4

Methodology

4.1 Feature standardization

Feature Scaling is a technique to standardize the independent variables of a dataset to a specific range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. In particular, the formula used for standardization is the following:

$$X' = \frac{X - \mu}{\sigma},$$

where X is the feature value and μ and σ are, respectively, the mean and the standard deviation of the feature values.

4.2 Techniques for handling imbalanced data

An imbalanced dataset refers to the situation where one class is significantly more prevalent than the other. The imbalance can lead to biased models that favour the majority class, underperforming on the other class. In order to solve this problem, different resampling techniques can be used, primarly **undersampling** and **oversampling**. In cases of extreme class imbalances, a combination of undersampling and oversampling can be used.

4.2.1 Undersampling

Undersampling involves the reduction of the number of instances in the majorrity class in order to rebalance the dataset. There are several undersampling techniques used in classification tasks to balance a dataset, the one that has been used in this thesis is the **Random Undersampling (RUS)**. Specifically, RUS involves randomly selecting a subset of the majority class while keeping all instances of the minority class. In this technique, a crucial parameter is **sampling strategy**, that is the proportion between the number of instances in the minority class and the number of instances in the majority class that is to be achieved. The advantages is the reduction of computational cost and the prevention of majority-class dominance. On the other hand, the disadvantages of this approach is the risk of losing useful information in the majority class.

4.2.2 Oversampling

Oversampling increases the number of instances in the minority class to balance the dataset. Similarly yo undersampling, it can be done by randomly duplicating existing minority class instances, but in a context such the one that is analyzed in this study it is quite riskful. In fact, if the minority class is very low in numbers, there is a high risk of overfitting due to repeated data points. So, the approach used in this thesis is a synthetic approach called **SMOTE** (Figure (4.1)), that stands for Synthetic Minority Over-Sampling Technique. Specifically, SMOTE generates new synthetic minority samples instead of duplicating existing ones. The steps to follow are:

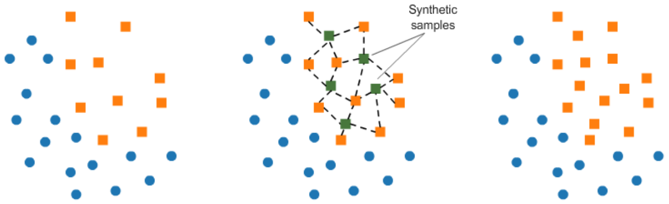

Figure 4.1. SMOTE algorithm

1. To select an instance of the minority class and to find its **k nearest neighbors** (a typical value is $k = 5$).
2. To choose one random neighbor and to generate a new sample along the line connecting them.
3. To repeat until the desired class balance is achieved.

Unlike random oversampling, using this algorithm can introduce diversity in the minority class and prevent overfitting.

4.2.3 Cost-sensitive learning approach

Cost-sensitive learning consists in adjusting the training process by assigning higher penalties to misclassifications of the minority class. Thanks to this technique, the model is able to pay greater attention to the minority class during the learning phase.

In fact, when dealing with imbalanced datasets, standard loss functions tend to be dominated by the majority class. Due to this fact, the models may achieve high accuracy while performing poorly on the minority class. To overcome this problem, it is introduced a weighting strategy where the loss incurred for misclassifying an instance from the minority class is increased.

This approach offers a complementary solution to data resampling techniques by directly modifying the training process rather than the dataset. In the thesis, models such as Random Forest, XGBoost, SVM, and MLP (that will be explained in the following section) were configured with class weighting strategies to counteract the adverse effects of class imbalance. This is useful to enhance the recognition of the minority class.

4.3 Classification models

Supervised classification represents one of the fundamental techniques in machine learning. This approach is based on the use of data labeled by a class value (1 in the case of positive class and 0 in the case of negative class) to build predictive models capable of assigning a class to new observations not seen during the training phase. The main goal is to learn a function $f(x)$ that associates the vector of independent variables x with a label y by exploiting an example data set known as a training set.

In the following sections, the supervised classification algorithms used in this thesis work will be presented, paying attention to the advantages and disadvantages of each.

4.3.1 SVM (Support Vector Machine)

A support vector machine (SVM) is a type of supervised machine learning algorithm designed to classify data by identifying the optimal line or hyperplane that separates different classes in an N-dimensional space, maximizing the margin between them. SVMs were introduced in the 1990s by Vladimir N. Vapnik and his collaborators. This kind of algorithm works by identifying the hyperplane that best separates two classes while maximizing the margin, that is the distance between the nearest data points from each class and the hyperplane. The dimensionality of the input data determines whether this hyperplane is represented as a line in 2D space, a plane in 3D space, or a hyperplane in higher dimensions. Since there can be multiple possible hyperplanes to separate classes, SVM selects the one that maximizes the margin, leading to better generalization and improved prediction accuracy on new data. The points that lie closest to the hyperplane, which influence its position and orientation, are called support vectors. For a better understanding, in Figure (4.2) SVM is represented in the 2-dimensional case.

From a mathematical point of view, in the SVM algorithm the data points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T$ must be divided in the two classes (the classes y_i, where y_i can be 0 or 1), maximizing the amplitude of the separation margin. Assuming the data are linearly separable and denoting by $\Pi_{\boldsymbol{w},b}$ a generic hyperplane in $\mathbb{R}^n$ defined by the normal vector $\boldsymbol{w} \in \mathbb{R}^n$ and the parameter $b \in \mathbb{R}$, i.e., $\Pi_{\boldsymbol{w},b} := \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{w}^\top \boldsymbol{x} + b = 0\}$,, the optimization problem for finding the separating hyperplane can be formulated as follows:

$$\begin{cases} \min_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} \\ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 \,, \quad \forall\ i = 1, \dots, T \end{cases} \tag{4.1}$$

where T indicates the size of the training set.

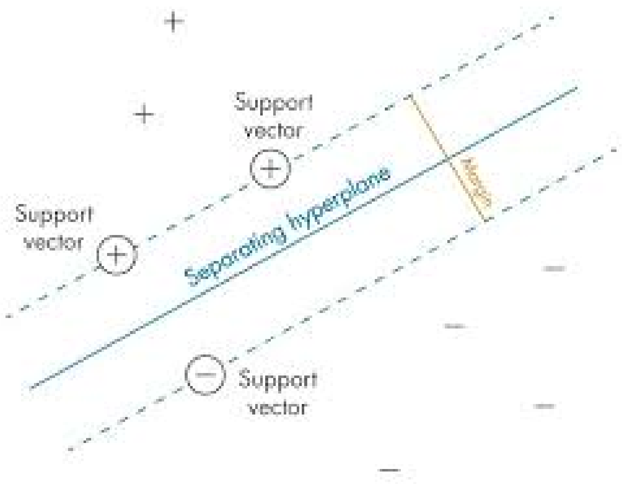

Figure 4.2. SVM in the 2-dimensional case

By using the duality theory for optimization problems [11], the minimization problem for SVMs described in (4.1) is solved by considering its dual problem:

$$\begin{cases} \min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^\top Q\boldsymbol{\alpha} - \sum_{i=1}^{T} \alpha_i \\ \sum_{i=1}^{T} \alpha_i y_i = 0 \\ \alpha_i \geq 0\,, \quad \forall\ i = 1, \dots, T \end{cases}, \tag{4.2}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^T$ and the matrix $Q \in \mathbb{R}^{T\times T}$ is defined as:

$$Q = (q_{i,j}) = \left(y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j\right) \quad i, j = 1, \dots, T\,. \tag{4.3}$$

Given the solution $\boldsymbol{\alpha}^*$ of the dual problem, it determines the solution $(\boldsymbol{w}^*, b^*)$ of the primal problem. In particular, the following observations are important:

1. If $\boldsymbol{x}_i$ is NOT a support vector for the optimal separating hyperplane $\Pi \boldsymbol{w}^*, b^*$, then the i-th element of $\boldsymbol{\alpha}^*$ is zero, i.e., $\alpha_i^* = 0$;

2. If the i-th element of $\boldsymbol{\alpha}^*$ is non-zero, i.e., $\alpha_i^* \neq 0$, then $\boldsymbol{x}_i$ is a support vector for the optimal separating hyperplane $\Pi \boldsymbol{w}^*, b^*$.

Two classes are not always perfectly linearly separable. Due to noise in the data, it may happen that classes that are theoretically linearly separable are not so in practice. In such cases, it may be necessary to allow for data points that are closer to the hyperplane than the margin (or, in the worst case, are on the wrong side of the hyperplane). The model that allows for such classification is called the Soft Margin SVM, which relaxes the Hard Margin SVM (the model described so far) to achieve a more flexible form. Thus, it becomes necessary to introduce a certain tolerance for margin violations for the vectors $\boldsymbol{x}_i$.

This "relaxation" of the margin conditions translates into the introduction of slack variables ξ_i in the primal problem, associated with a cost C for each margin violation.

The primal problem becomes:

$$\begin{cases} \min_{\boldsymbol{w}} \frac{1}{2}\left(\boldsymbol{w}^\top \boldsymbol{w} + C\sum_{i=1}^{T} \xi_i^2\right) & \\ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i\,, & \forall\, i = 1, \ldots, T \\ \xi_i \geq 0\,, & \forall\, i = 1, \ldots, T \end{cases}, \tag{4.4}$$

The parameter $C \in \mathbb{R}^+$ is a regularization hyperparameter:

- $C \to 0$ increases the "softness" of the margin, allowing the vectors $\boldsymbol{x}_i$ to violate it without bound;
- $C \to +\infty$ increases the "hardness" of the margin, allowing the vectors $\boldsymbol{x}_i$ to violate it only imperceptibly.

The SVM algorithm is widely applied in machine learning due to its ability to handle both linear and nonlinear classification problems. In fact, in many real-world problems data points are not linearly separable in their original feature space. However, if they are mapped to a higher-dimensional space, we may find that the data becomes linearly separable.

This transformation is typically achieved using a feature map:

$$\Phi : \mathbb{R}^n \to \mathbb{R}^m,$$

where $m > n$. The goal is to apply the SVM optimization in this higher-dimensional space. However, explicitly computing $\Phi(x)$ for every data point in the high-dimensional space can be computationally expensive, especially if m is very large. This is where the so-called "kernel trick" becomes valuable. The kernel trick avoids the need to explicitly compute the transformation $\Phi(x)$. Instead, it relies on a kernel function K that computes the inner product between two transformed data points directly in the high-dimensional space:

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j), \qquad i, j = 1, ..., T \tag{4.5}$$

By replacing all inner products in the SVM formulation defined in matrix Q in (4.3) with the expression in (4.5) the problem can be solved in the original space without ever explicitly performing the transformation Φ. This significantly reduces the computational complexity.

Kernel types

Several kernel functions are commonly used, depending on the nature of the data:

1. Linear Kernel:
$$K(x_i, x_j) = x_i^T x_j.$$
This corresponds to no transformation and is used when the data is already (approximately) linearly separable.

2. Polynomial Kernel:
$$K(x_i, x_j) = (x_i^T x_j + c)^d.$$
This maps the data to a higher-dimensional space with polynomial features. The degree d controls the flexibility of the decision boundary.

3. Radial Basis Function (RBF) or Gaussian Kernel:
$$K(x_i, x_j) = exp\Big(-\frac{||x_i - x_j||^2}{2\sigma^2}\Big),$$
where σ is a parameter that controls the amplitude of the Gaussian function. This kernel is widely used for problems with complex, non-linear decision boundaries.

4. Sigmoid Kernel:
$$K(x_i, x_j) = tanh(\alpha x_i^T x_j + c),$$
where $\alpha > 0$ is a parameter that controls the amplitude of the sigmoid function.

4.3.2 MLP classifier

The Multilayer Perceptron (MLP) is an artificial neural network that is used in supervised learning tasks as binary classification problems (Figure (4.3). It is a feedforward network, meaning that data flows through the network from the input layer to the output layer, without any feedback connections. The MLP is composed of multiple layers of neurons, including an input layer, one or more hidden layers, and an output layer. It is a fully connected network, meaning that each neuron in a given layer is connected to every neuron in the subsequent layer. The input layer of the MLP receives the features from the dataset, and specifically each feature corresponds to one neuron in the input layer. Regarding the

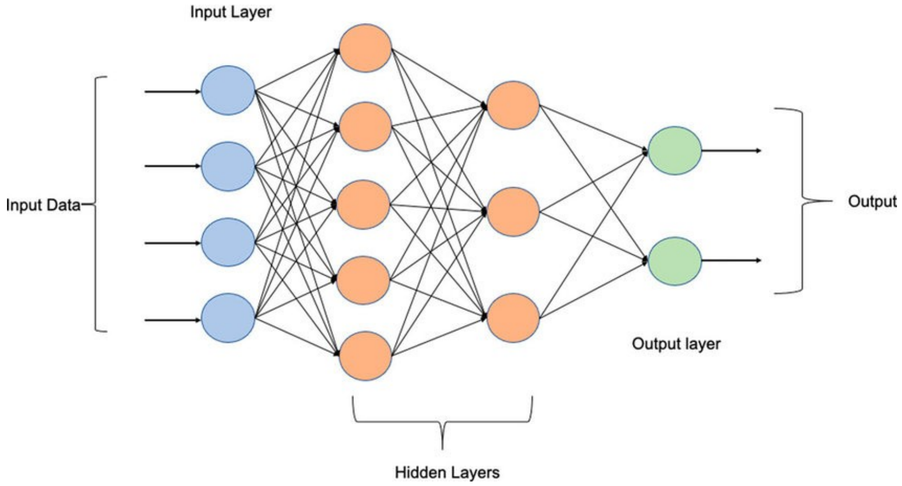

Figure 4.3. MLP architecture

hidden layers, they are responsible for learning complex connections of the data. Each neuron in a hidden layer calculates a weighted sum of its inputs, adds a bias term, and then applies a non-linear activation function. This non-linearity is crucial as it allows the MLP to model complex relationships in the data. In mathematical formulas, this is done by computing:

$$f(z) = f\biggl(\sum_{i=1}^{n} w_i x_i + b\biggr),$$

where f is the activation function, x_i is one of the inputs, w_i is one of the weights and b is the bias term. The activation functions that can be used are:

- ReLU function:
  $$f(z) = \max\{0, z\}$$
  It is the most common activation function used in the MLP.
- Identity function:
  $$f(z) = z.$$
- Hyperbolic tanh function:
  $$f(z) = tanh(z)$$
- Softmax function:
  $$f(z) = \frac{1}{1 + e^{-z}}.$$
  This function is used in the output layer (the last layer) to produce a single probability value, which is interpreted as the likelihood of belonging to the positive class.

The training process of an MLP involves:

1. Forward propagation: during this phase, input data flows through the network, and the output is calculated.
2. Loss computation: a loss function is defined, and it measures the difference between the predicted and actual labels. For binary classification, the binary cross-entropy loss is commonly used.
   $$\text{Cross Entropy Loss} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i}),$$
   where y_i is the true label of the i-th sample, $\hat{y_i}$ is the predicted label for the i-th sample and n is the number of samples.
3. Backpropagation: it is employed to calculate the gradients of the loss function with respect to the network's weights and biases, in order to minimize this loss function. These gradients indicate the direction in which the weights should be adjusted to reduce the loss. An optimization algorithm, such as Stochastic Gradient Descent (SGD) or Adam, is used to update the weights iteratively.

4. Weight updates: the update rule for a weight w is:

$$w \leftarrow w - \eta \frac{\partial \mathcal{L}}{\partial w},$$

   where η is the learning rate, controlling the step size of the update, while $\mathcal{L}$ is the loss function

The MLP is particularly suited for classification tasks because it can learn complex patterns in the data through its multiple layers and non-linear activation functions. However, it is sensitive to hyperparameters, such as the number of layers, the number of neurons in each layer, and the learning rate. These parameters must be carefully tuned to achieve optimal performance.

4.3.3 Ensemble methods

The aim of ensemble methods is to improve the performance of the predictive models by composing multiple base models. The main idea is that collecting diverse models can outperform any individual model. Specifically, ensemble methods leverage the concept of "wisdom of the crowd", where aggregating decisions from multiple sources yields more robust and accurate prediction, reducing errors due to **bias**, **variance** and **noise**. The three sources of errors just mentioned will now be analyzed:

1. Bias: it refers to errors concerning simplistic assumptions in the model, that bring to a low capability to capture the complexity of the data in certain situations.

2. Variance: it measures the sensitivity of a model to fluctuations in the training data. High-variance models may overfit the training data and generalize poorly to unseen data.

3. Noise: it represents the error caused by random variations in the data, such as measurement errors in the process being modeled.

In the world of machine learning, there are several approaches to implement ensemble methods. The two types used in this thesis are the following:

- Bagging (Bootstrap aggregation): it reduces the variance by training multiple models on different bootstrap samples (i.e. randomly drawn subsets with replacement) of the training data. Subsequently, predictions from single models are aggregated using averaging for regression tasks or voting for classification tasks. One example of this method is **Random Forest**, which combines multiple decision trees trained on bootstrap samples with random feature selection.

- Boosting: it reduces bias by sequentially training models, where each model focuses on correcting the errors of its predecessor. One example of this kind of methods is **XGBoost**, trained with using decision trees as single models.

Decision trees

Decision trees (Figure (4.4)) are a supervised learning algorithm used for both classification and regression tasks. While these models are not directly used in this thesis, it is important to understand them because they are the building blocks for the two ensemble methods used, Random forests and XGBoost. Specifically, the algorithm recursively partitions the dataset into subsets based on feature values to create hierarchy of decisions. At the core of a decision tree is the **root node**, which represents the entire dataset and is responsible for the first split. This split is determined by selecting the feature that best separates the target variable. As the tree grows, for each **decision node** a decision based on a specific feature value is carried out, effectively dividing the data into smaller subsets. Finally, in the **leaf nodes** there are the terminal points of the tree and they contained the predicted values or classes.

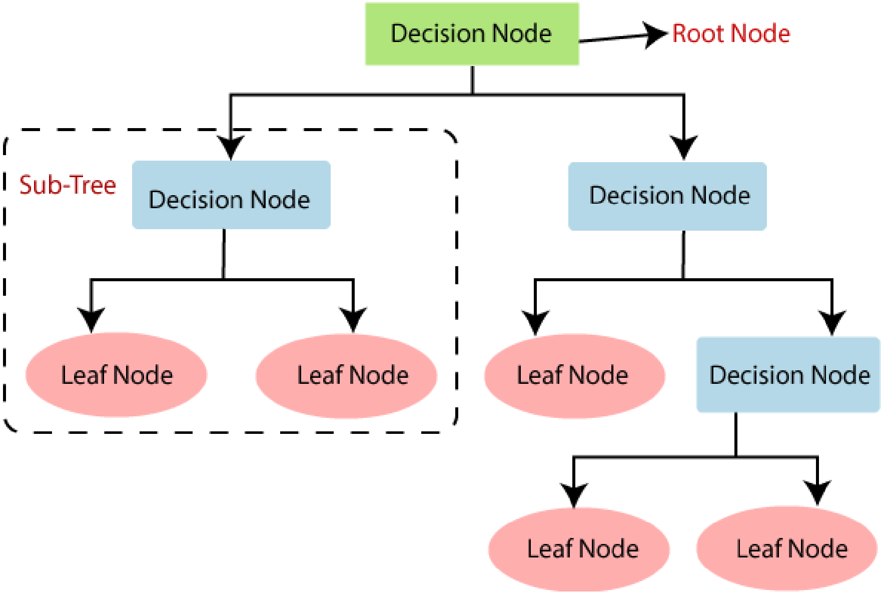

Figure 4.4. Decision tree

Decision trees are highly interpretable and provide a clear visual representation of the decision-making process. They are versatile, capable of handling both numerical and categorical data, and require minimal preprocessing, such as feature scaling. However, decision trees are prone to overfitting, especially when they grow too deep and start capturing noise instead of general patterns. They can also be unstable, as small changes in the training data can lead to significant alterations in the tree structure. Moreover, their standalone performance may not be optimal due to high variance or bias.

Random Forests

Random Forests (Figure (4.5) are and ensemble learning method constructed upon the simple model of decision trees. The method exploits two techniques:

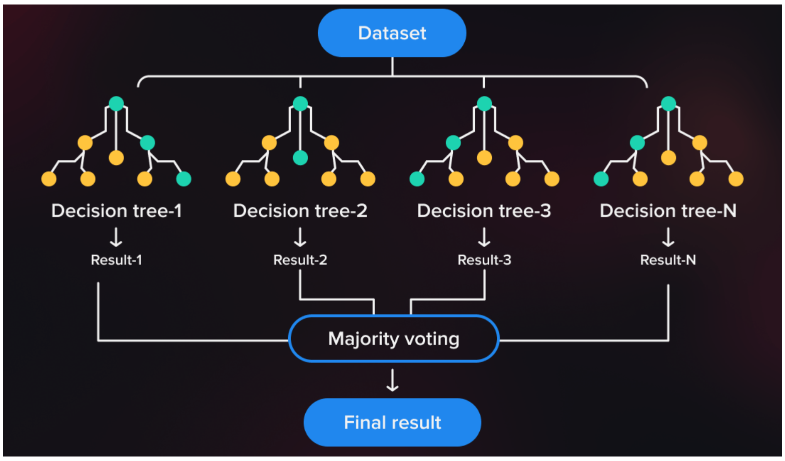

Figure 4.5. Random Forest

1. Bootstrap aggregation: as anticipated in the previous section, each tree is trained on a random sample of the training data, drawn with replacement. This guarantees that different trees are trained and exposed to different subsets of the data.

2. Feature Randomness: at every split in a tree, it is considered only a subset of features for splitting. This technique decorrelates the trees and prevents overfitting.

There are some hyperparameters to consider when creating the random forest. The most important ones are given below:

- Number of estimators: this parameter specifies the number of trees in the forest. Increasing the number of trees generally reduces variance and improves the performances. However, using a number of trees too large can bring to high computational costs.

- Max depth: this parameter defines the maximum depth of each tree in the forest. If it is too low the model can underfit, failing to capture sufficient patterns in the data. On the other hand, if the value is too high the model may overfit.

- Minimum samples to split: it controls the minimum number of samples required to split a decision node.

- Minimum samples to form a leaf node.

In a binary classification context, Random Forests can be useful to have in output class probabilities in addition to predicted class labels. Specifically, the probability of an instance belonging to a specific class is calculated as the proportion of trees in the forest that predict that class, following a similar approach to that proposed in the EFAS ERIC model.

Another advantage of using Random Forest is their ability to compute the so-called **feature importance**, which provides insights into the relative contribution of each feature to the model's prediction. This can be exploited in order to discharge the features that are less useful to make prediction in the model.

XGBoost

XGBoost (eXtreme Gradient Boosting) is an ensemble learning method that builds upon the principles of gradient boosting. By optimizing both speed and accuracy, XGBoost is one of the most popular machine learning models for classification and regression tasks.

Specifically, XGBoost (4.6) builds a series of decision trees sequentially. This means that in XGBoost every tree is created to correct the mistakes of the previous ones. The boosting process focuses on minimizing a specific loss function by adding new models that predict the residuals (errors) of the previous models. Over successive iterations, the ensemble converges to a highly accurate prediction. The most important peculiarities of XGBoost are the following facts:

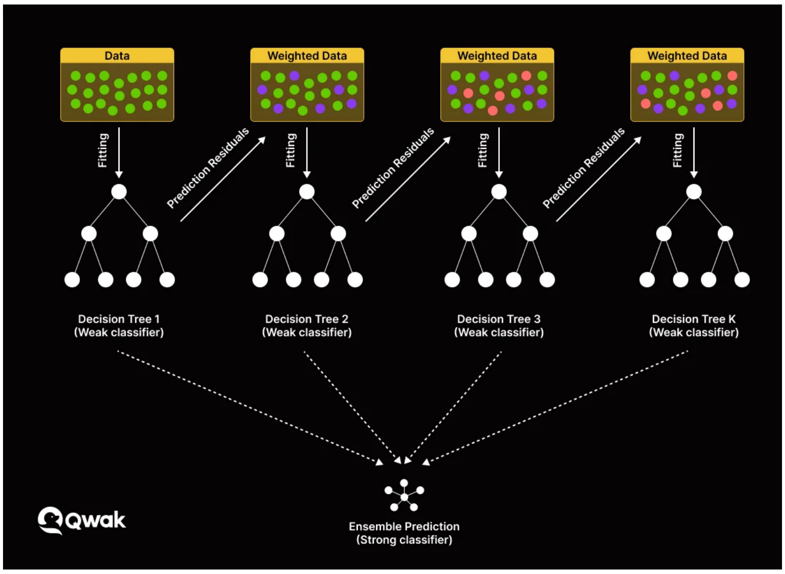

Figure 4.6. XGBoost

- XGBoost uses both the first derivative (gradient) and the second derivative (Hessian) of the loss function for more precise updates.
- it incorportates L1 and L2 regularization to control the complexity of the model and prevent overfitting.

- an important fact is about **shrinkage**. Specifically, it concerns a learning parameter that scales the contribution of each tree, ensuring that the model learns gradually and avoiding overfitting.

The iterative nature of boosting and the use of residuals highlight why XGBoost is a boosting method: each iteration builds upon and improves the predictions of the previous ones, focusing on areas where the model performs poorly.

There are some important hyperparameters to consider when creating the XGBoost model. The most important ones are given below:

- Number of estimators: this parameter specifies the number of trees in the model. Increasing the number of trees generally reduces variance and improves the performances. However, using a number of trees too large can bring to high computational costs.
- Max depth: this parameter defines the maximum depth of each tree in the XGBoost model. If it is too low the model can underfit, failing to capture sufficient patterns in the data. On the other hand, if the value is too high the model may overfit.
- Learning rate: this hyperparameter is about shrinkage. Specifically, it controls the contribution of each tree to the overall model. Smaller values slow down the learning process, allowing for more precise adjustments.

4.4 Performance measures

A metric is a function that gives information about the model performances. Specifically, it compares the predicted class label to the expected class label. In a binary classification task, first of all the confusion matrix must be defined. It is a fundamental tool for understanding the performance of a binary classifier. In this confusion matrix (Figure (4.7)), a tabular representation that compares the predicted labels with the true labels of a dataset is created. The key components of this tabular representations are:

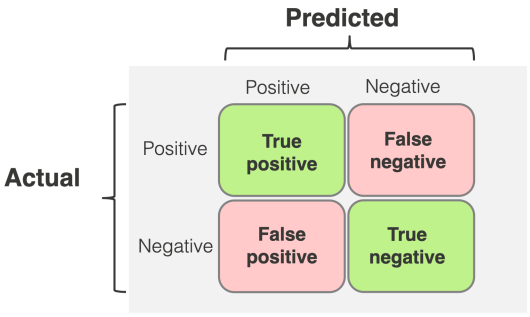

Figure 4.7. Confusion matrix for a binary task

- True Positive (TP): the number of instances where the model correctly predicted the positive class.
- True Negative (TN): the number of instances where the model correctly predicted the negative class.
- False Positive (FP): the number of instances where the model incorrectly predicted the positive class, while the true label was negative.
- False Negative (FN): the number of instances where the model incorrectly predicted the negative class, while the true label was positive.

Using these quantities, some useful performance metrics can be defined.

Accuracy is one of the simplest and most commonly used metrics in classification tasks. It measures the proportion of correct predictions out of the total number of instances. In a binary classification task, the formula is the following:

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

While accuracy provides a general sense of model performance, it can be misleading in scenarios where the dataset is imbalanced, as it may disproportionately reflect the majority class.

Precision is measured for each class. Taking the positive class as example, precision measures the proportion of true positive predictions among all instances predicted as positive. Specifically, the formula for the positive class is the following:

$$\text{Precision of positive class} = \frac{TP}{TP + FP}.$$

For the negative class, the idea is the same but with negative predictions.

Recall is also measured for each class. For the positive class, it quantifies the proportion of true positives identified out of all actual positive instances. The formula for the positive class is the following:

$$\text{Recall of positive class} = \frac{TP}{TP + FN}$$

For the negative class, the idea is the same but with negative predictions. High recall is crucial in scenarios where it is critical to minimize false negatives, such as identifying potential diseases or rare events.

F1-Score is the harmonic mean of precision and recall, providing a single metric that balances these two aspects. The F1-score is particularly useful when the dataset is imbalanced, as it considers both precision and recall in its calculation. Moreover, a parameter β to weight the importance of recall relative precision is introduced in this metric. The formula is:

$$\text{F1-Score} = (1 + \beta^2)\frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

So, when $\beta = 1$ equal importance is given to precision and recall, while using $\beta > 1$ recall is more important than precision and vice versa for $\beta < 1$.

4.5 ROC Curves

A Receiver Operating Characteristic (ROC) curve is a graphical representation of a classifier's performance. Two important quantities for understanding these curves are the true positive rate and the false positive rate. Specifically:

- True positive rate (TPR): it is defined as

  $$TPR = \frac{TP}{TP + FN}$$

  and it measures the percentage of positive data points that are correctly classified. On the ROC curves it is located on the Y-axis.

- False positive rate (FPR): is is defined as

  $$FPR = \frac{FP}{FP + TN}$$

  and it measures the percentage of negative data points that are misclassified by the model. On the ROC curves it is located on the X-axis.

Before explaining better the ROC curve, it is crucial to understand the concept of the discrimination threshold: specifically, many binary classifiers output a probability score indicating the likelihood that a given instance belongs to the positive class. At this point, the discrimination threshold comes into play. It is the cutoff value used to convert the predicted probability into a binary classification. For example, if the threshold is set at 0.5:

- An instance with a predicted probability ≥ 0.5 is classified as positive.
- An instance with a predicted probability < 0.5 is classified as negative.

By varying the discrimination threshold from 0 to 1, one can trace out the ROC curve, which illustrates the trade-off between TPR FPR across all possible thresholds. Specifically, looking at Figure (4.8), the point (0,0) corresponds to a very high discrimination threshold, where every instance is labelled with the negative class; on the other hand, the point $(1, 1)$ corresponds to a very low discrimination threshold, where every instance is labelled with the positive class.

A very important scalar metric is represented by the Area Under the ROC Curve (AUC-ROC). Specifically, it summarizes the overall performance of a classifier across all threshold settings. While the ROC curve provides a graphical representation of the trade-off between the TPR and the FPR while changing the discrimination threshold, the AUC-ROC condenses this information into a single numerical value.

Since the ROC curve is located in a unit square bounded by the vertices (0,0) and (1,1) the maximum possible area under the curve is 1. Specifically, an AUC-ROC of 1 indicates a perfect classifier that consistently ranks all positive instances higher than negative ones, regardless of the chosen discrimination threshold. Conversely, an AUC-ROC of 0.5 suggests that the classifier performs no better than random guessing, as

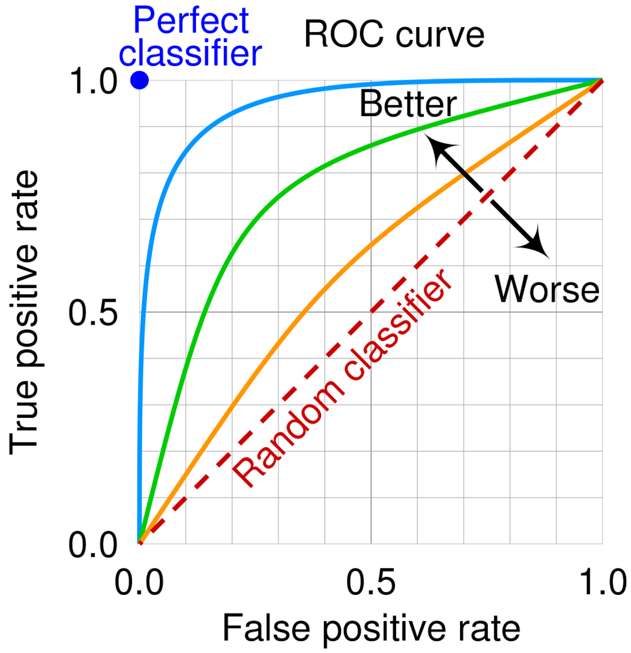

Figure 4.8. ROC curve

the ROC curve in this case would lie along the diagonal of the unit square. One possible interpretation is that the AUC-ROC can be interpreted in a probabilistic way: it represents the probability that a randomly selected positive instance will be assigned a higher score than a randomly selected negative instance. Thus, the higher the AUC-ROC value, the better the classifier is at distinguishing between the two classes.

While the ROC curve and the AUC-ROC metric provide useful information about the performances of the binary classifier, they may not be appropriate in cases where the dataset is highly imbalanced. In fact, in these scenarios the ROC curve and the AUC may seem good, even in situations in which only the majority class is predicted in the right way. In these cases, Precision-Recall (PR) curves and their associated area under the curve (AUC-PR) offer a more informative alternative.

A Precision-Recall curve (Figure (4.9)) is a graphical representation that illustrates the trade-off between recall (on the X-axis) and precision (on the Y-axis) across different classification thresholds.

The key difference between ROC and PR curves is that PR curves focus exclusively on the performance of the model with respect to the positive class, making them particularly useful when the positive class is rare. In fact, in imbalanced datasets a high false positive rate (FPR) may not impact in a significant way the ROC curve, as the large number of true negatives (TN) can dominate the denominator in the FPR calculation. However, in such cases, precision becomes a more meaningful metric, as it directly accounts for false positives. Finally, in a similar way to the AUC-ROC, the area under the Precision-Recall

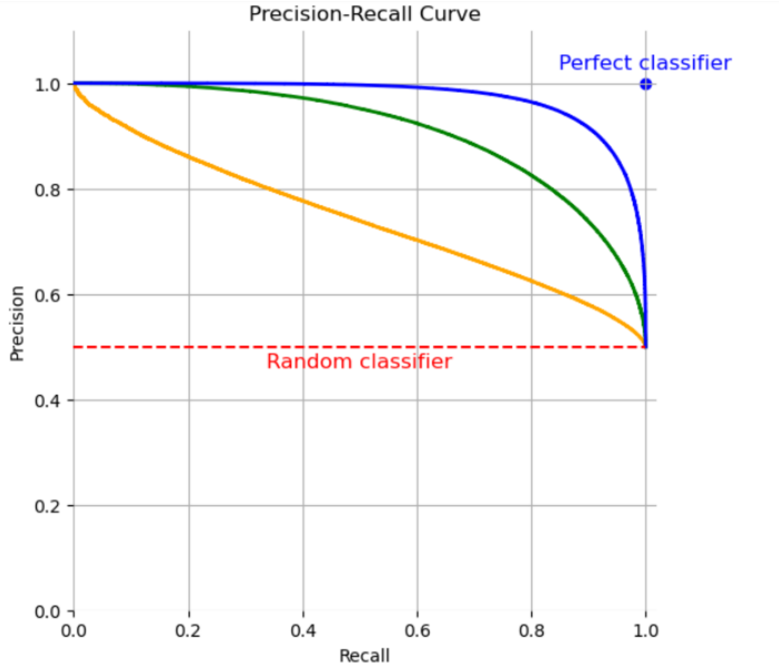

Figure 4.9. PR curve

curve (AUC-PR) serves as a single scalar metric to summarize classifier performance varying across different discrimination thresholds. Even in this case, due to the fact that the curve is bounded by the square of vertices (0,0) and $(1,1)$, the maximum AUC is still one for a perfect classifier, while a random classifier corresponds to a PR curve that is horizontal, with AUC equal to 0.5.

Chapter 5

Experimental results

In this chapter details on model implementation and experimental results of supervised classification algorithms will be presented. Specifically, all code was executed on AWS (Amazon Web Services) Sagemaker using Python. Alongside the standard libraries such as pandas, Numpy, Scipy, and Scikit-learn, they were utilized several geospatial libraries, such as Geopandas, Rasterio, Shapely and Xarray. These libraries were crucial, given the environmental and urban context of the problem. The chapter is organized as follows. Section 5.1 focuses on the choice of hyperparameters and the methodologies used for their optimization. Finally, Section 5.2 presents the classification results, comparing the performance of different models using various evaluation metrics.

5.1 Hyperparameters setting

As a first operation, the dataset was split into training set (80%) and test set (20%), in order to ensure a robust evaluation of the classification models. Due to the fact that the classification problem is imbalanced, a stratified sampling approach was adopted. Specifically, this method is adopted because it preserves the original class distribution in both subsets, preventing the minority class from being underrepresented in the training or test sets. This kind of operation is widely used technique in imbalanced classification problems, as it helps models learn from a representative distribution of the data while ensuring a fair evaluation during testing.

At this point, it is necessary to set the hyperparameters explained in Chapter 4 for every classification algorithm. In order to do this operation, the technique of **K-fold cross validation** (Figure (5.1)) is introduced. This technique is applied on the training set and it consists of dividing the training set in a number K of equal subsets or folds (in this thesis K has been chosen equal to 5). At this point there are K iterarions for the algorithm. For each of the K iterations, the model is trained on $K-1$ folds and validated on the remaining fold. This rotation ensures that each fold is used exactly once for validation. By systematically evaluating the model's performance across different hyperparameter combinations on each validation set, the best set of hyperparameters is identified. In every case a performance metric is chosen to evaluate the performances

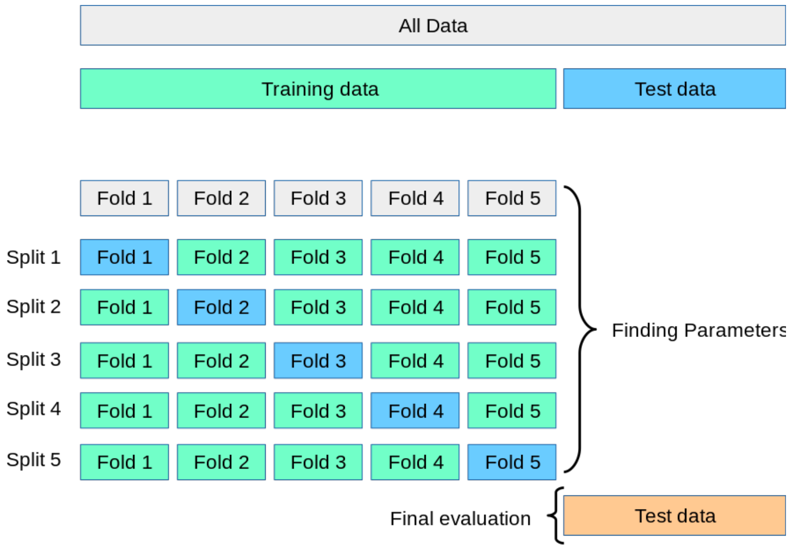

Figure 5.1. Cross validation process

of the models across different hyperparameter combinations. For this thesis, due to the imbalanced classes, the f1-score of the minority class has been chosen as the performance metric to be considered in the cross validation process. In fact, in imbalanced classification problems, accuracy can be a deceptive metric because a model might achieve high accuracy by simply predicting the majority class. On the other hand, the f1-score, which combines precision and recall, is more informative for this kind of context. So, this metric was chosen because it not only penalizes false positives and false negatives but also ensures that the final model maintains a balanced performance in both dimensions. F1-score is then computed for each iteration, and the average f1-score across all folds serves as the key metric for hyperparameter selection.

At this point, after the best hyperparameters were determined through K-fold cross validation on the training set, the final model was trained using these parameters on the entire training set. The final evaluation was then conducted on the test set, which was not involved in the cross validation process. This step was essential to ensure that the performance metrics, particularly the f1-score, accurately reflected the model's ability to generalize to unseen data (test set).

Finally, depending on the specific classification algorithm and its computation time, several combinations of hyperparameters has been tried. These hyperparameter combinations were validated on the dataset using 5-fold cross validation applying solely undersampling techniques as well as a combination of undersampling and oversampling techniques. Fortunately, both experiments yielded identical results, making the performance metrics of the classification models easily comparable regardless of whether only undersampling or both undersampling and oversampling were employed. With reference to the hyperparameters introduced in Chapter 4, the results of the choice of hyperparameters are listed below:

1. **Random forest**:
   - Maximum depth of trees: None;
   - Minimum samples for leaves: 1;
   - Minimum samples for split: 2;
   - Number of estimators: 100.
2. **SVM**:
   - Kernel: rbf;
   - C: 100.
3. **XGBoost**:
   - Maximum depth of trees: 9;
   - Number of estimators: 300.
   - Learning rate: 0.2
4. **MLP**:
   - Hidden layer sizes: (100,50);
   - Activation function: tanh;
   - α: 0.0001;
   - Learning rate: constant;
   - Maximum number of iterations: 200;
   - Solver: Adam.

5.2 Results of classification algorithms

The results obtained from the experiments with binary classification models will now be presented. Specifically, as discussed in previous chapters, the models that have been used are Random Forest, SVM, XGBoost and MLP, and every model has been trained using the cost-sensitive learning approach in order to deal the unbalancing of the dataset. In fact, the primary challenge of the dataset is the significant class imbalance, which necessitates careful handling during both training and evaluation phases. For this reason, the results will be presented dividing two different solutions for handling imbalanced data. The first one is the presentation of the results using undersampling techniques only, while the second one is the combination of undersampling and oversampling techniques. The third approach of applying only oversampling techniques has not been taken into account because it would have made the training set too large and thus training the models would have required too much computational effort.

By presenting the results from these two described approaches, this section aims to provide a comprehensive comparison of data balancing techniques in the context of imbalanced binary classification.

5.2.1 Undersampling results

Using random undersampling approach provides an advantage from the point of view of model run time and memory occupied because the model is trained on a smaller training set. In Table (5.1) are presented the models performances for the minority class only. This choice will be done for every result that will be presented, because due to the unbalance of the dataset the models have not difficulties in the prediction of the majority class. The results using the undersampling approach show high recall values (ranging from 0.96 to 0.99), which is particularly important for flood detection. In fact, this metric indicates that very few flood events are missed (i.e., a low number of false negatives). However, while recall remains high across models, precision and F1-score values vary considerably. For instance, the Random Forest model achieves a strong precision (0.89) and F1-score (0.93), whereas the SVM model exhibits very low precision (0.49) and a corresponding F1-score (0.65). One possible reason for this drop in precision and F1-score is the poor ability of the SVM model to handle outliers and extreme values, which in this case occur in the rare flood events that are part of the positive class. This could lead the model to overfit the rare positive examples, generating a decision threshold that favors positive classification even for negative samples. The other two models obtain quite good results. Specifically, XGBoost achieves a precision of 0.77 with a correspondig F1-score of 0.87, while MLP achieves a precision of almost 0.7 with a corresponding F1-score of 0.80. Finally, all the 4 models have a very high AUC regarding the Precision-Recall curve.

Model	**Recall**	**Precision**	**F1-score**	**AUC-PR**
Random Forest	0.98	0.89	0.93	0.98
SVM	0.97	0.49	0.65	0.85
XGBoost	0.99	0.77	0.87	0.98
MLP	0.96	0.68	0.80	0.94

Table 5.1. Performance metrics of binary classification models using undersampling approach

5.2.2 Combination of undersampling and oversampling results

In this section, the performance metrics obtained when combining undersampling with oversampling are presented. This hybrid approach is designed to leverage the computational benefits of undersampling while mitigating its potential drawbacks through oversampling. As detailed in Table (5.2), this combination yields a slight improvement in recall for all models from about 0.98 to 0.99 in most cases thereby enhancing the detection of flood events. Additionally, improvements in precision and F1-score for models such as XGBoost and SVM suggest that this strategy not only captures more positive instances but also refines the overall predictive balance.

When combining undersampling with oversampling, every model experiences a slight improvement in recall from approximately 0.98 to 0.99 for most models. Although the increase might seem marginal, in practical applications such as flood prediction, even a

Model	**Recall**	**Precision**	**F1-score**	**AUC-PR**
Random Forest	0.99	0.81	0.89	0.97
SVM	0.99	0.23	0.37	0.74
XGBoost	0.99	0.66	0.79	0.97
MLP	0.98	0.64	0.78	0.94

Table 5.2. Performance metrics of binary classification models using a combination of undersampling and oversampling approach

0.01 improvement in recall can lead to the detection of several additional flood events in a test set comprising hundreds of samples.

However, using oversampling adds noise in the data, and this leads to a slight worsening of precision and F1-score compared to the previous subsection. Specifically:

- XGBoost: Precision worsens from 0.77 to 0.66, and the correspondig F1-score worsens from 0.87 to 0.89.
- SVM: If precision was low before, now it is even worse. In fact, it passes from 0.49 to 0.23, conferming the hypothesis that this model is not suitable for this kind of problem.
- Random Forest: Even for this model the precision slightly worsens from 0.89 to 0.81.

Looking at the AUC for the Precision-Recall curves, it can be seen how this models have a very good performance, similarly to the previous case.

Overall, while undersampling alone already yields high recall—critical for ensuring that flood events are not overlooked—the combined strategy of undersampling and oversampling further enhances the model's recall itself. This improvement is significant for flood detection applications, where the cost of a missed event can be high. The trade-off between a slight reduction in precision and a marginal increase in recall is justified by the critical need to identify as many flood occurrences as possible. However, more models will now be presented with the aim of improving accuracy again, yet maintaining a recall like the one just seen.

5.2.3 Feature Selection and Model Refinement

Based on the experimental results presented in the previous sections, SVM and MLP will no longer be considered in this section due to their lower performances. So, the focus will be only on Random Forest and XGBoost, as they have demonstrated superior predictive capabilities. An additional advantage of considering only these models is their ability to provide probabilistic predictions in a similar way of the ERIC model from EFAS. In fact, these models can estimate the probability of an area experiencing flooding within a given timeframe. This probabilistic output provides a confidence level associated with each prediction, enhancing risk assessment. This approach is particularly useful because, it was noted that in many cases where the classifier makes incorrect predictions, the confidence level is not so high, with probabilities not close to 1. This indicates that individual decision trees within the ensemble are less certain, and therefore, providing probability scores can be more informative than a simple binary output. By analyzing these probabilities, decision-makers can better assess the flood risk and do different actions based on the level of uncertainty of the models.

Considering Random Forest and XGBoost, in this subsection feature importance analysis will be leveraged for these models to enhance performance further. As explained in Chapter 4, feature importance is a crucial tool in machine learning that helps identify the most relevant features for a given predictive task. Specifically, by selecting only the features with an importance score greater than 0.02, more efficient models have been built with several advantages:

- Computational Efficiency: Reducing the number of input features decreases training time and memory usage, making the models faster and more scalable.
- Improved Generalization: Removing less relevant features helps prevent overfitting, leading to better performance on unseen data.
- Enhanced Interpretability: A model with fewer but meaningful features is easier to analyze and understand, aiding in explaining predictions.

Using the threshold of 0.02 as it was just presented, different subsets of features were selected based on their respective feature importance scores:

- Random Forest Features:
  1. Total precipitation;
  2. Mean temperature;
  3. Mean skin reservoir content;
  4. Mean evapotranspiration;
  5. Count of verified floods in the past;
  6. Mean SAR CPR Index;
  7. Mean SAR RVI Index.
- XGBoost:
  1. Total precipitation;
  2. Mean skin reservoir content;
  3. Count of verified floods in the past;
  4. Mean NDWI.

In Table (5.3) results are presented, divided as in the previous subsections based on the sampling techniques utilized.

As in the previous case, it can be noticed how the use of combined undersampling and oversampling approach led to a slight improvement in recall, which is beneficial for

Sampling Strategy	Model	Recall	Precision	F1-score	AUC-PR
Undersampling	Random Forest	0.98	0.93	0.96	0.97
	XGBoost	0.98	0.94	0.96	0.97
Undersampling + Oversampling	Random Forest	0.99	0.90	0.94	0.96
	XGBoost	0.99	0.80	0.88	0.94

Table 5.3. Best models using feature importance

identifying positive cases. On the other hand, the inclusion of synthetic data through SMOTE introduces some noise, leading to a decrease in precision compared to using only undersampling.

To provide a clear comparison, the Precision-Recall curves for the two best-performing models (Random Forest with undersampling and oversampling and XGBoost with only undersampling) are presented (Figure (5.2) and Figure (5.3)), showcasing their trade-offs in precision and recall.

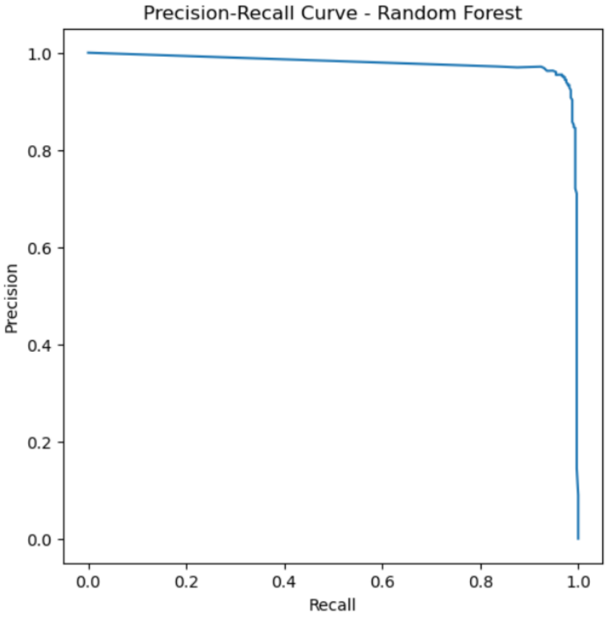

Figure 5.2. PR curve for Random Forest using undersampling and oversampling and feature importance

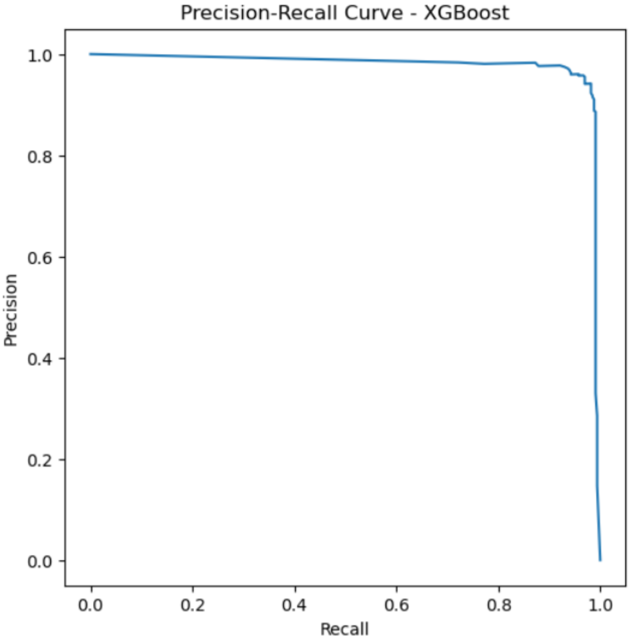

Figure 5.3. PR curve for XGBoost using undersampling and feature importance

5.2.4 Results using higher return period for floods

In this section, the flood prediction analysis by considering floods with higher return periods is presented. Specifically, the return period has been increased from 1.5 years to 10 years. The choice of a 10-year return period is based on data availability, as no past flood events beyond this threshold exist for training the model in the city of Turin. Using a dataset with such high return periods further exacerbates the class imbalance problem. While the dataset was already highly imbalanced in previous analyses, this scenario results in an even more extreme imbalance, with only about fifty instances belonging to the positive class. However, from a logical perspective, the separation between the two classes should improve since the flood events considered are more extreme. This suggests that the associated features may be more distinctly different between flooded and non-flooded cases, potentially leading to a clearer decision boundary for classification. Despite the limited number of positive instances, we proceeded with training the models. This decision was supported by the results obtained using a return period of 1.5 years, where class distributions were less imbalanced, and the models demonstrated validity. However, the models were not trained directly on a dataset with a 10-year return period, the reliability of the models could have been questioned due to the very small test set, so the passage of the return period of 1.5 was necessary: by first validating the models on a more balanced dataset, we ensured that any observed trends with higher return periods were built upon a foundation of reliable performance.

In Table (5.4) results for the models with return period of 10 years are shown. In this case, similarly to previous section, only Random Forest, XGBoost classifiers are shown, because these kind of method are the ones that offer the best predictions, as it was seen. Moreover, also in this section the results are divided according to the sampling technique that has been used. Again, the same pattern present of the 1.5-year return period models can be seen. In fact, there is a very high recall, a symptom of the fact that the models succeed well in identifying true positives from false negatives. In particular, if in the models with 1.5-year return time there was still a few rare cases where false negatives were present, having more extreme flood situations (and therefore also more extreme associated features such as precipitation, humidity, spectral indices and SAR indices) this happens less and less by increasing the return time, and it is a promising result in case we want to use the model to study even more catastrophic events. Also, similar to the results

Sampling Strategy	**Model**	**Recall**	**Precision**	**F1-score**
Undersampling	Random Forest	1.00	0.90	0.95
	XGBoost	1.00	0.69	0.82
Undersampling + Oversampling	Random Forest	1.00	0.90	0.95
	XGBoost	1.00	0.65	0.79

Table 5.4. Random forest and XGBoost for return periods of 10 years with different sampling techniques

commented on in the previous sections, the use of overssampling improves recall (or leaves it the same in this case since it was already maximal with the use of undersampling alone) but worsens performance for precision because of the noise introduced by SMOTE. Thus, in this case, the best model both from the point of view of performance and from the point of view of computational time and memory occupied turns out to be the Random Forest with only the use of undersampling as the sampling strategy. To appreciate again the importance of feature importance, this last model was again trained by considering only features with feature importance ≥ 0.07:

- Total precipitation;
- Mean skin reservoir content;
- Mean SAR VV backscatter value;
- Mean evapotranspiration;
- Count of verified floods in the past;
- Mean ARI index.

Using these features and training again the model, performances are perfect, with 0 cases of false negatives and false positives. Of course, given the dimension of the test set of just 10 elements this cannot be considered as the results obtained in the 1.5-year return period models, but it is another prove of the fact that the followed approach is valid for this kind of study.

Chapter 6

Conclusions

This study has explored the use of various data sources and machine learning techniques for flood prediction in the city of Turin. The research required integrating satellite data from Sentinel-1 and Sentinel-2 to compute SAR and multispectral indexes, meteorological data from ERA5-Land, topographic data from Digital Elevation Models (DEM), and river discharge information along with flood thresholds from the EFAS dataset. Additionally, flood hazard maps for the city of Turin were incorporated, and Voronoi areas populated with the data just discussed were created for four key rivers: Po, Dora Riparia, Sangone, and Stura.

Given the nature of the data, binary classification models were developed to handle the severe class imbalance using cost-sensitive learning, undersampling, and oversampling techniques. The models achieved excellent results, demonstrating strong predictive performance and providing an encouraging foundation for the continuation of this research and project.

6.1 Limitations

The main limitations of this study stem from the availability and frequency of satellite data. While SAR and multispectral indices have proven to be valuable tools for flood prediction, the dataset's 6-hour temporal resolution would ideally require more frequent index updates, which is not feasible due to the limitations of satellite imagery acquisition.

Sentinel-2 has an average revisit time of approximately 6 days, which significantly limits the frequency of multispectral index updates. The situation is even more challenging when considering cloud coverage restrictions: by selecting only images with less than 10% cloud cover, the effective revisit time increases to around 12 days on average. Moreover, this revisit time is not uniform throughout the year during summer, when cloud cover is lower, more images are available, whereas in colder months, fewer images are acquired, leading to longer gaps between observations. This irregularity poses a challenge in maintaining a consistent and timely dataset for flood prediction modeling.

6.2 Next steps and further works

There are some steps to be followed in order to continue this work and improve it for real-world application:

- First and foremost, discussions with the Civil Protection of Turin highlighted the potential interest in extending this type of study to the hilly area of Turin, located in the eastern part of the city. This area is characterized by numerous secondary streams that frequently overflow, thereby increasing the risk of landslides. In general, it was suggested that a study encompassing these secondary streams could be even more valuable.

- A possible further work regards SAR images. In fact, Sentinel-1 images are good but it is possible to use other SAR images like COSMO-SkyMed ones from ASI (Agenzia Spaziale Italiana). In fact, these SAR images provide a better spatial resolution (3m x 3m instead of 10m x 10m), and they are for free for Italian users. Moreover, using the combination of Sentinel-1 and COSMO-SkyMed can improve also the temporal resolution, with more images available for the study. Finally, it can be explored also the use of SAR images that are not freely available, in order to construct SAR indexes that are more precise and more frequently updated.

- For the same reason, it can be exploited the possibility of using different multispectral images in addiction to the Sentinel-2 ones. Similarly to the previous point, this can bring to an improvement in temporal and spatial resolution, even if there is the complication about cloud coverage for this kind of images.

- For the meteorological data, one step is to use forecasted data. This of course will add uncertainty to the model, so the next step would be to treat this uncertainty in order to have good performances of the models. This improvement can bring to change the time of the prediction. In fact, for now an immaginary user could run the model and take a decision only 6 hours before the time that he is trying to focus on, because before of having all the data he has to wait until this time. Instead, using forecasted data can open to the possibility of predict the flood even before these 6 hours, of course with an amount of uncertainty that grows going back in time.

- A futher work is to use these models in other areas. In fact, there is no visibile limitation of using the approach used in Turin also for other cities, so it could be interesting to see the results for cities that are more usual to flood risk.

Moreover, in addiction to these more concrete steps, it can be analyzed the possibility to use model for anomaly detection, that can be more suitable for dataset with this umbalancing of the dataset. In fact, a first phase of model search and attempt was carried out but with poor results on performance, however, the type of dataset lends itself very much to the use of this type of modeling. In addiction to that, many other undersampling and oversampling techniques for binary classification problems will be considered and tested.

In conclusion, the proposed approach is able to provide good performances and good representation of the flood risk. Meteorological data are essential for this type of data, but satellite data like SAR or Multispectral provide useful information in understanding when weather conditions will result in flooding or not.

Bibliography

[1] ADBPO. Autorità di bacino distrettuale del fiume po. https://www.adbpo.it/.

[2] ADBPO. Aggiornamento e revisione delle mappe di pericolosità e del rischio di alluvione redatte ai sensi dell'art. 6 del d.lgs. 49/2010 attuativo della dir. 2007/60/ce – ii ciclo di gestione. 2020.

[3] Thomas P. Ager. *The Essentials of SAR.* 2023.

[4] Britannica. Spectroscopy. https://www.britannica.com/science/spectroscopy.

[5] ZHENG Ze-zhong CAO Yun-gang, YAN Li-juan. Extraction of information on geology hazard from multi-polarization sar images.

[6] Kim L. Boyer Cem Unsalan. *Multispectral Satellite Images Understanding.* Springer, 2011.

[7] CFI. Ensemble methods. https://corporatefinanceinstitute.com/resources/data-science/ensemble-methods/#:~:text=Ensemble%20methods%20are%20techniques%20that,accuracy%20of%20the%20results%20significantly.

[8] Jakob van Zyl Charles Elachi. *Introduction to the Physics and Techniques of Remote Sensing.* WILEY-INTERSCIENCE, 2006.

[9] Oscar A. Ishizawa Daniel B. Wright, Fernando Ramirez-Cortés. Methods in flood hazard and risk assessment. 2016.

[10] Ashesh Das. Oversampling to remove class imbalance using smote. *Medium*, 2019.

[11] Anetor Clement Dr. Rowland Jerry Ekeocha, Uzor Chukwunedum. The use of the duality principle to solve optimization problems. 2018.

[12] ESA. Rvi documentation. https://documentation.dataspace.copernicus.eu/APIs/openEO/openeo-community-examples/python/RVI/RVI.html.

[13] ESA. S1 products. https://sentiwiki.copernicus.eu/web/s1-products.

[14] ESA. S1 products. https://sentiwiki.copernicus.eu/web/s2-products.

[15] ESA. Sentinel-1. https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1/Introducing_Sentinel-1.

[16] ESA. Sentinel-2. https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1/Introducing_Sentinel-2.

[17] ESA. Sentinel-2 rs indices. https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/indexdb/.

[18] ECA Europe. Floods directive. 2018.

[19] Geeks for geeks. Feature engineering: Scaling, normalization, and standardization. https://www.geeksforgeeks.org/ml-feature-scaling-part-2/.

[20] Geopillole. Geometrie radar. https://www.nicoladeinnocentis.it/le-geometrie-radar.

[21] GSP. History of remote sensing. https://gsp.humboldt.edu/olm/Courses/GSP_216/online/lesson1/history.html.

[22] IBM. Machine learning topics. https://www.ibm.com/think/topics.

[23] ICC. Electromagnetic spectrum. https://www.icc.dur.ac.uk/~tt/Lectures/Galaxies/Images/Infrared/Windows/irwindows.html.

[24] ISPRA. Rapporto sulle condizioni di pericolosità da alluvione in italia e indicatori di rischio associati. 2021.

[25] George Joseph. *FUNDAMENTALS OF REMOTE SENSING.* Universities Press, 2005.

[26] LEGAMBIENTE. Alluvioni in italia: i nuovi dati dell'osservatorio città clima e gli interventi urgenti che servono al paese. https://www.legambiente.it/comunicati-stampa/alluvioni-in-italia-i-nuovi-dati-citta-clima-e-interventi-Urgenti/#:~:text=Preoccupante%20anche%20il%20dato%20complessivo,Marche%20e%206%20in%20Umbria.

[27] Farzin Shabani Mahyat Shafapour Tehrany, Lalit Kumar. A novel gis-based ensemble technique for flood susceptibility mapping using evidential belief function and support vector machine: Brisbane, australia. 2019.

[28] Machine Learning Mastery. How to use roc curves and precision-recall curves for classification in python. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/.

[29] Lily Lisa YEVUGAH Michael NYOAGBE, John AYER. Flood prediction using machine learning and gis. 2023.

[30] NASA. Remote sensing. https://www.earthdata.nasa.gov/learn/earth-observation-data-basics/remote-sensing.

[31] NOAA. Noaa. https://www.nssl.noaa.gov/education/svrwx101/floods/forecasting/.

[32] Government of Canada. Viewing geometry and spatial resolution. https://natural-resources.canada.ca/maps-tools-publications/satellite-elevation-air-photos/viewing-geometry-spatial-resolution.

[33] WORLD METEOROLOGICAL ORGANIZATION. Floods. https://wmo.int/topics/floods.

[34] European Parliament. Direttiva 2007/60/ce del parlamento europeo e del consiglio del 23 ottobre 2007 relativa alla valutazione e alla gestione dei rischi di alluvioni. 2007.

[35] Arpa Piemonte. Arpa piemonte. https://www.arpa.piemonte.it/.

[36] THE NATIONAL ACADEMIES PRESS. *Tying Flood Insurance to Flood Risk for Low-Lying Structures in the Floodplain.* THE NATIONAL ACADEMIES PRESS, 2015.

[37] Copernicus Programme. Efas. https://www.copernicus.eu/en/european-flood-awareness-system.

[38] Federico Rubiano. Considerazione sui costi e benefici delle opere di mitigazione del rischio alluvionale. 2022.

[39] UNISDR. Flood hazard and risk assessment. 2017.

[40] Bauer-Marschallinger B. Vreugdenhil M., Navacchi C. Sentinel-1 cross ratio and vegetation optical depth: A comparison over europe. *Remote Sens.*, 2020.

[41] Wikipedia. Evaluation of binary classifiers. https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers.