POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Politecnico di Torino

Master's Degree Thesis

Machine Learning for Predicting Grape Quality Using Spectral Imaging Techniques

Supervisors

Prof. Paolo GARZA

Candidate

Luca PESCE

December 2024

Abstract

The production of high-quality wine is strongly impacted by the quality of grape clusters, especially their sugar level, which is a key factor in deciding the best time to harvest. Destructive sample techniques that are labor-intensive, time-consuming, and not scalable are traditionally used to evaluate grape maturity. This thesis addresses these limitations by investigating the non-destructive prediction of grape sugar concentration utilizing machine learning algorithms in conjunction with modern imaging technologies, with a primary focus on hyperspectral imaging (HSI). The study starts with a thorough overview of the principles of viticulture, covering grape biology, development phases, and the importance of sugar buildup during maturation. After that, it discusses the fundamentals of imaging techniques, including RGB, multispectral, and hyperspectral imaging, with a focus on how these technologies record and communicate important information about the product. This fundamental knowledge lays the groundwork for talking about how several imaging modalities, each with a different spectral resolution, can be used to forecast crucial quality metrics like the Brix Index, which measures the amount of sugar in grapes. The main focus of the thesis is the use of hyperspectral imaging in conjunction with machine learning models, like Partial Least Squares Regression (PLSR), to forecast grape sugar concentration. Preparing imaging data for predictive modeling required a large portion of the study in order to improve the signal-to-noise ratio and lower dimensionality. The results analysis is enhanced by philosophical thoughts on the models' underlying assumptions and their suitability for various viticultural situations. An additional exploration was planned, potentially using RGB datasets to test whether simpler, more accessible imaging methods could provide comparable results to the more sophisticated hyperspectral approach. The results of this thesis demonstrate the potential of imaging-based techniques for viticulture's non-destructive quality evaluation. This work lays the groundwork for future precision agriculture research targeted at enhancing the effectiveness and precision of grape quality monitoring by critically analyzing the methods employed and considering the types of data collected by various imaging technologies.

Acknowledgements

I wish to express my sincere gratitude to Prof. Paolo Garza for his support and encouragement in my time of need.

A special thanks goes to my parents, who have always given me even what they did not have, and from whom I have learned, directly or indirectly, the values that are the foundation of my being. It is thanks to their example that every morning I wake up with the desire to be kind to others.

I sincerely thank my siblings, Carolina and Alessandro, whose voices resonate within me, while their best qualities emerge in times of need, like a reflection of the precious bond that unites us.

A special thought goes to my brother Francesco, with whom I share a unique understanding made of unspoken words but perfectly understood. He taught me to be a friend before being an older brother, always reminding me of the value of this bond.

My deepest gratitude goes to my girlfriend Margherita, who has been a delicate touch in the most difficult moments. She knows a part of me that no one else does, and with her, I have been able to confide my fears during challenging times.

Finally, my sincerest thanks go to my friends, who have been, consciously or unconsciously, a beacon of light in dark moments. From each of them, I have learned valuable lessons that I will carry with me throughout my life.

"Inspired by the book Change*"*

Table of Contents

List of Tables

List of Figures

Acronyms

RGB

Red Green Blue

VIS-NIR

Visible and Near-Infrared

NIR

Near-Infrared

NDVI

Normalized Difference Vegetation Index

ANN

Artificial Neural Network

TSS

Total Soluble Solids

HSI

Hyperspectral Imaging

SPA-MLR

Successive Projections Algorithm-Multiple Linear Regression

GAPLS-LS-SVM

Genetic Algorithm Partial Least Squares-Least Squares Support Vector Machine

SSC

Soluble Solids Content

pH

Potential of Hydrogen

PLSR

Partial Least Squares Regression

NN

Neural Network

RMSE

Root Mean Square Error

CNN

Convolutional Neural Network

RR

Ridge Regression

TB

Tinta Barroca

TF

Touriga Franca

TN

Touriga Nacional

RMSEP

Root Mean Square Error of Prediction

SVR

Support Vector Regression

RMSECV

Root Mean Square Error of Cross-Validation

PLS-DA

Partial Least Squares Discriminant Analysis

TA

Total Anthocyanin Content

SNV

Standard Normal Variate

MLR

Multiple Linear Regression

VIP

Variable Importance in Projection

RMSEV

Root Mean Square Error of Validation

RoBoost-PLSR

Robust Boosting Partial Least Squares Regression

MLP

Multilayer Perceptron

3D-CNN

Three-Dimensional Convolutional Neural Network

AlexNet

AlexNet Convolutional Neural Network Architecture

ResNet

Residual Network

DCAE

Deep Convolutional Autoencoder

FCAE

Fully Connected Autoencoder

PLS

Partial Least Squares

LSSVM

Least Squares Support Vector Machine

AdaBoost

Adaptive Boosting

GA

Genetic Algorithm

NaCl

Sodium Chloride

SAM

Spectral Angle Mapper

LED

Light Emitting Diode

LabVIEW

Laboratory Virtual Instrument Engineering Workbench

FW

Fresh Weight

PCR

Principal Component Regression

PCA

Principal Component Analysis

PC

Principal Component

Chapter 1

Fundamentals of viticulture and grape ripening

1.1 Introduction

Viticulture and enology constitute two distinct sciences and each has its own specific literature. Viticulture represents the set of agronomic techniques involving the cultivation of vines (for table and wine), thus being able to consider itself as a branch of arboriculture[1]. Oenology is the science that studies the transformation of grapes into wine, the grapes suitable for its production (the microbiology, chemistry and sensory characteristics), but also the production process itself, thus the techniques related to it (e.g., filtrations, pressing, pumping over)[2]. Despite this distinction, these two sciences communicate with each other, and the key to their communication is the grape. In fact, grape quality is the result of winemaking practices and the study of viticulture and is the key factor in excelling in wine production and obtaining a distinctive, quality product[3]. When we talk about wine quality, we first refer to the quality of grapes at the time of harvest, which will constitute the raw material for wine production. Grape quality, in turn, is a complex and not uniquely defined factor and is closely related to the degree of maturity of the grape at the time of harvest. Grape maturity, which is reached at the end of the ripening process, is defined as a set of properties that depend on the physical and chemical characteristics of the grape. Finding and designing systems that support winemakers in predicting the degree of grape maturity is one of the greatest challenges for research in viticulture, this is because there are so many variables involved that are constantly changing and interacting. If we want to understand how mathematical models, and machine learning algorithms in particular, can fit in as key tools and evolve this field, we cannot fail to take an informed look at the complex domain we are interfacing with. In fact, modeling a

physical system, in this case the growth of a plant and its fruit, means first of all studying and understanding its evolution over time with a look at the variables at play that interest us most, in this case ripening. Only in this way can we generate informed assumptions that will allow us to better understand why models work or do not work, how they might do so, and answer questions such as, "Can there be a general methodology that solves problem X ?" and "Can a model trained on a D1 dataset work on a D2 dataset ?" In this chapter, I will explain the complex system "Grapevine and its Grapes" at a level sufficient to understand the explanations and reasoning in the next chapters, with an eye toward the variables that will interest us most. All the explanations and insights on grapes in this chapter come from the study of books [3], [4], [5], and [6].

1.2 The vine

Making wine is a long and complex process that requires years of dedication, knowledge and passion. It all begins with planting young vines on carefully selected soils, carefully evaluating the soil and climate to determine which varieties will thrive best. For example, in regions like Napa Valley, Cabernet Sauvignon prefers rocky soils and warm temperatures, while Sauvignon Blanc finds ideal conditions in cooler microclimates and sandy soils. Once planted, the vines go through a growth phase that lasts several years before producing quality grapes. During this period, the plants require constant care: careful watering, precise pruning and protection from pests and diseases. Pruning, in particular, is essential to ensure an optimal balance between shoots and buds, directly affecting the quantity and quality of grapes produced. With the arrival of spring, vines emerge from dormancy and begin budding, followed by flowering about a month later. This time is crucial, as weather conditions can significantly affect fruit formation. Late frosts or heavy rains can affect flower fertilization, reducing the amount of grapes that will be produced. During the summer, we witness fruit set and, subsequently, veraison. The latter is a key stage in the vine life cycle: the berries begin to change color and accumulate sugars through photosynthesis. In red grapes, the berries change from green to red-purple, while in white grapes they take on golden-yellow hues. Veraison indicates the beginning of ripening, a period when the balance between sugars and acids develops, directly affecting the organoleptic characteristics of the grapes. Determining the ideal time for harvesting is not simple and requires careful evaluation of several factors. In addition to the analysis of sugar content-measured in degrees Brix-winemakers consider the development of acids, tannins and aromas. Weather conditions play a key role: warm days accelerate ripening, while cooler temperatures slow it down, allowing a more balanced development of grape compounds. The harvest represents the culmination of this long journey.

Harvesting can be done manually or mechanically, but in either case it is essential to act at the right time to preserve the quality of the grapes. White grapes are generally harvested earlier than red grapes, and those destined for sparkling wine production are the absolute first, as they require lower sugar content. Once harvested, the grapes begin their journey toward becoming wine. Red grapes are destemmed and crushed, allowing the must to ferment in contact with the skins to extract color, aromas and tannins. White grapes, on the other hand, are often pressed immediately to avoid prolonged skin contact, thus preserving their freshness and acidity. Alcoholic fermentation transforms the sugars in the must into alcohol and carbon dioxide through the action of yeasts. Next, some wines go through malolactic fermentation, which helps soften acidity and develop more complex flavor profiles. Aging takes place in different types of containers, oak barrels, steel tanks or amphorae, depending on the style of wine desired. This long and fascinating cycle, from the planting of the vine to the ripening of the grapes, highlights how important it is to fully understand each stage of the process. In the next chapter, we will delve into grape ripening, focusing on veraison and photosynthesis, to understand how sugar content becomes one of the key parameters in deciding the ideal time for harvest. Only through this knowledge can we appreciate how viticultural practices influence wine quality and how innovative tools can support winemakers in this complex process [7].

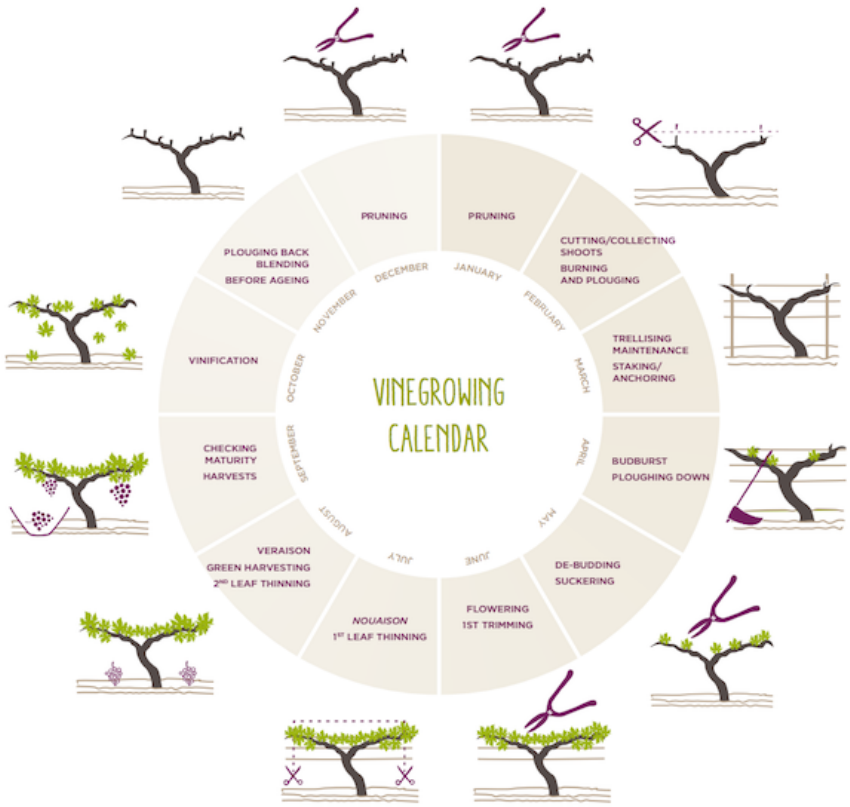

Figure 1.1: Annual phases of vine growth, from planting to harvest [8]

1.3 The grape, the harvest and the importance of sugar

As far as viticulture is concerned, the climax of the vintage is sanctioned by the time of harvest. As written in the book [3] "the harvest is the irrevocable step that connects enology and viticulture." At this time, winemakers and winegrowers must cooperate closely to decide the exact day of harvest with the goal of picking grapes at the correct degree of ripeness. The desired degree of ripeness depends strictly on the quality of the grapes planted and the environmental conditions to which the grape variety was subjected during the vintage. As mentioned earlier, grapes constitute the raw material for wine production, and their degree of maturity is certainly one of the factors that most impact wine quality. This is the result of the physiological and biochemical processes that affect grapes throughout their life cycle and will be the cause of their physicochemical composition at harvest. Unlike

other fruits, the capillary study of grapes presents several problems related to the fact that the growth of this fruit is the result of a long and complex reproduction cycle.

1.3.1 Bunch morphology and growth phase

Before delving into the stages of growth that most interest this analysis, we need to understand at least on a superficial level the morphology and terminology of grapes. In fact, when we talk about grapes in general, we may want to talk about the vine in general, the grape cluster, the grape berry, or any of its other components. What we are interested in illustrating in this study is that grape berries are organized into clusters. The cluster consists of a stalk (or rasp) and numerous berries (also called grains, or more properly berries), small in size and light in color (yellowish-green, yellow, golden yellow) in the case of white grapes, or dark in color (pink, purple or bluish violet) in the case of black grapes. The stalk, or rachis, is the central axis of the cluster, branched into racemes and then pedicels, which bear the flowers and later the fruit, the berries. In addition, the stalk also has a very important function indeed: it performs the task of dragging nutrients into the berry through the plant's lymphatic network. The stalk, being a woody element, is inedible and in the course of winemaking is normally removed because it may contain unpalatable compositional elements that would give an unpleasant taste to the wine[9].

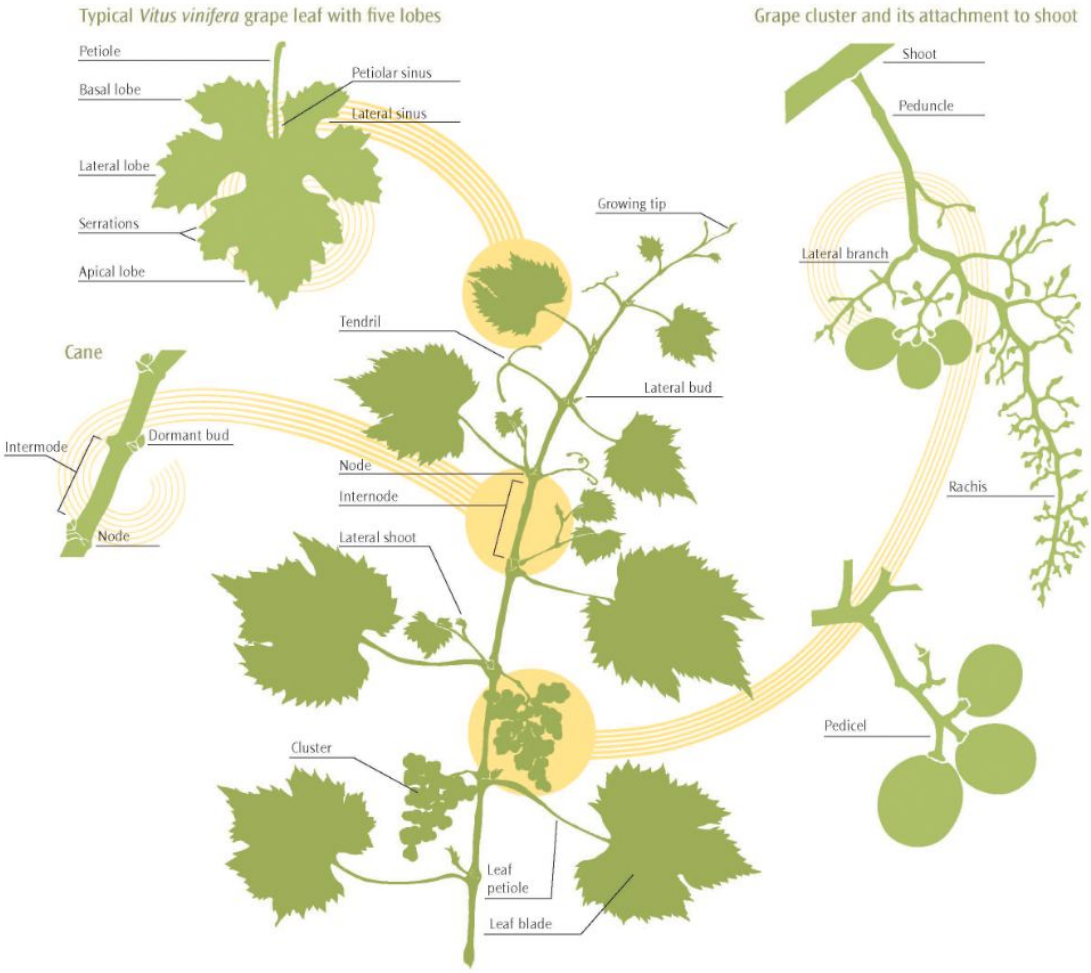

Figure 1.2: Bunch morphology
[10]

In the context of predicting the degree of grape maturity, we find ourselves considering the individual grape berries that make up the cluster. In the course of its development, the grape follows an evolution that includes all berries and is divided into three stages, taking into consideration the diameter, weight and volume of the berries:

1. Rapid initial growth, which, depending on environmental conditions, lasts 45 to 65 days and begins during the period of fertilization and fruit set. During this period the berry has intense metabolic activity characterized by high respiratory activity and rapid acid accumulation.

2. A slower growth phase during which one of the most important and relevant phenological stages for the study of berry evolution, veraison, occurs. This

event is characterized by the appearance of reddish color in red grapes and a translucent skin in white varieties.

3. A final, rapid growth phase corresponding to berry maturation. During this phase, cell growth resumes and important physiological changes are observed. Respiratory activity of the berry decreases, while some enzymatic activities increase significantly. This phase, lasting about 35-55 days, is characterized by the accumulation of simple sugars, cations such as potassium, amino acids and phenolic compounds, while there is a reduction in malic acid and ammonium. The final size of the berry depends largely on these accumulation processes, as well as on the number of cells present. There is a close relationship between the size of the berry and the number of seeds it contains.

Since in this study we focus on the prediction of sugar content in ripe grapes, in the following explanations we will mainly focus on the last stage of development. This will allow us to lay the foundation for a comprehensive view of the phenomenon, which is essential for a meta-understanding of the context, premises and limitations associated with the application of the models.

1.3.2 Grape morphology and composition at maturity

The berry, the fruit of the grape, is characterized by its typical round or tending to oval shape. Depending on the complexity of the phenomenon we are going to study, it is possible to visualize its composition in different ways. For our analysis, it is sufficient to know that the berry is composed of skin, pulp and seeds (pips), more precisely:

- the skin, the outermost part of the berry, has the function of protecting and containing the pulp so that water does not evaporate and external agents (insects, fungi, etc.) do not penetrate. This is also very important for its pigments, colored substances, which give color to the berries, making them white or black depending on the type of grape variety;

- the pulp, which contains within it the grape seeds, is composed, when ripe, of water (70-85%), sugars (15-20%), proteins and various nitrogenous substances, organic acids and some colloidal substances, including pectins;

- finally, the grape seeds, in addition to being the seeds of the plant, are very important because they contain tannins, which can characterize the final composition of the wine[11].

In ripe grapes, seeds account for 0 to 6 percent of grape weight and are a key resource of phenolic compounds during red wine production. Depending on the

variety, they contain between 20 and 55 percent of the total polyphenols in the berry. The skin, depending on the variety, accounts for between 8 and 20 percent of the berry's weight. It occurs as a heterogeneous tissue and its importance depends on the extraction method used during wine production. The importance of the skin depends more on the fact that it contains significant amounts of compounds such as polyphenols (benzoic and cinnamic acids, flavonols and tannins) and aromatic substances. In red grapes, other very important phenolic compounds called anthocyanins occur, which give red grapes their color. The pulp makes up most of the weight fraction of the berry, it is around 75 to 80 percent. The vacuolar content, which is the liquid present inside the vacuoles of the pulp cells, consists mostly of the must and a small solid part(1%). Must appears as a cloudy liquid and has a high density derived from the many chemicals within it. Sugars, mainly glucose and fructose, account for most of these substances. The concentration of sugar in ripe grapes, which is one of the basic parameters by which the level of ripeness is derived, ranges from 150 g/l to 240 g/l. As can be imagined, this paragraph has not mentioned all the substances that make up the grape berry, but an analysis of the most important ones has been made by casting an eye over the study in question.

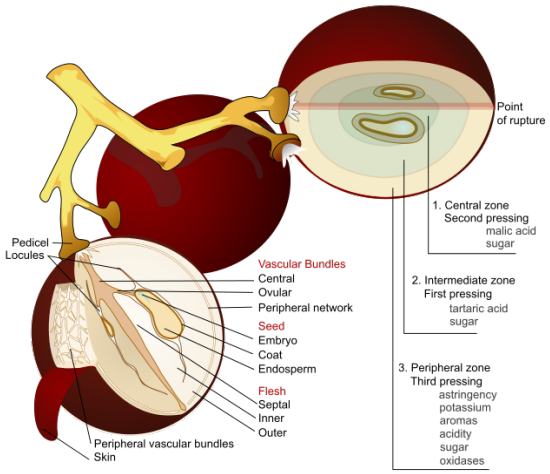

Figure 1.3: Cross section of a grape showing its components [12]

1.3.3 Maturation

The veraison, which follows the fruit set, that is, the fertilization of the flowers and the subsequent appearance of the grape berries, is the phenological stage that marks the beginning of the ripening phase. After this event we see a real physiological change in the grape, from this moment onwards the internal mechanisms of the

grape change the way they function. During the ripening stage, the grape changes from being a hard, sour, green fruit to a soft, colorful fruit rich in sugar and aroma. During this period, the berry begins to accumulate significant amounts of sugar. At the beginning of ripening, the grape berry imports comparable amounts of water and sugar, but the amount of water tends to decrease going forward with time. This phenomenon can also be verified by the fact that the percentage of solid material continues to increase, so more solute is transported than water. Sugar accumulation, which is one of the most spectacular phenomena of ripening, mainly affects seed and pulp growth and is caused by the berry's high demand for the products of photosynthesis. In addition, during ripening, there is a rapid accumulation of phenolic compounds, including anthocyanins, pigments responsible for coloration in red grapes. These pigments are by-products of sugar metabolism, and their synthesis is closely linked to sugar accumulation in the berry. In fact, the appearance of anthocyanins begins about two weeks before the color is externally visible, and their concentration gradually increases during ripening, peaking and then stabilizing or declining slightly by the time of full ripeness. The relationship between sugar accumulation and anthocyanin synthesis is crucial, as it indicates that increased sugar content not only contributes to the alcohol potential of the future wine, but also influences the development of phenolic compounds that determine the color and some of the organoleptic characteristics of the wine itself. This correlation is particularly significant because color changes in berries, due to the presence of anthocyanins, can be observed and measured through imaging techniques, such as RGB or multispectral imaging. Understanding this link allows us to hypothesize that it is possible to predict the sugar content of grapes by analyzing color changes during ripening. Accordingly, machine learning models applied to grape images could be developed to nondestructively estimate the sugar level in grapes.

1.3.4 Definition of maturity and the moment for harvesting

The environmental conditions to which the vine is subjected and various other factors strongly influence the processes described in the previous explanations in a complex and nonlinear manner. For this reason these are not necessarily simultaneous and do not evolve at the same rate. In some cases physiological changes in the vine do not occur in the same order from year to year even in the same grape variety. Ripening, unlike veraison, which is a well-defined event on a physiological and biochemical level, does not represent a precise stage. Different parts of the grape reach the stage of physiological maturity at different times. For example, the seeds reach this stage in the veraison period, while the pulp and skin continue their ripening process for many more weeks. Consequently, the definition of maturity, and therefore the time of harvest, varies depending on what

our goal is and the wine we are going to produce. In enology, skin ripeness is consequent to the maximum concentration of phenolic compounds and aromatic substances, while flesh ripeness is defined by the optimal ratio of sugars to acids. To simplify the problem, we tend to focus on increasing sugar concentration and decreasing acidity. However, it is also crucial to consider the accumulation of aromas in white grapes and phenolic compounds, such as anthocyanins, in red grapes. A quality wine-growing area promotes harmonious ripening, in which all these transformations reach their peak simultaneously at harvest time. For this reason, predicting the sugar content of grapes becomes crucial, as it is one of the key indicators for determining the ideal time of harvest. I will explore this aspect in more detail in my paper.

1.3.5 Maturity index and sampling

"Nothing is more heterogeneous than grapes from the same vineyard at a given time, even when considering the same variety." This sentence, taken from the book [4], and the explanations that follow, represent one of the most important factors in understanding the significance of my reviews and study, this is because it is of paramount importance to understand the inherent and nonlinearly controllable variability of the physical and chemical characteristics of grapes. If we go to analyze a single bunch of grapes during the ripening stage, we will observe that its berries will grow in weight and change in color one after another, that these processes will start at different times and evolve at different speeds. This causes that, at the time of harvest, even the same bunch of grapes will show some variability in the physical and chemical characteristics of its berries. This fact leads to the verification that in the same vine, at the hypothetical time of harvest, different clusters will have different levels of maturity. For this reason it becomes very risky to determine the day of harvest by making an analysis of the clusters of a single vine. The most common method to date is to collect, through the use of shears, the fragments of three or four bunches from 100 different vines. In doing so, special attention must be paid to the variability of environmental and soil conditions to which the vines were subjected during the cycle. It will therefore be necessary from the same vine to take samples from clusters under the most light-exposed leaves and under the least light-exposed leaves, and it will be useful to take into account that in the more compact clusters often the berries located inside are less ripe than the others. It will also be necessary to take samples of bunches from all areas of the vineyard because of the variability of the soil and environmental conditions at different heights in the vineyard. After being harvested, the grapes are brought to the laboratory and the berries are separated, counted and weighed. The juice is extracted using a small hand press or centrifugal separator, and the volume obtained is measured and expressed in liters of must. After this is done,

the concentrations of sugars and acids in the juice are determined. Instruments such as the refractometer and hydrometer are used to measure the sugar content of the must. The refractometer measures the density of the wort. The unit of measurement used here can be gram per liter or the Brix degree, which expresses the amount of sugars in the wort in grams per 100 g of solution, that is, as a percentage of solid content. However, it is important to note that measuring in degrees Brix is reliable only from a certain level of grape maturity, around 15° Brix; before that stage, other compounds such as organic acids and amino acids can interfere with the measurement because their refractive indices are similar to those of sugars. The hydrometer, on the other hand, measures the specific gravity of the must, which increases with the concentration of sugars present. However, the types of analysis described above have some significant limitations. First, they are destructive methods, as they require the harvesting and processing of berries, preventing continuous, noninvasive monitoring directly in the field. In addition, referring to what we wrote earlier, the inherent variability of grapes, influenced by several factors such as sun exposure, soil conditions, and microclimatic differences in the vineyard, hampers the possibility of obtaining a representative sample of the entire harvest. This could lead to incorrect estimates of sugar content and, consequently, influence the decision on the best time to harvest. The limitations mentioned above have led research over the years to move toward the study of alternative, non-destructive methods for measuring the sugar content of grapes. What one would ideally like are mobile sensors within the vineyard that would be able to accurately estimate the sugar and polyphenol content of the grapes without too much waste using techniques and algorithms that could be the ones analyzed in this study. At any given time, such methods would show grape growers the state of ripeness of the grapes and improve the accuracy of the predictions, thus enabling more accurate decisions about when to harvest. What is of paramount importance to note is that during the ripening stage, the increase of sugars in the grape berry is very rapid, with values that can grow by an average of 2.2 to 3.2 degrees Brix in one week, with variations from 0.6 to 4.8 degrees Brix [6]. Counting that the residual sugar content of ripe grapes varies from 150 g/l ($\approx$ 15 degrees Brix) and 240 g/l ($\approx$ 24 degrees Brix), the latter can grow by as much as 20 percent per week. These data will be useful in the reflections in the next chapters.

Chapter 2

Imaging techniques and cultures

In the previous chapter, I recounted how scholars in recent years are oriented toward finding reliable, non-destructive and low-cost systems to achieve the goal of estimating grape quality parameters in order to decide on the optimal day of harvest. When we talk about predicting the physical and chemical characteristics of fruit, we first refer to remote sensing techniques and machine learning algorithms applied to computer vision. What we in fact want to do ideally, and what would be sufficient, is to estimate quality parameters through the use of images, that is, digital representations of the entity we are studying. The most frequently used images are RGB and VIS-NIR images, the reasons for which will be explained later in the paragraphs. In this chapter I will give a sufficient explanation of the concepts behind image interpretation, the goal is to present all the necessary tools to understand how and why machine learning algorithms can achieve excellent performance in computer vision tasks and in particular in the estimation of fruit quality parameters. In doing so, I will try to present what assumptions we voluntarily or involuntarily make when trying to train a machine learning algorithm in computer vision. In particular, the focus will be on the concept of information, trying to explain the importance of being aware of what raw information is encapsulated within an image. The raw information will in fact be that from which more complex and higher-level information will be deduced by experience that will lead us to trying to solve a certain task, through a certain machine learning algorithm, by means of a certain input. After doing this I will focus on and present some studies using RGB, multispectral or hyperspectral images for the prediction of physical and chemical characteristics of fruit. The focus of this chapter will not be solely and exclusively on grapes, with the goal of giving a comprehensive view and then going into more detail on grapes and the study in the next chapter. The considerations on imaging

techniques are the result of the study and reworking of the information contained within the book[13].

2.1 Remote sensing

"A picture is worth a thousand words" is a saying often used when one wants to introduce the concept of an image and the information it contains. This phrase actually refers to the fact that within an image are contained a myriad of data that, if understood and interpreted in the correct way, can tell us about the image. Remote sensing is the science of deriving information about an object through measurements at a distance from it, without having direct contact with it[13]. The quantity most often measured is the electromagnetic energy emitted by the object. Seismic waves, sound waves, and gravitational force are some of the other entities that can be measured through the use of these instruments, however, our focus will be on systems that measure electromagnetic energy. Remote sensing originated as a discipline for analyzing images and thus its origins coincide with those of photography, in the early 1800s. However, this discipline began to attract much interest during World War I, when aerial photography began to be used for military reconnaissance and surveillance. However, it was during World War II that this discipline began its greatest acceleration, which would not stop. In fact, at that time remote sensing began to be used to study the earth and battle zones also by means of electromagnetic waves outside the visible spectrum. To understand what imaging is and, in particular, RGB imaging and in the VIS-NIR spectrum, which are the ones of most interest to this study, we need to start with the concepts of electromagnetic wave, electromagnetic spectrum and spectral properties of matter. Having a clear understanding of these concepts will allow one to make consistent assumptions and reasoning to explain the results and analyses present in this study.

2.1.1 Electromagnetic waves

Why do we see ? This question, which might seem trivial, when thoroughly analyzed, allows us to be aware of some concepts that we do not dwell on and ask questions about in our daily lives. These concepts are the basis of the theory on image interpretation and are therefore fundamental to our purpose. The first thing that comes to mind to answer the previous question is the fact that we see because we have eyes. This is true, but what exactly do we perceive ? The eyes, like all the other senses, are instruments that can measure some form of magnitude derived from the outside world and instantly inform the brain about the measurement. The brain's response and task, once informed, is to make us perceive. The magnitude that the eyes can measure and thus the one from which we can extract information to see is the electromagnetic wave, specifically a small

subset of electromagnetic waves. Once measured, this magnitude will be perceived as color and thus give us the ability to see the outside world. The origin of these electromagnetic waves can be traced back to the sun. More precisely, nuclear reactions in the sun produce electromagnetic waves that propagate at the same speed to the earth without undergoing any major changes along the way. The sun is said to produce a full spectrum of electromagnetic waves. A practical interpretation is that the sun produces an infinite amount of electromagnetic waves at all possible wavelengths. The electromagnetic spectrum is the range of possible wavelengths. Every object, depending on its shape and physical chemical composition, inherently has the property of absorbing or reflecting some or other electromagnetic waves depending on their wavelength. The human eye is sensitive to a small portion of the electromagnetic spectrum, particularly to length waves in the 400-700 nm range. So the reason we see is that every object reflects a certain portion of the spectrum between 400 and 700 nm, the reflected electromagnetic waves will reach our eye, which will receive the raw information, send it to the brain, and the brain will make us perceive colors. From the color information is then derived several other pieces of information such as shapes, sizes, spaces, and everything related to sight as we know it.

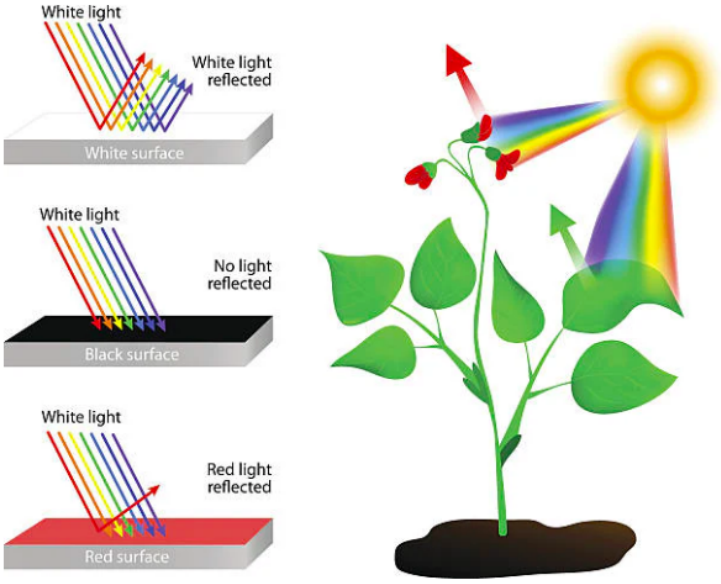

Figure 2.1: Illustration of how objects absorb and reflect different wavelengths of light

[14]

The first to investigate the properties of the visible spectrum was Isaac Newton, who through some experiments came to the conclusion that visible light is divisible

into 3 segments, today we know these segments as "primary colors" and they are respectively the waves from 400 to 500 nm (blue), 500 to 600 nm (green) and 600 to 700 nm (red). The fundamental property of each primary color is that none of them can be derived by mixing the other two in any proportion and that all possible colors can be derived by mixing the 3 primary colors in the proper proportions. In practice what we do automatically is to perceive waves with lengths from 400 to 500 nm as blue, from 500 to 600 nm as green, and from 600 to 700 nm as red. Depending on the intensity of the radiation perceived in these spectra we will perceive all other colors. When we see a blue object, it means that it is absorbing radiation with wavelengths from 500 to 700 nm and is reflecting that from 400 to 500 nm. The fact that you can also see the individual colors blue, green and red in different ways is that we are talking about continuous ranges. Within each range there are infinite wavelengths, and each object in turn can reflect some with greater intensity and some with less intensity. The fact that we perceive only the electromagnetic waves in the visible spectrum does not mean that there are only those. In fact, the visible spectrum is only a small part of the electromagnetic spectrum. Objects, depending on their characteristics, will also absorb or reflect electromagnetic waves in the rest of the spectrum, and consequently the hypothetical measurement of their intensity can in turn tell us about the characteristics of matter. In its totality, the spectrum is divided, by convention, in this way (see Table 2.1). In the next

Table 2.1: Division of the Electromagnetic Spectrum

Region	**Wavelength**	**Frequency**
Radio Waves	$> 1\,\text{mm}$	$< 300\,\text{MHz}$
Microwaves	$1\,\text{mm} - 1\,\mu\text{m}$	$300\,\text{MHz} - 300\,\text{GHz}$
Infrared (IR)	$1\,\mu\text{m} - 700\,\text{nm}$	$300\,\text{GHz} - 430\,\text{THz}$
Visible Light	$700\,\text{nm} - 400\,\text{nm}$	$430\,\text{THz} - 750\,\text{THz}$
Ultraviolet (UV)	$400\,\text{nm} - 10\,\text{nm}$	$750\,\text{THz} - 30\,\text{PHz}$
X-Rays	$10\,\text{nm} - 0.01\,\text{nm}$	$30\,\text{PHz} - 30\,\text{EHz}$
Gamma Rays	$< 0.01\,\text{nm}$	$> 30\,\text{EHz}$

section I will explain what images are from the perspective of the information they encapsulate and how we can make informed choices in creating image-based prediction models. In particular, I will focus on images that capture information in the visible spectrum (RGB) and those that also capture information in the near infrared (VIS-NIR).

2.1.2 RGB, multispectral and hyperspectral images

For a long time in the history of remote sensing, images have been recorded physically in the form of photographic images. A photographic image is a physical record that through the use of chemical coatings records a scene. This record will be precisely the representation of the scene and will be a more or less similar representation of what we see. In this section we will deal with another type of images, which are easier to analyze, visualize, transfer and save; digital images. Unlike the physical format, the digital format represents the image as an array of many individual values called pixels. Each pixel represents in some way a portion of the image. By convention the image is dissected by a number of horizontal and vertical segments thus dividing it into small squares, these squares are the pixels and are associated with a value. RGB images are generated using sensors designed to replicate the human eye's response to visible light. These sensors capture three spectral bands (red, green, and blue), which correspond to the sensitivities of the three types of cones in the retina, allowing colors to be represented similarly to how we perceive them visually. For each pixel in an RGB image, there are three numerical values representing the intensity of light in the red, green and blue bands. Although the spectra of red, green and blue are continuous, each pixel retains a single value for each of these bands, which corresponds to an estimate of the average intensity of light detected in that spectral range. Thus, the color of each pixel is determined by combining these three values in different proportions. Sensors that record data across a far larger range of wavelengths than typical RGB photos are used to create multispectral and hyperspectral images. Multispectral images record information across many spectral bands, frequently encompassing not just the visible (VIS) but also the near infrared (NIR), whereas RGB images are restricted to three bands (red, green, and blue) intended to mimic human vision. A small number of spectral bands, such as the visible and a few particular near-infrared bands, are represented by values in each pixel of multispectral photographs. The intensity of light reflected in each of these bands is represented by these values, which provide a more thorough understanding of the phenomenon being viewed. For instance, because it can reveal details like plant moisture, chlorophyll content, and crop health that are not evident in the visible bands, NIR is very significant in applications pertaining to viticulture and agriculture. The continuous spectrum from the visible to the near-infrared and beyond is covered by hyperspectral imaging, which records data in many more bands, often hundreds. A hyperspectral image has a whole range of values in each pixel, which corresponds to a spectral signature specific to the surface or reflective material. It would be challenging to identify between identical materials and situations using RGB or even multispectral bands alone, but these spectral characteristics allow for this. Hyperspectral images are very helpful in the agricultural setting for closely examining the physiological

traits of crops, including their nutritional content, level of water stress, and other attributes linked to plant health. Pixel values in both systems reduce the continuous spectrum to a collection of discrete values by estimating the average intensity of light reflected in each band. Advanced assessments of vegetation, chlorophyll content, and crop health, all crucial elements in agricultural and viticultural contexts, are made possible by the combination of various bands, especially those in the VIS and NIR, which offer significantly more information than an RGB image.

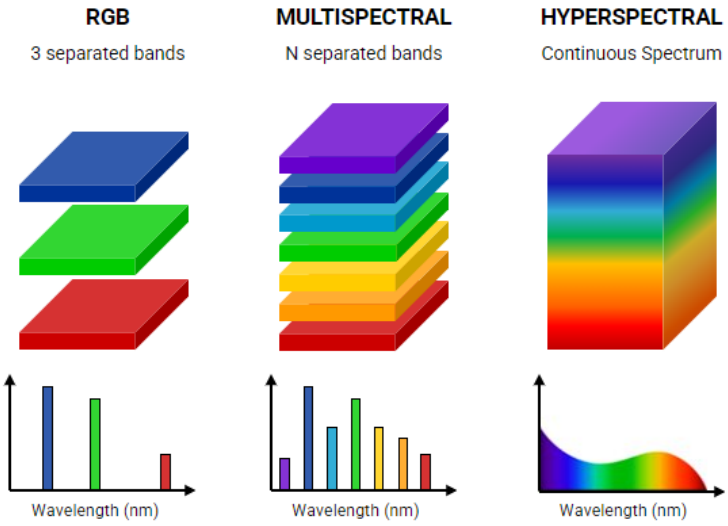

Figure 2.2: Comparison of RGB, Multispectral, and Hyperspectral Imaging [15]

2.1.3 Information

In the last chapter I gave a detailed overview of the types of images that will be analyzed in the study and review of the scientific literature. In this chapter I will make some considerations about the raw information that digital images contain and how, through experience, higher-level, interpretable information can be derived from that information. Finally I will make some considerations about the maximum amount of information that an image can contain. When we try to create a machine learning model that classifies a photograph according to whether a dog, a cat, or neither is present and train it through the use of input examples (dog and cat pictures) and output examples (dog class, cat class, neither class) we are from a certain point of view cheating. The reason I say this lies in the fact that through our experience, we already know from looking at an image that we know how to interpret what it depicts. Since an image is a digital, albeit simplified, representation of what we see, we are already certain that the image contains all

the information necessary to classify what is within it. Whenever such awareness is present, model building becomes much easier, since the starting assumptions are already present in our experiential background. In other words, we are already sufficiently convinced that from a given X we can derive x1 information, from which we can derive x2 information and going on until we derive the xn information we are interested in. We can therefore say with some confidence that a machine learning algorithm trained with sufficient data and of sufficient quality can learn to predict an information xn from an entity X. Going into the details of RGB images we are aware of basically two things:

1. The raw information contained in RGB images corresponds to the intensity of light reflected in the spectra of blue, red and green, for each pixel.

2. The RGB image is a simplified digital representation of what is received by our eye. In fact, our eye receives the intensity of electromagnetic waves in the visible spectrum and through an exchange of information with the brain this is perceived by us as color. We can therefore say that the information contained in an RGB image is a proxy for color.

It is important to note that in the raw information contained in an RGB image there is actually no reference to colors, yet it is as if it were present because it is directly derivable information from the raw information. The concept is that since we are innately capable through the use of sight, and therefore colors, to distinguish different objects, group them, see their depth, perceive distance and shapes, and since color is directly correlated with the electromagnetic radiation received by the eye, a sufficiently intelligent algorithm will be able to perform computer vision tasks well enough. The algorithm will then learn from the raw information to derive patterns in the data and more complex, higher-level information to derive the final information, similar to how we do it by going through colors, then from shapes etc. The issue becomes more complicated, however, when we attempt to apply the same way to forecast the amount of a chemical molecule in a fruit. In contrast to visual categorization, we cannot be as sure that an RGB image has all the information required to identify a particular chemical property. This is dependent on whether and how the chemical property of interest is connected to the reflected radiation. For instance, let's consider the perfect example of grapes: we know that anthocyanins, which give the berries their dark hue, rise during the ripening period as the berries develop and acquire more vibrant hues. In this instance, we may speculate that the color information in the picture can be used to determine the anthocyanin content. In fact, a machine learning model trained using RGB photos may be able to correlate color with anthocyanin content because anthocyanins are strongly linked to the shift in berry hue. The primary challenge is that, unlike object recognition, we cannot determine a fruit's chemical composition only from

visual observation since our experience prevents us from properly interpreting that information. Multispectral and hyperspectral photographs are useful in this situation. Because these images can record reflectances over a wide variety of spectral bands, including some that are invisible to the human eye, such near-infrared, they are able to capture a lot more information than RGB photographs. We may be better equipped to forecast the composition of particular chemical compounds as a result of this increase of information. It is nevertheless crucial to comprehend the experience of what data these extra spectral bands record and how it connects to the chemicals of interest. We can only develop machine learning models that can accurately predict features beyond visual interpretation if we have this knowledge and interpretation. Furthermore, the number of spectral bands and their coverage along the electromagnetic spectrum greatly affects the quantity of information that is available when looking at multispectral and hyperspectral photographs. Consider a hypothetical image A that only records the blue spectrum's reflectance value and another hypothetical image B that has a far greater spatial resolution but is still only able to record the blue spectrum. Since all of the bands are coupled within the same range, picture B adds no discernible spectral information variety while having higher resolution. To put it another way, we are merely making the blue spectrum more detailed without increasing the amount of information that can be gleaned from the visible spectrum. As the phrase "bands that are close to one another tend to replicate the information in their adjacent region of the spectrum"[13] describes, this limitation draws attention to a crucial idea: spectral bands that are close to one another tend to replicate the information of adjacent regions of the spectrum. This implies that including more adjacent spectral bands might just make the information already collected more redundant rather than greatly increasing the diversity of information contained in the image. Now, if we look at another situation where we have a hyperspectral image that includes several visible and near-infrared (VIS-NIR) bands, we can record a lot more information about color as well as physiological traits like the amount of water or chlorophyll. Since each successive band in VIS-NIR adds a new dimension of potentially helpful information, this kind of image can in fact give a far more comprehensive portrayal of the fruit or plant. For instance, a fruit's water content, which cannot be determined by the visible bands alone, may be identified using a near-infrared band. However, it is important to realize that the capacity to analyze the correlations between various bands and how these relate to the traits we aim to assess is equally essential to obtaining the most valuable information. Not every band has the same level of information; some can be quite redundant, while others might offer crucial information for calculating particular chemical components. In the end, the diversity of these bands throughout the electromagnetic spectrum determines how much information an image can hold in addition to the number of bands and spatial resolution. While multispectral

and hyperspectral images with bands evenly distributed throughout the VIS-NIR spectrum can offer a substantially higher amount of pertinent information for comprehending the internal characteristics of a fruit, an image with numerous highly correlated bands may not always be valuable. Even yet, it is still essential to have the knowledge needed to relate the unprocessed data recorded by the various bands to the chemical or physiological characteristics we wish to forecast; this process calls for ongoing experience and a thorough comprehension of the phenomenon.

2.2 Spectral behavior of plants

Understanding how to use multispectral and hyperspectral pictures to evaluate crop maturity and its physicochemical characteristics requires an understanding of the spectral behavior of plants. A distinct "spectral signature" that can be used for agricultural study is produced when electromagnetic waves and plant structure interact to determine reflectance at various spectrum areas. Depending on the wavelengths taken into consideration and the physiological state of the plant, different plants, including grapevines, reflect light in different ways. For instance, chlorophyll is essential in the visible portion of the electromagnetic spectrum. In order to perform photosynthesis, it mostly absorbs red and blue light, reflecting green light, giving the leaves their green appearance to the human eye. This behavior is also seen in grape berries, where the anthocyanin content, which gives the berries their color, varies according to the ripeness level and helps to provide a distinctive spectral signal for red and white grapes. Since reflectance in the near-infrared (NIR) band is not affected by pigmentation but rather by the interior structure of the leaves and berries, such as cell arrangement and water content, it is especially intriguing for spectral study of plants. Near-infrared technology can be used in agriculture to provide vital information on the health of plants, the amount of water present, and the maturity of fruit, particularly grape berries. The relationship between ripeness and spectral changes is especially important for grapes. For example, as grapes develop, the concentration of other pigments, including anthocyanins, rises while the amount of chlorophyll in the skins drops. This alteration causes a "red shift" that can be utilized as a ripeness indicator by decreasing absorption in the red band and increasing reflectance in the near-infrared. Numerous crop studies have documented this phenomena, which has led to the development of vegetation indices like the Normalized Difference Vegetation Index (NDVI), which may also be used in winemaking to track the growth of plants and berries. In order to create predictive models that can assess grape quality from photos, these spectral factors are crucial. We can develop stronger ideas regarding the relationship between the observed spectral features and chemical attributes,

including the amount of sugar and anthocyanins, by knowing how the plant reflects various wavelengths. The key to learning how to use imaging technology to make better agronomic decisions, such when to harvest, is ultimately tracking the spectral activity of grapes from the veraison phase to full ripeness.

2.3 Imaging techniques in agriculture

As we mentioned in earlier explanations, imaging technologies have advanced significantly as a result of the need for accurate, efficient, and non-destructive ways to evaluate the quality of agricultural products. These technologies, which enable the evaluation of quality indicators without causing harm to the product, such as RGB, multispectral, and hyperspectral imaging, have emerged as crucial instruments in precision agriculture. This section lays the groundwork for the machine learning applications that will be discussed in the following chapter by examining the various imaging modalities' uses in agriculture, with an emphasis on their usage in assessing the quality of different fruits.

2.3.1 Overview of imaging modalities

RGB imaging

In agriculture, RGB imaging is frequently used to evaluate the external characteristics of fruits, including their color, size, and form. RGB imaging is a common option for preliminary quality evaluations due to its affordability and ease of use. For example, using Artificial Neural Networks (ANNs) to forecast oil and phenol content and to evaluate overall fruit quality, RGB imaging has been used to identify faults in olive fruits and classify olives according to maturity[16]. However, because RGB imaging only records visible light and thus provides limited spectral information, it is limited in its capacity to assess interior attributes.

Multispectral imaging

A limited number of distinct spectral bands, frequently encompassing visible and near-infrared (NIR) ranges, are captured via multispectral imaging. This makes it especially helpful for evaluating characteristics like water status and chlorophyll levels that are associated with different wavelengths. Fruit firmness and Total Soluble Solids (TSS) can be efficiently characterized using multispectral imaging, which offers important information on the internal quality and maturity of the product. Although it still lacks the full spectrum resolution provided by hyperspectral imaging, the ability to target particular spectral bands enables higher precision than RGB imaging.

Hyperspectral imaging(HSI)

A potent technique for assessing the internal quality of agricultural products is hyperspectral imaging (HSI), which blends spectral and spatial data. HSI provides a comprehensive spectral fingerprint of every pixel by capturing hundreds of contiguous spectral bands, in contrast to RGB or multispectral photography. A more thorough examination of internal quality characteristics like sugar content, acidity, and phenolic chemicals is made possible by this capacity. Using sophisticated models such as SPA-MLR (Successive Projections Algorithm-Multiple Linear Regression) and GAPLS-LS-SVM (Genetic Algorithm Partial Least Squares-Least Squares Support Vector Machine), studies on kiwifruit have shown how well HSI predicts firmness, soluble solids content (SSC), and pH[17]. In a similar vein, mango hardness, TSS, and titratable acidity have all been evaluated using hyperspectral imaging, which maps internal changes during the course of ripening[18].

2.3.2 Application in fruit quality assessment

Kiwifruit: When it comes to kiwifruit, HSI has been utilized to forecast important quality parameters like pH, firmness, and SSC. In order to create predictive models that were highly accurate in calculating these values, the study examined a variety of spectral bands, including the visible and NIR ranges. The ability of HSI to non-destructively assess the interior quality of kiwifruit was demonstrated by the application of techniques like SPA-MLR and GAPLS-LS-SVM[17].

Mango: Hyperspectral imaging has also been used to map the physicochemical properties of mangoes. Firmness, TSS, and titratable acidity were all predicted using HSI, which shed light on how these characteristics change within the fruit as it ripens. Determining the ideal harvest time and guaranteeing fruit quality depend on a thorough examination of these changes, which is made possible by HSI's capacity to record extensive spectral data[18].

Olive: When it comes to olives, RGB imaging in conjunction with ANNs has been used to identify surface flaws, classify olives according to age, and evaluate crucial quality attributes like oil and phenol content. This method shows how even more basic imaging methods can yield useful information about fruit quality when paired with machine learning models[16].

Strawberry and Other Fruits: In order to identify physical characteristics including color, size, and surface flaws, image processing techniques were utilized in the effective application of RGB imaging to strawberries. In spite of its drawbacks, RGB imaging can be a very useful tool for external quality assessment when backed by strong computational models, as the study's excellent accuracy in quality assessment shows[19].

Banana and Grapes: Other fruits, such bananas and grapes, have also been imaged using hyperspectral and RGB techniques. By examining spectral properties,

HSI was utilized to assess the maturity and interior quality of bananas[20]. RGB imaging was used to evaluate the physical characteristics of grapes, such as size and color, which are markers of quality and maturity[21].

2.3.3 Advantages and challenges of imaging technique

Advantages: Every imaging modality has special benefits. RGB imaging is appropriate for external quality evaluations when color and shape are the main factors because it is affordable and simple to use. With its capacity to target particular spectral bands, multispectral imaging offers more details about particular quality features, such moisture content or chlorophyll. Because of its extensive spectral coverage, hyperspectral imaging is unique in that it may identify minute variations in chemical composition that are invisible using conventional techniques. Because of this, HSI is especially effective in evaluating internal characteristics that are crucial for judging the quality of grapes and other fruits, such as sugar and acid concentration.

Challenges and limitations: Every imaging method has drawbacks despite its benefits. Since RGB imaging is limited to visible wavelengths, it is unable to reveal interior quality factors that are not apparent from the outside. Despite providing more information than RGB, multispectral imaging is still constrained by the amount of spectral bands it can record. Despite being incredibly informative, hyperspectral imaging has several drawbacks, such as high processing demands, massive data quantities, and the requirement for specialist hardware. Additionally, variations in fruit attributes like size, shape, and maturity can make analysis more difficult and necessitate sophisticated preprocessing and calibration methods in order to get accurate results[22]. Furthermore, standardized protocols and better data handling techniques are required to make hyperspectral imaging more dependable and accessible for broad agricultural use, according to a systematic review of recent applications([23], [24]).

2.3.4 Conclusion link to machine learning application

In conclusion, imaging methods are essential for the non-destructive evaluation of fruit quality since they provide a variety of instruments for analyzing both internal and exterior characteristics. While multispectral and hyperspectral imaging offer increasingly more detailed and informative data, RGB imaging is appropriate for fundamental evaluations. Building on these imaging methods, the next chapter will concentrate on how machine learning models, specifically, regression techniques, can use the abundant data from hyperspectral imaging to forecast crucial quality indicators, such as grape sugar content. With the ultimate goal of optimizing harvest time and enhancing grape quality assessment through sophisticated data

analytics, this shift from imaging technology to predictive modeling will demonstrate the full potential of these tools in precision viticulture.

Chapter 3

Machine learning for sugar grape prediction

3.1 Introduction

In the contemporary wine industry, grape quality is a crucial aspect that directly affects the organoleptic characteristics of the wine and its market value. Chapters 1 and 2 emphasize that the sugar concentration in grape berries is a crucial factor for establishing the ideal harvest time, affecting the winemaking process and the wine's final attributes. The evaluation of grape maturity and sugar concentration has conventionally been conducted by damaging and labor-intensive techniques that necessitate hand sampling and laboratory examination. These methods, while precise, possess considerable constraints regarding time, money, and representativeness, hence complicating broad and continual vineyard monitoring. Chapter 2 examined advanced imaging techniques, specifically multispectral and hyperspectral imaging, which provide novel avenues for the non-destructive assessment of qualitative crop metrics. We have emphasized that these techniques, utilized in precision agriculture, can yield comprehensive data on the chemical composition of fruits, facilitating more efficient and sustainable oversight. This chapter is to critically analyze the current research that has utilized machine learning techniques on multispectral images to forecast the sugar concentration in grapes. We will conduct a comprehensive examination of the used methodology, utilized datasets, and achieved results to ascertain the factors contributing to the success or limitations of specific strategies. Additionally, we will examine the issues associated with viticultural data, including genetic and environmental variability, as well as the consequences of mathematical assumptions in prediction models. This chapter will establish a robust basis for the experimental work detailed in Chapter 4, in which diverse machine learning approaches will be employed and evaluated for predicting the Brix index from

multispectral pictures.

3.2 Understanding the choices

As already discussed in previous reflections, this thesis primarily aims to serve as a means to deeply understand all the concepts, choices, and assumptions, whether voluntary or involuntary, that underlie the methodologies integrating machine learning for predicting grape quality parameters, particularly sugar content. In fact, although the foundation is always training a machine learning model, the methodologies can be very different. It will then be the chosen methodology that will determine the usability, reliability, and performance of the system. In doing so, it is important to keep in mind that the generally pursued objective is to find a system that effectively estimates the average sugar content of the grapes in a vineyard to help winemakers decide on the day of the harvest. To provide a more concrete idea, let's use two cases that could be real and compare them. In both cases, we have a machine learning model that requires a multispectral vector as input and returns a value corresponding to the sugar content expressed in Brix degrees as output. To understand which entities correspond to the input multispectral vector and the predicted sugar content, we need to add some information, which will be different in the two cases. In the first case, a sampling of grapes from a single variety v1 was taken from a single vineyard t1 in the year a1. 20 grapes were collected each day for 20 days, in different parts of the vineyard that are more or less sunny and at different altitudes, during the period from the beginning of the ripening phase until the day of the harvest. At the end of the procedure, we will have a sample of 400 grape berries that will ideally cover a wide range of sugar content. After sampling, through appropriate methodologies and instrumentation, an average hyperspectral vector in the range of 400 nm - 1000 nm (VIS-NIR) is extracted from each berry and its sugar content is measured. The method by which the average multispectral vector is obtained will be analyzed in more depth in the following sections; for now, it is sufficient to know that, through the use of specific instruments, the multispectral image of the grape is obtained, which is then averaged to derive a multispectral vector. This multispectral vector tells us how the grape berry reflects incident light at different wavelengths in the reference spectrum. As already mentioned in Chapter 2, we know that these reflectances contain information regarding some characteristics, both visible and invisible, of the grape. Finally, with the obtained data, a machine learning model M is trained to predict the sugar content of the grape berry using a multispectral vector, achieving performance P1. In the second case, a sampling of grapes from 5 different varieties v21, v22, v23, v24, v25 was taken from two different vineyards t21 and t22, in two different years a21 and a22. The grapes of the v21, v22, and v23 varieties

are red and were sampled from the t21 vineyard, while the grapes of the v24 and v25 varieties are green and were sampled from the t22 vineyard. In this case, the sampling occurs differently. For each grape variety, 10 bunches are harvested per day from different parts of the vineyard over the 20 days leading up to the harvest day, exactly as in case 1. After sampling, 6 berries are selected from each cluster, from different parts of the cluster, representative of it. With a system analogous to case 1, the multispectral images of the 6 berries are obtained and then averaged to derive a multispectral vector that represents the entire bunch. This multispectral vector is associated with the average sugar content of the 6 berries. Finally, with these data, the same machine learning model M is trained, achieving performance P2. Now that we have provided context, we will separately consider the differences between cases 1 and 2 in terms of model generalization, usability, complexity, and the intrinsic loss or risk of loss of information.

Generalization and Complexity

Regarding this aspect, the two cases are very different from each other and are based on more or less strong assumptions. In the first case, we are training the model to predict the sugar content of a single grape variety from a single vintage. On the contrary, in the second case, the goal is to generalize the model to more grape varieties, both red and white, across different vintages. What can we expect from the two models and what are the conceptual differences underlying the two systems? When we train a model to predict a dependent variable Y from an independent variable X, the general and basic assumption we are making is that Y is a random variable generated by a distribution $f(X, w) + \varepsilon$ (random error), where X is also a random variable. What the discipline of machine learning does is assume a function f and estimate the parameters w of the function with the real observations of X and Y. Having this aspect clear, we can understand the fundamental differences between the two cases. In case 1, I am practically assuming that the sugar content of the grapes of variety v1 can be estimated based on its spectral vector with a certain margin of error. In the second case, I am making the same assumption but simultaneously for 5 different varieties, collected over two years in 5 different vineyards. This in practice amounts to assuming that if I take two samples from two different years, in two different vineyards, and of two different qualities, their sugar content can be estimated from their spectral vectors using the same function, always with a certain margin of error. The intention of these reflections is not to delve into the truthfulness or validity of these assumptions, but rather to understand how they plausibly influence the performance of the models and, above all, their usability. To give a more concrete idea, it is useful to think that in the first case I am assuming, without knowing it a priori, that the grapes of the variety v1, harvested in the vintage a1, constitute a group, and that the

elements of this group (all the individual berries or bunches) are somehow similar to each other. In the second one, I am making the same assumptions about a much larger group. Although no one can know the truth of the assumptions, since by their nature they are taken as true without being verified, we can still try to reason probabilistically about their degree of truthfulness. We can indeed assert that the probability of there being parameters and indicators that, when compared, reveal that the entities of the group are actually similar to each other is greater in the first case than in the second. This is because in the second case we increased the degree of complexity by bringing together a greater number of entities that, based on our experience, have a certain degree of dissimilarity (different vintages, different varieties, different vineyards). All these reflections lead to the conclusion that the assumptions underlying model 1 are stronger and have a higher degree of truthfulness than those underlying model 2. For this reason, we expect model 1 to perform better than model 2, meaning it will be able to better predict the sugar content of the type of grape it was trained on. At this point, keeping in mind that the ultimate goal is to estimate the average sugar content of the grapes in a vineyard, it is important to understand how applicable the two models are in reality. If it is true that model 1 is likely more accurate in its estimates on data similar to those on which it was trained, it is also true that it manages a much lower degree of variability and that a hypothetical system estimating the harvest day by integrating model 1 will be less universal. On the contrary, although it is true that model 2 is likely less performant, a system built on the basis of this model will be far more usable because it is intrinsically designed to handle variability regarding regions, years, and grape varieties. The objective of the previous reflections is to draw attention to the importance of carefully evaluating all factors directly and indirectly involved in processes of this type, without underestimating the complexity. It is of fundamental importance to understand that the quality of the grape is subject to a large number of variables such as soil quality, climate, plant health, and the variety itself, and that these variables can be very different if certain conditions are changed such as the vintage, the vineyard's location, the position of the individual vine, and the position of the individual bunch. It will therefore be essential to find a compromise between the robustness of the model, its practical usability, and the quality of its estimates. In the next section, we will focus on the concept of "information loss".

Information loss

Another aspect that the two cases allow us to evaluate is related to the difference in the significance that my data represents within the dataset. In both cases, in fact, the single spectral vector within the dataset and the corresponding sugar index represent two different entities. In the first case, we said, the spectral vector is

obtained by averaging all the spectral vectors of the pixels in the hyperspectral image of the berry; in the second case, the spectral vector is obtained by averaging the spectral images of 6 different berries belonging to the same bunch, taken from different parts of the bunch to ensure the greatest possible variability. Although in both cases the data with which we train the model have the same structure, they actually represent very different entities and therefore do not capture the same type of information. In the first case, the spectral vector and the sugar index represent a specific grape berry, while in the second case, they represent the entire bunch. In light of these facts, it is natural to wonder what kind of predictions the machine learning model is intrinsically structured to make if it is trained with one dataset or the other. Encoding an entity through data is, in itself, a process that leads to a certain loss of information depending on the type of data we use and the entity to be encoded. The lost information will carry more or less weight depending on how connected it is to the final information we want to estimate, in our case the sugar content of the grape. Understanding these arguments leads us to the idea that when I represent an entire bunch of grapes with a single spectral vector, I am losing more information compared to when I represent a single grape. This is because in the second case, I am materially averaging 6 different spectral images and 6 different sugar degree values to construct a single data point related to the bunch. I am therefore performing two distinct operations, namely using 6 berries from the bunch as a proxy for the bunch itself and then averaging the respective spectral images and sugar contents. Each of these operations carries a certain loss of information. In the first case, the loss of information will definitely be smaller, because I am averaging vectors related to the pixels of the same grape. The variability of these vectors will be relatively low, and therefore the average vector will be a more accurate representation of the entire image. All these considerations are fundamental when we test models on new data and interpret their results, because they already tell us what to expect and what the model can and cannot do. Returning to the two hypothetical cases we are analyzing, the first model will be trained to predict the sugar content of a grape berry from its spectral vector, so it will be a model that from a certain point of view we can define as pointwise. The second model, on the other hand, will be inherently trained to predict an average sugar content related to multiple entities, and therefore cannot be used to accurately predict the sugar content of a single grape. This means that if I take a grape, calculate its spectral vector, and estimate the sugar content using model 2, I cannot expect that to be an accurate estimate of the sugar content of the grape, but rather an estimate of the average sugar content of the bunch from which the grape comes. If we now imagine a real scenario, where we have a mobile system moving through a vineyard and integrating a camera capable of capturing spectral images, model 2 will certainly be the most suitable for estimating the average sugar content of the vineyard. This is because we can imagine that

the mobile system will capture images of entire clusters, which can then be used to estimate their average content and finally the average of all the clusters. If we used model 1, we would likely encounter a greater error, because we would estimate the sugar content of the bunches in a precise manner as if they were individual berries, and we would not have a representative value for the entire bunch. If we imagine having more time and resources to collect a representative sample of grapes from the entire vineyard and to individually acquire their spectral vectors, the integration of model 1 into the system would certainly be the most appropriate. We have therefore seen how the initial choices and assumptions are inextricably linked with the integration of models into real processes. Even in this case, it is important to reach compromises between performance, costs, time, data quantity, and specific objectives. These reflections highlight the intricate balance between model performance, generalization, and practical applicability in the context of grape sugar content prediction. Understanding the impact of methodological choices on the usability and reliability of machine learning models is crucial. To further explore how these considerations manifest in real-world applications, we will now review existing studies that have applied hyperspectral imaging and machine learning techniques in viticulture.

3.3 Critical Review of the Literature

3.3.1 State-of-the-Art Research Review

This section offers a thorough critical analysis of previous research that has used machine learning and hyperspectral imaging to forecast grape sugar content and other quality metrics. Understanding the methods used, the presumptions made, the results obtained, and the applicability of the models across various vintages, varieties, and environmental circumstances are the main points of emphasis.

The use of hyperspectral imaging and machine learning algorithms to forecast enological parameters including grape sugar, anthocyanin, and flavonoid content has been investigated in a number of research. Among these, Gomes et al. [25] carried out an extensive investigation contrasting several machine learning techniques, such as Neural Networks (NN) and Partial Least Squares Regression (PLSR), for forecasting the sugar content of Port wine grape berries. Using hyperspectral data obtained in reflectance mode from samples consisting of just six whole berries, they created prediction models. Data from the 2012 vintage was used to train the models, and samples from the 2012 and 2013 vintages were used for testing. A rare consideration in the literature, this method assesses the model's generalizability to vintages not used in model building.

With R^2 values of 0.93 and 0.92, respectively, the findings demonstrated that the RMSE values for the test set containing 2012 samples were 0.94 °Brix for PLSR

and 0.96 °Brix for NN. The RMSE values for PLSR and NN climbed to 1.34 °Brix and 1.35 °Brix, respectively, when test data with 2013 samples were used. This suggests that performance declined while predicting on a new vintage. Nonetheless, both models maintained a high degree of correlation between predicted and actual values, as indicated by the R^2 values, which stayed high at 0.95 for PLSR and 0.92 for NN.

The RMSE is a standard way to measure the error of a model in predicting quantitative data, calculated as the square root of the average of squared differences between predicted and observed values. The R^2 value indicates the proportion of variance in the dependent variable that is predictable from the independent variables, ranging from 0 to 1, with higher values indicating better model fit.

In a later study, Gomes et al. [26] extended their investigation by including other machine learning techniques, like Convolutional Neural Networks (CNN) and Ridge Regression (RR), to forecast the amount of sugar in Port wine grape berries. The vast dataset included samples from three grape varietals that are commonly utilized in the production of Port wine: Tinta Barroca (TB), Touriga Franca (TF), and Touriga Nacional (TN). Due to variations in terroir, climate, and grape maturation stages, samples were gathered from 2012 to 2018 for TF and from 2013 to 2017 for TN and TB. This allowed for the capture of a broad range of variability.

The models were tested on separate test sets that included TN and TB samples (different varieties), as well as TF samples (same variety as training), after being trained on TF samples from all vintages. The CNN model performed better than the other approaches, according to the results, obtaining the lowest RMSEP values in both the test and validation sets. In particular, CNN's RMSEP for the TF independent test set was 0.97 °Brix, whereas NN's, RR's, and PLSR's were 1.14 °Brix, 1.45 °Brix, and 1.47 °Brix, respectively. With RMSEP values of 1.15 °Brix for TN and 1.31 °Brix for TB, which were much lower than those attained by the other models, the CNN model once again demonstrated superior generalization ability when tested on various grape varietals (TN and TB).

Silva et al. [27] examined the prediction of anthocyanin concentration, pH index, and sugar content in whole grape berries using Support Vector Regression (SVR) in conjunction with hyperspectral imaging. Three grape varieties, TF, TN, and TB, collected throughout several vintages were included in the samples. Reflectance mode was used to gather hyperspectral data in the 380—1028 nm range. To evaluate the SVR models' capacity for generalization, they were evaluated on TF samples as well as the other two kinds after being trained on TF samples. On the TF test set, the SVR model's R^2 value for sugar content prediction was 0.96, and its RMSE was 0.80 °Brix. The model performed well when evaluated on TN and TB samples, with R^2 values of 0.90 and 0.89 and RMSE values of 3.19 °Brix for both types. These findings imply that SVR in conjunction with hyperspectral imaging can yield reliable forecasts for a range of grape types and vintages.

The viability of employing hyperspectral imaging to track grape maturity in the field under natural lighting conditions was investigated by Benelli et al. [28]. Thirteen separate days during the pre-harvest and harvest periods were analyzed for the study, which concentrated on the 'Sangiovese' grape variety. In the visible and near-infrared (Vis/NIR) region (400–1000 nm), hyperspectral data were gathered, and a portable digital refractometer was used to determine the soluble solids content (SSC). With a R^2 value of 0.77 and an RMSECV of 0.79 °Brix, PLSR was used to predict SSC. PLS-DA was also utilized to classify the samples into 'ripe' (SSC $\geq$ 20 °Brix) and 'not-ripe' (SSC $<$ 20 °Brix) classes, with 86% to 91% correct classification rates. The study showed how hyperspectral imaging may be used for non-destructive in-field grape maturity monitoring, offering a useful method for choosing the best time to harvest.

The potential of hyperspectral imaging for defining table grapes according to their sugar content (Total Soluble Solids or TSS), total flavonoid content (TF), and total anthocyanin content (TA) was assessed by Gabrielli et al. [29]. Seven table grape varietals, including both white and red grapes, were used to get hyperspectral pictures in the visible and short-wave near-infrared range (411–1000 nm). To improve the quality of the spectral data, a number of data preprocessing techniques were used, including white and dark adjustments, first and second derivatives, and Standard Normal Variate (SNV). To predict TSS, TF, and TA, PLSR models were created utilizing the entire spectral range.

Gabrielli et al. employed regression coefficients (β-coefficients) and Variable Importance in Projection (VIP) scores to choose the best wavelengths in order to decrease the complexity of the data and increase computational efficiency. These chosen wavelengths were then used to build Multiple Linear Regression (MLR) models. The findings showed that both the calibration and validation sets' R^2 values were high for the PLSR models. The SNV-pretreated PLSR model for TSS prediction, for example, had an RMSEV of 1.1 g/100 g and R^2_{cal} and R^2_{val} of 0.94 and 0.91, respectively. With smaller data sets and lower processing demands, the MLR models that used the best wavelengths also demonstrated strong predictive performance. This study demonstrates how well hyperspectral imaging predicts important quality metrics in table grapes when paired with suitable preprocessing and variable selection techniques.

In order to forecast the sugar content of grape berries for agronomic applications, particularly grape maturity monitoring, Courand et al. [30] examined the use of a robust regression approach, RoBoost-PLSR. Three grape varieties, Syrah, Fer-Servadou, and Mauzac, were the subject of the investigation. Densimetric baths were used to quantify reference sugar levels, and hyperspectral pictures were obtained in the VIS-NIR region. Predictive performance may suffer from traditional PLSR models' sensitivity to outliers. In order to limit the impact of outliers in the calibration set, the authors used the RoBoost-PLSR technique, which combines

boosting methods with PLSR.

According to the findings, the RoBoost-PLSR models performed better than the conventional PLSR models for every grape variety. The RoBoost-PLSR model produced an RMSEP of 3.14 g/L and a R_p^2 of 0.990 for the Syrah variety, while the regular PLSR model produced an RMSEP of 5.36 g/L and a R_p^2 of 0.971. Similar enhancements were noted for the Mauzac and Fer-Servadou types. For practical applications in viticulture, where data quality can vary, the study highlights the value of robust regression techniques in boosting model dependability and prediction model accuracy in the presence of outliers.

A new ground truth multispectral image dataset of grape berries was presented by Navarro et al. [31]. It included weight, anthocyanin content, and Brix index values. 1,238 multispectral photos of five different grape varieties, Autum Royal, Crimson, Itum4, Itum5, and Itum9, make up the dataset. Detailed ground truth data is included with every image, making it an invaluable tool for creating and evaluating machine learning algorithms for use in agricultural applications. In order to categorize grape types based on multispectral photos, the authors used this dataset to train machine learning models, such as Multilayer Perceptron (MLP) and three-dimensional Convolutional Neural Networks (3D-CNN). Both models demonstrated the dataset's efficacy for supervised learning tasks by achieving 100% accuracy.

Attempts to create regression models to forecast continuous variables like anthocyanin content and Brix index were less effective, despite the excellent classification accuracy. Accurate predictive model development was hampered by the unequal distribution of these factors and the small sample numbers for some classes. The potential of extensive datasets to advance machine learning applications in agriculture was emphasized by Navarro et al. The availability of the dataset makes it easier to create sophisticated algorithms for classifying fruits and evaluating their quality, which supports precision farming methods. The study also emphasizes the need for larger and more balanced datasets by highlighting the difficulties in forecasting continuous variables in agricultural datasets.

3.3.2 Other relevant studies

To broaden the scope of this review, we briefly highlight several additional studies that apply advanced imaging and machine learning methods for grape quality assessment without delving into detailed analysis, as the previously reviewed works already cover a comprehensive range of approaches and insights.

- **3DeepM: An Ad Hoc Architecture Based on Deep Learning Methods for Multispectral Image Classification** by Navarro et al. [32]. This work presents a deep learning model named *3DeepM*, specifically developed for the

categorization of multispectral grape images. Employing 3D convolutional layers, 3DeepM effectively extracts spatial-spectral features from multispectral data, attaining 100% classification accuracy for grape varieties while significantly reducing the parameter count relative to models such as AlexNet and ResNet, thus rendering it appropriate for real-time agricultural applications.

- **In situ grape ripeness estimation via hyperspectral imaging and deep autoencoders** by Tsakiridis et al. [33]. The authors investigate non-destructive estimate of grape ripening by hyperspectral imaging and deep autoencoders to mitigate the effects of fluctuating illumination in field environments. The research, concentrating on predicting sugar content for selective harvesting, reveals that deep convolutional autoencoders surpass fully connected autoencoders in accuracy for ripeness estimate.

- **Developing deep learning based regression approaches for prediction of firmness and pH in Kyoho grape using Vis/NIR hyperspectral imaging** by Xu et al. [34]. This study use stacked autoencoders to forecast grape firmness and pH from Vis/NIR hyperspectral pictures, surpassing conventional techniques like PLS and LSSVM. The research illustrates the efficacy of stacked autoencoders in handling high-dimensional hyperspectral data and improving non-destructive quality evaluation for post-harvest applications.

- **Determination of anthocyanin concentration in whole grape skins using hyperspectral imaging and adaptive boosting neural networks** by Fernandes et al. [35]. This study introduces a novel application of AdaBoost neural networks for the non-destructive assessment of anthocyanin concentration in grape skins, attaining a moderate correlation ($R^2 = 0.65$) and demonstrating potential for quality control applications utilizing hyperspectral data.

- **Brix, pH, and anthocyanin content determination in whole Port wine grape berries by hyperspectral imaging and neural networks** by Fernandes et al. [36]. This research utilizes hyperspectral imaging and neural networks to concurrently forecast essential enological parameters, attaining high accuracy (R^2 values of 0.73 for pH, 0.92 for sugar content, and 0.95 for anthocyanins) and showcasing the method's effectiveness for swift, non-destructive evaluation of grape quality.

- **Soluble solids content and pH prediction and variety discrimination of grapes based on visible–near infrared spectroscopy** by Cao et al. [37]. This study utilized Vis–NIR spectroscopy alongside genetic algorithms for feature selection, attaining a 96.58% accuracy rate in differentiating grape varieties and producing reliable predictions for SSC and pH, underscoring the

efficacy of Vis–NIR spectroscopy as a non-destructive, cost-efficient quality assessment instrument.

These studies collectively underscore the diverse potential of hyperspectral and multispectral imaging, deep learning, and machine learning techniques in viticulture, from real-time classification in field conditions to non-destructive measurement of quality metrics and varietal discrimination. The breadth of methodologies and success across various grape quality parameters reinforce the applicability of these technologies in advancing precision agriculture and grape quality management.

3.3.3 Techniques Employed

The papers under consideration used a variety of approaches that combined machine learning algorithms with hyperspectral imaging. Gomes et al. [25, 26] and Silva et al. [27] used machine learning models such as CNN, NN, SVR, RR, and PLSR in conjunction with hyperspectral imaging in reflectance mode. They concentrated on evaluating the models' capacity to generalize across various grape varietals and vintages. To increase computational efficiency, Gabrielli et al. [29] selected variables using regression coefficients and VIP scores. The RoBoost-PLSR method was created by Courand et al. [30] to successfully handle outliers. Benelli et al. [28] used PLSR and PLS-DA for prediction and classification tasks while doing in-field hyperspectral imaging in natural light.

3.3.4 Hyperspectral Imaging Techniques and General Sampling Methods

Hyperspectral image acquisition and grape berry sampling are crucial procedures in the reviewed studies that have a big impact on the caliber and usefulness of the prediction models created. Gaining knowledge of these techniques helps one to better understand the difficulties in obtaining representative data as well as the possible variability that may be introduced during data collecting. Choosing berries or clusters that best reflect the variety found in a vineyard or experimental setup is the usual procedure for grape sampling. To guarantee that the samples reflect the variability required for reliable model building, factors including grape variety, maturity stage, vineyard location, and environmental circumstances are taken into account. As an example, Gomes et al. [25, 26] gathered samples from several grape types and vintages, resulting in a broad range of variability brought about by variations in terroir and climate. In a similar vein, Silva et al. [27] evaluated the generalization capacity of their models by incorporating samples from three wine types gathered across several vintages. Using specialized cameras and sensors that record spectral data across a broad range of wavelengths is

necessary to acquire hyperspectral images. These studies' hyperspectral imaging devices work in the visible to near-infrared (Vis/NIR) spectrum, usually between 380 and 1028 nm or 400 and 1000 nm. As demonstrated by Benelli et al. [28], the imaging procedure can be carried out immediately in the field in natural light or in laboratory settings under controlled lighting conditions. Hyperspectral data must be preprocessed in order to improve data quality and lower noise. First and second derivatives, calibration using white and dark references, and Standard Normal Variate (SNV) correction are examples of common preprocessing methods. These procedures aid in adjusting for baseline shifts, scattering effects, and other environmental or instrumental factors that could have an impact on the spectral data. According to Gabrielli et al. [29], preprocessing data is crucial for enhancing the prediction capabilities of their models. Following the acquisition and preprocessing of the hyperspectral pictures, spectral characteristics are retrieved for use as machine learning model input. Principal Component Analysis (PCA) and other dimensionality reduction techniques are sometimes used to focus on the most informative spectral bands and minimize the number of variables. In order to improve computational efficiency and model interpretability, variable selection techniques are also used to pick wavelengths that contribute most significantly to the prediction models. These techniques include regression coefficients and Variable Importance in Projection (VIP) ratings. Interpreting the findings of these investigations and evaluating the generalizability of the created models require an understanding of the sampling strategies and hyperspectral imaging techniques. Model performance may be impacted by the variability introduced during sampling and data collection, underscoring the necessity of rigorous experimental design and uniform procedures in subsequent studies.

3.3.5 Model Generalization Analysis

The assessment of model generality over various vintages, grape types, and climatic circumstances is a crucial component of these investigations. CNN models had the best generalization ability, according to Gomes et al. [25, 26], who showed that models could generalize to new vintages and different grape varietals. Good generalization was demonstrated by Silva et al. [27], who demonstrated that SVR models trained on one grape variety could accurately predict sugar content in other varieties. Benelli et al. [28] demonstrated that accurate predictions could still be made using hyperspectral imaging in changeable field circumstances. These results imply that hyperspectral imaging in conjunction with machine learning can provide prediction models that generalize effectively across various situations, provided that the right data preprocessing and model selection are implemented.

3.3.6 Implications for Real-World Uses

Together, the reviewed research highlight the potential of machine learning and hyperspectral imaging for quick, non-destructive evaluation of grape quality criteria. According to Gabrielli et al. [29], cutting down on wavelengths helps preserve high accuracy while increasing computing efficiency, which is essential for industrial applications. According to Courand et al. [30], robust regression techniques are crucial for managing data variability and enhancing model reliability. Despite difficulties in predicting continuous variables because of problems with data distribution, Navarro et al. [31] offered a useful dataset for creating and evaluating machine learning algorithms.

Silva et al. [27] and Benelli et al. [28] highlighted the usefulness in real-world situations. Benelli et al. proved that in-field measurements are feasible, which is essential for making harvest decisions in real time. Silva et al. demonstrated how models may generalize between vintages and kinds, minimizing the requirement for retraining. These results have real-world applications in the grape and wine industries for quality control, sorting procedures, and harvest scheduling considerations.

However, factors like data requirements, processing complexity, and the demand for calibration under various climatic conditions must be taken into account. Future studies should concentrate on streamlining pipelines for data collection and processing, growing datasets to encompass more vintages and varieties, refining data preprocessing methods, and investigating cutting-edge machine learning approaches that can manage imbalance and unpredictability in agricultural data.

3.3.7 Conclusion

Significant progress has been made in forecasting grape quality metrics using hyperspectral imaging and machine learning, as demonstrated by the examined literature. Despite advancements, there are still difficulties in creating models that accurately represent various vintages, cultivars, and environmental circumstances. In order to improve the technology' practical application in the agricultural industry, future research should focus on addressing these issues.

3.4 Mathematical and Methodological Assumptions in Machine Learning Models

For applications like grape sugar content prediction, where environmental variability and grape phenotypic diversity present particular challenges, choosing the right machine learning algorithm is essential to obtaining accurate and generalizable results in the field of predictive modeling using hyperspectral imaging data. With an

emphasis on Partial Least Squares Regression (PLSR) and its connection to Multiple Linear Regression (MLR), this section explores the mathematical underpinnings of the main regression models utilized in the examined studies. To give a thorough grasp of their suitability, advantages, and disadvantages in viticultural applications, we also investigate other models such as Support Vector Regression (SVR), Ridge Regression (RR), and neural networks, especially Convolutional Neural Networks (CNNs).

3.4.1 Multiple Linear Regression (MLR)

A fundamental statistical technique for modeling the linear relationship between a dependent variable Y and several independent variables $X_1, X_2, \ldots, X_p$ is multiple linear regression (MLR). The expression for the MLR model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{3.1}$$

The intercept term is represented by β_0, the regression coefficients that show how each independent variable affects Y are represented by $\beta_1, \beta_2, \ldots, \beta_p$, and the error term is represented by ε, which is assumed to be normally distributed with zero mean and constant variance, $\varepsilon \sim N(0, \sigma^2)$.

Linearity, residual independence, homoscedasticity, residual normality, and the lack of multicollinearity among predictors are among the presumptions of MLR. When used to hyperspectral data in viticulture, where spectral bands show strong collinearity because of the tiny wavelength intervals, these assumptions, while simple, become restrictive. Large variations in the estimated regression coefficients and, as a result, reduced interpretability and predictive power might result from this multicollinearity problem, which can destabilize MLR. Furthermore, hyperspectral datasets frequently have a large number of predictors, which results in a high-dimensional environment where there may be more predictors than observations. This situation decreases computing efficiency and raises the possibility of overfitting, in which the model learns noise in the data instead of actual signal patterns.

MLR's incapacity to manage collinearity well frequently leads to biased predictions in applications such as grape sugar content prediction, where high-dimensional and collinear hyperspectral data are common. Therefore, more reliable techniques like Partial Least Squares Regression (PLSR) are favored since they can extract crucial predictive data while addressing multicollinearity-related problems.

3.4.2 Partial Least Squares Regression (PLSR)

By combining the advantages of Principal Component Analysis (PCA) and MLR, PLSR overcomes the drawbacks of MLR in high-dimensional, collinear datasets. For hyperspectral data applications like grape sugar content prediction, where

spectral data from a variety of wavelengths needs to be compressed into useful predictive characteristics, this hybrid approach is especially well-suited.

In order to maximize the covariance between $\mathbf{X}$ and $\mathbf{Y}$, PLSR projects the predictors $\mathbf{X}$ and the response $\mathbf{Y}$ onto a new set of latent variables (or scores). PLSR breaks down a responsive matrix $\mathbf{Y} \in \mathbb{R}^{n \times q}$ and a predictor matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\top + \mathbf{E} \tag{3.2}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^\top + \mathbf{F} \tag{3.3}$$

where $\mathbf{T} \in \mathbb{R}^{n \times a}$ is the matrix of latent variables (scores), $\mathbf{P} \in \mathbb{R}^{p \times a}$ and $\mathbf{Q} \in \mathbb{R}^{q \times a}$ are the loading matrices for $\mathbf{X}$ and $\mathbf{Y}$ are the residual matrices.

Weight vectors $\mathbf{w}_i$ and $\mathbf{c}_i$ that maximize the covariance between $\mathbf{X}_{i-1}\mathbf{w}_i$ and $\mathbf{Y}_{i-1}\mathbf{c}_i$ are found iteratively by the PLSR algorithm. After computing the latent scores $\mathbf{t}_i$ and $\mathbf{u}_i$, loadings $\mathbf{p}_i$ and $\mathbf{q}_i$ are calculated. The procedure is repeated for every component as the residual matrices $\mathbf{X}_i$ and $\mathbf{Y}_i$ are updated. The following formula yields the final regression coefficients:

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^\top\mathbf{W})^{-1}\mathbf{Q}^\top \tag{3.4}$$

where the weight matrix for each component is represented by $\mathbf{W}$. After that, the response predictions are acquired by:

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B} \tag{3.5}$$

PLSR offers a number of benefits, such as interpretability through latent variables, computational efficiency, dimensionality reduction capabilities, and multicollinearity handling. Nevertheless, resilience may be impacted by its drawbacks, which include its linearity assumption and sensitivity to outliers. In viticultural applications, these factors are crucial since sugar content forecasts must take into consideration variations in grape types, maturation stages, and weather circumstances.

Because it may simplify hyperspectral data by detecting latent structures that contain crucial spectral information linked to sugar concentration, PLSR is frequently chosen in grape quality prediction experiments, including those covered in Section 3.3. To further increase PLSR's adaptability for grape quality monitoring, Courand et al. [30] showed that even little adjustments to conventional PLSR, such RoBoost-PLSR, might greatly increase robustness against outliers.

3.4.3 PLSR and MLR Comparison

Despite being linear models designed to predict a response variable, PLSR and MLR handle predictors very differently. Due to multicollinearity, MLR struggles in

high-dimensional environments characteristic of hyperspectral data and necessitates separate predictors. In contrast, predictors are converted into latent variables using PLSR in order to optimize their covariance with the response. PLSR's coefficients are linked to latent variables, which may make interpretation more difficult but offer a more trustworthy predictive framework in collinear datasets than MLR's, which explicitly depict correlations between predictors and the answer.

PLSR has been widely used in the reviewed studies due to its effectiveness in managing high-dimensional, collinear data in grape analysis. Scholars such as Gomes et al. [25] emphasized that PLSR outperformed MLR in sugar level prediction because of its structure, which enables it to take advantage of intricate spectral band covariations that are missed by more straightforward MLR models.

3.4.4 Alternative Models for Nonlinear Relationships

Although PLSR gets around a lot of MLR's drawbacks, it still makes the assumption that latent variables and the response have a linear connection. Other models, such as Support Vector Regression (SVR) and neural networks, provide extra possibilities in viticulture, where sugar concentration may be dependent on nonlinear relationships among spectral parameters.

Support Vector Regression (SVR)

When there is nonlinearity in the interactions between predictors and the response, Support Vector Regression (SVR) offers a reliable substitute for linear models. By mapping input data into a higher-dimensional space where linear regression is applied using kernel functions, SVR allows the model to identify intricate patterns that linear approaches might overlook. The SVR optimization method improves robustness to noise and outliers by minimizing a loss function while maintaining predictions within an epsilon-insensitive margin.

When data variability is substantial or there is a nonlinear relationship between spectral characteristics and sugar content, SVR is useful for grape sugar prediction. SVR's performance is dependent on hyperparameters like the kernel function and regularization constant, which need to be adjusted for every dataset, and it can be computationally demanding, especially for big hyperspectral datasets.

Neural Networks: Focus on Convolutional Neural Networks (CNNs)

A subclass of neural networks called Convolutional Neural Networks (CNNs) has become popular in hyperspectral imaging because of its ability to represent intricate, nonlinear relationships. Because CNNs can identify spectral and spatial patterns across several bands, they are especially useful for image data. CNNs are capable of processing large amounts of spectral data for grape quality prediction, identifying

complex wavelength correlations that more straightforward models like PLSR can miss.

1. **Data Requirements:** Given the complex patterns found in hyperspectral data, CNNs require sizable datasets to prevent overfitting. Getting enough labeled data across grape varietals and vintages is a challenge in viticulture. Data restrictions can be addressed with methods like data augmentation or transfer learning from related applications.

2. **Computational Complexity:** CNNs' use in real-time vineyard monitoring is frequently limited by the high-performance computational resources needed for training.

3. **Interpretability:** CNNs function as "black-box" models, which makes it more difficult to interpret certain spectral contributions than PLSR. Although they don't offer precise, quantitative interpretations, methods like Grad-CAM can be useful.

According to research by Gomes et al. [25], CNNs show good generalization in situations with a variety of data, despite these difficulties. CNNs require more resources and specialist handling, but they are particularly useful for complex spectral-spatial interactions.

Ridge Regression (RR)

Another linear technique for managing multicollinearity is Ridge Regression (RR), which lessens collinearity problems by minimizing big coefficients by adding a penalty term. It does not, however, have the dimensionality reduction capabilities of PLSR.

3.4.5 Model Selection Considerations

Data properties, interpretability requirements, and computing limitations all influence the model selection:

- **PLSR and RR:** Ideal for high-dimensional data with linear relationships. While RR provides simplicity, PLSR is chosen for interpretability.

- **SVR:** It needs to be tuned and can be computationally demanding, but it works best for mild nonlinearity.

- **CNNs:** Ideal for highly variable, nonlinear, complicated data, but at the expense of interpretability and resource requirements.

3.4.6 Conclusion

For high-dimensional data, PLSR is still useful because it strikes a balance between interpretability and prediction accuracy. Nonlinearities are handled by SVR and CNNs, although CNNs demand a significant amount of resources. For managing collinearity, Ridge Regression offers a more straightforward option. Particularly in viticulture, where grape type, weather circumstances, and spectral patterns change, model selection should take into account data complexity, interpretability, and resource availability.

3.5 Conclusion

This chapter included an in-depth analysis of hyperspectral imaging integrated with machine learning methods for forecasting grape sugar content and additional quality metrics. By critically reflecting on methodological choices and conducting a thorough literature analysis, we emphasized the significance of sampling strategies, data preprocessing, and model selection in the development of effective predictive systems. The mathematical analysis of PLSR, SVR, and neural networks elucidates the foundational assumptions, advantages, and constraints of each model. Comprehending these factors is essential for choosing the suitable modeling methodology according to the particular environment and goals. Building upon these insights, in the next chapter, we implement and evaluate these machine learning algorithms using two real-world datasets of multispectral grape images. We specifically investigate the challenges of predicting the Brix index, considering the issues of data variability, potential information loss, and model generalization discussed herein. By tackling these challenges, we aim to contribute to the advancement of precision viticulture and develop effective methods for grape quality assessment.

Chapter 4

Methodology and deep analysis

4.1 Introduction and Objective

As anticipated in previous chapters, this chapter focuses on the study of grape sugar content prediction using multispectral and hyperspectral vectors. The main objective is to compare three methodologies widely adopted in machine learning applied to chemometrics, analyzing their behavior in predicting sugar content and discussing their limitations. Throughout the analysis, detailed explanations of the results obtained will be provided, with a focus on the mathematical theory underlying each model. The three approaches examined are Principal Component Regression (PCR), Partial Least Squares Regression (PLSR) and its robust method, RoBoost PLSR. To conduct this analysis, I used two separate datasets. The first dataset proved to be particularly effective in highlighting the potential of the models considered, allowing significant prediction of sugar content. In contrast, the second dataset raised some difficulties, providing an opportunity to explore the limits of application of these techniques and of the dataset itself. The main focus will be on the first dataset, which provides the ideal context for implementing and evaluating a personal contribution: the implementation of RoBoost PLSR in Python, which is not available in standard libraries. In the course of the analysis, not only will the performance of the predictive models be examined, but also the relationships between the variables will be explored in depth through various analytical tools. Among them, correlation matrices will be used to identify key connections between spectral bands, principal components extracted from the models and sugar content, highlighting how some wavelengths are more influential than others. VIP scores calculated for the PLSR and RoBoost PLSR models will allow identification of the most significant bands in the prediction and show differences with the components

extracted from PCA. A crucial role will be played by the graphical representation of the data. In particular, graphs will be presented showing the average spectra of different grape varieties, divided into red grapes, white grapes and for specific varieties. These graphs will make it possible to observe the differences in reflectances at different wavelengths, highlighting the physical phenomena that distinguish red and green grapes, discussed in previous chapters. Other graphs, such as heatmaps of correlations between bands, will be used to understand how indeed features of multispectral vectors are strongly correlated with each other, particularly those adjacent to each other and belonging to the same segments of the electromagnetic spectrum, highlighting that there is a limit to the amount of information present in a multispectral vector even if the wavelengths represented are many but fall in the same range. Finally, model results on datasets divided into subgroups, such as red and white grapes or specific varieties, will be discussed, and comparative graphs showing how performance changes depending on data segmentation will be included. These comparisons will not only make clear the differences between the subgroups, but also allow the results to be linked to physical phenomena related to the chemical composition and spectral structure of the grapes.

4.2 Datasets description

In this section, we provide a detailed description of the two datasets utilized in our study. The first dataset was instrumental in achieving successful predictive modeling, while the second dataset presented challenges that impacted its utility in our analysis. For each dataset, we outline the nature of the data collected, the methodologies employed in data acquisition, and the characteristics that define their variability and composition.

4.2.1 First Dataset: Hyperspectral Imaging and Sugar Content Measurements

The first dataset, titled *"Dataset containing spectral data from hyperspectral imaging and sugar content measurements of grapes berries in various maturity stages"* [38], comprises hyperspectral reflectance spectra and corresponding sugar content measurements of grape berries at different maturity levels. This dataset was curated to explore the feasibility of using hyperspectral imaging for monitoring grape berry maturity, particularly focusing on the prediction of sugar content, an essential parameter in viticulture.

Data Acquisition and Sample Preparation

A total of 274 samples were collected during the summer of 2020 from the experimental vineyard Domaine Expérimental Viticole Tarnais, located in Gaillac, France. The samples represent three grape varieties: two red varieties (*Syrah* and *Fer Servadou*) and one white variety (*Mauzac*). Table 4.1 summarizes the number of samples per variety.

Table 4.1: Number of samples per grape variety in the first dataset.

Variety	**Syrah**	**Fer Servadou**	**Mauzac**
Number of Samples	126	63	85

Grape berries were harvested approximately once a week, starting one or two weeks after véraison (the onset of ripening) until just before harvest. In the laboratory, berries were carefully detached at the pedicel to maintain the integrity of the fruit. To ensure homogeneity in maturity levels within samples, berries were sorted using densimetric sodium chloride (NaCl) baths of increasing concentrations, ranging from 70 to 190 g/L. This sorting technique leverages the correlation between berry density and sugar content, effectively grouping berries with similar ripeness based on their flotation in the NaCl solutions.

Each sample consisted of 100 berries of similar maturity, collectively placed on a tray for hyperspectral imaging. Prior to imaging, the sugar content of each sample was measured using a refractometer (HI-96816, Hanna Instruments) on the must obtained from the 100 berries.

Hyperspectral Imaging Procedure

Hyperspectral images were acquired using a Specim IQ hyperspectral camera (Specim, Finland) covering the visible to near-infrared (VIS-NIR) spectral range from 400 nm to 1000 nm with a spectral resolution of 7 nm. The camera was positioned 1.5 m above the sample tray. Illumination was provided by a halogen lamp (Arrilite 750 Plus ARRI, Munich, Germany) with consistent angles of illumination maintained at $-50°$ and $50°$ relative to the camera axis to ensure uniform lighting conditions.

To account for instrument and illumination non-uniformities, a certified reflectance standard (Labsphere, SRS-40-010) was included in each image as a reference. Reflectance spectra ($R_s(\lambda)$) were calculated for each pixel using the following equation:

$$R_s(\lambda) = \frac{I_s(\lambda) - I_b(\lambda)}{I_o(\lambda) - I_b(\lambda)}, \tag{4.1}$$

where $I_s(\lambda)$ is the measured intensity reflected from the sample, $I_b(\lambda)$ is the dark current image, and $I_o(\lambda)$ is the intensity reflected from the reference standard.

Data Processing and Spectral Data Extraction

Image processing was performed using MATLAB (The MathWorks, Natick, MA, USA). Segmentation of grape berry pixels was achieved using the Spectral Angle Mapper (SAM) method, which compares the spectral similarity between each pixel and predefined reference spectra for each grape variety. This approach effectively isolates berry pixels from the background by calculating the spectral angle α between the pixel spectrum $\mathbf{x}$ and the reference spectrum $\mathbf{y}$:

$$\alpha = \arccos\left(\frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||}\right). \tag{4.2}$$

By applying a spectral similarity threshold, pixels corresponding to grape berries were identified and extracted. For each sample, an average reflectance spectrum was computed by averaging the spectra of all berry pixels within the image. This resulted in 204 mean reflectance spectra, each associated with a corresponding sugar content measurement.

Dataset Characteristics

The dataset includes reflectance spectra spanning from 400 nm to 1000 nm at 7 nm intervals, yielding a total of 204 spectral bands per sample. Each row in the dataset represents a single sample, which corresponds to an average spectrum of 100 grape berries of similar maturity and their associated sugar content. The sugar content values range from approximately 100 to 300 g/L, covering various stages of grape maturation.

This dataset is well-suited for chemometric analyses and modeling techniques aimed at predicting sugar content based on spectral data. The inclusion of multiple grape varieties and maturity stages enhances the variability and robustness of the dataset, making it valuable for testing regression methods and variable selection techniques.

4.2.2 Second Dataset: Multispectral Imaging with Weight, Anthocyanins, and Brix Index Measures

The second dataset, titled *"A novel ground truth multispectral image dataset with weight, anthocyanins, and Brix index measures of grape berries tested for its utility in machine learning pipelines"* [31], consists of multispectral images of individual grape berries along with associated measurements of weight, anthocyanin content,

and Brix index. This dataset was developed to facilitate the application of machine learning techniques in viticulture, particularly for tasks involving classification and regression based on multispectral imaging data.

Data Acquisition and Sample Preparation

A total of 1,283 grape berries were collected from five seedless table grape varieties: *Autumn Royal*, *Crimson*, *Itum4*, *Itum5*, and *Itum9*. The berries were harvested from a commercial vineyard in Alhama de Murcia, Spain, during the fully ripe stage suitable for marketing and export, approximately 3 to 4 weeks after véraison. Table 4.2 details the sample distribution among the varieties.

Table 4.2: Number of samples per grape variety in the second dataset.

Variety	**Autumn Royal**	**Crimson**	**Itum4**	**Itum5**	**Itum9**
Number of Samples	199	401	84	504	95

From each bunch, berries were sampled from three distinct regions (top, middle, bottom) to capture intra-bunch variability. The berries were cleaned, weighed, and individually labeled prior to imaging.

Multispectral Imaging Procedure

The imaging was conducted using a custom-built multispectral chamber designed to capture images across a broad spectrum of wavelengths. The chamber components include:

- **Illumination System**: A multispectral LED illumination system covering wavelengths from 450 nm to 970 nm.
- **Imaging System**: Two Photonfocus snapshot mosaic multispectral cameras were used. One camera (MV1-D2048x1088-HS03-96-G2) captured 12 bands in the visible range (488 nm to 625 nm), and the other (MV1-D2048x1088-HS02-96-G2) captured 25 bands in the red-infrared range (676 nm to 952 nm), resulting in a total of 37 spectral bands per image.
- **Software Control**: A LabVIEW-based application controlled the illumination and acquisition parameters.

Each berry was imaged alongside a 1 cm^2 reference marker to facilitate spatial calibration and size measurements. The imaging process produced raw multispectral images, which were subsequently calibrated using dark and white reference images to correct for sensor noise and illumination inconsistencies.

Data Processing and Feature Extraction

Due to the variability in berry reflectance, particularly among darker varieties like *Autumn Royal*, a custom image processing algorithm was developed for grape segmentation. The algorithm involves edge detection, adaptive thresholding, and template matching to accurately isolate the berry from the background in each spectral band.

Once segmented, the multispectral images of each berry were associated with their corresponding weight, anthocyanin content, and Brix index measurements. Anthocyanin content was quantified using spectrophotometric methods, measuring absorbance at 530 nm and 657 nm, and calculated using the formula:

$$Q_{\text{total anthocyanin}} = \frac{A_{530} - 0.25 \times A_{657}}{\text{FW}}, \tag{4.3}$$

where A_{530} and A_{657} are the absorbance values at the respective wavelengths, and FW is the fresh weight of the sample. The Brix index, indicative of sugar content, was measured using a digital refractometer (ATAGO PAL-1) on the juice extracted from each berry.

Dataset Characteristics

The dataset comprises 1,283 multispectral image arrays, each with 37 spectral bands ranging from 488.38 nm to 952.76 nm. Each row in the dataset represents an individual grape berry, including its multispectral image data and associated measurements of weight, anthocyanin content, and Brix index.

The dataset's extensive variety coverage and detailed measurements make it a valuable resource for developing and testing machine learning algorithms for regression and classification tasks in viticulture. However, challenges such as the complexity of image segmentation and potential variability in imaging conditions may impact the dataset's utility in predictive modeling.

4.2.3 Comparison of the Datasets

The two datasets differ significantly in their structure, content, and potential applicability:

- **Sample Composition**:
  - *First Dataset*: Each sample represents an average of 100 grape berries of similar maturity, providing a collective spectral signature and sugar content measurement.

  - *Second Dataset*: Each sample corresponds to an individual grape berry, offering detailed multispectral images and multiple associated measurements (weight, anthocyanins, Brix index).

- **Spectral Data**:
  - *First Dataset*: Hyperspectral data with 204 spectral bands covering 400 nm to 1000 nm at 7 nm intervals.
  - *Second Dataset*: Multispectral data with 37 spectral bands ranging from 488.38 nm to 952.76 nm.

- **Varietal Coverage**:
  - *First Dataset*: Includes three grape varieties with both red (*Syrah*, *Fer Servadou*) and white (*Mauzac*) grapes.
  - *Second Dataset*: Encompasses five seedless table grape varieties, all of which are red or white, but focuses on table grapes rather than wine grapes.

- **Measurement Parameters**:
  - *First Dataset*: Focuses on sugar content as the primary measurement associated with spectral data.
  - *Second Dataset*: Provides additional measurements such as weight and anthocyanin content alongside the Brix index.

The differences in sample composition and data structure impact the applicability of each dataset in predictive modeling. The first dataset's aggregation of berries into samples may smooth out individual variability, facilitating more robust sugar content predictions. In contrast, the second dataset's focus on individual berries introduces greater variability and complexity, which may present challenges in modeling but also offers a more detailed analysis at the berry level.

4.2.4 Implications for Analysis

The first dataset proved to be effective for our modeling purposes due to its controlled sample composition, consistent imaging methodology, and focus on a key predictive parameter (sugar content). The homogeneity within samples and the comprehensive spectral coverage enhanced the reliability of the regression models developed.

Conversely, the second dataset posed challenges, possibly due to the high variability among individual berries, the complexity of image segmentation, and potential inconsistencies in imaging conditions. These factors may have contributed

to less satisfactory modeling results, highlighting the importance of dataset structure and quality in predictive analysis.

4.3 Analytical Tools

As previously discussed, when we are faced with the challenge of predicting the sugar content of grapes using a multispectral vector as a predictor, we encounter data that present multiple difficulties. Firstly, as in the dataset we will analyze in this study, the number of observations is often quite limited. Even when many berries are sampled, they are frequently aggregated based on their maturity level. For example, in the dataset described in the previous section, a single multispectral vector is produced by averaging the spectral vectors of 100 berries. The result is a dataset of only 274 spectral vectors.

Secondly, multispectral vectors, and even more so hyperspectral vectors, contain a very large number of features, which can often exceed 200. As already discussed, each feature represents the intensity of reflected light at a specific wavelength, which is strongly correlated with the intensities at preceding and succeeding wavelengths. The closer the wavelengths are, especially within the same segment of the electromagnetic spectrum, in our case, BLUE, GREEN, RED, and NIR, the stronger this correlation becomes. Therefore, the datasets used to train our machine learning models are characterized by few observations and many features that are highly correlated with each other. In the dataset used for this study, the number of features of the spectral vectors (204) are very similar to the number of observations (274). If we do not perform preliminary operations to reduce the number of features in the dataset, this will pose a problem for the convergence of machine learning algorithms, which would have too many parameters relative to the observations available for training.

In general, the most commonly used machine learning models when dealing with such data are regression models because they have a reduced number of parameters to train compared to more complex models like neural networks. In this scenario, classical multiple linear regression (MLR) does not produce good results. This happens for two reasons:

1. **Lack of sufficient data**: As previously discussed, the limited number of observations hampers the model's ability to generalize effectively.

2. **Collinearity among features**: One of the assumptions of linear regression is that the predictor variables are independent of each other. Having variables that are highly correlated can lead to problems because the design matrix becomes ill-conditioned, making it difficult to estimate the regression coefficients accurately. Multicollinearity inflates the variance of the coefficient estimates,

making them unstable and unreliable. This can result in overfitting and poor predictive performance on new data.

Ideally, before training a regression model, we would like to reduce the dimensionality of the dataset while retaining the information it contains and creating new features that are uncorrelated with each other. This is what Principal Component Regression (PCR) achieves by performing a Principal Component Analysis (PCA) before training the regression model. PCA is one of the most widely used techniques in multivariate statistical analysis and is based on the fact that, if a dataset has high dimensionality but the variables within it are highly correlated, it can be projected onto a lower-dimensional hyperplane while retaining most of the information within the data [39]. More precisely, PCA seeks directions $\mathbf{w}$, orthogonal to each other, onto which to project the data points of the dataset. The optimization problem is as follows:

$$\begin{aligned} \mathbf{w}_k &= \arg\max_{\mathbf{w}} \left(\mathbf{w}^\top \mathbf{S} \mathbf{w} \right), \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{w} = 1, \\ & \mathbf{w}^\top \mathbf{w}_j = 0 \quad \text{for} \quad j < k, \end{aligned} \tag{4.4}$$

where $\mathbf{S}$ is the covariance matrix of the data, and $\mathbf{w}_k$ is the k-th principal component.

By solving this problem, we obtain a set of orthogonal components that capture the maximum variance in the data. The number of principal components to retain is determined by the amount of variance we wish to preserve. Typically, we select the smallest number of components that account for a desired percentage (e.g., 95%) of the total variance.

In PCR, after reducing the dataset's dimensionality through PCA, we train a linear regression model using the principal components.

While this system is powerful and consistent, it can present problems in some cases. The reason lies in the fact that PCA reduces the dimensionality of the data $\mathbf{X}$ without taking into account the target variable $\mathbf{Y}$. The principal components are created with the objective of retaining the variance in $\mathbf{X}$, without considering which directions most influence the variable $\mathbf{Y}$. PCR can therefore struggle in regression problems where the target variable is strongly associated with directions in the data that have low variance[40]. These directions might not be adequately captured by PCA.

For this reason, in chemometrics, where problems often involve predictor variables with very high dimensions, increasing the risk that important variables have low variance, one of the most used methods is Partial Least Squares Regression (PLSR)[41]. PLSR is in some ways similar to PCA, as it reduces the dimensionality of the dataset and performs regression, but it does so differently by taking into

account the target variable. Specifically, it sequentially searches for directions to project the dataset in order to maximize the covariance between $\mathbf{Xw}$ and $\mathbf{Y}$. The optimization problem is:

$$\mathbf{w}_k = \arg\max_{\mathbf{w}} \left(\mathrm{Cov}(\mathbf{Xw}, \mathbf{Y})\right), \tag{4.5}$$

where $\mathrm{Cov}(\mathbf{Xw}, \mathbf{Y})$ represents the covariance between the projected data and the target variable.

After finding each direction $\mathbf{w}_k$, a deflation step is performed to remove the information explained by that component from $\mathbf{X}$. This is necessary because the directions captured by PLSR are not orthogonal to each other. The deflation process ensures that subsequent components capture new information that is not redundant with the previously extracted components. This sequential extraction and deflation continue until a sufficient number of components have been extracted to model the relationship between $\mathbf{X}$ and $\mathbf{Y}$ effectively.

However, even PLSR can be sensitive to outliers, which, if not properly managed, can adversely affect the model's predictive performance. Therefore, robust methods have been developed to handle outliers without excluding samples that significantly contribute to the quality of the model. As noted in [42], "These methods must be parsimonious so as not to exclude major samples who contribute strongly to the good predictive quality of the model. According to Ref. [16], 'For high-dimensional data this would result in a severe loss of information as long as the outliers still contain some valuable information, and thus intelligent robust methods adapt the weights according to the outlyingness or inconsistency of the observations.'"

RoBoost PLSR (Robust Boosted Partial Least Squares Regression) is particularly useful in this context. It not only enhances the robustness of the model against outliers but also aids in their identification. Often, outliers are detected by examining only one dimension, which is insufficient. An observation that may seem anomalous in a single variable, such as the target, can be generated by legitimate combinations of the predictor variables. Similarly, in the context of spectral analyses, it is possible for a sample to have normal values of $\mathbf{Y}$ and $\mathbf{X}$ separately but together form a multivariate outlier.

RoBoost PLSR addresses this by adaptively weighting observations based on their consistency with the model, reducing the influence of outliers without discarding them outright. This allows the model to remain robust while still leveraging valuable information contained in the data.

These considerations will be crucial in the next section, where we will apply these methods to our data and demonstrate how RoBoost PLSR can enhance model robustness and improve predictive performance in the presence of outliers.

4.4 Exploratory data analysis

4.4.1 Description of features and target variable

As previously described in the data presentation, we have a dataset consisting of 274 samples, each representing a group of 100 grape berries of similar maturity. In addition to the 204 columns indicating reflectance at specific wavelengths in the 400 nm – 1000 nm range (VIS-NIR), there are three additional columns that indicate the variety, color, and sugar content of the grapes, expressed in g/L. The varieties considered are three: two red (Syrah and Fer) and one white (Mauzac).The average sugar content, calculated across all varieties, is 189 g/L, with a minimum value of 101 g/L, a maximum value of 283 g/L, and a standard deviation of 35 g/L (Figure 4.1). As can be observed in the histogram in Figure 4.2, the sugar content shows an approximately normal distribution, which is favorable for the convergence of machine learning algorithms.

Statistic	Value
count	274.00
mean	189.09
std	35.50
min	100.98
25%	163.25
50%	180.92
75%	213.74
max	282.74

Figure 4.1: Summary of the sugar content

Regarding the 204 spectral columns, their means range from 0.05 to 0.47 and all have a low standard deviation (<0.1). This suggests the need for data standardization to center and increase the variability of the columns before proceeding with model training (Figure 4.3).

4.4.2 Spectral Analysis by Color and Variety

Spectral plots of the mean spectra were generated separately by color and variety, using five random samples for each subset. Along with the spectral plots, histograms of the sugar content for each subset were also generated. This approach allowed for the analysis of spectral differences between white grapes, red grapes, and specific varieties.

The spectral plots highlight differences in reflectance at certain wavelengths between green grapes and red grapes. In particular, green grapes reflect more at specific wavelengths. As one might expect, green grapes have higher reflectance in

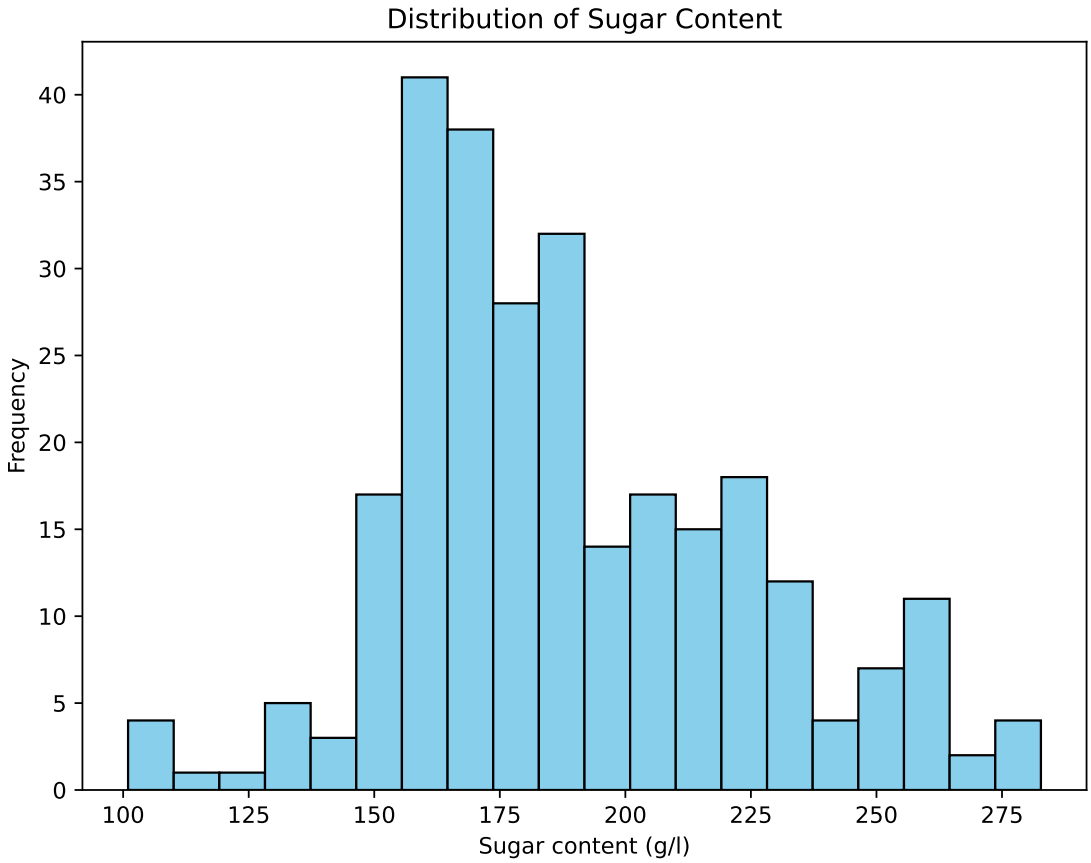

Figure 4.2: Distribution of Sugar Content in the Dataset

Sugar content (g/l)	Variety	Color	x.400.2	x.498.8	x.598.6	x.699.6	x.801.79	x.905.18
144.738	SYRAH	red	0.12199017	0.04019466	0.04083202	0.14994469	0.3913785	0.42456901
156.519	SYRAH	red	0.13368159	0.0395067	0.04113579	0.15329032	0.44918122	0.51333185
176.715	SYRAH	red	0.12741836	0.04317339	0.0420696	0.1128589	0.43540566	0.52841731
175.032	MAUZAC	white	0.12161104	0.06439081	0.13859836	0.25191399	0.43459635	0.43820844
210.375	SYRAH	red	0.12223633	0.04273176	0.0407927	0.1078726	0.3964465	0.48791351
163.251	MAUZAC	white	0.12315952	0.06421663	0.12227819	0.2397786	0.43936886	0.44784781
131.274	FER	red	0.13759702	0.05116224	0.05047972	0.15353956	0.45341595	0.51594504
164.934	FER	red	0.12365252	0.04775364	0.0448	0.11545025	0.40476499	0.47870894
104.346	MAUZAC	white	0.13754324	0.07484834	0.15983984	0.27633848	0.48013057	0.49409832
190.179	FER	red	0.1094622	0.0406918	0.03830987	0.1074722	0.3856914	0.44987139

Figure 4.3: Summary Statistics of Sugar Content in the Dataset.

the spectral range of green and yellow-orange from 500 to 620 nm, and also in the near-infrared region. Understanding that the color and chemical composition of an object directly affect its reflection characteristics, these results are comprehensible and suggest that separate models for green grapes and red grapes could perform better (see Figure 4.4).

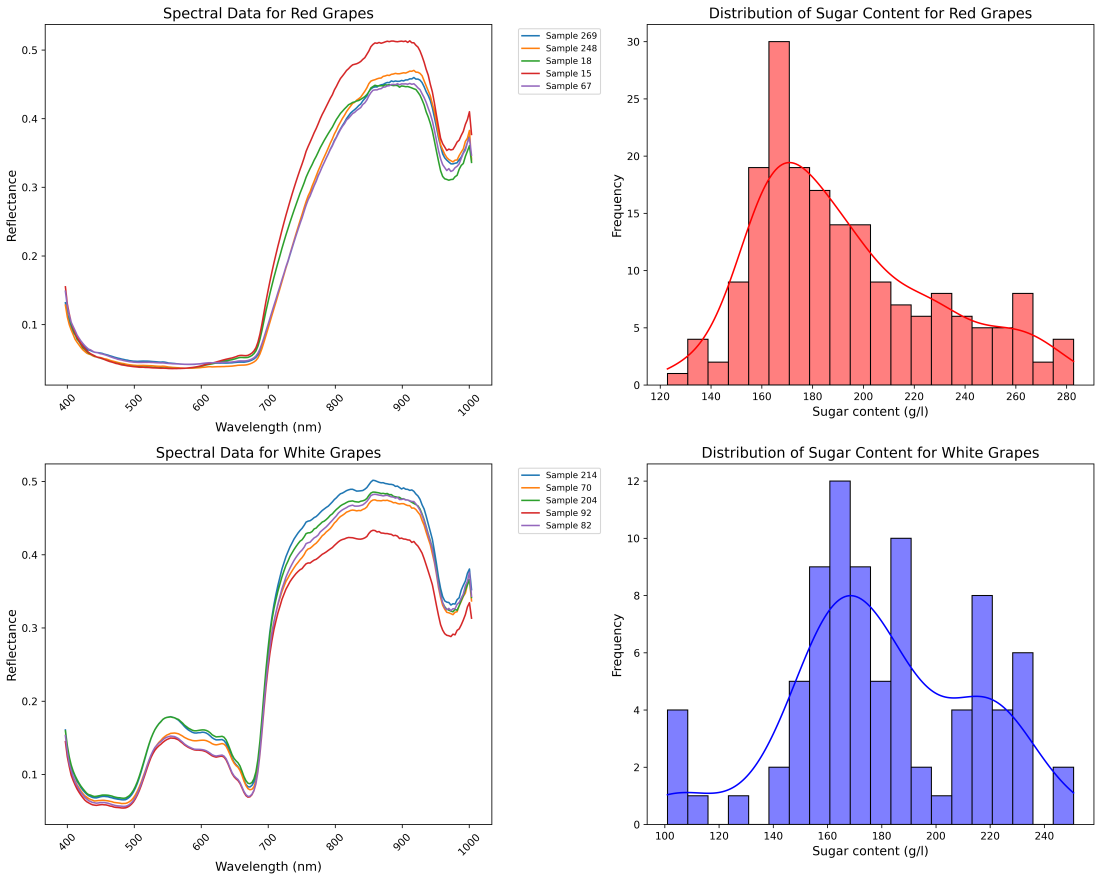

Figure 4.4: Spectral plots and histograms for red and green grapes.

Subsequently, a correlation matrix of the entire dataset was calculated and visualized for all 204 spectral bands. The result is consistent with the explanations in previous chapters regarding the strong correlation and redundancy of information between adjacent bands and those belonging to the same segments of the spectrum. The heatmap in Figure 4.5 clearly shows how the bands belonging to various segments of the spectrum are strongly correlated with each other. These results highlight the possibility of reducing the dimensionality of the dataset while retaining its inherent information.

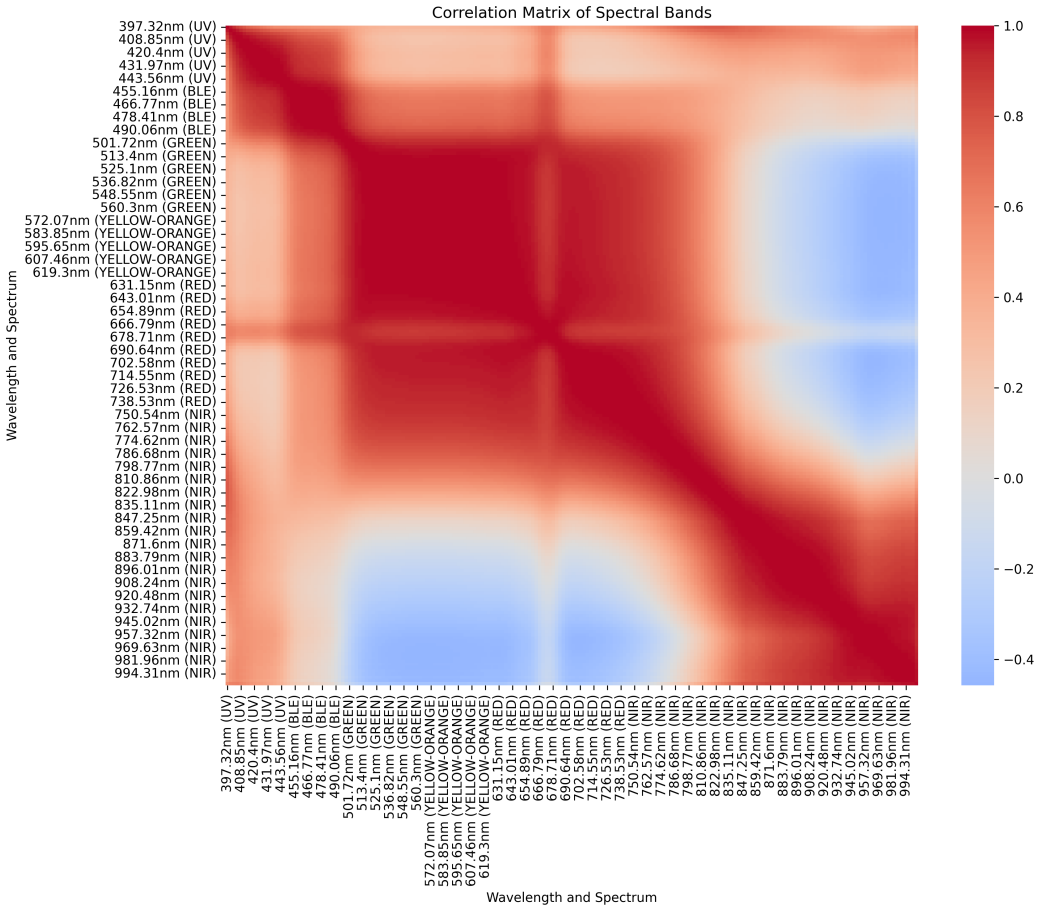

Figure 4.5: Correlation matrix heatmap of the 204 spectral bands.

4.4.3 Principal Component Analysis (PCA)

A Principal Component Analysis (PCA) was applied using the scikit-learn library. A threshold value of 95% was set, meaning that the minimum number of components was selected to explain at least 95% of the dataset's variance. This allowed us to reduce the dataset from 204 features to 3 principal components. Specifically, PC1, PC2, and PC3 retain 57%, 32%, and 10% of the dataset's variance, respectively (see Figure 4.6).

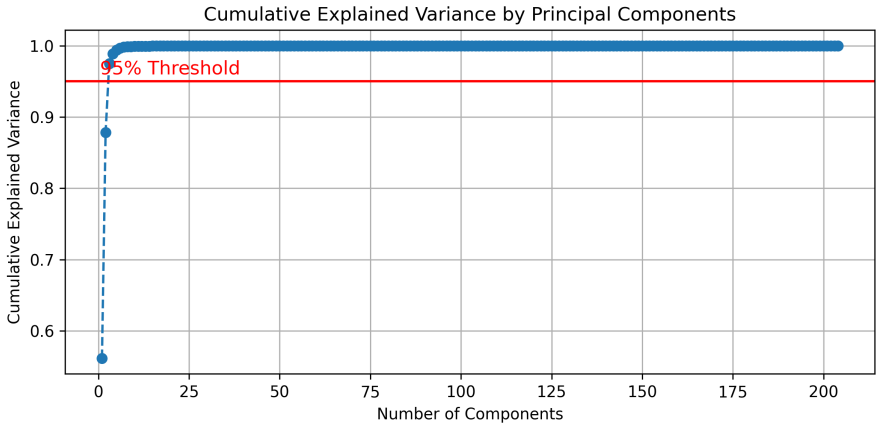

Figure 4.6: Variance explained by the first three principal components.

By analyzing the loadings of the components and using a heatmap, we visualized how each spectral band contributes to the creation of each principal component. What emerges is a parallel with the correlation matrix of the spectral bands. Spectral bands that are correlated with each other contribute uniformly to the generation of the same principal component. This is consistent with the fact that if two variables are correlated, the direction of their correlation will be the best projection to maximize the variance of the projection itself. From this, we learn that the directions along which a dataset is projected tend to be connected with the directions along which there is a strong correlation between variables (see Figure 4.7).

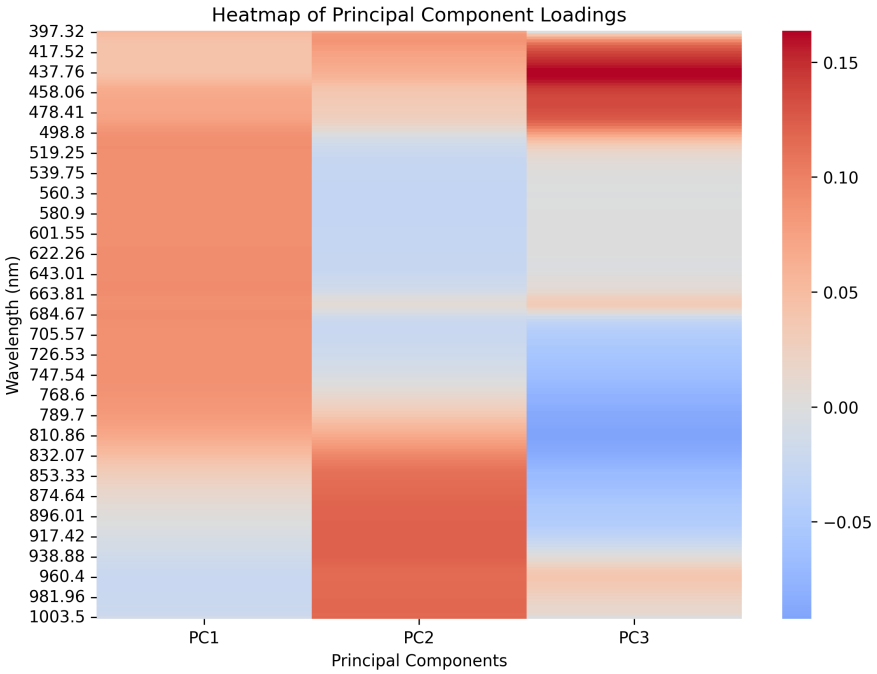

Figure 4.7: Variance explained by the first three principal components.

Subsequently, a scatter plot of the first two principal components was generated, differentiating the observations by color and variety to search for patterns or clusters that might group some of the subgroups. What is observed is a clear distinction between white and red grapes, which are visibly organized into two subgroups when plotted against the first two principal components. This suggests once again the difference between white and red grapes and implies the necessity of separating the models to improve performance (see Figure 4.8).

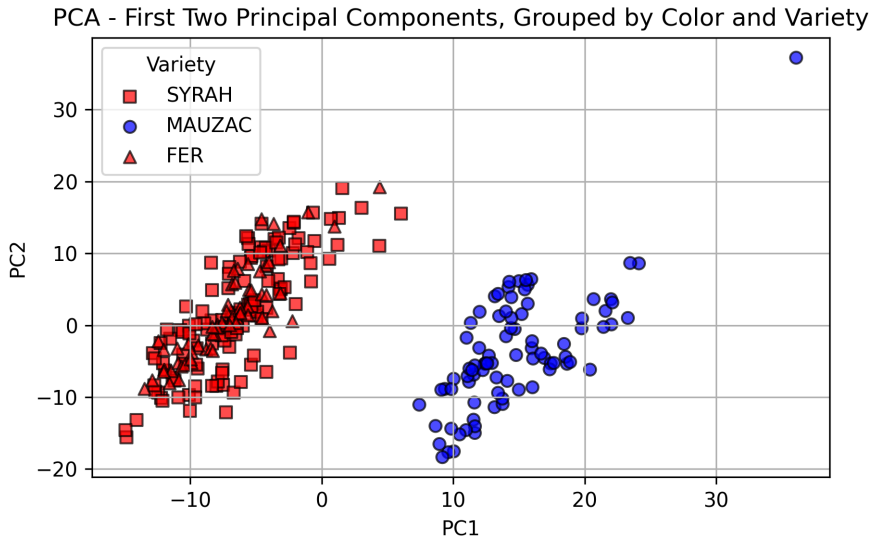

Figure 4.8: Variance explained by the first three principal components.

4.5 Regression's models implementation

As previously mentioned, three different models were tested: PCR, PLSR, and a custom implementation of RoBoost PLSR. The models were initially trained and tested on the entire dataset, without considering differences between variety and color, and subsequently on specific subsets to verify if performance improved on these. For PCR and PLSR, the `PCA`, `LinearRegression`, and `PLSRegression` classes from the `scikit-learn` library [43] were used. However, since `scikit-learn` and other Python libraries do not provide an implementation of RoBoost PLSR, it was necessary to develop a custom algorithm in Python, the algorithm of which is provided in Appendix A.

The algorithm implements the RoBoost PLSR approach based on the reference article, adapting it to the specific needs of the analyzed dataset. The implementation involves defining several functions to manage the adaptive weights of observations, using robust weighting functions like Tukey's bisquare function. Functions were created to calculate weights based on residuals, to center the data, and to compute the weighted mean, ensuring correct manipulation of the data matrices. The core of the algorithm consists of a function that implements the weighted NIPALS PLS algorithm, adapted to handle a single response variable. This allows the extraction of latent components while considering the weights assigned to the observations, improving the model's robustness against outliers.

Additionally, a `RoBoostPLSR` class was developed to manage the entire training

and prediction process of the model. The class includes methods for parameter initialization, iterative model training with weight updates, and generating predictions on new data. During training, convergence criteria are calculated and updated, and latent components and regression coefficients are stored. The implementation requires careful management of matrix operations and weights, ensuring numerical stability and convergence of the algorithm. Controls were implemented to avoid divisions by zero and to handle situations where weights become zero, thus ensuring the robustness of the algorithm.

The complete implementation algorithm is provided in Appendix A, where detailed comments explain the individual parts of the algorithm. This custom implementation allowed the effective application of the RoBoost PLSR method to the dataset in question, overcoming the limitations of existing libraries and providing a model robust to anomalous values.

For training, the dataset was divided into two subsets using the `train_test_split` function, obtaining a training set containing 80% of the observations and a test set with the remaining 20%. Before training, the explanatory variables ($\mathbf{X}$) were standardized using `StandardScaler`. This operation was necessary because regression algorithms, such as those implemented in PCR, PLSR, and RoBoost PLSR, assume that the explanatory variables have a centered distribution (zero mean) and similar variance among the different features.

The standardization process was carried out with particular attention to the separation of training and test data. The scaler was fitted exclusively on the training set ($\mathbf{X}_{\text{train}}$), calculating the mean and standard deviation for each feature. Subsequently, these same transformations were applied to both the training set and the test set ($\mathbf{X}_{\text{test}}$), ensuring that the test data remained independent of the training process. It is important to emphasize that the target variable ($\mathbf{Y}$) was not standardized. This is because the scale of the dependent variable does not directly affect the tested models, which focus on the relationship between input features ($\mathbf{X}$) and output values ($\mathbf{Y}$). This approach ensured proper data preparation, minimizing the risk of contamination between training and test sets and allowing a more reliable evaluation of the models' performance.

For the PCR model, a pipeline was defined that first implements PCA, using the first three components that cumulatively explain 97.5% of the variance, followed by a linear regression. For hyperparameter tuning for PLSR and RoBoost PLSR, cross-validation with 5 folds was performed. In particular, the parameter optimized for PLSR is the number of components to extract, while for RoBoost PLSR, three additional parameters were optimized: α, β, and γ, which regulate the handling of outliers within the model's algorithm.

Regarding training on the complete dataset, PCR achieved on the test set an $R^2 = 0.54$ and an RMSE = 26.07. PLSR improved performance, with $R^2 = 0.71$ and RMSE = 20.54. RoBoost PLSR outperformed both, obtaining $R^2 = 0.81$ and

RMSE = 10.53. Cross-validation identified the best number of components for PLSR as 10, while for RoBoost PLSR, the optimal hyperparameters were $\alpha = \infty$ and $\beta = \gamma = 4$.

Table 4.3: Model Performance on Complete Dataset

Model	R^2	**RMSE**
PCR	0.54	26.07
PLSR	0.71	20.54
RoBoost PLSR	0.81	10.53

The same models were then trained on specific subsets, obtaining the results shown in Table 4.4.

Table 4.4: Model Performance by Subset

Subset	**PCR RMSE**	**PCR** R^2	**PLSR RMSE**	**PLSR** R^2	**RoBoost RMSE**	**RoBoost** R^2
White Grapes	23.26	0.42	20.97	0.53	19.01	0.61
Red Grapes	13.72	0.82	13.03	0.84	12.24	0.86
SYRAH	8.97	0.92	8.67	0.93	8.37	0.94
MAUZAC	23.26	0.42	23.24	0.53	19.01	0.61
FER	32.68	0.17	30.82	0.25	29.1	0.35

As expected, the models' performance varies when trained on specific subsets. PLSR consistently showed better performance than PCR, while RoBoost PLSR achieved the best results in all subsets, confirming its robustness, especially in the presence of outliers.

For white grapes, the models showed a slight decrease in performance, with lower R^2 values and higher MSE compared to the complete dataset. This decline could be due to less variability in the sugar content within the white grape subset, limiting the models' ability to generalize effectively.

In the case of red grapes, performance improved, with PLSR achieving an $R^2 = 0.84$ and RoBoost PLSR obtaining $R^2 = 0.86$. This improvement can be attributed to the distinctive spectral characteristics of red grape varieties, which seem to provide more useful information for predicting sugar content.

For the SYRAH variety, all models performed exceptionally well, with RoBoost PLSR reaching an $R^2 = 0.94$ and an MSE of 70.00. This high performance can be linked to the relatively large number of observations for this variety and its specific spectral properties.

Conversely, for the FER variety, performance was significantly lower, with PLSR

obtaining an $R^2 = 0.25$ and RoBoost PLSR an $R^2 = 0.35$. The reduced sample size for this variety (only 63 observations) likely contributed to the lower performance, as the models had fewer data to learn from.

In summary, while the general trends align with expectations, the variation in performance among subsets underscores the importance of data quantity and variability in achieving robust predictive models. Further investigations into the characteristics of subsets with lower performance, such as FER and white grapes, could provide valuable insights for future studies.

4.6 Interpretation of Results

4.6.1 Correlations

After confirming the superior performance of the PLSR model over the PCR model, and of the RoBoost PLSR model over the other two, we conducted further analyses and generated graphs to critically examine the results and verify that the behavior of the models is consistent with the theoretical understanding of their functioning.

Firstly, we generated plots of the first three principal components for PCR and the first three latent variables for PLSR and RoBoost PLSR. These plots show the loadings of each spectral band on the first three components. This illustrates, separately for each of the three components, how much each of the 204 predictive variables, corresponding to the spectral bands, contributes to the construction of the components. As expected, we observe that the three models extract components differently, so the importance that each model assigns to each variable for prediction is different (see Figure 4.9).

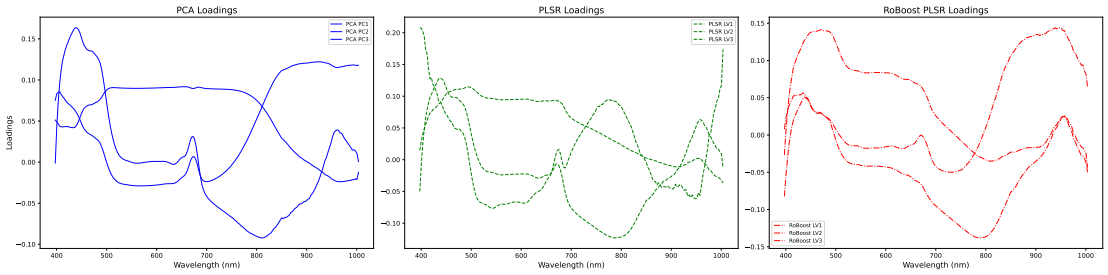

Figure 4.9: Loading's Comparison

Next, we calculated the correlation matrices between the extracted components and the sugar content, separately for each model. These matrices indicate how much each component is correlated with the target variable. Regarding the principal components from PCA, we note that the only component correlated with the target variable is the third one, with a correlation coefficient of 0.69. Very different results

are shown by the correlation matrices of the other two models. For the PLSR and RoBoost PLSR models, we see that almost all the extracted components are connected with the target variable (see Figure 4.10).

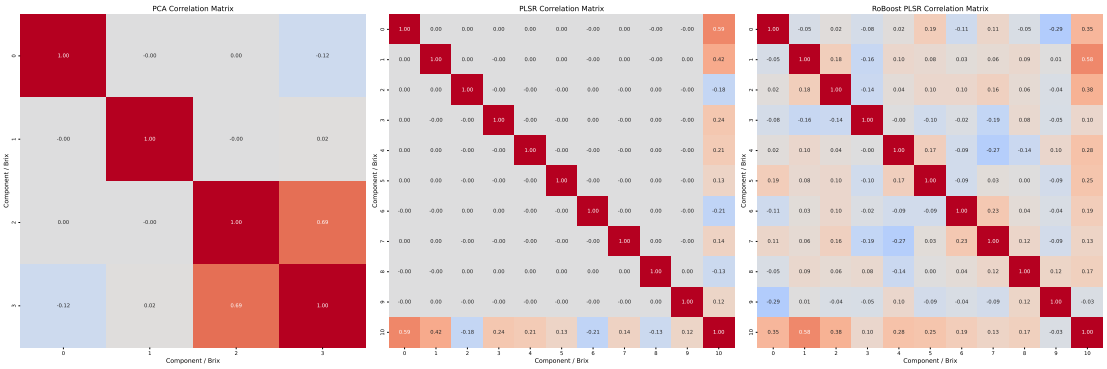

Figure 4.10: Matrix Correlations

The findings from these analyses explain why the PCR model has lower performance compared to the other two models. By extracting only one latent component correlated with the target variable through PCA, the model can derive information related to the prediction of the target variable only from that single component (as shown by the PCR regression coefficients). In contrast, the PLSR and RoBoost models can extract information from multiple components, leading to more accurate results.

These results align with the explanations provided in Section 4.3. It is evident that, during component extraction, PCA does not consider the relationship with the target variable. By projecting the data into a new dimensional space with the aim of preserving a certain percentage of variance, PCA loses information that links low-variance variables with the target variable. The PLSR algorithm, on the other hand, extracts each component with constant consideration of how it is connected with the target variable, resulting in less information loss and latent components more correlated with the target variable.

Finally, the graph in Figure 4.11 shows the variance explained by the first three components for the three models separately. In this graph, it is clearly seen how the variance explained by the first three components of PLSR and RoBoost PLSR is significantly lower compared to those extracted by PCA, illustrating once again how the focus and objectives of the two methods are different. PLSR extracted the first three components in that way even though their explained variance of the predictor data ($\mathbf{X}$) was very low, effectively recognizing that, despite the low variability of those directions, they were strongly correlated with the target variable.

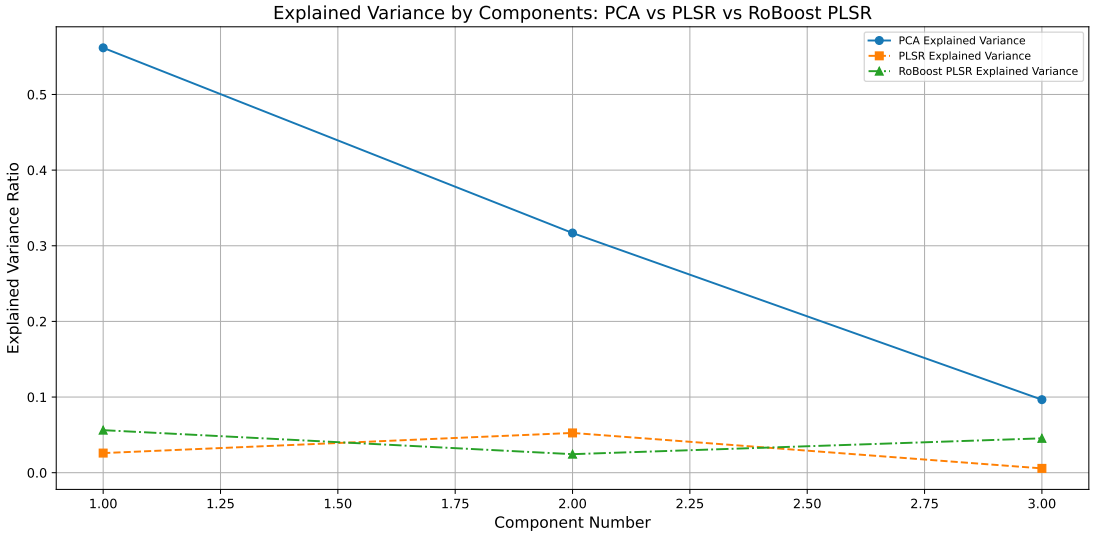

Figure 4.11: Explained variance comparison

4.6.2 VIP Scores Analysis

Subsequently, an in-depth analysis of the *VIP Scores* (Variable Importance in Projection) was conducted to assess the relative importance of each predictor variable in the PLSR and RoBoost PLSR models. The VIP Scores are metrics that quantify the contribution of each variable in the construction of the PLS model, considering both the ability to explain the independent variables and the dependent variable. A high VIP Score indicates that a particular variable plays a significant role in predicting the sugar content of grapes.

To calculate the VIP Scores in the PLSR and RoBoost PLSR models, the standard formula that combines the weights of the latent components with the explained variance was used. Specifically, the VIP score for the j-th variable is calculated as:

$$\mathrm{VIP}_j = \sqrt{p \times \frac{\sum_{a=1}^{A} \left(w_{ja}^2 \times \mathrm{SSY}_a \right)}{\mathrm{SSY}_{\mathrm{total}}}}$$

where:

- p is the number of predictor variables,
- w_{ja} is the weight of the j-th variable in the a-th latent component,
- SSY_a is the explained sum of squares by the a-th component for the dependent variable,
- $\mathrm{SSY}_{\mathrm{total}}$ is the total sum of squares of the dependent variable.

This approach allows the identification of the spectral variables (the different wavelengths) that most influence the model predictions. The VIP Scores were then normalized using the Min-Max scaling technique to facilitate comparison between the different models.

Similarly, for the PCR model, a comparable analysis was performed using the regression coefficients of the principal components. Although PCA does not directly provide VIP Scores, it is possible to evaluate the importance of variables by analyzing the regression coefficients associated with the principal components used in the linear regression model.

The analysis revealed that there are variables that are important in the PLSR model but not adequately represented in the PCR model. In particular, some spectral bands have high VIP Scores in the PLSR model but show negligible importance in the PCR model. This further confirms how PLSR is able to identify and exploit relationships between the predictor variables and the target variable that PCA, focusing only on maximizing the variance of the independent variables, may overlook.

The VIP Scores plots shown in Figure 4.12 highlight the differences in the importance assigned to variables by the different models. It is observed that PLSR and RoBoost PLSR assign importance to a greater number of variables compared to PCR, and in particular, some specific wavelengths are identified as particularly relevant for predicting sugar content.

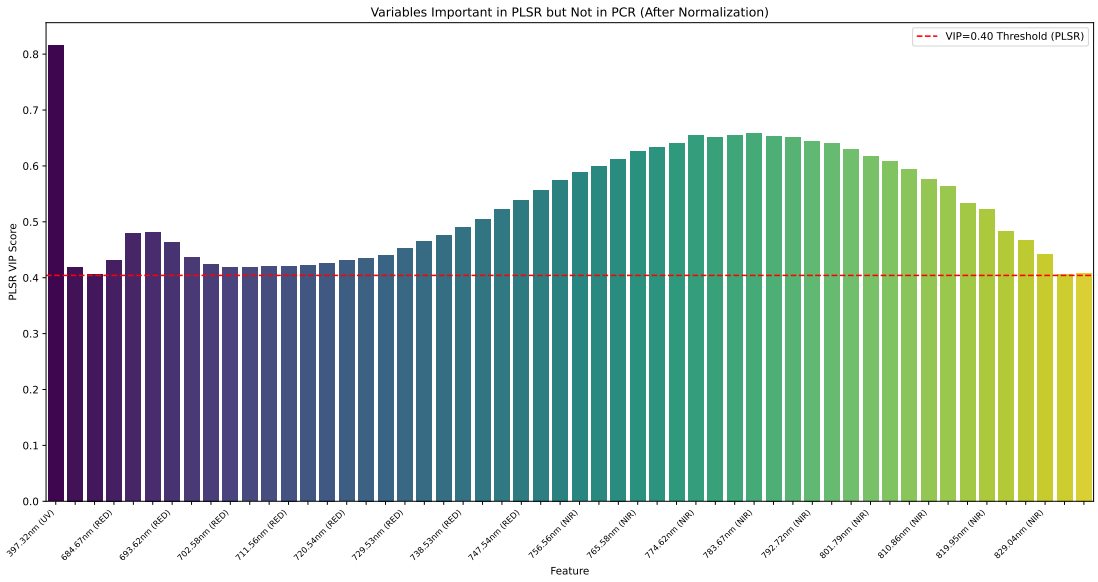

Figure 4.12: Vip scores confront between PCR and PLSR

These results underscore the effectiveness of VIP Scores as a tool to interpret PLSR and RoBoost PLSR models and to identify key variables in the prediction process. Moreover, they highlight the limitations of PCR in identifying important variables that do not contribute significantly to the total variance but are strongly correlated with the target variable.

The analysis of the VIP Scores, along with the previous evaluations of the latent components and correlations with the target variable, provides a deeper understanding of the models' functioning and why PLSR and RoBoost PLSR achieve better performance compared to PCR. It confirms that PLSR models, by focusing on the relationships with the target variable during component extraction, are more effective in capturing the relevant information for predicting the sugar content of grapes.

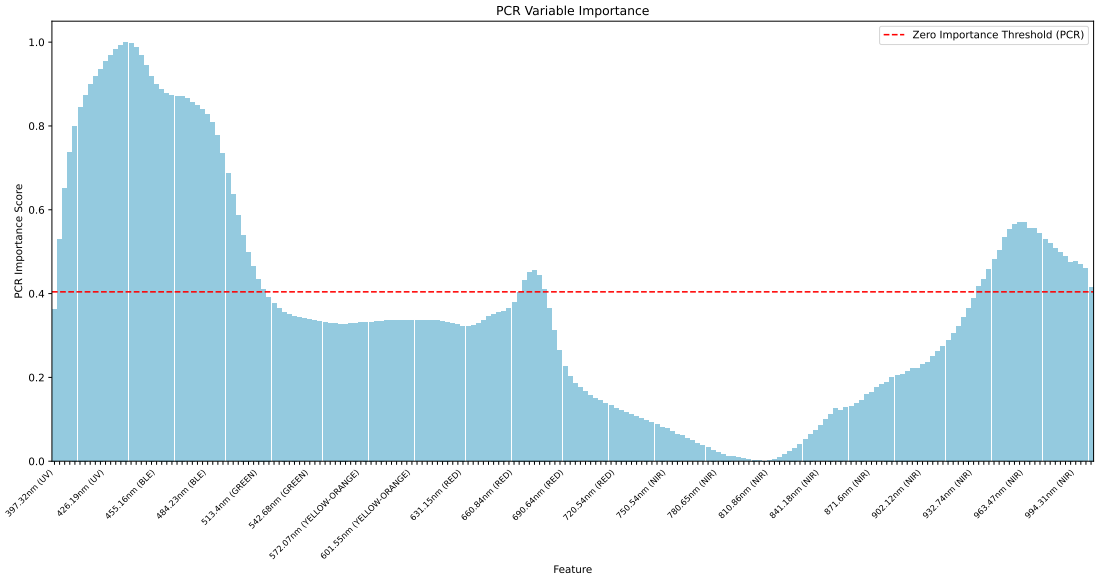

Figure 4.13: Importance of the variables in the PCR model

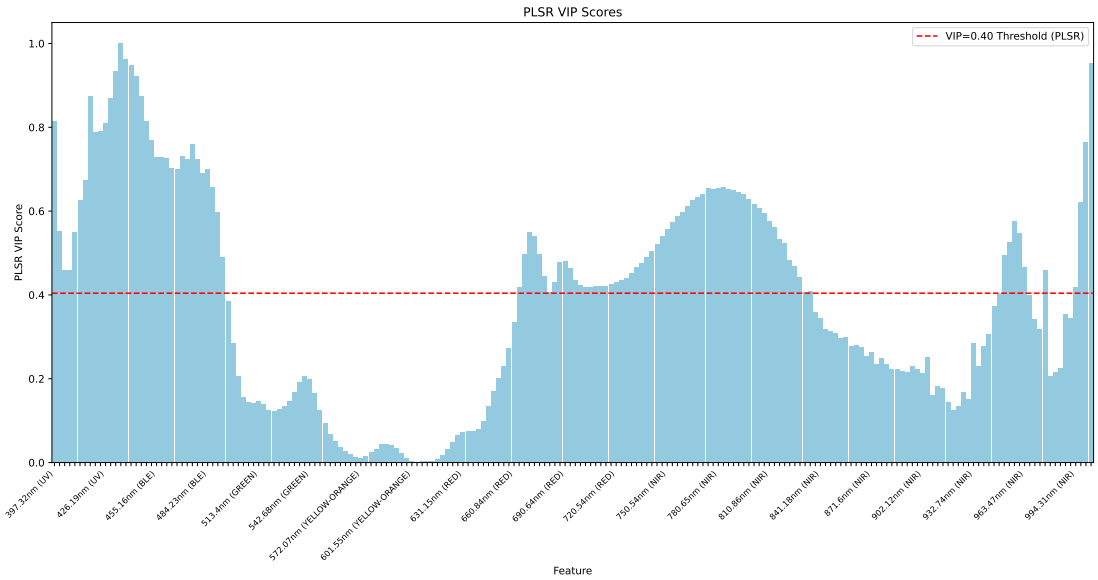

Figure 4.14: Importance of the variables in the PLSR model

4.7 Predictive Analysis of Sugar Content in the Syrah Variety with Outlier Identification

In this section, we focus on the predictive modeling of sugar content specifically for the Syrah grape variety. The Syrah dataset consists of 126 samples, each representing an average spectrum of 100 berries of similar maturity from the Syrah variety. This subset allows us to analyze the performance of the models in a more controlled setting, where varietal characteristics are consistent, and to delve into the identification and impact of outliers on model performance.

4.7.1 Data Preprocessing and Exploration

We filtered the original dataset to include only the samples corresponding to the Syrah variety. This resulted in a dataset of 126 observations. Missing values were checked and none were found, ensuring data integrity for subsequent analysis. The target variable, sugar content (expressed in g/L), showed an approximately normal distribution with a mean of 194 g/L, a minimum of 123 g/L, a maximum of 283 g/L, and a standard deviation of 37 g/L (see Figure 4.15). The histogram indicates a relatively symmetric distribution, which is favorable for regression modeling.

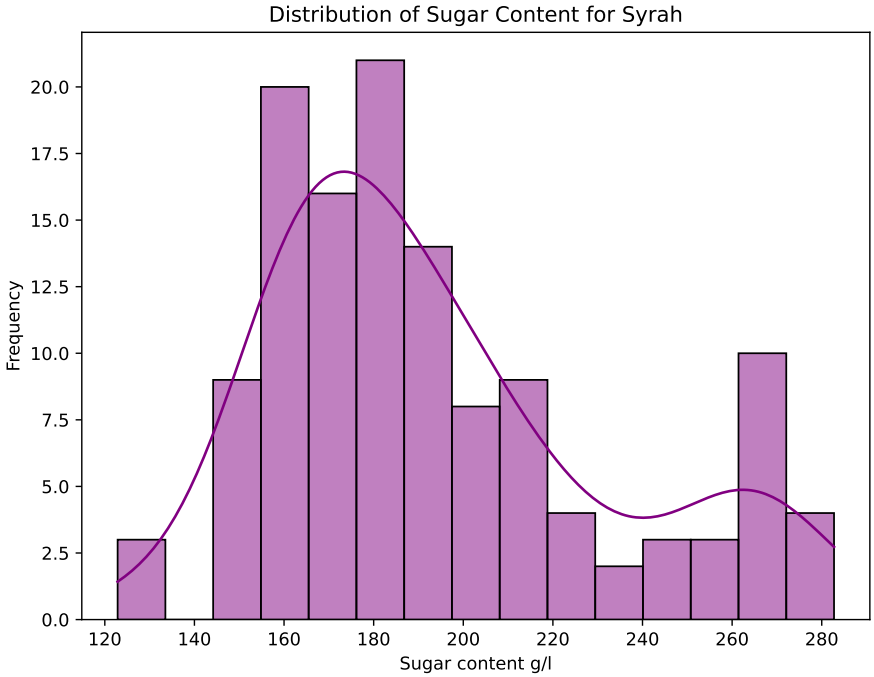

Figure 4.15: Distribution of Sugar Content in Syrah Variety

We standardized the spectral features ($\mathbf{X}$) using the `StandardScaler` from scikit-learn, fitting the scaler on the training data and transforming both training and test sets accordingly. The target variable (Y) was not standardized.

4.7.2 Model Implementation and Hyperparameter Optimization

We applied the three models, PCR, PLSR, and RoBoost PLSR, to the Syrah dataset. To ensure fair comparison and robust evaluation, we used 5-fold cross-validation to optimize hyperparameters for each model.

Principal Component Regression (PCR)

For PCR, we constructed a pipeline consisting of PCA followed by linear regression. We performed grid search cross-validation over the number of principal components, ranging from 1 to 10. The optimal number of components was found to be 5, which explained approximately 95% of the variance in the predictor variables.

Partial Least Squares Regression (PLSR)

For PLSR, we similarly performed grid search cross-validation over the number of components, also ranging from 1 to 10. The optimal number of components was determined to be 6.

RoBoost Partial Least Squares Regression (RoBoost PLSR)

For RoBoost PLSR, we performed cross-validation over the number of components (from 5 to 15) and the hyperparameters α, β, and γ, which control the robustness of the model to outliers. The optimal parameters were found to be $n_{\text{components}} = 6$, $\alpha = 4.685$, $\beta = 4.685$, and $\gamma = \infty$.

4.7.3 Model Performance and Comparison

The models were evaluated on a test set comprising 20% of the data. The performance metrics, including the coefficient of determination (R^2) and Root Mean Squared Error (RMSE), are summarized in Table 4.5.

Table 4.5: Model Performance on Syrah Test Set

Model	**Optimal Components**	R^2	**RMSE (g/L)**
PCR	5	0.92	9.0
PLSR	6	0.93	8.6
RoBoost PLSR	6	**0.97**	**5.5**

As expected, PLSR outperformed PCR due to its ability to consider the relationship between predictors and the target variable during component extraction. RoBoost PLSR achieved the best performance, with an R^2 of 0.97 and an RMSE of 5.5 g/L, indicating its effectiveness in handling outliers and enhancing predictive accuracy.

4.7.4 Outlier Identification and Analysis

An important aspect of RoBoost PLSR is its capacity to identify and down-weight outliers during model training. We extracted the final observation weights assigned by the RoBoost PLSR model to each sample. Lower weights indicate observations considered as outliers. Upon analyzing the distribution of these weights, we identified a subset of observations with significantly lower weights, specifically those below the 10th percentile. These observations are considered outliers by the model (see Figure 4.16).

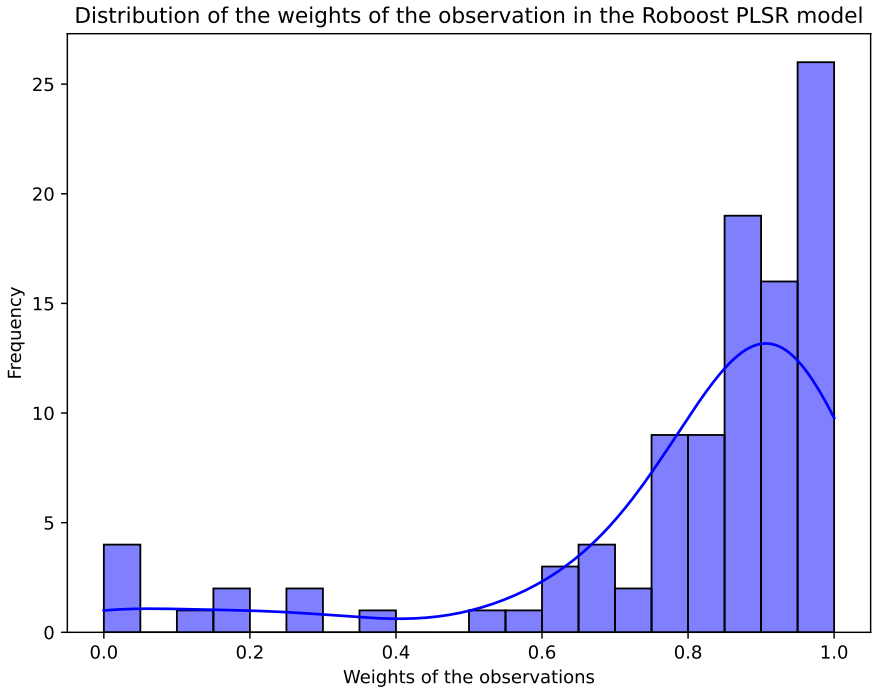

Figure 4.16: Distribution of Observation Weights in RoBoost PLSR for Syrah

Interestingly, when we examined the sugar content values of these outliers, we found that they were distributed across the central range of the sugar content distribution (see Figure 4.17). This implies that these outliers cannot be easily identified using classical univariate methods based solely on the target variable Y, as they do not exhibit extreme values in sugar content. Instead, the outliers are likely multivariate anomalies, exhibiting unusual combinations of spectral features and sugar content.

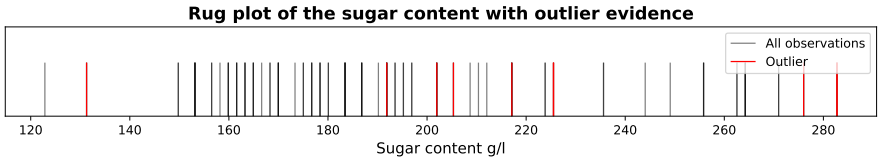

Figure 4.17: Distribution of Sugar Content with Outliers Indicated (Syrah)

4.7.5 Impact of Outliers on Model Performance

To assess the impact of the identified outliers on model performance, we retrained the PLSR model after removing the outliers detected by the RoBoost PLSR model.

The performance of the PLSR model improved significantly, with R^2 increasing from 0.93 to 0.98 and RMSE decreasing from 8.6 g/L to 5.3 g/L (see Table 4.6).

Table 4.6: PLSR Performance Before and After Outlier Removal

PLSR Model	R^2	**RMSE (g/L)**
Original PLSR	0.93	8.6
PLSR without Outliers	0.98	5.3

This improvement indicates that the RoBoost PLSR model effectively identified observations that negatively impacted the PLSR model's performance. By removing these outliers, the PLSR model becomes more robust and achieves better predictive accuracy.

4.7.6 Discussion on Outlier Detection

The inability of classical methods to detect these outliers is attributed to their focus on univariate distributions. Since the outliers have sugar content values within the normal range, they do not appear as anomalies when examining Y alone. However, in the multivariate space of spectral features, these observations exhibit atypical relationships between $\mathbf{X}$ and Y, which are effectively identified by the robust weighting mechanism of RoBoost PLSR.

This highlights the importance of considering multivariate outlier detection methods in chemometric analyses, as univariate approaches may overlook influential anomalies that affect model performance.

4.7.7 Conclusions

The analysis of the Syrah variety demonstrates the effectiveness of RoBoost PLSR in improving predictive performance and handling outliers. By identifying observations that traditional methods might overlook, RoBoost PLSR enhances model robustness and accuracy.

The outliers detected were not apparent when examining the sugar content alone, underscoring the necessity of multivariate outlier detection in chemometric modeling. The removal of these outliers led to improved performance in the PLSR model, validating the relevance of the outliers identified by RoBoost PLSR.

This case study emphasizes the importance of robust modeling techniques in viticulture and chemometrics, particularly when dealing with complex, high-dimensional data where outliers may not be evident through classical univariate analyses.

4.8 Additional Case Study

In this part, we provide another case study that uses a different dataset to estimate the Brix Index from hyperspectral vectors. The results were inadequate even if methods akin to those employed in the earlier analysis were applied. We look into the causes of this result, concentrating on the features and caliber of the dataset.

4.8.1 Dataset Description

Multispectral photos of individual grape berries with weight, anthocyanin content, and Brix Index measurements make up the dataset in question. Each image has dimensions of 140×200 pixels and is in the `.tif` format. The 37 spectral bands that make up each pixel record the intensity of reflected light at wavelengths between around 450 and 970 nm. *Autumn Royal*, *Crimson*, *Itum4*, *Itum5*, and *Itum9* are seedless table grape varieties that represent a range of maturity stages.

There are some significant differences between this dataset and the primary one that was previously examined. First of all, because berries naturally vary from one another, each sample is a single grape berry rather than an average of several berries, adding to the variability. Second, the dataset may have an impact on the spectral features and their relationship to the Brix Index because it concentrates on seedless table grapes rather than wine grape varietals. Thirdly, the imaging method uses 37-band multispectral imaging instead of the 204-band hyperspectral imaging that was previously employed. Finally, in order to filter and compress the images into spectral vectors that can be used for analysis, more preprocessing processes are needed.

4.8.2 Model Application

A number of preparation procedures were carried out in order to get the data ready for modeling. The grape fruit pixels were first separated from the backdrop using image segmentation. A threshold-based approach was used to do this, keeping pixels in channel 22 with intensity levels higher than 25. A single spectral vector of length 37 was produced for each grape berry by averaging the spectral values across all berry pixels in each image to determine the mean reflectance spectrum.

To comprehend the data structure and possible connections between spectral properties and the Brix Index, an exploratory data analysis was carried out. To reduce dimensionality and show the data, Principal Component Analysis (PCA) was used. More than 90% of the variance was explained by the first three main components. Correlation analyses, however, showed weak connections between the principal components and the Brix Index and low correlations between particular

spectral properties and the Brix Index. This suggested that there were not enough linear relationships for regression models to take advantage of.

As in the previous experiments, we used Partial Least Squared Regression. To find the ideal amount of components, cross-validation was done. The model's performance was subpar in spite of these efforts. High prediction errors were indicated by the Root Mean Squared Error (RMSE), which was similar to the Brix Index standard deviation. Given the poor coefficient of determination (R^2) values, it appears that the model only partially accounted for the variation in the Brix Index.

4.8.3 Discussion on Differences

The models' poor performance on this dataset seems to be related to the poor quality of the dataset itself, in particular the Brix Index measurment. With comparable spectral ranges covering important wavelengths related to grape composition and good correlation among spectral bands, as anticipated in hyperspectral data, the dataset looks similar when analyzed only based on the spectral data. This resemblance implies that the primary problem is not the spectral data per se. It's possible that the Brix Index values weren't gathered consistently or precisely enough. Any model's capacity to discover significant connections between the predictors and the target variable is directly hampered by inaccurate ground truth measurements. The most likely reason for the poor predictive performance is the carelessness with which the Brix Index measurements were collected, since the spectral data shares characteristics with the successful dataset. Inconsistencies in the Brix Index data or measurement errors may have a major impact on the model's capacity to generate reliable predicted correlations.

4.8.4 Conclusions

This case study emphasizes how crucial data quality is to predictive modeling, particularly for the target variable. Models cannot correct for errors in basic data, even with the right preprocessing and analytical methods. Developing successful predictive models in chemometrics and related domains requires high-quality, dependable measurements. Future research could enhance model performance by reassessing the Brix Index measurements, perhaps by collecting data again with more accurate techniques, and putting in place uniform procedures for data gathering. Furthermore, taking into account sophisticated preprocessing methods or different modeling strategies could aid in identifying intricate patterns in the data.

Chapter 5

Conclusion

The research conducted in this thesis once again highlights how the use of VIS-NIR spectroscopy is a valid and effective methodology for developing models that estimate fruit quality parameters. Although the results obtained are in line with the existing literature, the fundamental contribution of this thesis lies in its very structure. All aspects of this research are indeed analyzed by posing the main question as "why" before "how." Starting from some fundamental, almost philosophical questions regarding ML and system modeling, the reader is guided through a process that first leads them to discover the fundamentals of viticulture and imaging, which serve to build the assumptions at the base of the problem-solving, before moving on to practical implementations. The critical analysis and interpretation of the results, conducted in a meticulous manner, represents another innovative aspect of this work. The perspective from which the results are analyzed is always that of information, leading the reader to clearly understand the fundamentals of each model used and to question how the information conveyed by the result is related to what we know about the domain in question. To appreciate this thesis, it is fundamentally important to truly understand what the greatest difficulty a novice faces when approaching a problem in a specific domain they are unfamiliar with. One might indeed think that the greatest lack is the *knowledge* of the domain. This is true to some extent, in the sense that what is actually lacking is the part of domain knowledge that is strictly necessary to interface with the problem (reading an entire book on the use of pesticides in viticulture may help, but it is not strictly necessary to tackle and understand how to estimate the sugar content of grapes). The real problem is that most of the time, one does not know a priori what the necessary subset of knowledge is that provides the context to tackle the problem. What we are faced with, therefore, is a lack of *meta-knowledge*, that is, the kind that allows us to group, categorize, and conceptualize new knowledge through language. The only way to fill this gap is to be aware of it, look at the problem from the outside, and act methodically, always considering that there

might be something that has been overlooked. In this context, the value of this thesis is primarily the method with which it was conceived.

The limitations encountered are mainly related to the lack of public datasets that are numerous and well-provided. The future perspective is that more and more companies will start collecting and publishing their data in order to advance research in a more solid and faster manner. It is indeed by looking at new types of data that new methodologies and approaches to solving problems can more easily come to mind. Finally, having a vast amount of heterogeneous data, concerning different grape varieties, varying degrees of ripeness, and originating from areas with variable soils and climates, could allow for understanding whether universal methodologies exist and whether the models developed are transferable from one grape to another. This would not only promote the development of more robust and generalizable models but also allow for the exploration of the possibility of applying common approaches to different contexts, improving the prediction of grape quality parameters on a global scale. In this way, new perspectives for research would open up, facilitating the adoption of innovative techniques in the wine sector and promoting greater collaboration between researchers and industry professionals.

Appendix A

RoBoost PLSR Algorithm and Weighted NIPALS

In this appendix, we provide high-level pseudocode descriptions of the RoBoost PLSR algorithm and the weighted NIPALS algorithm used in our analysis.

A.1 RoBoost PLSR Algorithm

Algorithm 1 RoBoost PLSR Algorithm

Require:
- Dataset $\mathbf{X} \in \mathbb{R}^{n \times p}$ (predictor variables)
- Response variable $\mathbf{Y} \in \mathbb{R}^n$
- Number of components A
- Maximum iterations N_{iter}
- Tuning parameters α, β, γ
- Convergence threshold θ

Ensure: • Model parameters: weights $\mathbf{W}$, loadings $\mathbf{P}$, scores $\mathbf{T}$, regression coefficients $\mathbf{b}$

Initialize weights $\mathbf{d} \leftarrow \mathbf{1}_n$ (Vector of ones)
Center $\mathbf{X}$ and $\mathbf{Y}$ using weighted means with weights $\mathbf{d}$
for $a = 1$ to A **do**
  Set convergence criterion cor $\leftarrow 0$, iteration counter $f \leftarrow 1$
  while cor $< \theta$ **and** $f \leq N_{\text{iter}}$ **do**
    Compute weighted PLS component:
      Use weighted NIPALS algorithm (**Algorithm 2**) to compute weights $\mathbf{w}_a$, scores $\mathbf{t}_a$, loadings $\mathbf{p}_a$, and coefficient c_a with weights $\mathbf{d}$
      Update residuals:
        $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}_a \mathbf{p}_a^\top$
        $\mathbf{Y} \leftarrow \mathbf{Y} - c_a \mathbf{t}_a$
    Compute residuals:
      $\mathbf{r}_Y \leftarrow$ residuals of $\mathbf{Y}$
      $\mathbf{r}_X \leftarrow$ residuals of $\mathbf{X}$
      $\mathbf{r}_T \leftarrow$ residuals of $\mathbf{T}$
    Update weights using robust weight functions:
      $\mathbf{w}_Y \leftarrow F_\beta(\mathbf{r}_Y)$
      $\mathbf{w}_X \leftarrow F_\alpha(\mathbf{r}_X)$
      $\mathbf{w}_T \leftarrow F_\gamma(\mathbf{r}_T)$
    Update overall weights:
      $\mathbf{d} \leftarrow \mathbf{w}_Y \circ \mathbf{w}_X \circ \mathbf{w}_T$ (Element-wise multiplication)
    Normalize weights: $\mathbf{d} \leftarrow \mathbf{d} / \sum_{i=1}^n d_i$
    Compute convergence criterion cor
    Increment iteration counter: $f \leftarrow f + 1$
  end while
  Store component parameters $\mathbf{w}_a$, $\mathbf{t}_a$, $\mathbf{p}_a$, c_a
end for
Compute regression coefficients:
$\mathbf{R} \leftarrow \mathbf{W} \left(\mathbf{P}^\top \mathbf{W} \right)^{-1}$
$\mathbf{b} \leftarrow \mathbf{R}\mathbf{C}$ =0

A.2 Weighted NIPALS Algorithm

Algorithm 2 Weighted NIPALS Algorithm for PLS

Require:
- Centered predictor matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$
- Centered response vector $\mathbf{Y} \in \mathbb{R}^{n}$
- Number of components A
- Observation weights $\mathbf{d} \in \mathbb{R}^{n}$

Ensure:
- Weights $\mathbf{W}$, loadings $\mathbf{P}$, scores $\mathbf{T}$, regression coefficients $\mathbf{C}$

Initialize residual matrices: $\mathbf{X}_0 \leftarrow \mathbf{X}$, $\mathbf{Y}_0 \leftarrow \mathbf{Y}$
for $a = 1$ to A **do**
  Compute weight vector:
    $\mathbf{w}_a \leftarrow \mathbf{X}_{a-1}^{\top}(\mathbf{d} \circ \mathbf{Y}_{a-1})$
    Normalize: $\mathbf{w}_a \leftarrow \mathbf{w}_a / \|\mathbf{w}_a\|$
  Compute score vector:
    $\mathbf{t}_a \leftarrow \mathbf{X}_{a-1}\mathbf{w}_a$
  Compute regression coefficient:
    $c_a \leftarrow (\mathbf{d} \circ \mathbf{Y}_{a-1})^{\top}\mathbf{t}_a / (\mathbf{t}_a^{\top}(\mathbf{d} \circ \mathbf{t}_a))$
  Compute loading vector:
    $\mathbf{p}_a \leftarrow \mathbf{X}_{a-1}^{\top}(\mathbf{d} \circ \mathbf{t}_a) / (\mathbf{t}_a^{\top}(\mathbf{d} \circ \mathbf{t}_a))$
  Update residuals:
    $\mathbf{X}_a \leftarrow \mathbf{X}_{a-1} - \mathbf{t}_a\mathbf{p}_a^{\top}$
    $\mathbf{Y}_a \leftarrow \mathbf{Y}_{a-1} - c_a\mathbf{t}_a$
  Store component parameters $\mathbf{w}_a$, $\mathbf{t}_a$, $\mathbf{p}_a$, c_a
end for=0

Bibliography

[1] Wikipedia contributors. *Viticoltura — Wikipedia, The Free Encyclopedia*. Accessed: 2024-11-26. 2024. URL: https://it.wikipedia.org/wiki/Viticoltura (cit. on p. 1).

[2] Wikipedia contributors. *Enologia — Wikipedia, The Free Encyclopedia*. Accessed: 2024-11-26. 2024. URL: https://it.wikipedia.org/wiki/Enologia (cit. on p. 1).

[3] Roger B. Boulton, Vernon L. Singleton, Linda F. Bisson, and Ralph E. Kunkee. *Principles and practices of winemaking*. New York: Chapman & Hall, 1996 (cit. on pp. 1, 2, 4).

[4] Pascal Ribéreau-Gayon, Denis Dubourdieu, Bernard Donèche, and Aline Lonvaud. *Handbook of Enology, Volume 1: The Microbiology of Wine and Vinifications*. Trans. by Jr. Jeffrey M. Branco and Christine Rychlewski. 2nd. Chichester, England: John Wiley & Sons, 2006. ISBN: 978-0-470-01034-1 (cit. on pp. 2, 10).

[5] Bruce W. Zoecklein, Kenneth C. Fugelsang, Barry H. Gump, and Fred S. Nury. *Wine Analysis and Production*. Originally published by Chapman & Hall, 1995. New York, NY: Kluwer Academic/Plenum Publishers, 1999. ISBN: 0-8342-1701-5 (cit. on p. 2).

[6] Raffaele Guzzon, Fulvio Mattivi, Maurizio Ferrari, and Alberto Menta. *Enologia e biotecnologie vitivinicole*. National relevance. Bologna, Italy: Zanichelli, 2023. ISBN: 9788808553409 (cit. on pp. 2, 11).

[7] *Wine Cooler Direct*. Accessed: 2024-11-28. 2024. URL: https://winecoolerdirect.com (cit. on p. 3).

[8] Evineyard. *vineyard calendar*. https://www.evineyardapp.com/blog/wp-content/uploads/2016/02/vineyard-calendar.png. Accessed: 02/12/2024 (cit. on p. 4).

[9] *Commento di Roberto Zironi sulla viticoltura*. Rubrica radiofonica TgrRaiFVG Vita Nei Campi. Accademico ordinario dell'Accademia Italiana della Vite e del Vino. 2024 (cit. on p. 5).

[10] Lodigrowers. *Bunch Morphology*. https://www.lodigrowers.com/wp-content/uploads/2014/05/Capture1.jpg. Accessed: 02/12/2024 (cit. on p. 6).

[11] *Morfologia della Vite*. Accessed: 2024-11-28. 2024. URL: https://www.agraria.org/viticoltura-enologia/morfologia-della-vite (cit. on p. 7).

[12] Daywen Vineyard. *Grape Morphology*. https://daywen-vineyard-grape juice-and-organically-grown-grapes.co.nz/wp-content/uploads/2015/12/grape-cross-section.png. Accessed: 02/12/2024 (cit. on p. 8).

[13] James B. Campbell and Randolph H. Wynne. *Introduction to Remote Sensing*. 5th. New York, NY: The Guilford Press, 2011. ISBN: 978-1-60918-176-5 (cit. on pp. 13, 19).

[14] Mapir. *Light reflection of objects*. https://www.mapir.camera/en-gb/pages/what-is-reflectance-calibration. Accessed: 02/12/2024 (cit. on p. 14).

[15] Nireos. *Comparison of RGB, Multispectral, and Hyperspectral Imaging*. https://nireos.com/application/what-is-hyperspectral-imaging/. Accessed: 02/12/2024 (cit. on p. 17).

[16] Giuseppe Montanaro, Angelo Petrozza, Laura Rustioni, Francesco Cellini, Antonio Carlomagno, and Vitale Nuzzo. «Measuring fruit quality traits in olive through RGB imaging and artificial neural networks: opportunities and limitations». In: *2023 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*. 2023, pp. 685–688. DOI: 10.1109/MetroAgriFor58484.2023.10424226 (cit. on pp. 21, 22).

[17] H. Zhu, B. Chu, Y. Fan, et al. «Hyperspectral Imaging for Predicting the Internal Quality of Kiwifruits Based on Variable Selection Algorithms and Chemometric Models». In: *Scientific Reports* 7 (2017). Accessed: 2024-11-28, p. 7845. DOI: 10.1038/s41598-017-08509-6. URL: https://doi.org/10.1038/s41598-017-08509-6 (cit. on p. 22).

[18] Parika Rungpichayapichet, Marcus Nagle, Pasinee Yuwanbun, Pramote Khuwijitjaru, Busarakorn Mahayothee, and Joachim Müller. «Prediction mapping of physicochemical properties in mango by hyperspectral imaging». In: *Biosystems Engineering* 159 (2017). Accessed: 2024-11-28, pp. 109–120. ISSN: 1537-5110. DOI: 10.1016/j.biosystemseng.2017.04.006. URL: https://www.sciencedirect.com/science/article/pii/S1537511016300836 (cit. on p. 22).

[19] C Nugroho, Makhmudun Ainuri, and Mohammad Falah. «Physical quality determination of fresh strawberry (Fragaria x ananassa var. Osogrande) fruit in tropical environment using image processing approach». In: *IOP Conference Series: Earth and Environmental Science* 759 (May 2021), p. 012020. DOI: 10.1088/1755-1315/759/1/012020 (cit. on p. 22).

[20] A. R. Mesa and J. Y. Chiang. «Multi-Input Deep Learning Model with RGB and Hyperspectral Imaging for Banana Grading». In: *Agriculture* 11.8 (2021), p. 687. DOI: 10.3390/agriculture11080687. URL: https://doi.org/10.3390/agriculture11080687 (cit. on p. 23).

[21] Camilla Menozzi, Rosalba Calvini, Giovanni Nigro, Paola Tessarin, Domenico Bossio, Marco Calderisi, Veronica Ferrari, Giorgia Foca, and Alessandro Ulrici. «Design and application of a smartphone-based device for in vineyard determination of anthocyanins content in red grapes». In: *Microchemical Journal* 191 (2023), p. 108811. ISSN: 0026-265X. DOI: 10.1016/j.microc.2023.108811. URL: https://www.sciencedirect.com/science/article/pii/S0026265X23004290 (cit. on p. 23).

[22] Shuang Tian and Hui Xu. «Nondestructive Methods for the Quality Assessment of Fruits and Vegetables Considering Their Physical and Biological Variability». In: *Food Engineering Reviews* 14 (2022), pp. 380–407. DOI: 10.1007/s12393-021-09300-0. URL: https://doi.org/10.1007/s12393-021-09300-0 (cit. on p. 23).

[23] Iylia Adhwa Mazni, Samsul Setumin, Mohamed Syazwan Osman, Khusairi Osman, and Mohd Subri Tahir. «Systematic Literature Review Approach on the Fruit Quality Assessment Based on Fruit Imaging Techniques». In: *Journal of Electrical and Electronic Systems Research* 21 (2022), pp. 139–147. DOI: 10.24191/jeesr.v21i1.019. URL: https://doi.org/10.24191/jeesr.v21i1.019 (cit. on p. 23).

[24] Shanmuga Sundaram Anandan P. Pathmanaban B.K. Gnanavel. «Recent application of imaging techniques for fruit quality assessment». In: *Trends in Food Science & Technology* 94 (2019), pp. 32–42. ISSN: 0924-2244. DOI: 10.1016/j.tifs.2019.10.004. URL: https://www.sciencedirect.com/science/article/pii/S0924224418307374 (cit. on p. 23).

[25] V. M. Gomes, A. M. Fernandes, A. Faia, and P. Melo-Pinto. «Comparison of different approaches for the prediction of sugar content in new vintages of whole Port wine grape berries using hyperspectral imaging». In: *Computers and Electronics in Agriculture* 140 (2017), pp. 244–254 (cit. on pp. 30, 35, 36, 40, 41).

[26] V. Gomes, M. S. Reis, F. Rovira-Más, A. Mendes-Ferreira, and P. Melo-Pinto. «Prediction of Sugar Content in Port Wine Vintage Grapes Using Machine Learning and Hyperspectral Imaging». In: *Processes* 9.7 (2021), p. 1241 (cit. on pp. 31, 35, 36).

[27] R. Silva, V. Gomes, A. Mendes-Faia, and P. Melo-Pinto. «Using Support Vector Regression and Hyperspectral Imaging for the Prediction of Oenological Parameters on Different Vintages and Varieties of Wine Grape Berries». In: *Remote Sensing* 10.2 (2018), p. 312 (cit. on pp. 31, 35–37).

[28] A. Benelli, C. Cevoli, L. Ragni, and A. Fabbri. «In-field and non-destructive monitoring of grapes maturity by hyperspectral imaging». In: *Biosystems Engineering* 207 (2021), pp. 59–67 (cit. on pp. 32, 35–37).

[29] M. Gabrielli, V. Lançon-Verdier, P. Picouet, and C. Maury. «Hyperspectral Imaging to Characterize Table Grapes». In: *Chemosensors* 9 (2021), p. 71 (cit. on pp. 32, 35–37).

[30] A. Courand, M. Metz, D. Héran, C. Feilhes, F. Prezman, E. Serrano, R. Bendoula, and M. Ryckewaert. «Evaluation of a robust regression method (RoBoost-PLSR) to predict biochemical variables for agronomic applications: Case study of grape berry maturity monitoring». In: *Chemometrics and Intelligent Laboratory Systems* 221 (2022), p. 104485 (cit. on pp. 32, 35, 37, 39).

[31] P. J. Navarro, L. Miller, M. V. Díaz-Galián, and M. Egea-Cortines. «A novel ground truth multispectral image dataset with weight, anthocyanins and Brix index measures of grape berries tested for its utility in machine learning pipelines». In: *GigaScience* 11 (2022), giac052 (cit. on pp. 33, 37, 46).

[32] Pedro J. Navarro, Leanne Miller, Alberto Gila-Navarro, María Victoria Díaz-Galián, Diego J. Aguila, and Marcos Egea-Cortines. «3DeepM: An Ad Hoc Architecture Based on Deep Learning Methods for Multispectral Image Classification». In: *Remote Sensing* 13.4 (2021), p. 729. DOI: 10.3390/rs13040729. URL: https://doi.org/10.3390/rs13040729 (cit. on p. 33).

[33] Nikolaos L. Tsakiridis, Nikiforos Samarinas, Stylianos Kokkas, Eleni Kalopesa, Nikolaos V. Tziolas, and George C. Zalidis. «In situ grape ripeness estimation via hyperspectral imaging and deep autoencoders». In: *Computers and Electronics in Agriculture* 212 (2023), p. 108098. ISSN: 0168-1699. DOI: 10.1016/j.compag.2023.108098. URL: https://www.sciencedirect.com/science/article/pii/S0168169923004866 (cit. on p. 34).

[34] Min Xu, Jun Sun, Kunshan Yao, Qiang Cai, Jifeng Shen, Yan Tian, and Xin Zhou. «Developing deep learning based regression approaches for prediction of firmness and pH in Kyoho grape using Vis/NIR hyperspectral imaging». In: *Infrared Physics & Technology* 120 (2022), p. 104003. ISSN: 1350-4495. DOI: 10.1016/j.infrared.2021.104003. URL: https://www.sciencedirect.com/science/article/pii/S1350449521003753 (cit. on p. 34).

[35] Armando Manuel Fernandes, Paula Oliveira, João Paulo Moura, Ana Alexandra Oliveira, Virgílio Falco, Maria José Correia, and Pedro Melo-Pinto. «Determination of anthocyanin concentration in whole grape skins using hyperspectral imaging and adaptive boosting neural networks». In: *Journal of Food Engineering* 105.2 (2011), pp. 216–226. ISSN: 0260-8774. DOI: 10.1016/j.jfoodeng.2011.02.018. URL: https://www.sciencedirect.com/science/article/pii/S0260877411000823 (cit. on p. 34).

[36] Armando M. Fernandes, Camilo Franco, Ana Mendes-Ferreira, Arlete Mendes-Faia, Pedro Leal da Costa, and Pedro Melo-Pinto. «Brix, pH and anthocyanin content determination in whole Port wine grape berries by hyperspectral imaging and neural networks». In: *Computers and Electronics in Agriculture* 115 (2015), pp. 88–96. ISSN: 0168-1699. DOI: 10.1016/j.compag.2015.05.013. URL: https://www.sciencedirect.com/science/article/pii/S0168169915001490 (cit. on p. 34).

[37] Fang Cao, Di Wu, and Yong He. «Soluble solids content and pH prediction and varieties discrimination of grapes based on visible–near infrared spectroscopy». In: *Computers and Electronics in Agriculture* 71 (2010), S15–S18. ISSN: 0168-1699. DOI: 10.1016/j.compag.2009.05.011. URL: https://www.sciencedirect.com/science/article/pii/S0168169909000957 (cit. on p. 34).

[38] Maxime Ryckewaert, Daphné Héran, Carole Feilhes, Fanny Prezman, Eric Serrano, Aldrig Courand, Silvia Mas-Garcia, Maxime Metz, and Ryad Bendoula. «Dataset containing spectral data from hyperspectral imaging and sugar content measurements of grapes berries in various maturity stage». In: *Data in Brief* 46 (2023), p. 108822. ISSN: 2352-3409. DOI: 10.1016/j.dib.2022.108822. URL: https://www.sciencedirect.com/science/article/pii/S2352340922010253 (cit. on p. 44).

[39] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Second. Sebastopol, CA: O'Reilly Media, Inc., 2019. ISBN: 978-1-492-03264-9 (cit. on p. 51).

[40] Scikit-Learn Developers. *Principal Component Regression (PCR) vs Partial Least Squares Regression (PLS)*. Accessed: 2024-11-29. 2024. URL: https:

//scikit-learn.org/dev/auto_examples/cross_decomposition/plot_pcr_vs_pls.html (cit. on p. 51).

[41] Svante Wold, Michael Sjöström, and Lennart Eriksson. «PLS-regression: a basic tool of chemometrics». In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130. ISSN: 0169-7439. DOI: 10.1016/S0169-7439(01)00155-1. URL: https://www.sciencedirect.com/science/article/pii/S0169743901001551 (cit. on p. 51).

[42] Maxime Metz, Florent Abdelghafour, Jean-Michel Roger, and Matthieu Lesnoff. «A novel robust PLS regression method inspired from boosting principles: RoBoost-PLSR». In: *Analytica Chimica Acta* 1179 (2021), p. 338823. ISSN: 0003-2670. DOI: 10.1016/j.aca.2021.338823. URL: https://www.sciencedirect.com/science/article/pii/S0003267021006498 (cit. on p. 52).

[43] Fabian Pedregosa et al. «Scikit-learn: Machine Learning in Python». In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 59).