FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Proximal sensing of plants in an Android context using spectroradiometric sensors

Mariana Achando Coimbra



Master in Electrical and Computer Engineering

Supervisor: Prof. Filipe Neves dos Santos

Jul 31, 2026

© Mariana Achando Coimbra, 2026

Proximal sensing of plants in an Android context using spectroradiometric sensors

Mariana Achando Coimbra

Master in Electrical and Computer Engineering

Approved in oral examination by the committee:

President: Prof. Vitor Manuel Gomes Correia
Examiner: Prof. João Carlos Sampaio Maldonado Costa Mendes
Supervisor: Prof. Filipe Neves dos Santos

Jul 31, 2026

Sumário

A agricultura de precisão requer informação oportuna, precisa e acessível sobre o estado fisiológico das plantas, para apoiar uma tomada de decisão eficaz no terreno. No entanto, os agricultores têm falta de ferramentas acessíveis e dos conhecimentos técnicos necessários para avaliar a saúde das plantas diretamente no local. Esta tese aborda esta lacuna, investigando plataformas de diagnóstico baseadas em smartphones, integradas com sensores espectrorradiométricos de baixo custo, para a monitorização proximal do estado das plantas.

O trabalho desenvolve e avalia modelos de aprendizagem computacional, capazes de estimar parâmetros agronómicos chave, teor de água, teor de clorofila, teor de nitrogénio e inflorescências a partir de dados de refletância espectral. Foram seguidas duas linhas experimentais complementares: um estudo de *benchmarking* algorítmico utilizando conjuntos de dados espectrais publicados (abacate, oliveira e vinha) sob restrições de comprimento de onda, comparando algoritmos de classificação e regressão, estratégias de normalização e técnicas de redução de dimensionalidade; e o BudScan, um protocolo original de recolha de dados no terreno e em estufa, concebido para relacionar a refletância espectral dos botões da videira com o desenvolvimento reprodutivo, incluindo abordagens de classificação multiclasse e de classe única.

Os resultados mostram que o teor de água foi o parâmetro previsto de forma mais fiável em todos os conjuntos de dados, com a Regressão Logística (LR) e a Máquina de Vetores de Suporte (SVM) a apresentarem um desempenho consistentemente superior ao de outros modelos após a otimização dos hiperparâmetros, e que os intervalos espectrais ótimos variam significativamente consoante a espécie e o alvo. No caso do BudScan, a simplificação da tarefa de classificação de três para duas classes melhorou substancialmente a fiabilidade, enquanto os desafios relacionados com a oclusão do tecido dos botões, a sensibilidade à luz solar e o desequilíbrio entre classes foram identificados como fatores limitantes. Um sensor multiespectral simplificado (CropSign Lite) foi também validado em comparação com uma referência hiperespectral, alcançando uma precisão comparável para vários indicadores.

Em geral, este trabalho demonstra a viabilidade de combinar *hardware* espectral de baixo custo, aprendizagem automática e diagnóstico de plantas em tempo real, contribuindo para um caminho prático rumo a um apoio à decisão acessível e baseado em dados para a agricultura de precisão.

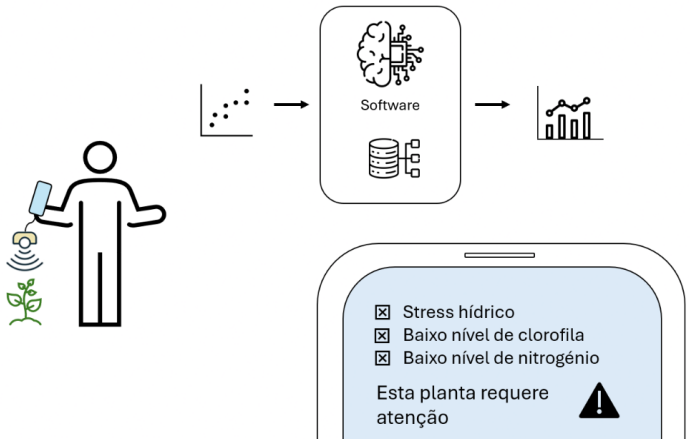


Figure 1: Sumário gráfico.

Abstract

Precision agriculture requires timely, accurate, and accessible information on crop physiological status to support effective decision-making in the field. However, farmers often lack affordable tools and the technical expertise needed to assess plant health directly on-site. This dissertation addresses this gap by investigating smartphone-based diagnostic platforms integrated with low-cost spectroradiometric sensors for proximal monitoring of plant condition.

The work develops and evaluates machine learning models, capable of estimating key agronomic parameters, water content, chlorophyll content, nitrogen content, and inflorescences from spectral reflectance data. Two complementary experimental tracks were pursued: an algorithmic benchmarking study using published spectral datasets (avocado, olive, and vineyard) under wavelength constraints, comparing classification and regression algorithms, normalization strategies, and dimensionality reduction techniques; and BudScan, an original field and greenhouse data-collection protocol designed to relate grapevine bud spectral reflectance to reproductive development, including multi-class and one-class classification approaches.

Results show that water content was the most reliably predicted parameter across datasets, with Logistic Regression (LR) and Support Vector Machine (SVM) consistently outperforming other models after hyperparameter optimization, and that optimal spectral ranges highly species and target. For BudScan, simplifying the classification task from three to two classes substantially improved accuracy, while challenges related to bud tissue occlusion, sunlight sensitivity, and class imbalance were identified as limiting factors. A simplified multispectral sensor (CropSign Lite) was also validated against a hyperspectral reference, achieving comparable accuracy for several indicators.

Overall, this work demonstrates the feasibility of combining low-cost spectral hardware, machine learning, real-time plant diagnostics, contributing a practical pathway toward affordable, data-driven decision support for precision agriculture.

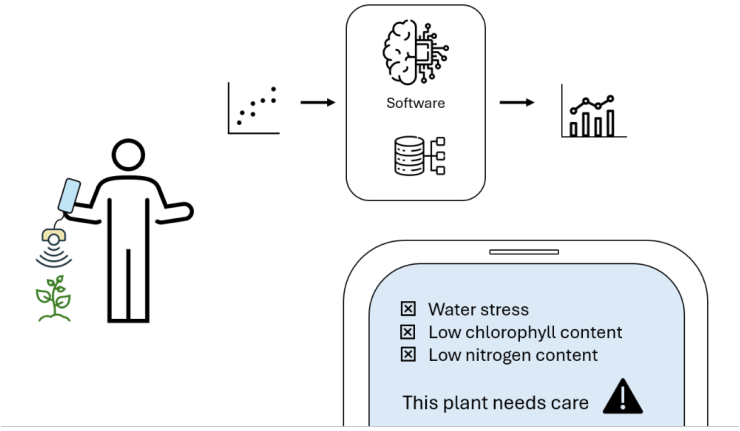


Figure 2: Graphical abstract.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework for achieving a better and more sustainable future for all. It includes 17 goals[1] to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

The specific Sustainable Development Goals mentioned have the following names:

SDG 2 End hunger, achieve food security and improved nutrition and promote sustainable agriculture

SDG 12 Ensure sustainable consumption and production patterns

SGD	Target	Contribution	Performance Indicators and Metrics
2	2.3	Integrating smartphone application with low-cost proximal sensing tool and decision making support.	Increase the access to agricultural productivity enhancing technology.
	2.4	Monitor plant physiological status (chlorophyll, nitrogen, water, influrescences) in real time, supporting data-driven decisions on fertilization and irrigation, optimizing resource allocation	Increase sustainable food production systems and implement resilient agricultural practices.
12	12.2	Proximal sensing enables proactive intervention before visual damage occurs, preventing crop loss, resource miscalculation, and input waste. Not only that, but it leads to precise allocation eliminating over-fertilization and over-watering, reducing cost and environmental runoff.	Deviation in resource consumption compared to conventional uniform treatment models.

These SDGs call for more productive and resilient agricultural systems, technological innovation, efficient use of natural resources as part of the 2030 Agenda.

[1] https://sdgs.un.org/goals

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Filipe Santos, and my advisor at INESC TEC, Pedro Moura, for their guidance, support, and valuable knowledge throughout the development of this work. To Germano Moreira for your collaboration and availability which contributed to the development of this disseratation. To all the members of the TRIBE Laboratory for providing a good working environment.

I would like to thank, Pedro Beirão, for supporting me throughout the process of writing my dissertation from day one, for encouraging me to carry on, and for always believing in me.

To all my friends from faculty, my deepest appreciation for the companionship and support shown in all stages, to my friends who have watched me grow to become an engineer. I am thankful to my family, particularly my sister and my parents for their support throughout my time in university.

Mariana Coimbra

"Start where you are. Use what you have. Do what you can."

Arthur Ashe

Declaration of honour

I declare that this dissertation is my original work and has not previously been submitted or assessed in any other course or curricular unit, whether at this or any other institution. References to other authors (statements, ideas, thoughts), as well as the use of published works fully complies with the rules of attribution and are appropriately cited in the text and bibliography, in accordance with reference standards. I am aware that plagiarism and self-plagiarism constitute academic misconduct.
I further declare that I have used Artificial Intelligence tools to rewrite parts of this dissertation and the results were carefully reviewed for the detection of errors, inaccuracies, unfounded attributions, or other forms of content and argument bias, as well as for the determination of authorship.
I further declare that I have included studies carried out with multiple co-authors, having specified, with transparency and rigour, the contribution of each co-author, provided information on the possible existence of any other dissertation in which the same article has been included, reproduced the work in full with the permission of the journal in which it was published, and obtained the explicit written consent of all co-authors for the inclusion of the article.

Mariana Achando Coimbra

Contents

Acknowledgements v

1 Introduction 1
- 1.1 Context . . . 1
- 1.2 Objectives . . . 1
- 1.3 Motivation . . . 2
- 1.4 Dissertation Structure . . . 3

2 Background and Fundamental Aspects 5
- 2.1 Precision agriculture . . . 5
- 2.2 Spectrometry . . . 6
  - 2.2.1 Reflectance, transmittance and emission phenomena . . . 6
  - 2.2.2 Ultraviolet-Visible-Near-Infrared (UV-VIS-NIR) Spectroscopy . . . 6
  - 2.2.3 Fluorescence Spectroscopy (FS) . . . 7
  - 2.2.4 Raman Spectroscopy (RaS) . . . 7
- 2.3 Data collection, visualization and local processing . . . 7
  - 2.3.1 Preprocessing and Scaling Methods . . . 7
  - 2.3.2 Feature engineering and feature selection . . . 9
  - 2.3.3 Predictive Modelling Strategies . . . 9
  - 2.3.4 Model Performance Metrics . . . 12
- 2.4 Built-in Smartphone Sensors and Hardware Components . . . 13

3 Bibliography review 15
- 3.1 Plant and fruit monitoring . . . 16
- 3.2 Sensors for plant and fruit monitoring . . . 16
  - 3.2.1 Sensors application . . . 18
  - 3.2.2 Measurements . . . 21
- 3.3 Problems associated with model transferability . . . 21
- 3.4 Smartphone-Based Agricultural Monitoring Systems . . . 21
- 3.5 Spectral datasets for plant and fruit monitoring . . . 22
- 3.6 Analysis of Machine Learning (ML) approaches for crop spectral measurements . . . 24
  - 3.6.1 Federated learning in Android applications and agricultural contexts . . . 26
- 3.7 Framing the dissertation problem in the state-of-the-art . . . 26

4 Methodology 28
- 4.1 Characteristics of the devices being tested . . . 28
  - 4.1.1 CropSign Lite . . . 28
  - 4.1.2 CropSign Pro . . . 29

4.2 Algorithmic Benchmarking Under Wavelength Constraints . . . 31
4.2.1 Tools . . . 32
4.2.2 Classification Procedure . . . 32
4.2.3 Regression Procedure . . . 33
4.3 Data collection protocol: BudScan . . . 36
4.3.1 Experimental Setup . . . 36
4.4 Data analysis: BudScan . . . 40
4.4.1 Initial Processing . . . 40
4.4.2 Data Splitting and Model Training . . . 40

5 Results and Discussion **43**
5.1 Algorithmic Benchmarking Under Wavelength Constraints . . . 43
5.1.1 Classification Algorithms . . . 43
5.1.2 Classifier Discussion . . . 65
5.1.3 Regressor Algorithms . . . 66
5.1.4 Regressor Discussion . . . 72
5.1.5 Dataset Augmentation . . . 73
5.2 BudScan Results . . . 75
5.2.1 Classifier Algorithm . . . 75
5.2.2 One-Class Classification (OCC) . . . 82
5.2.3 One-Class Classification (OCC) Discussion . . . 84

6 Conclusion **85**
6.1 Conclusions . . . 85
6.2 Future Work . . . 86

References **87**

A Algorithmic Benchmarking Under Wavelength Constraints **95**
A.1 Classification algorithms . . . 95
A.1.1 Water Content in Avocado . . . 95
A.1.2 Chlorophyll Content in Avocado . . . 97
A.1.3 Nitrogen Content in Avocado . . . 100
A.1.4 Water Content in Olive . . . 103
A.1.5 Chlorophyll Content in Olive . . . 105
A.1.6 Nitrogen Content in Olive . . . 107
A.1.7 Water Content in Vineyard . . . 109
A.1.8 Chlorophyll Content in Vineyard . . . 111
A.1.9 Nitrogen Content in Vineyard . . . 112
A.2 Regressor Algorithms . . . 114

List of Figures

1 Sumário gráfico. . . . ii
2 Graphical abstract. . . . iii

1.1 CropSign mobile application with a commercial spectral sensor soldered onto a PCB. . . . 3
1.2 CropSign with implemented control logic to receive instructions and send spectral data via USB connection to a smartphone. . . . 3

2.1 Precision Agriculture (PA) [6]. . . . 5
2.2 Reflectance, transmittance and emission phenomena [9]. . . . 7

3.1 Pipeline of the decision making system [28]. . . . 15
3.2 Wavelengths of Multispectral and Hyperspectral [34]. . . . 17
3.3 Relation between Wavelength and Reflectance [13]. . . . 18
3.4 Methods to detect diseases in each phase [13]. . . . 19
3.5 Relation between Normalized Difference Vegetation Index (NDVI), Near-Infrared (NIR) and RGB. . . . 22
3.6 ML algorithms to detect plant stress [45]. . . . 25

4.1 Typical Light-Emitting Diode (LED) Spectra Used in Final Test of AS7341 [62]. 29
4.2 Measured Spectral Responsivity Relative to F8 [62]. . . . 29
4.3 Spectra of an RGB LED measured with the Open-source miniature spectrophotometer (OSMS) (C12880MA) and the commercial system Com. [63]. . . . 30
4.4 Normalized Spectral Responsivity [64]. . . . 31
4.5 Schematic flow diagram of methodology adopted for the Classification Algorithm. 34
4.6 Schematic flow diagram of methodology adopted for the Regression Algorithm. . 35
4.7 Samples' location points in Centro de Estudos Vitivinícolas do Dão. . . . 36
4.8 BudScan system. . . . 38
4.9 In field BudScan measurements. . . . 38
4.10 Sample labeling. . . . 39
4.11 Labeled Canes in the Greenhouse. . . . 39
4.12 DD-SIMCA pipeline diagram. . . . 42

5.1 Wavelengths vs Reflectance by Class: Water Content Avocado. . . . 45
5.2 Confusion Matrices: Water Content Avocado. . . . 45
5.3 Wavelengths vs Reflectance by Class: Chlorophyll Content Avocado. . . . 47
5.4 Confusion Matrices: Chlorophyll Content Avocado. . . . 47
5.5 Wavelengths vs Reflectance by Class: Nitrogen Content Avocado. . . . 49
5.6 Confusion Matrices: Nitrogen Content Avocado. . . . 49
5.7 Wavelengths vs Reflectance by Class: Water Content Olive. . . . 51

5.8 Confusion Matrices: Water Content Olive. 51
5.9 Wavelengths vs Reflectance by Class: Chlorophyll Content Olive. 53
5.10 Confusion Matrices: Chlorophyll Content Olive. 53
5.11 Wavelengths vs Reflectance by Class: Nitrogen Content Olive. 55
5.12 Confusion Matrices: Nitrogen Content Olive. 55
5.13 Wavelengths vs Reflectance by Class: Water Content Vineyard. 57
5.14 Confusion Matrices: Water Content Vineyard. 58
5.15 Wavelengths vs Reflectance by Class: Chlorophyll Content in Vineyard. 60
5.16 Confusion Matrices: Chlorophyll Content Vineyard. 60
5.17 Wavelengths vs Reflectance: Nitrogen Content in Vineyard. 62
5.18 Confusion Matrices: Nitrogens Content Vineyard. 62
5.19 Vegetation Indices (VIs) Hyperspectral sensor VS VIs Multispectral sensor (Lite Discrete): Avocado. 64
5.20 VIs Hyperspectral sensor VS VIs Multispectral sensor (Lite Discrete): Olive. . . 64
5.21 VIs Hyperspectral sensor VS VIs Multispectral sensor (Lite Discrete): Vineyard. 65
5.22 Wavelengths vs Reflectance: Water Content in Avocado. 67
5.23 Ground Truth vs Predictions. 67
5.24 Wavelength Reflectance: Chlorophyll Content in Avocado. 69
5.25 Ground Truth vs Predictions. 69
5.26 Wavelength Reflectance. 71
5.27 Ground Truth vs Predictions. 71
5.28 Light and Dark Mean Representations 75
5.29 Spectra BudScan 76
5.30 Data-Driven Soft Independent Modelling of Class Analogy (DD-SIMCA) BudScan Results. 83
5.31 One-Class Support Vector Machine (SVM). 83

A.1 Importances of Water Content in Avocado. 95
A.2 Optimized Logistic Regression (LR) Confusion Matrix for NIR Chlorophyll Content Avocado. 97
A.3 Importances from Random Forest (RF) model NIR. 98
A.4 Importances of Chlorophyll Content in Avocado. 98
A.5 Optimized SVM Confusion Matrix for NIR Nitrogen Avocado. 100
A.6 Importances from RF model NIR. 101
A.7 Importances of Nitrogen Content in Avocado. 101
A.8 Importances of Water Content in Olive. 103
A.9 Importances of Chlorophyll Content in Olive. 105
A.10 Confusion Matrix: Nitrogen Content in Olive 107
A.11 Importances of Nitrogen Content in Olive. 107
A.12 Importances of Water Content in Vineyard. 109
A.13 Importances of Chlorophyll Content in Vineyard. 111
A.14 Importances of Nitrogen Content in Vineyard. 112
A.15 Importances 114

List of Tables

2.1 Confusion matrix for binary classification evaluation. 12

3.1 Summary of Experimental Conditions and Sensor Specifications. 20

5.1 Performance Metrics for Wavelength Water Content Avocado. 46
5.2 Performance Metrics for Wavelength Chlorophyll Content Avocado. 48
5.3 Performance Metrics for Wavelength Nitrogen Content Avocado. 50
5.4 Performance Metrics for Wavelength Water Content Olive. 52
5.5 Performance Metrics for Wavelength Chlorophyll Content Olive. 54
5.6 Performance Metrics for Wavelength Nitrogen Content Olive. 56
5.7 Performance Metrics for Wavelength Water Content Vineyard. 59
5.8 Performance Metrics for Wavelength Chlorophyll Content Vineyard. 61
5.9 Performance Metrics for Wavelength Nitrogen Content Vineyard. 63
5.10 Performance Metrics for Wavelength Water Content Avocado Regressor. 68
5.11 Performance Metrics for Wavelength Chlorophyll Avocado Regressor. 70
5.12 Performance Metrics for Wavelength Nitrogen Avocado Regressor. 72
5.13 Performance Metrics for Wavelength Chlorophyll Avocado Augmentation. 74
5.14 Performance Metrics for BudScan Three classes. 77
5.15 LR, Decision Tree (DT) and K-Nearest Neighbour (KNN) Best Hyperparameters using GridSearchCV BudScan Three Classes. 78
5.16 SVM Best Hyperparameters using GridSearchCV BudScan Three Classes. 78
5.17 Performance Metrics for BudScan Two Classes. 80
5.18 LR, DT and KNN Best Parameters using GridSearchCV BudScan Two Classes. . 81
5.19 SVM Best Parameters using GridSearchCV BudScan Two Classes. 81

A.1 LR, DT and KNN Best Hyperparameters using GridSearchCV Water Content Avocado. 96
A.2 SVM Best Hyperparameters using GridSearchCV Water Content Avocado. . . . 96
A.3 LR, DT and KNN Best Hyperparameters using GridSearchCV Chlorophyll Content Avocado. 99
A.4 SVM Best Hyperparameters using GridSearchCV Chlorophyll Content Avocado. 99
A.5 LR, DT and KNN Best Hyperparameters using GridSearchCV Nitrogen Content Avocado. 102
A.6 SVM Best Hyperparameters using GridSearchCV Nitrogen Content Avocado. . . 102
A.7 LR, DT and KNN Best Hyperparameters using GridSearchCV Water Content Olive. 104
A.8 SVM Best Hyperparameters using GridSearchCV Water Content Olive. 104
A.9 LR, DT and KNN Best Hyperparameters using GridSearchCV Chlorophyll Content Olive. 106
A.10 SVM Best Hyperparameters using GridSearchCV Chlorophyll Content Olive. . . 106

A.11 LR, DT and KNN Best Hyperparameters using GridSearchCV Nitrogen Content Olive. 108
A.12 SVM Best Hyperparameters using GridSearchCV Nitrogen Content Olive. 108
A.13 LR, DT and KNN Best Hyperparameters using GridSearchCV Water Content Vineyard. 110
A.14 SVM Best Hyperparameters using GridSearchCV Water Content Vineyard. . . . 110
A.15 LR, DT and KNN Best Hyperparameters using GridSearchCV Chlorophyll Content Vineyard. 111
A.16 SVM Best Hyperparameters using GridSearchCV Chlorophyll Content Vineyard. 112
A.17 LR, DT and KNN Best Hyperparameters using GridSearchCV Nitrogen Content Vineyard. 113
A.18 SVM Best Hyperparameters using GridSearchCV Nitrogen Content Vineyard. . . 113

List of Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Networks
CNN	Convolutional Neural Network
CSV	Comma Separated Values
DD-SIMCA	Data-Driven Soft Independent Modelling of Class Analogy
DE	Detrend
DS	Direct Standardization
DSS	Decision Support System
DT	Decision Tree
FELIDS	Federated Learning-based Intrusion Setection System
FL	Federated learning
FS	Fluorescence Spectroscopy
GIS	Geographic Information System
IDS	Intrusion Detection System
IID	Independent and Identically Distributed
InGaAs	Indium gallium arsenide
IoT	Internet of Things
KNN	K-Nearest Neighbour
LAI	Leaf Area Index
LASSO	Lasso Regularized Generalized Linear Models
LED	Light-Emitting Diode
LoRa	Long Range
LR	Logistic Regression
ML	Machine Learning
MSC	Multiplicative Scatter Correction
MSI	Multispectral Imaging
NDVI	Normalized Difference Vegetation Index
NFC	Near Field Communication
NIR	Near-Infrared
OCC	One-Class Classification

OSAVI	Optimized Soil Adjusted Vegetation Index
OSMS	Open-source miniature spectrophotometer
PA	Precision Agriculture
PC	Principal Component
PCA	Principal Component Analysis
PCB	Printed Circuit Board
PDS	Piecewise Direct Standardization
PLS	Partial Least Squares
PRI	Photochemical Reflectance Index
PSMICGS	Portable Snapshot Multispectral Imaging Crop-Growth Sensor
PV	Pseudo-validation
RaS	Raman Spectroscopy
RBF	Radial basis function
RDVI	Renormalized Difference Vegetation Index
RE	Red Edge
RF	Random Forest
RFR	Random Forest Regression
RGB	Red Green Blue
RMSE	Root Mean Square Error
SDN	Software-Defined Networking
SG	Savitzky–Golay
SIF	Solar-induced chlorophyll fluorescence
SNV	Standard Normal Variate
SPAD	Soil Plant Analysis Development
SVM	Support Vector Machine
SWIR	Short-Wave Infrared
USB	Universal Serial Bus
UV	Ultraviolet
UV-VIS-NIR	Ultraviolet-Visible-Near-Infrared
VI	Vegetation Indice
VIS	Visible
VRT	Variable Rate Technologies

Chapter 1

Introduction

1.1 Context

Proximal sensing plays a central role in Precision Agriculture (PA), enabling real-time data collection on the physiological state of plants. The advancement of Android mobile devices and low-cost sensors enables the creation of portable and affordable solutions for monitoring vegetation in the field [1].

Modern agriculture increasingly relies on accurate and timely information about crop health to support effective decision-making. However, farmers often face significant challenges in obtaining reliable in-field measurements, as they may lack both the technical means required to assess whether plants are in a healthy physiological state. This limitation hinders the early detection of stress conditions and reduces the effectiveness of crop management strategies [2].

To address this challenge, this study proposes the use of a low-cost sensor-equipped devices coupled to a smartphone, which provides the interface for data acquisition, processing, and visualization as a practical and accessible solution for in-field crop monitoring. By leveraging the widespread availability of smartphones, these systems enable non-specialist users to collect and analyze plant data directly in the field [3]. To collect data, spectroradiometric sensors will be used. To ensure measurement reliability, Machine Learning (ML) techniques will be applied to improve plant condition assessment. The proposed approach focuses on estimating key agronomic parameters, including water content, chlorophyll content, nitrogen content and fruitfulness. The diagnostic results are intended to be communicated to the farmer through a user-friendly mobile application, providing intuitive and actionable feedback to support agricultural decision-making.

1.2 Objectives

This dissertation proposes the development and evaluation of ML models for integration into an Android application capable of interfacing with spectral sensors to enable real-time assessment of plant metrics. The ultimate goal is to support agronomic decision-making through applications

such as fertilization management, irrigation scheduling, and detecting fruit production potential at an early stage.

Accordingly, this work is structured around these primary objectives:

- To identify the most suitable predictive algorithms for assessing water content, nitrogen content, chlorophyll content and bud fruitfulness;
- To determine if the same model can have consistently good results across different species;
- To evaluate the wavelength range which conducts to better predictions;
- To assess the influence of normalization and optimization algorithms in predictions;
- To validate the accuracy and robustness of a simplified multispectral sensor;
- To investigate the relationship between grapevine bud spectral reflectance and reproductive development, assessing the effects of bud position and measurement conditions on inflorescence prediction.
- To evaluate the feasibility of integrating the predictive models into a mobile application that presents results in an intuitive manner without exposing users to unnecessary data complexity.

1.3 Motivation

The need for sustainable and efficient agricultural practices drives the demand for accessible, real-time crop monitoring solutions. Although proximal sensing technologies are effective for assessing crop status, their use is often limited by cost and uses destructive techniques [4]. Smartphones provide a widely available platform with processing, communication, imaging, and georeferencing capabilities, making them suitable interfaces for portable sensing systems. Integrating smartphones with low-cost spectral and visual sensors enables practical, in-field plant health assessment and supports data-driven decision-making in PA [3]. The goal is to develop ML models that can be subsequently integrated into the CropSign mobile application with a spectral-based sensing hardware. The Figure 1.1 presents the CropSign Android application with a commercial spectral sensor soldered onto a Printed Circuit Board (PCB). Figure 1.2 depicts the CropSign system with implemented control logic capable of receiving instructions and transmitting spectral data to a smartphone through a Universal Serial Bus (USB) connection.

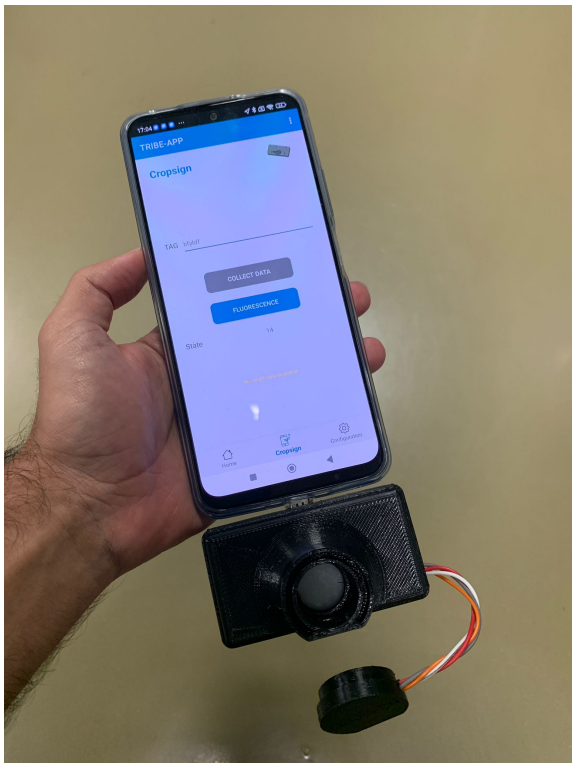


Figure 1.1: CropSign mobile application with a commercial spectral sensor soldered onto a PCB.

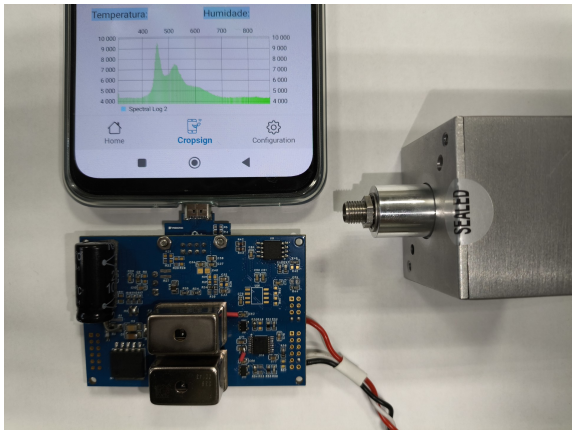


Figure 1.2: CropSign with implemented control logic to receive instructions and send spectral data via USB connection to a smartphone.

1.4 Dissertation Structure

This dissertation is structured as follows. Chapter 2 presents the Background and Fundamental Aspects, introducing the theoretical concepts and foundational knowledge necessary to understand the scientific and engineering context of this work. This includes an overview of relevant sensing technologies, data processing techniques, and ML principles applied in PA.

Chapter 3 reviews the state of the art on plant and fruit monitoring, covering sensing technologies, spectral datasets, ML approaches, smartphone-based agricultural systems, and the challenges of model transferability. The chapter concludes by framing the dissertation within the existing literature and identifying the research gap addressed by the proposed methodology.

The Chapter 4 describes the experimental procedures and methodology adopted throughout this work. It begins with a characterization of the sensing devices under evaluation, namely CropSign Lite and CropSign Pro. The chapter then presents the methodology used to select the most suitable ML algorithms for prediction tasks, including both classification and regression approaches on a published dataset. Finally, it details the data collection protocol developed for the BudScan case study, covering the experimental setup, data preprocessing and standardization procedures, model training and evaluation,

The Chapter 5 presents and discusses the experimental results obtained throughout this work. It evaluates the performance of the proposed ML models under wavelength constraints, compares classification and regression approaches, analyzes the impact of data augmentation, and assesses the performance of the developed BudScan system, including both multi-class and one-class classification methods.

Finally, Chapter 6 presents the final considerations, summarizing the main conclusions regarding the problem under study and states possible future work.

Chapter 2

Background and Fundamental Aspects

2.1 Precision agriculture

PA refers to the integrated use of sensing technologies, data analysis, geospatial information systems, and variable-rate application equipment to manage spatial and temporal variability in agricultural systems. Its core principle is to apply the right input, at the right time, in the right amount, and in the right place, thereby optimizing productivity, resource efficiency, and environmental sustainability [5].

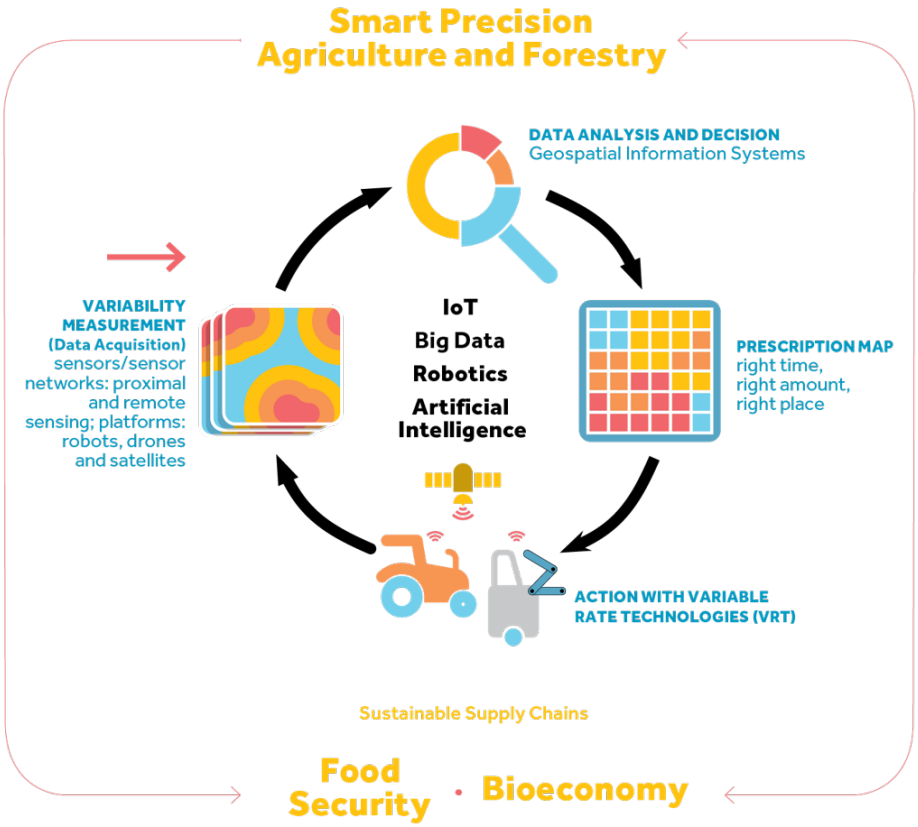


Figure 2.1: PA [6].

As it can be seen in Figure 2.1, PA follows a closed data-to-action loop, consisting of:

1. Variability Measurement (Data Acquisition) using proximal and remote sensing platforms such as robots, drones, satellites, and sensor networks;
2. Data Analysis and Decision Support through Geographic Information System (GIS) and advanced modelling tools that generate prescription maps;
3. Action and Actuation using Variable Rate Technologies (VRT) that apply inputs (e.g., water, nutrients, plant protection, biomass removal) according to site-specific needs.

When implemented at scale, PA contributes to sustainable supply chains, improved food security, and the emerging bioeconomy, while reducing environmental impact and dependence on labour [7].

2.2 Spectrometry

Recent research demonstrates extensive applications of spectrometry in PA. For crop quality assessment, spectroscopy techniques enable rapid, non-destructive analysis of soluble solids, water content, and ripeness indicators in agricultural products, facilitating real-time quality control during harvest and processing [4]. Modifications in the plants biophysical, molecular, and structural characteristics lead to observable changes in the plants' spectral behaviour (e.g., reflectance, transmittance, and emittance) [8].

2.2.1 Reflectance, transmittance and emission phenomena

Spectrometry predominantly use reflectance/transmittance techniques operating in the Visible (VIS) (400-700 nm), Near-Infrared (NIR) (700–1300 nm) and Short-Wave Infrared (SWIR) (1300–2500 nm) regions of the electromagnetic spectrum [4].

Reflectance measures the quantity of light reflected from a surface, assessing both specular and diffuse radiation. Transmittance measures the ratio of light that falls on a sample and passes through it. Emission captures light radiated by a sample due to electrons transitioning from a high energy state to a lower one. The way these phenomena work is represented in Figure 2.2.

2.2.2 Ultraviolet-Visible-Near-Infrared (UV-VIS-NIR) Spectroscopy

UV-VIS-NIR spectroscopy is a technique that encompasses a broad spectral range, covering the Ultraviolet (UV)(200–400 nm), VIS (400–700 nm), and NIR (700–1300 nm) regions. The combination of UV-VIS-NIR techniques, such as measuring both reflectance and transmittance, is beneficial as it allows for the simultaneous retrieval of both external and internal quality characteristics of the plant [4].

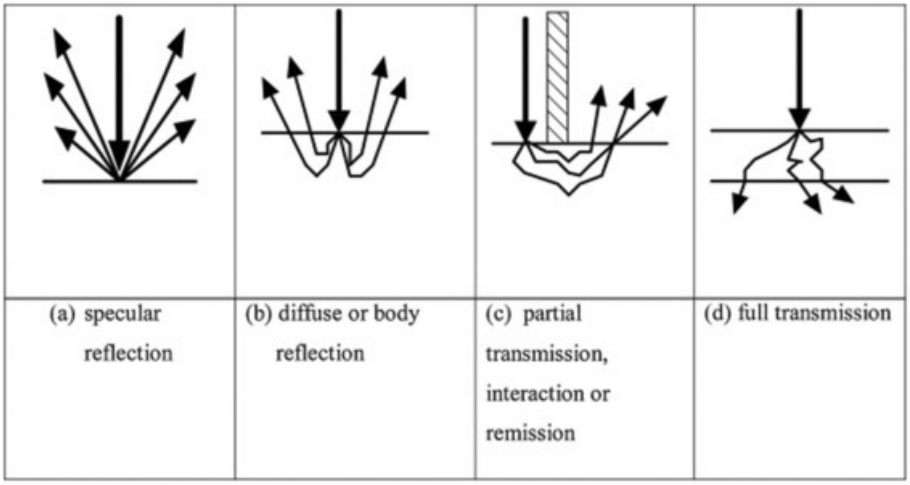


Figure 2.2: Reflectance, transmittance and emission phenomena [9].

2.2.3 Fluorescence Spectroscopy (FS)

Fluorescence Spectroscopy is a sensitive, non-invasive, and non-destructive UV-VIS-NIR technique that operates based on emission. It assesses the light emitted (fluorescence) from a sample after it absorbs excitation radiation (e.g., UV or short-wavelength visible light). Plants typically exhibit two types of fluorescence: blue-green fluorescence (400–600 nm) and chlorophyll fluorescence (650–800 nm). Disturbances to photosynthesis, often caused by disease, can be observed through the plant's fluorescence emission (specifically at 450–550 nm and 690–720 nm) [4].

2.2.4 Raman Spectroscopy (RaS)

Raman Spectroscopy is a vibrational spectroscopy technique used to analyse light scattered from a sample to identify its chemical composition and molecular structure. RaS senses the inelastic scattering of photons by molecular vibrational modes, carrying information about the chemical composition. RaS uses a laser, typically in the VIS or NIR range (frequently 532, 785, or 1064 nm) to excite a sample [4].

2.3 Data collection, visualization and local processing

Data handling and modelling are fundamental steps after data acquisition. Processing involves sequential preparation steps aimed at reducing noise, managing high dimensionality, and preparing data for predictive modelling.

2.3.1 Preprocessing and Scaling Methods

The initial step of preprocessing aims to identify and handle missing values and outliers, along with denoising and smoothing methods [10].

Zahir *et al.* [9] and Li *et al.* [11] state that preprocessing is a crucial step to clean spectra and handle scatter and instrument differences before applying predictive modelling. It removes noise, baseline drift, and scattering for the plant stress signals become clearer.

As Rashmi and Shantala [12] stated, feature scaling, also known as standardization, is a vital step in data preprocessing, ensuring data normalization within a specific range.

i. `StandardScaler()` is used to transform the data distribution by subtracting the mean and dividing by the standard deviation. This process normalizes the data around zero and scales it to have a standard deviation of 1. The formula is as it follows

$$z = \frac{x - \mu}{\sigma} \tag{2.1}$$

where, z represents the scaled value, x is the original value, μ is the mean of the feature's values, σ is the standard deviation of the feature's values.

ii. `MinMaxScaler()` is used to rescale the data to a specific range, typically between 0 and 1. It transforms the data by linearly scaling each feature to a given range using the minimum and maximum values. The formula is as it follows

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{2.2}$$

where, z represents the scaled value, x is the original value, $min(x)$ is the minimum value of the feature, $max(x)$ is the maximum value of the feature.

iii. `Normalizer()` is used to normalize samples independently. It scales each sample (each row in the dataset) to have a unit norm (a magnitude of 1) in the specified norm (either L1 or L2). L1 norm is Manhattan norm, L2 is Euclidean norm. If L2 is considered, the formula is as follows

$$z = \frac{x}{\sqrt{x_1 + x_2 + ... + x_n}} \tag{2.3}$$

where, z represents the normalized value, x is the original sample, x_1, x_2, x_n are the values within the sample.

Several preprocessing techniques have been reported as having the capability of enhancing spectral data quality and improve model performance. Zahir *et al.* [9] applied Min-max normalization to reduce baseline shifts and intensity variability, thereby improving the signal-to-noise ratio. Standard normal variate (Standard Normal Variate (SNV)) corrects for overall intensity differences and scattering effects caused by variations in light source and sample transmittance. Similarly, multiplicative scatter correction (Multiplicative Scatter Correction (MSC)) is used to reduce scatter effects and improve spectral peak shape. General normalization techniques are commonly employed to mitigate global intensity variations between samples, while detrending (Detrend (DE)) is applied to correct baseline drift. Furthermore, first-order and second-order derivative transformations are used to resolve overlapping spectral bands and remove background

or baseline effects [11]. Savitzky–Golay (SG) smoothing fits low-order polynomials over a moving window to suppress random noise while preserving relevant spectral features [10].

2.3.2 Feature engineering and feature selection

Since data usually exhibits high dimensionality and redundancy, feature engineering techniques simplify relationships, decrease computational cost, and enhance model performance.

Principal Component Analysis (Principal Component Analysis (PCA)) is frequently used as a dimensionality reduction and exploratory data analysis tool in predictive modelling. By transforming correlated variables into a reduced set of uncorrelated principal components that maximize variance, PCA reduces data dimensionality while minimizing information loss. This transformation is achieved through eigenvalue decomposition of the covariance matrix, and the resulting components are frequently used for classification or visualization in plant disease detection studies [13]. The new k-dimensional orthogonal features are also known as principal components [14].

Regarding the feature selection, Vegetation Indices (VIs) represent one of the most used approaches in spectral analysis. VIs, also known as spectral indices, are a mathematical combination of two or more spectral bands, sometimes including three, four or more bands. They are developed with biophysical significance and are used to retrieve information related to plant traits contained in collected data by reducing the dimensionality of spectral data. VIs are widely employed due to their simplicity, fast processing, and computational inexpensiveness. VIs are typically linked to essential plant biophysical parameters such as structural molecules, photosynthetic pigments and water content [4].

2.3.3 Predictive Modelling Strategies

Following preprocessing, predictive modelling is applied to the data. Several predictive modelling techniques have been widely applied to spectral data analysis in PA.

Classification models are applied when the target variable is categorical, enabling qualitative discrimination between predefined classes. Classification approaches include single-class, binary, and multi-class models, which are often implemented using one-against-one or one-against-all strategies. In contrast, regression techniques are employed when the target variable is continuous and represented by numerical values, allowing the quantitative estimation parameters [4].

According to Olowe *et al.* [15], “Logistic regression is a statistical model used to predict the probability of a binary outcome based on one or more predictor variables.”. “The simplest form, binary logistic regression, is used when the dependent variable has two possible outcomes (e.g., 1 for disease presence, 0 for disease absence). In contrast, multinomial logistic regression is used when the outcome has more than two categories, such as different disease stages (mild, moderate, severe). At the core of logistic regression is the concept of odds and the odds ratio. The odds of an event occurring is the ratio of the probability that the event happens to the probability that it does not happen.”. “The odds ratio is used to quantify how much the odds of an event change with a

one-unit increase in the predictor variable. Logistic regression makes several assumptions, including: (1) the independence of observations, (2) a linear relationship between predictor variables and the logit of the outcome, (3) no multicollinearity among predictors, and (4) a large sample size to ensure the stability of parameter estimates". Logistic regression provides effective classification while avoiding the need for strong prior assumptions about the data, reducing the risk of errors associated with inaccurate assumptions. Its hyperparameters can be learning rate (α), which represents the gradient step size, max iterations and regularization (C), which is the inverse L2 strength. In addition, logistic regression is widely applicable and computationally efficient, making it well suited for practical applications. Its low computational requirements, fast execution, and reduced energy consumption enable the analysis of complex datasets while maintaining low processing costs and system overhead [16].

Support Vector Machine (SVM) and Decision Tree (DT) algorithms have been widely adopted as effective classification techniques in a variety of agricultural applications [17].

SVM is a supervised learning algorithm that operate on labelled datasets containing two or more classes. SVM classifies new samples by maximizing the margin between parallel hyperplanes defined by the training data. Due to their strong generalization capability and effectiveness with small datasets and high dimensionality, SVM methods are frequently adopted in the literature [13].

In particular, DT algorithms have gained popularity due to their simplicity, interpretability, and computational efficiency, making them suitable for addressing a broad range of problems across engineering, agriculture, economics, medicine, and market research domains [18]. The DT regression model segregates training data by recursively optimizing feature thresholds to maximize subset homogeneity until reaching constraints like maximum depth or minimum sample thresholds. The resulting tree structure maps feature decision boundaries at internal nodes and provides continuous numerical predictions at leaf nodes via sample averaging. A key limitation of this approach stems from its reliance on greedy learning algorithms; by optimizing for local rather than global solutions, the algorithm cannot ensure a globally optimal tree configuration, which may lead to overfitting challenges [19].

The K-Nearest Neighbour (KNN) algorithm is a nonparametric, instance-based classification method that assigns class labels based on the similarity between data points. Classification is performed by evaluating the proximity of a new sample to its K nearest neighbours in the feature space, typically using Euclidean distance, followed by a majority voting process [13]. KNN is particularly well suited for datasets with well-defined clusters and demonstrates adaptability across different feature types [20].

Partial Least Squares (PLS) regression is commonly employed for analysing high-dimensional NIR spectral datasets, as it effectively addresses multicollinearity and variability by projecting predictors onto a lower-dimensional latent space. However, because PLS is inherently a linear multivariate regression technique, its predictive performance may decrease when modelling nonlinear

relationships between spectral features and target variables [21].

Random Forest (RF) and Artificial Neural Networks (ANN) are effective ML algorithms for predicting soil water content, actual evapotranspiration, and biomass in irrigated agricultural systems. Both methods have demonstrated strong predictive performance, even when applied to limited or complex datasets, noise or non-linear agricultural datasets [22].

RF is an ensemble learning method that improves predictive accuracy and robustness by aggregating the outputs of multiple decision trees. Tree-based models are generally more resilient to outliers than linear models and do not require feature normalization to achieve strong performance [22]. This is true, because they require minimal feature scaling and preprocessing. Each tree independently performs classification or regression, with the final output determined by majority voting or averaging. RF models are resistant to overfitting, capable of handling high-dimensional data, and effective in capturing both linear and nonlinear relationships. Additionally, they provide feature importance measures that support feature selection and interpretability [20].

ANNs consist of interconnected processing nodes capable of learning both linear and nonlinear relationships from data. By modelling the interactions between environmental variables, such as temperature, humidity, soil type and wind speed to agricultural parameters, including soil water content, evapotranspiration, and crop biomass, ANNs can accurately predict key indicators of crop and soil conditions, making them particularly precise [22].

One-Class Classification (OCC) has become an important approach for anomaly detection in ML. Unlike conventional classification methods, which require labelled examples from multiple classes, OCC models are trained exclusively on samples belonging to a single target class. Their objective is to identify observations that deviate from the learned representation of the target class as anomalies.

Data-Driven Soft Independent Modelling of Class Analogy (DD-SIMCA) is OCC method based on PCA. It is primarily used for tasks like authentication and anomaly detection, where the goal is to identify if an object belongs to a specific target class without needing representative data from alternative classes. The model characterizes a target class by projecting the original data into a PCA subspace consisting of scores (the coordinates of the projected data) and loadings, leaving behind residual data. DD-SIMCA evaluates how well a data point fits the target class using two key metrics: score distance (h) and orthogonal distance (q). To prevent statistical errors associated with evaluating two distinct distances simultaneously (the multiple comparison problem), DD-SIMCA combines the score and orthogonal distances into full distance (f). The orthogonal and score distances theoretically follow chi-square distributions. Because the exact degrees of freedom can be hard to define empirically, DD-SIMCA estimates them dynamically based on the distance values of the data itself. To create a DD-SIMCA model, a global PCA model is trained using a representative set of samples that belong only to the target class. Before finalizing the model, the training data is analysed. The points are categorized into regular, extremes and outliers. The extremes are legitimate target-class objects that naturally fall outside the main cluster, while outliers are erroneous objects that distort the model. The outliers must be removed using robust parame-

ter estimates. Outliers must be removed one by one, rebuilding the model each time, to prevent them from "masking" one another. To validate the model, Procrustes cross-validation employs cross-validation resampling to create a new dataset, the Pseudo-validation (PV) set, which can be used as independent test set, providing all necessary results and statistics. Then, the number of optimal number of Principal Components (PCs) needs to be selected. The optimal number of PCs is the highest number where the observed sensitivity remains within the expected 95% confidence interval. Another way is to ensure the number of extreme objects matches statistical expectations in the extreme plot. After that a prediction is made. The prediction relies on comparing the full distance to a predefined acceptance threshold. If the full distance is below the threshold, the sample is accepted as a member of the target class ("in"); if it is above the threshold, it is rejected as a stranger ("out"). To access how well rejected objects fit a non-central chi-square distribution, the model can categorize them into two metaclasses: aliens (alternative objects that share similar, predictable statistical characteristics) and externals (true anomalies that do not fit the distribution).

Among these methods, the One-Class SVM is one of the most widely used. Rather than explicitly estimating the underlying data distribution, One-Class SVM learns a decision boundary that encloses the target class, enabling the detection of anomalous samples while remaining effective in high-dimensional feature spaces

2.3.4 Model Performance Metrics

To comprehensively evaluate the performance of the classification model, a confusion matrix is often considered. Table 2.1 maps the true target values against the labels predicted by the model, establishing the foundational counts for subsequent metric calculations. The model's predictive accuracy and error distribution is analysed using the confusion matrix outlined in Table 2.1, where TP = True Positive, TN= True Negative, FP = False Positive, FN = False Negative. The four distinct outcomes detailed in this matrix serve as the basis for calculating key performance criteria, including accuracy, precision, recall, and the F1-score.

Table 2.1: Confusion matrix for binary classification evaluation.

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Classifiers:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.4}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.5}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.6}$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP+FP+FN} \tag{2.7}$$

According to Bai *et al.* [23], the sensitivity represents the proportion of correctly categorized target samples to all target samples. The specificity indicates the proportion of samples classified as non-target to all non-target samples. These are important to interpret the DD-SIMCA model.

$$Sensivity = \frac{TP}{TP+FN} * 100\% \tag{2.8}$$

$$Specifity = \frac{TN}{TN+FP} * 100\% \tag{2.9}$$

Regressors:

The coefficient of determination (R^2) measures the proportion of variance in the observed data that is explained by the model [24]. It is commonly used to assess the strength of the relationship between predicted and observed values, with values closer to 1 indicating better model performance and greater predictive reliability [10], [24].

The Root Mean Square Error (RMSE) quantifies the average magnitude of the prediction error by measuring the deviation between predicted and observed values [10]. Lower RMSE values indicate higher prediction accuracy and a closer agreement between model outputs and actual observations [24]. Consequently, a high R^2 value combined with a low RMSE generally reflects a more accurate, reliable, and stable model [25].

2.4 Built-in Smartphone Sensors and Hardware Components

Smartphones possess the capability to perform data processing and control tasks, establish local communications, and transmit information to cloud-based platforms. They are also commercially available, affordable, and widely used daily by a large portion of the population [3]. Moreover, smartphones offer inherent sensing capabilities, particularly in imaging, as well as data georeferencing through integrated positioning systems. Cameras offer high-resolution imaging capabilities suitable for visual detection, colorimetric analysis, and surface characterization. Wireless communication infrastructure, specifically Bluetooth, Wi-Fi and cellular networks allows continuous, real-time data acquisition and seamless integration into broader Internet of Things (IoT) networks. To optimize device autonomy, Near Field Communication (NFC) modules are deployed to handle both energy harvesting and data transmission over a contactless interface, effectively facilitating the design of wireless, battery-free biosensors. The User Interface provides intuitive,

user-friendly environments for test execution, real-time result visualization, and digital data sharing, thereby enhancing accessibility for non-specialists. Computational power can execute computationally demanding algorithms for sensor data processing, such impedance spectrum analysis, Fourier transforms and real-time noise filtering, while matching the accuracy of standard benchtop laboratory systems [2].

However, smartphones alone cannot perform electrochemical sensing or several other sensing techniques addressed in this work. Instead, they serve as an efficient interface to external sensing devices, enabling data visualization, storage, processing, georeferencing, and communication.

Chapter 3

Bibliography review

PA involves the collection and analysis of complex datasets. Kamilaris *et al.* [26] highlight that the adoption of sustainable agricultural technologies is increasingly driven by the collection and analysis of large-scale datasets acquired from sensors, drones, and satellite platforms, forming the foundation of what is commonly referred to as "smart agriculture". With PA, it is possible to create Decision Support System (DSS) that facilitate farmers decisions, contributing to production optimization, resource usage minimization and product quality improvement [27]. Not only it can enhance production efficiency, but it can also preserve environmental integrity, providing sustainable solutions to ecological and economic challenges associated with excessive application of agrochemicals [1]. Figure 3.1 shows the pipeline of a DSS.

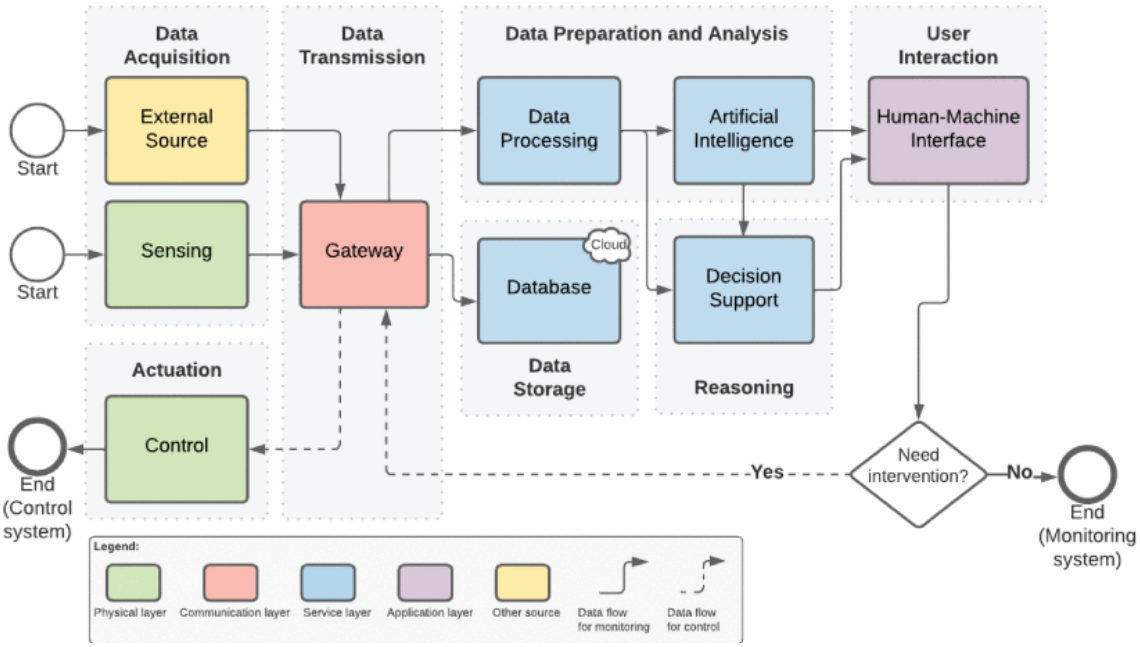


Figure 3.1: Pipeline of the decision making system [28].

3.1 Plant and fruit monitoring

Water stress is one of the primary factors limiting photosynthesis and plant productivity [29]. Although crop growth is influenced by multiple environmental and physiological factors, water availability is often the dominant constraint on plant development. Water deficits can lead to stomatal closure, reduced transpiration rates, and increased leaf temperature as a result of diminished evaporative cooling [30].

Stomatal conductance is a widely used indicator of plant water stress and can be assessed through several reflectance-based vegetation indices, including Photochemical Reflectance Index (PRI), Normalized Difference Vegetation Index (NDVI), Renormalized Difference Vegetation Index (RDVI), and Optimized Soil Adjusted Vegetation Index (OSAVI). Among these, NDVI, RDVI, and OSAVI are structural indices that evaluate water stress by analysing reflectance characteristics in the VIS and NIR spectral regions [31].

Chlorophyll is a fundamental pigment involved in photosynthesis, responsible for capturing and converting light energy into chemical energy required for plant growth. Consequently, chlorophyll concentration is closely associated with photosynthetic activity and overall plant health.

Nitrogen is an essential nutrient that directly influences chlorophyll synthesis and, therefore, photosynthetic performance. Adequate nitrogen availability promotes chlorophyll production and healthy plant development, whereas nitrogen deficiency leads to reduced chlorophyll content and decreased photosynthetic efficiency. Nonetheless, excessive nitrogen application may stimulate excessive vegetative growth, potentially resulting in adverse effects on plant health and productivity.

These physiological and biochemical indicators are interconnected. Water and nitrogen status play a crucial role in determining grapevine yield, as both water stress and nitrogen deficiency are major factors limiting productivity, particularly in Mediterranean climates. Dormant bud fruitfulness, associated with the number of inflorescence primordia or potential clusters per bud, is established during bud development. The fate of these primordia is strongly influenced by the vine's physiological status during the season preceding harvest, a process that occurs between flowering and the fruit-set stage, when reproductive differentiation is most sensitive to environmental and metabolic conditions. Consequently, severe water stress in the prior year to production can reduce the number of inflorescences formed, leading to fewer grape clusters harvested in the production year. Similarly, nitrogen deficiency shortly after flowering may decrease the number of inflorescences initiated for the following growing season. During the production year, inadequate nitrogen availability can also reduce fruit set, resulting in fewer developing berries [32].

3.2 Sensors for plant and fruit monitoring

Sensors primarily used in crop management, which monitor soil properties and plant status, generally rely on either remote or proximal sensing. Proximal sensors detect and measure physical properties of objects or environments in immediate proximity to the sensing device [7]. Its

primary advantage is the ability to gather highly precise data on plant status that lower-resolution, downward-looking remote platforms, such as satellites or drones, cannot capture [7]. Furthermore, while remote sensing observes crops strictly from above, proximal sensing allows for lateral observations at varying angles. This side-view capability is particularly crucial for perennial crops managed with foliar hedge architectures, where critical information like fruit development and disease symptoms resides primarily within the vertical trellising plane [33].

Non-destructive devices like Soil Plant Analysis Development (SPAD) meters enable fast estimation of leaf chlorophyll content, in contrast with traditional methods for chlorophyll determination, such as chemical analysis, which are frequently labor-intensive, destructive, and unsuitable for rapid field assessments [25]. Another effective tool for chlorophyll assessment is hyperspectral sensing. Hyperspectral sensors are widely employed for monitoring vegetation phenology and spectral characteristics They provide a rapid, accurate, and non-destructive alternative to conventional measurement techniques. These sensors typically acquire data across hundreds of narrow spectral bands, often exceeding 600 bands, with bandwidths ranging from less than 10 nm to sub-nanometre resolutions. Unlike point-based instruments such as the SPAD meter, hyperspectral sensors capture continuous spectral information over a broad wavelength range, enabling the integration of information from multiple bands while reducing the influence of confounding factors. Figure 3.2 illustrates the difference between hyperspectral and multispectral spectral signatures. Whereas multispectral sensors measure reflectance at a limited number of discrete wavelength bands, hyperspectral sensors provide a continuous reflectance spectrum across a broad spectral range.

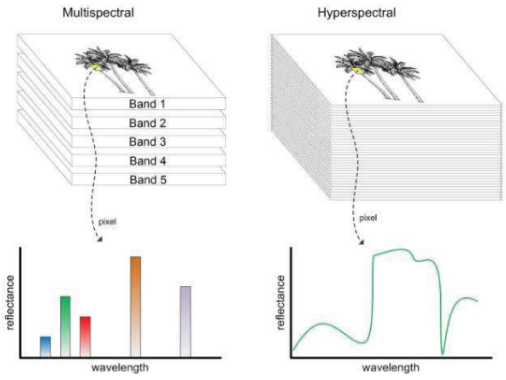


Figure 3.2: Wavelengths of Multispectral and Hyperspectral [34].

When deployed in the field, hyperspectral sensors measure radiation reflected by vegetation, which can be normalized by the incident downwelling light flux to obtain vegetation reflectance spectra [35].

The Figure 3.3 shows the classification of different wavelength intervals measured by the sensors and how the different wavelengths affects the reflectance. The reflectance can describe different phenomena. In the VIS, NIR and SWIR ranges have been observed different reflectance patterns [36].

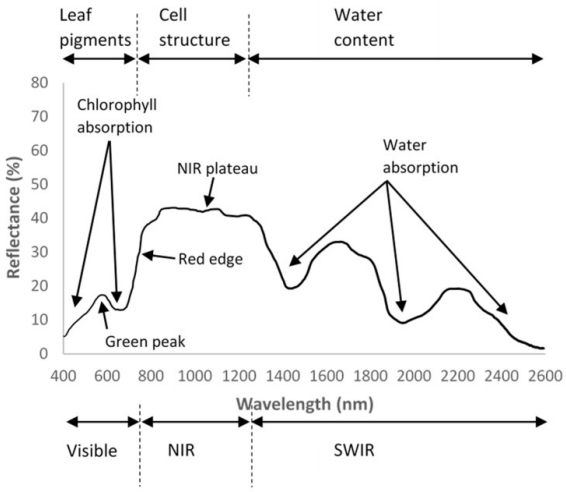


Figure 3.3: Relation between Wavelength and Reflectance [13].

3.2.1 Sensors application

The literature reports a wide range of sensing technologies applied in agricultural monitoring. Table 3.1, adapted from Laveglia *et al.* [1], presents a selection of sensors used in agriculture experiments. Additional examples of sensing technologies are discussed in the text below.

Diago *et al.* [37] details an investigation into the effectiveness of proximal NIR as a rapid and non-destructive approach for determining the water status variability within a vineyard. This research was motivated by the limitations of traditional measurements, which are typically time-consuming and destructive, hindering large-scale and real-time monitoring. Using an NIR sensor mounted on an all-terrain vehicle, researchers successfully measured the spectral properties of grapevine leaves and established strong correlations between the NIR data and physical parameters, such as stomatal conductance. The study concludes that the system is a highly effective, contact-less technology capable of capturing vine water status across high spatial and temporal resolutions. This success suggests future opportunities for integrating NIR systems directly into vineyard machinery to facilitate real-time irrigation scheduling and support advanced PA practices.

Wang *et al.* [38] documents the design and implementation of a Portable Snapshot Multispectral Imaging Crop-Growth Sensor (PSMICGS) intended for real-time, non-destructive field monitoring. This device utilizes an innovative optical system combining mosaic filters and dichroic mirrors to capture spectral information across a broad range of wavelengths necessary for assessing crop health. The PSMICGS integrates an embedded processor that enables the real-time interpretation of essential agronomic parameters directly in the field. Field experiments conducted on rice and wheat confirmed that the PSMICGS could accurately construct estimation models for Leaf Area Index (LAI), achieving validation determination coefficients (R^2) greater than 0.70.

Pereira [4] shows that using an integrated, in-situ, non-destructive proximal spectroscopy approach to assess bacterial diseases in crops. Hyperspectral models, whether based on reflectance or transmittance, are capable of distinguishing healthy from infected tissues, even before any visible symptoms appear. Fahey *et al.* [13] analyzes the role of advanced sensing techniques in PA,

emphasizing the necessity of early and accurate detection of biotic and abiotic crop stresses to improve yields and minimize excessive pesticide use. Figure 3.4 illustrates that certain sensing methods enable the detection of plant diseases at earlier stages compared to others. Early disease detection is critical, as it allows timely intervention, thereby facilitating effective treatment and limiting disease progression and spread.

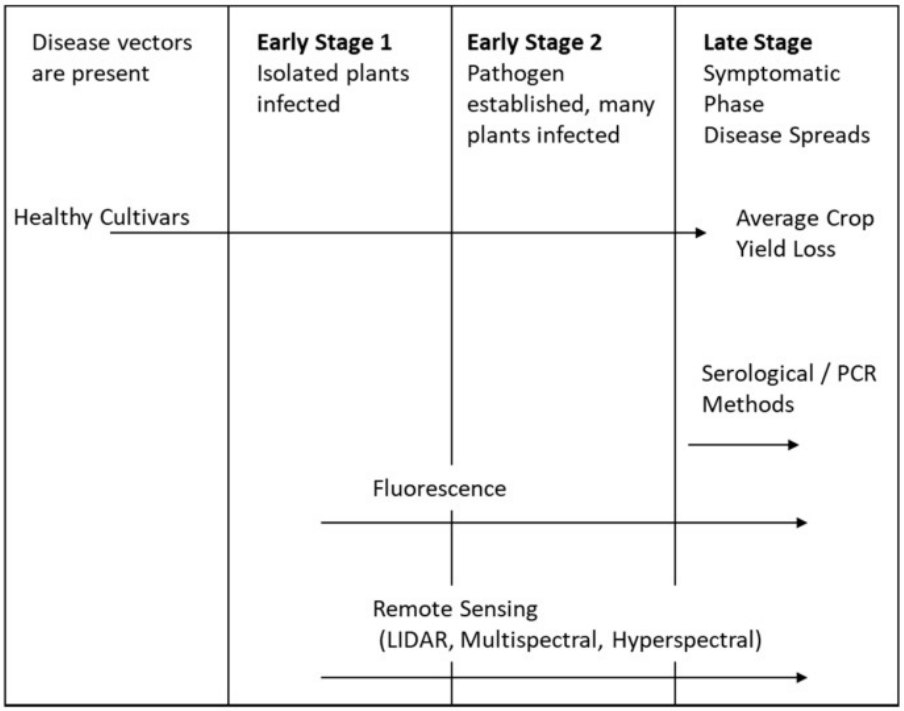


Figure 3.4: Methods to detect diseases in each phase [13].

The Multispectral Imaging (MSI) Crop Circle ACS-470 sensor operates across six spectral bands (blue: 0.43–0.47 µm; green: 0.53–0.57 µm; red: 0.63–0.67 µm and 0.66–0.68 µm; red-edge: 0.72–0.74 µm; and NIR: >0.76 µm), enabling the estimation of a broader range of vegetation indicators [1].

Table 3.1: Summary of Experimental Conditions and Sensor Specifications.

Objective	Experimental Condition	Crop type	Stress Evaluated	Sensor
Develop a multi--parametric system for variable rate nitrogen application [39].	Field	Winter wheat	Nitrogen fertilizers application	MSI N sensor: The Yara N-Sensor ALS 2: 670, 730, 740, and 770 nm
Design and evaluate an on-the-go VR fertilization system for the application of phosphate [40].	Field	Maize	Phosphorus fertilizer	Zeiss Corona 45 VIS-NIR (401-1135 nm $\pm$ 3.2 nm) and (1135-1663 nm $\pm$ 6 nm)
Develop a flexible robotic-based approach to detect *X. fastidiosa* on olive tree orchard [41].	Field	Olive tree orchard	Xylella fastidiosa infection	**CIR**: two DSLR modified cameras (EOS 600D, Canon), **MSI**: CMS-V, Silios Technologies (558–801 nm), **HyS imager**: Spectrograph Inspector V10 + camera uEye 5220CP, **Thermal camera**: A320, FLIR Systems
Develop a robotic disease detection system in [42]	Greenhouse	Bell peppers	RGB; MSI	**RGB:** (LifeCam NX-6000, Microsoft) 1600×1200; **MSI:** NIR-R-G multispectral camera (ADC Lite, Tetracam) 2048×1536, and single-laser-beam distance sensor (DT35, SICK)
Develop a robotic platform for single-plant fertilization [43].	Field	Organic vegetable	Single Plant Fertilization	Multispectral camera: Sequoia; Parrot Drones SAS, France

3.2.2 Measurements

A sensor is often modelled by a transfer function that maps a physical stimulus to an electrical output. The light source spectrum and the sensor responsivity are both wavelength dependent. To compare measurements across time, instruments, or illumination conditions, systems introduce a compensation factor or calibration factor that corrects for this combined gain so that the final value reflects the property of the sample, not the quirks of source and sensor [44]. Remote sensing and proximal sensing systems routinely calibrate their sensors against known reference targets (white panels, black bodies) to derive correction factors for spectral and radiometric response [34], [45], [46]. Integration time is the time over which the sensor accumulates charge [17]. Longer integration times increase signal level, but risk saturation or motion blur.

3.3 Problems associated with model transferability

Model transferability is a key challenge in the adoption of spectroscopy in industrial applications, particularly when applying models developed on one instrument to data acquired from another. Differences in instrument specifications, spectral resolution, wavelength range, detector type, and environmental conditions during data acquisition can significantly affect spectral measurements and model outputs. In particular, devices based on InGaAs and silicon detectors exhibit different sensitivity and noise characteristics, influencing the signal-to-noise ratio and, consequently, prediction accuracy during model transfer. As a result, models developed using high-resolution spectrometers may perform poorly on lower-resolution devices, as key spectral features may not be adequately captured [47].

Discrepancies may lead to inaccurate predictions if models are deployed without proper recalibration. In terms of signal correction, Direct Standardization (DS), Piecewise Direct Standardization (PDS), and combined DS–PDS are used to transfer calibration models between instruments, correcting systematic spectral differences from light source, detector, and optics. DS assumes a global linear relation and builds a single transfer matrix; PDS builds local transfer matrices on small wavelength intervals for more complex differences; DS–PDS applies DS globally then PDS locally [11].

3.4 Smartphone-Based Agricultural Monitoring Systems

Chung *et al.* [48] present smartphone near-infrared monitoring of plant stress, implementing NDVI measurement algorithms that quantify the contrast between chlorophyll absorption in the red spectral region and reflectance in the NIR region, which is primarily influenced by the internal cellular structure of plant tissues, as illustrated in Figure 3.5.

Araújo *et al.* [28] presents a smart agricultural system using proximal sensing, Long Range (LoRa) technology and Artificial Intelligence (AI) for vineyard management. The system revolves around a sensor node with an ESP32-S3 microcontroller connected to capacitive soil moisture

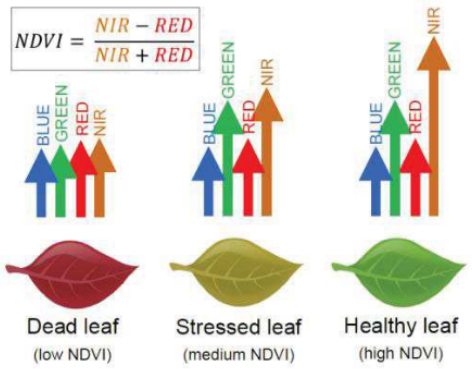


Figure 3.5: Relation between NDVI, NIR and RGB.

sensors, BME280 environmental sensors, and SI1145 sunlight sensors, demonstrating real-time environmental data collection.

Gupta *et al.* [49] demonstrate converting a personal smartphone camera to a spectrometer for proximal sensing of soil properties.

Rançon *et al.* [33] presents an in-depth analysis of a mobile proximal sensing camera acquisition system specifically engineered for applications within vineyards over an eight-year research period. This customizable system utilizes an RGB industrial camera and a synchronized xenon flash to ensure standardized, high-resolution imagery by mitigating the effects of natural sunlight. The technology has been successfully applied across four major viticulture projects, focusing on critical tasks such as in-field yield estimation and the early detection of various grapevine diseases. Researchers found that despite careful design, unavoidable environmental variability (including erratic framing and changing light conditions) poses the greatest challenge to data integrity. The authors conclude that advancing proximal sensing requires the development of more robust image analysis algorithms and the systematic creation of diverse, large-scale databases to accommodate real-world field complexity.

3.5 Spectral datasets for plant and fruit monitoring

Spectral datasets play a fundamental role in the development, calibration, and validation of sensing systems for PA. By combining spectral measurements with ground-truth information such as crop biomass, LAI, fruit ripeness, or physiological indicators, these datasets enable the evaluation of sensor performance and the development of predictive models for plant monitoring. The datasets reported in the literature vary considerably in terms of acquisition platforms, spectral resolution, measured variables, and intended applications, ranging from field-scale crop growth assessment to laboratory-based fruit quality evaluation and large-scale remote sensing studies. The following examples illustrate the diversity of available spectral datasets and their contribution to agricultural monitoring research.

The dataset on Wang *et al.* [50] consists of paired multispectral canopy images and destructive ground-truth measurements of aboveground biomass and LAI for wheat and rice, collected in field trials in Rugao, Jiangsu, China. It includes 190 wheat samples and 126 rice samples, covering multiple varieties, nitrogen levels, sowing dates or row spacings, and growth stages to ensure wide variability. For each plot, the portable multispectral sensor records reflectance in eight narrow bands plus derived vegetation indices, while plant sampling provides dried biomass and LAI, yielding a dataset suitable for developing and validating crop-growth prediction models.

The dataset in Scutelnic *et al.* [51] is composed of close-range spectral and multispectral measurements of fruit and leaf samples, primarily strawberries at different ripeness stages and kiwifruit with leaves, acquired under controlled laboratory conditions. High-resolution reflectance spectra in the UV-VIS-NIR (up to 1100 nm) were collected with two spectrometer systems, and these spectra were then used to select a reduced set of informative narrowband filters for multispectral imaging. Using a laboratory MSX camera and a compact 4-band MS5C camera, multispectral image stacks were recorded in the selected bands (e.g., around 532, 685, 740 nm), together with in-scene calibration targets, to build supervised multispectral datasets that link pixel intensities and simple spectral indices to known spectral signatures of fruit ripening and leaf reflectance. These data were mainly used to validate the proposed spectral band selection method and the hardware design. The experiments showed that the filters chosen through spectral analysis produced multispectral images and indices that clearly distinguished different ripeness stages and plant parts. These results demonstrate that the imaging system and the band-selection approach are effective for close-range PA applications.

Mowla *et al.* [52] provides theoretical background on how IoT-based wireless sensor networks generate agricultural datasets, but it does not build or release a single unified benchmark dataset itself. Instead, it surveys many existing systems that collect real-time measurements (e.g., soil moisture, temperature, humidity, nutrients, energy use) via heterogeneous sensors connected through ZigBee, Wi-Fi, SigFox, or LoRaWAN, and streams these data to cloud or edge platforms for analysis and decision support.

Mohammed *et al.* [53] is a broad review of 50 years of Solar-induced chlorophyll fluorescence (SIF) research and does not build or release a single, unified dataset of its own. Instead, it synthesizes many existing datasets across scales: laboratory leaf fluorescence measurements, canopy level tower and field spectroscopy, airborne imaging campaigns (e.g., HyPlant, CFIS, APEX), and multiple satellite SIF products from GOSAT, OCO-2, GOME-2, SCIAMACHY, TROPOMI, and future FLEX/FLORIS. These datasets are mainly described qualitatively and used as examples and references to illustrate retrieval methods, radiative transfer modelling, and applications (e.g., links between SIF and photosynthesis or stress), not as a validation benchmark for a new algorithm proposed in this paper. Validation is discussed conceptually: the authors review how other studies assess SIF retrieval accuracy using radiative transfer simulations and independent ground or airborne measurements, and they outline future needs for long-term SIF-physiology databases and multiscale Cal/Val networks for upcoming missions like FLEX, but they do not perform a dedicated validation experiment themselves.

3.6 Analysis of ML approaches for crop spectral measurements

Sharma *et al.* [54] examines the integration of AI and IoT technologies in PA, demonstrating how these technologies facilitate real-time crop monitoring and automated decision-making. The effectiveness of such approaches relies heavily on the availability of high-quality datasets capable of capturing the spectral and physiological variability of crops. In this context, ML and deep learning techniques have shown great potential for managing the high dimensionality and complexity of hyperspectral datasets, enabling the extraction of relevant features for agricultural monitoring applications [55]. In fact, spectral mapping is a strategy that can be used to evaluate the contribution of individual wavelengths and identify the spectral regions most relevant to the prediction process [4].

Chung *et al.* [48] showed that combining ML algorithms with spectroscopic methods has demonstrated significant efficacy in disease classification and stress detection. This study states that future work must prioritize elevating the spatial and spectral resolution of hyperspectral and multispectral sensors. Moreover, engineering algorithms capable of mitigating environmental and field variability remains critical for practical applications. Ultimately, the study posits that leveraging multi-sensor integration will be essential to overcoming current technological bottlenecks and improving diagnostic precision. The Figure 3.6 has some examples of ML algorithms to detect plant stress.

Navarro *et al.* [56] evaluates whether leaf level hyperspectral imaging combined with an ANN can distinguish early biotic (Fusarium wilt, Rhizoctonia rot) and abiotic (water deficit, salinity) stresses in wild rocket (Diplotaxis tenuifolia) grown in greenhouse pots. Plants underwater deficit showed the strongest impairment of water status, while Fusarium infection caused the greatest pigment loss and leaf discoloration; salinity had milder effects, consistent with the crop's salt tolerance. Hyperspectral indices related to vitality, chlorophyll and carotenoids reflected these physiological changes, and an ANN trained on VIS–NIR spectra identified stress type with good accuracy, highlighting key bands at 492–504, 540–568, 712–720, 855, 900–908 and 970 nm as most informative. The authors suggest these spectral regions and models could underpin compact sensors for early, non-destructive stress detection in precision management of leafy vegetables.

Wang *et al.* [50] investigates how ML can enhance precision nitrogen management for corn production in Northeast China. The authors specifically compare Random Forest Regression (RFR) and Lasso Regularized Generalized Linear Models (LASSO) regression against traditional simple models to predict crop nitrogen status and grain yield by integrating sensor data with environmental variables like soil properties and weather. Results demonstrate that combining active canopy sensor data with ancillary information significantly improves the accuracy of in-season diagnoses and yield forecasts. Leveraging these findings, the study introduces an innovative sidedress nitrogen recommendation strategy that simulates yield responses to various fertilizer rates. This computational approach successfully identifies optimum nitrogen application levels, offering a promising tool to increase nutrient use efficiency and reduce the environmental footprint of intensive farming.

Researchers have successfully deployed hyperspectral imaging to diagnose yellow rust and nutrient stress in wheat [57] and to discriminate between disease and insect damage in tea cultivars [45]. Utilizing a combination of spectral features and ML classifiers, these investigations yielded a precision rate of approximately 90% for stress and pest detection [45], [57].

Goyal *et al.* [55] concludes that a lightweight model is ideal for real-time deployment on mobile or IoT devices, where memory and processing power are limited.

Li *et al.* [11] addresses a critical challenge in NIR Spectroscopy: achieving accurate, universal models that work across different instruments without requiring expensive recalibration. The researchers applied three model transfer algorithms, direct standardization, piecewise direct standardization, and a combined approach of both to enable reliable cross-instrument detection of multiple highland barley quality parameters.

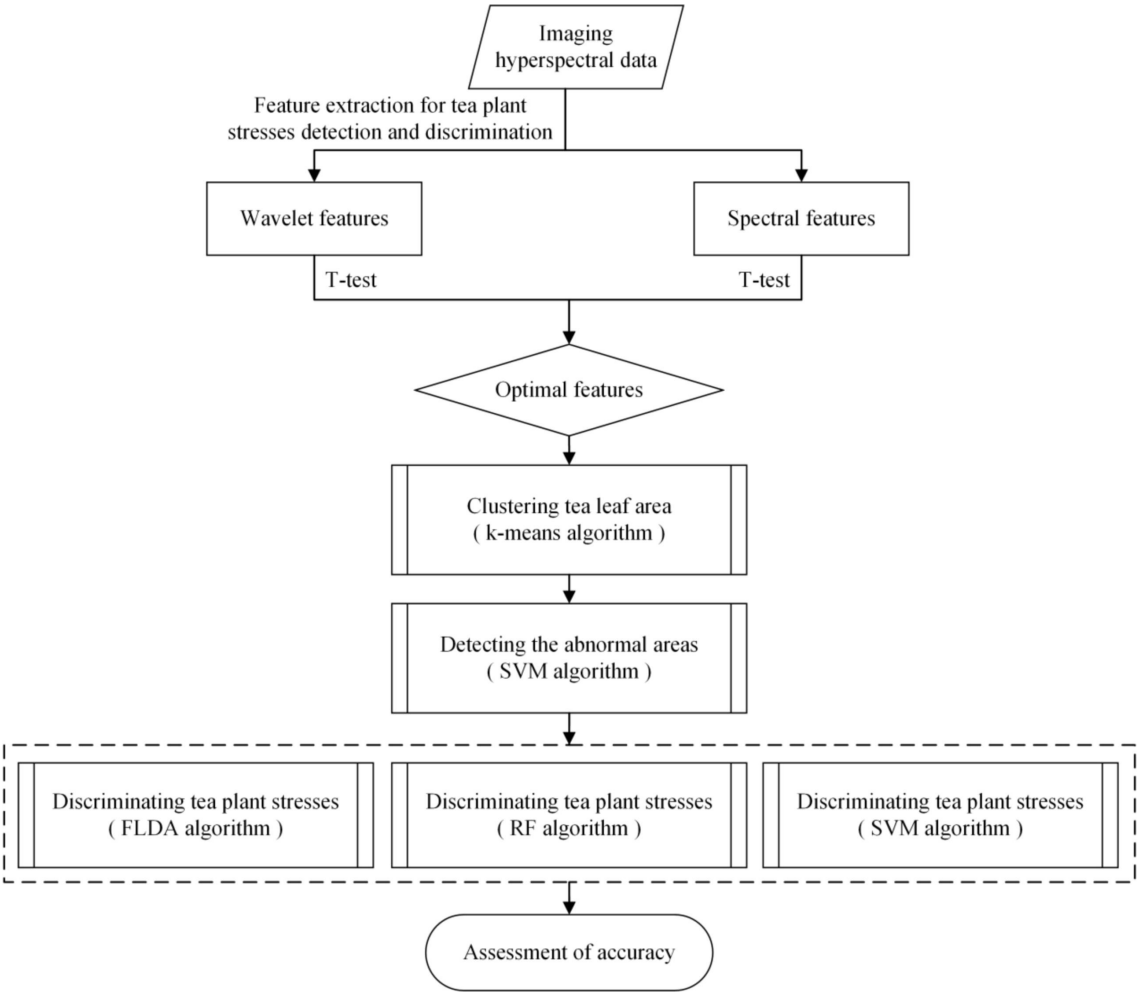


Figure 3.6: ML algorithms to detect plant stress [45].

3.6.1 Federated learning in Android applications and agricultural contexts

To address the challenge of unreliable or limited internet infrastructure in remote agricultural regions, federated learning offers a viable solution by significantly reducing communication costs. Localizing the model training process not only enhances operational efficiency, but also restricts the volume of continuous data exchange required across networks [58].

This approach allows models to be trained across multiple edge devices or servers while keeping data localized [52]. Friha *et al.* [59] proposes a Federated Learning-based Intrusion Setection System (FELIDS) designed to secure Agricultural IoT infrastructures while preserving data privacy. FELIDS supports the distributed training of deep learning models, such as convolutional neural networks (Convolutional Neural Networks (CNNs)), directly on edge nodes, which only share model updates with a central aggregator, and is evaluated using three recent network traffic datasets covering conventional, IoT protocol and Software-Defined Networking (SDN) based attacks. Results show that FELIDS often matches or surpasses centralized (non-federated) ML Intrusion Detection System (IDS) in detection accuracy, reaching up to about 99% on some SDN scenarios, while avoiding raw data centralization and keeping time and energy overheads at acceptable levels for edge deployment in Agriculture 4.0 environments.

Aggarwal *et al.* [58] details a research study on a lightweight Federated learning (FL) framework designed to classify rice leaf diseases while ensuring data privacy. By utilizing a distributed client-server architecture, the system allows local devices to train models without sharing raw images. The researchers established EfficientNetB3 as their baseline model, which achieved a high benchmark accuracy of 99% for identifying various fungal and bacterial infections. Experiments were conducted on both Independent and Identically Distributed (IID) and non-IID datasets and demonstrate that the FL approach maintains exceptional accuracy, reaching up to 99% and 95% respectively. Ultimately, this framework provides a secure and efficient solution for early disease detection in resource-constrained environments.

3.7 Framing the dissertation problem in the state-of-the-art

PA requires robust sensing technologies, advanced algorithms, and accessible interfaces for real-time decision support [1]. The literature demonstrates that proximal spectral sensing integrated with ML provides a cost-effective approach to plant health monitoring [4]. This section synthesizes key insights and establishes alignment between the proposed dissertation objectives and current research frontiers.

Existing spectral datasets for agriculture primarily focus on crop growth monitoring, fruit quality assessment, vegetation indices, or physiological parameters acquired from leaves, canopies, or mature fruits. While these datasets have demonstrated the potential of spectral sensing for PA, they rarely address the prediction of grapevine bud fertility before the growing season. Moreover, current datasets are generally collected under a single acquisition environment and are not designed to evaluate the robustness of predictive models across different conditions.

The dataset developed in this dissertation addresses these limitations by providing spectral measurements of grapevine buds from different grapevine varieties, acquired both in field conditions and after greenhouse development. Unlike previous datasets, the ground truth is based on the number of inflorescences produced by each bud, enabling the development of ML models for early fruitfulness prediction. This allows the evaluation of model performance under varying environmental conditions and contributes to precision viticulture by enabling early identification of productive and non-productive buds, supporting informed vineyard management decisions before visible reproductive development occurs.

The CropSign platform must operate reliably across distinct spectral modes. The combination of UV-VIS-NIR techniques, such as measuring both reflectance and transmittance, is beneficial as it allows for the simultaneous retrieval of both external and internal quality characteristics of the plant [4]. High-quality datasets must span diverse environmental conditions, plant varieties, and stress intensities to ensure model robustness. Chlorophyll assessment will pair spectral indices (red-edge reflectance) with handheld meter readings (SPAD units) [60]. Data validation protocols will identify outliers and sensor anomalies, ensuring dataset integrity [47].

The literature supports a hierarchical calibration approach: initial training on controlled conditions, then validation on independent field datasets under naturalistic variability. Transfer learning and domain adaptation techniques will enhance generalization across crop types and environments [11]. Although previous studies have demonstrated the effectiveness of ML for analysing spectral data, they typically evaluate a limited number of algorithms and often use full hyperspectral spectra. In contrast, this dissertation investigates the performance of several ML models using spectral information constrained to different wavelength ranges, representative of a compact multispectral sensor.

The evaluated models include Logistic Regression (LR), SVM, DT,KNN, PLS, and Ridge Regression.

Each model is evaluated across seven spectral wavelength ranges, allowing the influence of spectral coverage on prediction performance to be analyzed. This benchmarking identifies which learning algorithms remain robust when only a subset of wavelengths is available, an important consideration for low-cost multispectral devices such as CropSign Lite.

The comparison contributes to determining the trade-off between model complexity and spectral information, providing guidance for selecting algorithms that achieve reliable predictions while remaining suitable for deployment on portable agricultural sensing platforms.

This integrated approach addresses key research gaps: the absence of simplified ML models combined with accessible interfaces for spectral measurement; limited validation of the accuracy and robustness of a simplified multispectral sensor; and sparse mobile applications that presents results in an intuitive manner without exposing users to unnecessary data complexity.

Chapter 4

Methodology

One of the objectives of this study is to validate whether a multispectral sensor can produce predictions that are as reliable as those from a hyperspectral sensor. To this end, datasets collected with hyperspectral were used, and predictions called *Lite Discrete* were made using only the spectral bands of the multispectral sensor.

In order to investigate which ML models are capable of making accurate predictions based on spectral analysis, two experiments were carried out: one using a published dataset and one using a dataset collected in Nelas comprising field and greenhouse data.

4.1 Characteristics of the devices being tested

CropSign Lite device incorporates a multispectral sensor and Crop Sign Pro device integrates two hyperspectral sensors. Both devices comprise a hardware layer, consisting of a custom PCB integrated with a spectral sensor, and a firmware layer that implements the control logic required to receive instructions and transmit spectral data to a mobile application via a USB connection.

4.1.1 CropSign Lite

This device incorporates the AS7341 sensor. The spectrophotometric measurements are performed using the 10-channel light of the sensor. The AS7341 integrates silicon interference filters around of 415, 445, 480, 515, 555, 590, 630, and 680 nm wavelengths, enabling spectral detection across multiple discrete wavelength bands. In fact, it captures data across discrete bands with distinct spectral gaps. This multispectral sensor is equipped with adjustable gain and Light-Emitting Diode (LED) current settings. The gain can be varied from 1× to 512×, while the LED current can be adjusted between 1 and 100 mA. The Figures 4.1 and 4.2 illustrate the influence LED spectra and the spectral responsivity relative to the F8 channel [61], [62].

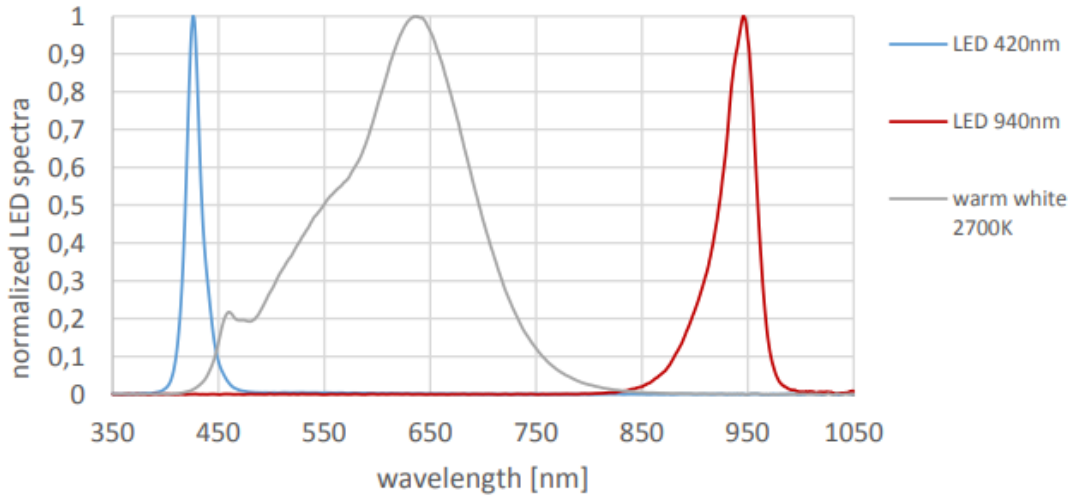


Figure 4.1: Typical LED Spectra Used in Final Test of AS7341 [62].

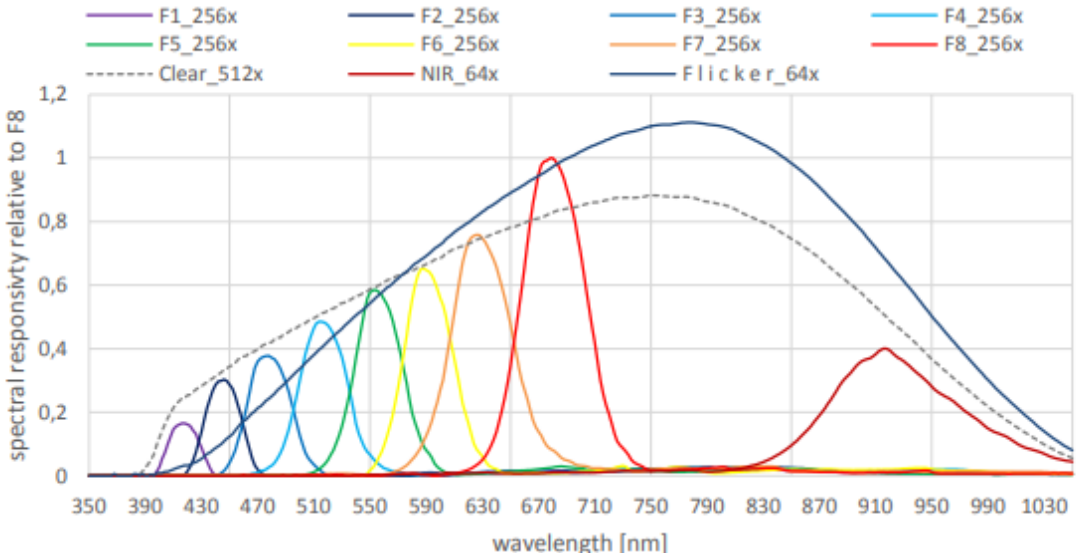


Figure 4.2: Measured Spectral Responsivity Relative to F8 [62].

4.1.2 CropSign Pro

It is a low-cost standalone portable spectrophotometer for fast and reliable measurement execution. This device is on average with moderate to high-cost spectrophotometers and outperforms low-cost ones [63]. It features more complex hardware based on two hyperspectral sensors, both with 288 channels and therefore a higher spectral resolution. The data collected in hundreds of narrow, contiguous bands and the full spectrum is registered at once. C16767MA has a wavelength range of 190 to 440 nm (UV) and 8 nm of spectral resolution. C12880MA has a spectral range between 340 and 850 nm (VIS-NIR) and 15 nm of spectral resolution [64]. The device consists of three main components: a fixed entrance slit, a reflective concave diffraction grating, and a CMOS sensor. The simplified optical design, which replaces multiple focusing mirrors with a

single concave diffraction grating, reduces stray light and enables a very small physical footprint while maintaining adequate sensitivity. However, a key limitation of such miniature spectrometer is its limited spectral resolution [65]. For the portability and the structural integrity, the board is inside a custom 3D-printed enclosure to house all the hardware components.

Figure 4.3 illustrates the spectra of an RGB LED with measured with the Open-source miniature spectrophotometer (OSMS) (C12880MA) and the commercial system Com. Each colour component turned on separately. It can be deduced that the OSMS has a similar measurement capability to the commercial system.

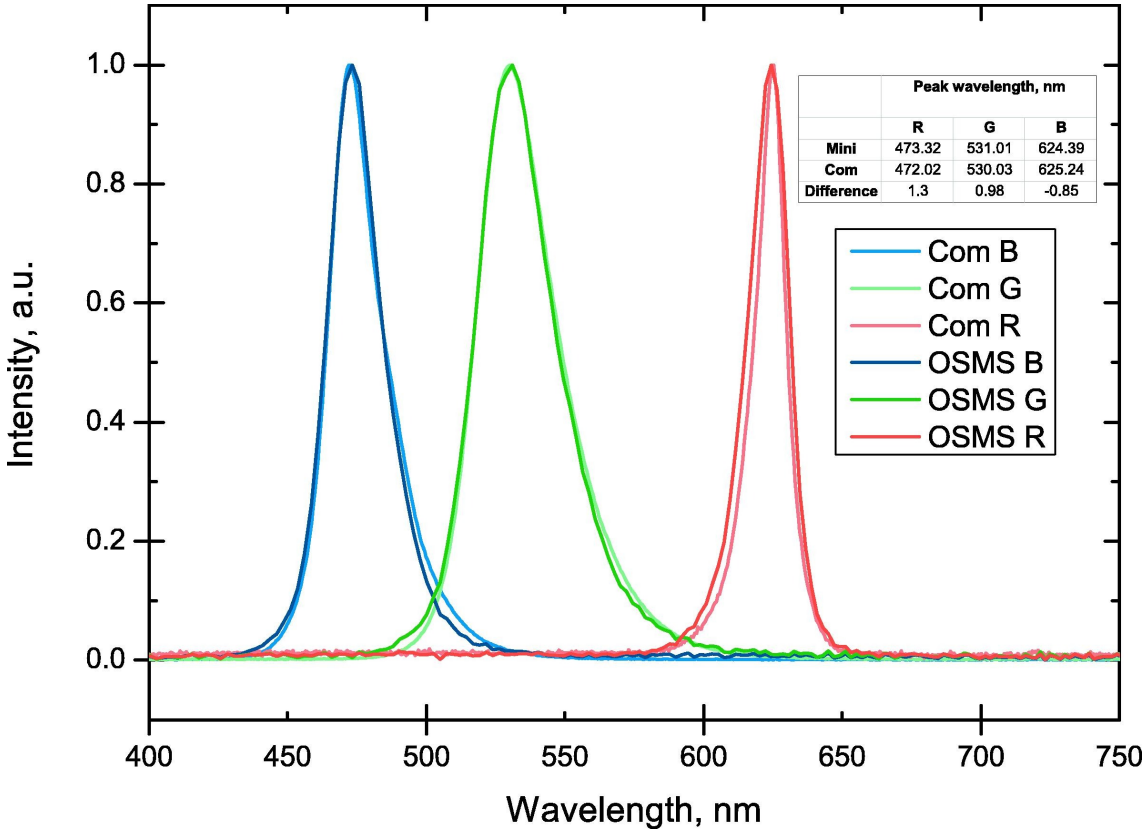


Figure 4.3: Spectra of an RGB LED measured with the OSMS (C12880MA) and the commercial system Com. [63].

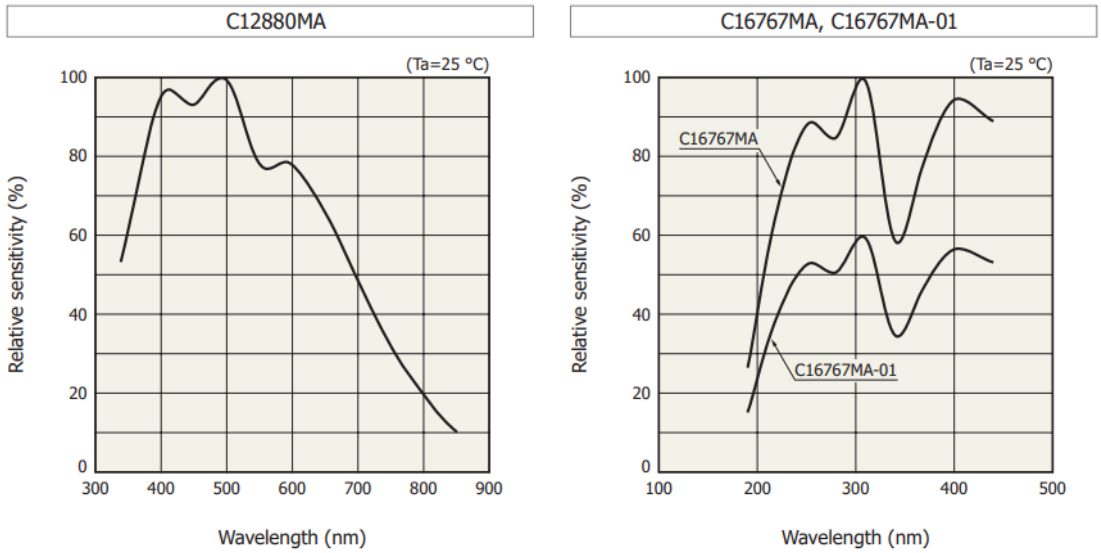


Figure 4.4: Normalized Spectral Responsivity [64].

Figure 4.4 shows the normalized spectral responsivity of the C12880MA and C16767MA. The relative sensitivity when the maximum sensitivity of each mini-spectrometer is 100 %. The relative sensitivity of the C16767MA depends on light incidence conditions.

4.2 Algorithmic Benchmarking Under Wavelength Constraints

This study aims to investigate which ML techniques can lead to a reliable prediction of key agronomic parameters, namely water, chlorophyll content and nitrogen content. For that the methodical steps were undertaken for data preparation and the development and evaluation of ML models. It will be presented a comparative analysis of different data processing models under specific wavelength ranges.

The "A multi-spectral and hyperspectral image dataset for evaluating chemical traits and the water status of avocado, olive and grape through leaf dehydration under laboratory conditions" was the dataset selected for this study [66]. It has data from 3 different species: Avocado, Olive and Vineyard. It consists of multi-spectral images (captured via a MicaSense RedEdge camera), hyperspectral reflectance data in the wavelength range from 350 to 2500 nm (captured via an ASD TerraSpec spectrometer) and physical and chemical measurements (leaf weight, chlorophyll content, and nitrogen content) across five distinct drying stages.

Regarding the measurement procedure, three individual readings were acquired from different regions of each leaf: one at the bottom, one in the middle, and one at the top. All measurements were taken across the leaf vein. The final spectral reflectance for each leaf was computed as the average of these three measurements.

To evaluate how does specific ranges impact the predictions, seven relevant ranges based on literature were selected to be analysed: VIS-NIR (350 to 850 nm), VIS (400 to 700 nm), NIR (700 to 1300 nm), SWIR (1300 to 2500 nm), RE (670 to 760 nm), Lite Continuous (350 to 1000 nm)

and Lite Discrete (415 nm, 445 nm, 480 nm, 515 nm, 555 nm, 590 nm, 630 nm, 680 nm, 910 nm). The Lite Discrete ranges represent the scope the sensor in CropSign Lite, which is going to be validated.

To assess the influence of normalization and optimization algorithms in predictions, the modelled was trained without scaling, with scaling and with the GridSearchCV model.

4.2.1 Tools

The experiments were conducted with Python 3 as the programming language. The main libraries used were NumPy for the manipulation of multidimensional arrays and matrices; Pandas for manipulation data structures and rich analysis tools; Scikit-learn for a wide range of implementation of both supervised and unsupervised ML algorithms and Seaborn and Graphviz for results' display.

4.2.2 Classification Procedure

The Classification Procedure is illustrated in the schematic flow diagram in Figure 4.5.

i. **Import data**: Data is extracted from .mat files to a data frame in order to create a dataset.

ii. **Preprocessing**: Since the metric has different levels are in different .mat files, the data frames need to be concatenated. Each measurement level is subsequently assigned to one of five predefined classes. To conclude the dataset establishment, verification whether there are any null values is performed. The augmented dataset is divided into seven subsets with different wavelength ranges, some of which partially overlap. Each dataset is then separated into features (wavelengths) and target (class).

iii. **ML**: Each dataset has its data split in 70% training data and 30% test data. The defined models are Logistic Regression (LR), Support Vector Classifier (SVM), Decision Tree (DT) Classifier and K-Nearest Neighbour (KNN) Classifier. The SVM without optimizations has a polynomial kernel. In order to see the impact of scaling data, the models are trained without scaling and with using a pipeline with the scaling and the model inside. To perform scaling, the `MinMaxScaler()` was used. Moreover, the pipeline is optimized using a GridSearchCV procedure, in which multiple combinations of hyperparameters are evaluated to identify the configuration that yields the best predictive performance. Furthermore, the Random Forest model ran using RandomizeSearchCV.

iv. **Feature Selection**: The following vegetation indices are computed: PRI550, OSAVI, NDVI, and RDVI. Since the feature selection process requires specific spectral bands that are not available in the Lite Discrete wavelength group, each VI is approximated using the closest lower and upper spectral bands relative to the required wavelength.

v. **Results**:

(a) For each of the seven datasets, the mean reflectance and corresponding standard deviation are plotted. This analysis is also performed separately for each class.

(b) After model training, the obtained results are presented using classification reports and confusion matrices.

(c) Following the identification of the best-performing Random Forest model, the relative feature importances contributing to the predictions are illustrated using bar plots.

(d) To determine whether each VI approximation should use the closest lower or upper neighbouring spectral band, both alternatives are evaluated, and their absolute differences are analysed through graphical representations.

4.2.3 Regression Procedure

The Regression Procedure can be visualized through the schematic flow diagram in Figure 4.6. The steps i, ii are the same as the 4.2.2.

iii. **ML**: Each dataset has is data split in 70% training data and 30% test data. The defined models are Ridge, Partial Least Squares (PLS) Regression, Support Vector Regressor (SVM) and KNN Regressor. The SVM without optimizations has a polynomial. In order to see the impact of scaling data, the models are trained without scaling and with using a pipeline with the scaling and the model inside. To perform scaling, the `StandardScaler()` was used. Moreover, the pipeline is optimized using a GridSearchCV procedure, in which multiple combinations of hyperparameters are evaluated to identify the configuration that yields the best predictive performance. Furthermore, the Random Forest Regression model is run using RandomizeSearchCV.

iv. **Results**:

(a) For each of the seven datasets, the mean reflectance and corresponding standard deviation are plotted. This analysis is also performed separately for each class.

(b) After model training, the obtained results are presented in a ground truth versus predictions graph. The determination coefficient (R^2) and RMSE were employed to comprehensively assess the model's accuracy.

(c) Following the identification of the best-performing Random Forest model, the relative Feature Importances contributing to the predictions are illustrated using bar plots.

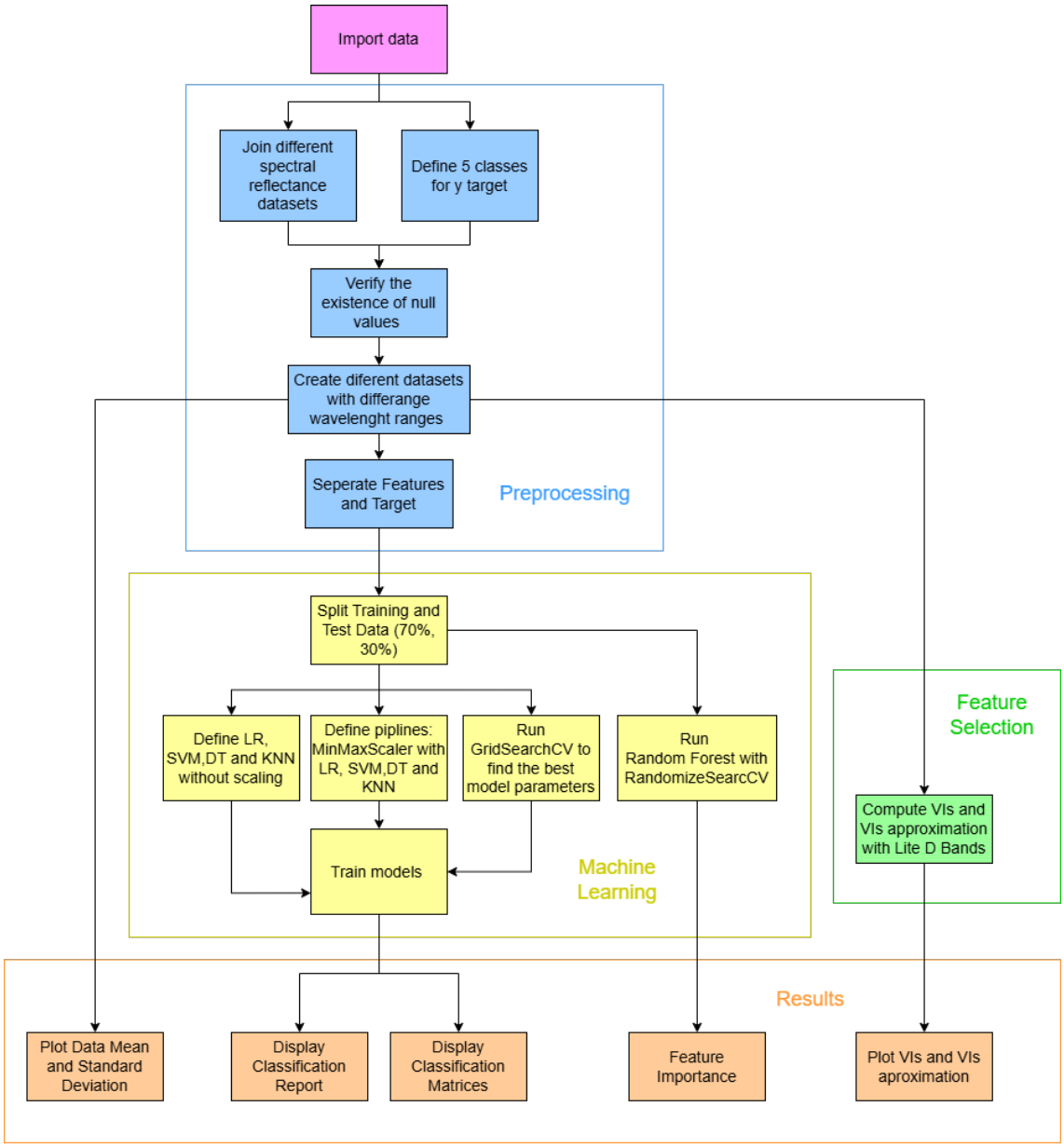


Figure 4.5: Schematic flow diagram of methodology adopted for the Classification Algorithm.

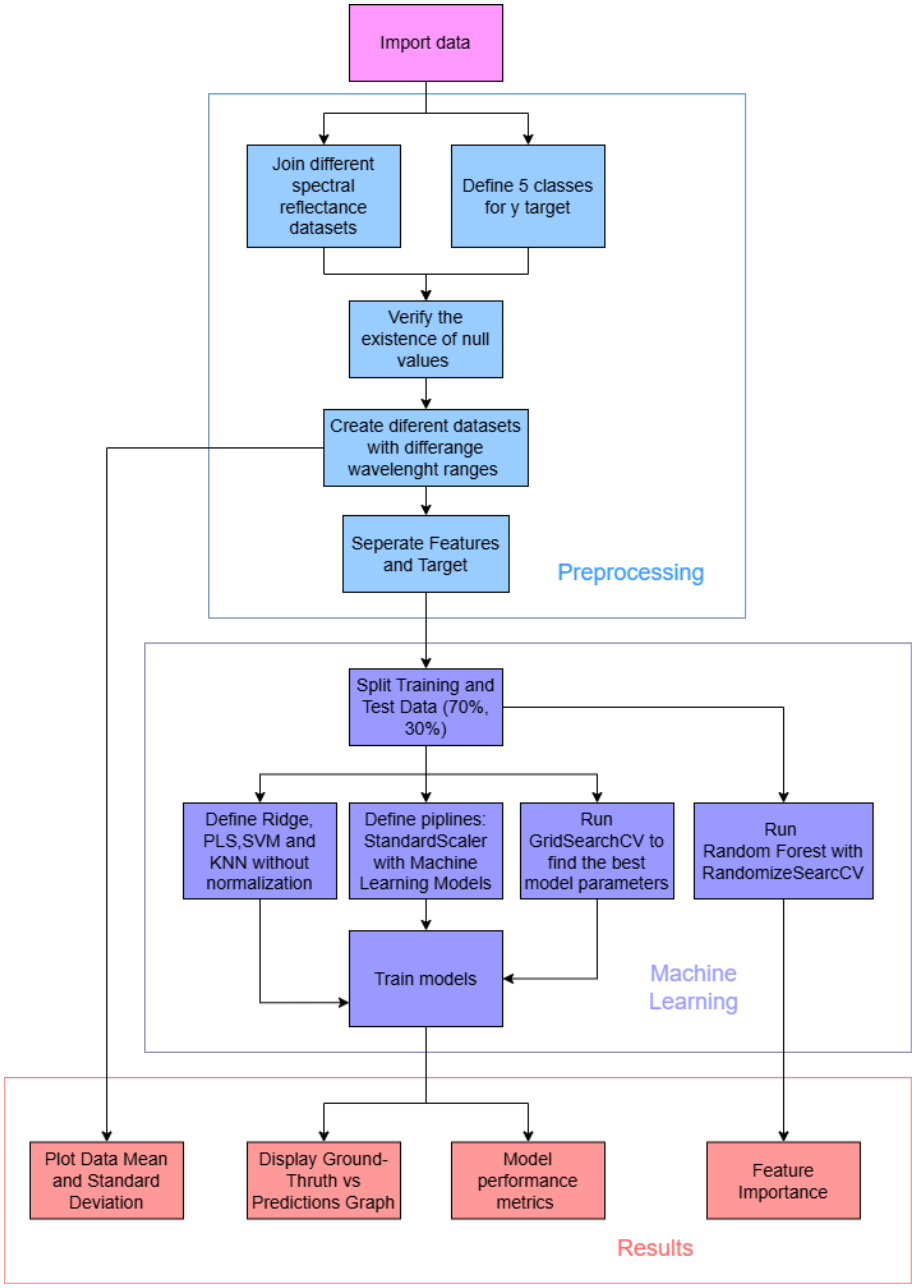


Figure 4.6: Schematic flow diagram of methodology adopted for the Regression Algorithm.

4.3 Data collection protocol: BudScan

The main goals of this protocol are:

- Explore the relationship between spectral reflectance of grapevine buds and their inflorescences.
- Evaluate the differences between predicting the inflorescences of buds placed in the greenhouse and the ones which stayed in the field.
- Access if the bud height in the cane interferes with predictions' accuracy.

4.3.1 Experimental Setup

The grapevines were planted in an open field environment at the *Centro de Estudos Vitivinícolas do Dão* (40°31'31.88"N, 7°51'22.87"W, 425 m above sea level), Nelas Innovation Hub, Viseu, Portugal. Its main areas of operation range from the recognition, characterization and conservation of Vitis genetic resources, vine selection, training systems, vineyard protection and irrigation, to advance production forecasting. Soil and climate conditions in the region heavily influence viticulture. Vineyards in this region are predominantly planted on granitic Cambisols (97%), featuring a light, porous, and permeable texture with low organic matter and poor water retention [67], [68]. The local Csb Köppen-Geiger climate is temperate and subhumid, marked by cold, rainy winters and hot, dry summers [69]. Crucially, distinct microclimatic variations across the region strongly influence final wine quality. The experiment was conducted in February - May 2026. Eleven grapevine varieties were used in the experiment: *Alvar* (A1, A2), *Arinto* (AR1, AR2), *Diagalves* (DI1, DI2), *Jaen* (J1, J5), *Sauvignon blanc* (SB1, SB2), *Semillon* (SE1, SE2), *Siria* (SI1, SI2), *Tinta Franciscana* (TF1, TF2), *Tinto Cão* (TC1, TC2), *Touriga Nacional* (T3, T4), *Trincadeira* (TR1, TR2).



Figure 4.7: Samples' location points in Centro de Estudos Vitivinícolas do Dão.

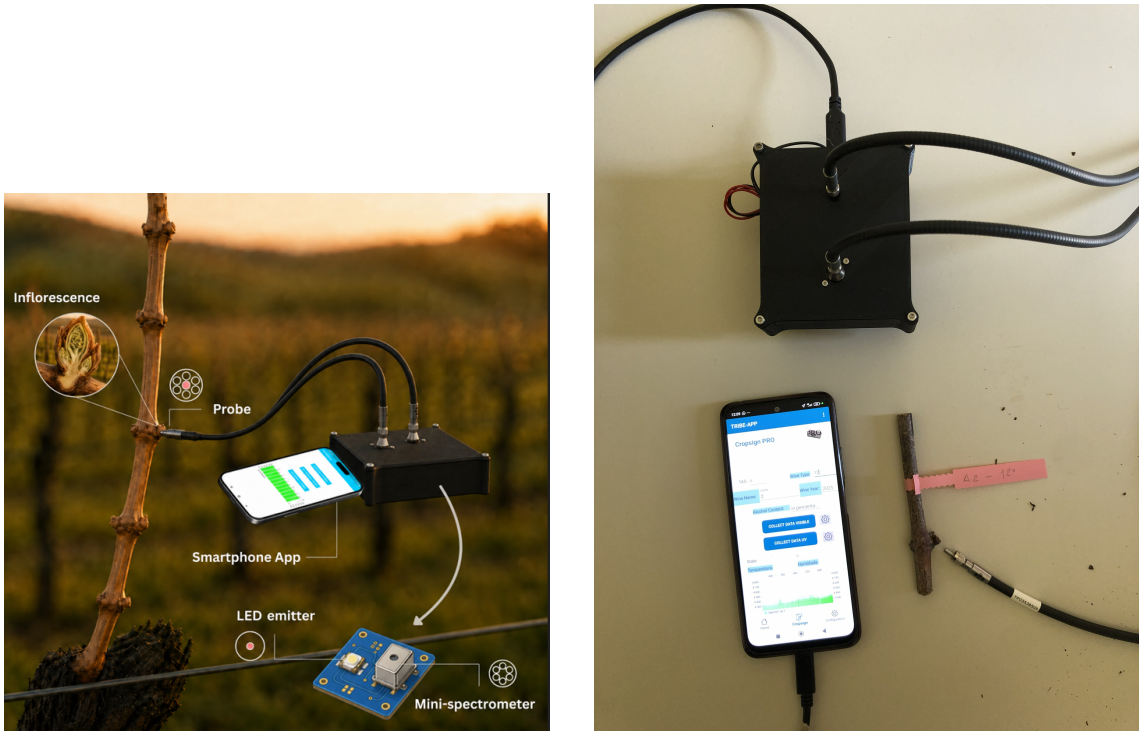
4.3.1.1 Data acquisition

The experimental measurements were conducted both in the field (as shown in Figure 4.7) and inside the laboratory. Measurements for grapevine fruitfulness assessment are acquired at the grapevine bud, using a probe connected to the board and the CropSign Pro (described in 4.1.2) via a Thorlabs RP21 bifurcated optical fibre. This is the description of the BudScan system, which is visually represented in Figure 4.8.

Before measuring the samples, the system was calibrated by placing the instrument probe over a standard white sheet to acquire the white-reference spectrum and inside a dark box to acquire the dark-reference spectrum.

Measurements taken around the bud at a 90º angle to the surface of the side chosen. Probe kept in contact with the bud to reduce external-light interference. This can be observed in Figure 4.9.

The data is transmitted from the board to the smartphone through a USB cable. Thus, the measurement is registered in the smartphone Crop Sign app, by labelling each sample with its tag, cane and bud, as shown in Figure 4.10. The samples are extracted from 4 different points in the grapevine cane. To determine whether bud scales affect the measurements or if the issue is rooted in the measuring devices, 12 buds were measured both with and without their scales. Three measurements are taken for each bud, at the top, on the right side and on the left side. A 4 ms integration time was used. After the samples are obtained, the buds measured inside the laboratory were placed in a greenhouse, Figure 4.11, for 3 months. In May 2026, the number of inflorescences was counted, to obtain a ground truth for the dataset regarding the fertility. In the laboratory, a budbreak forcing method was used, where cuttings are placed in a controlled temperature and humidity environment to accelerate the sprouting process and allow for earlier counting of inflorescences (4.11b). Since the ultimate goal is to have a portable device for taking field measurements, samples were collected in the field and the inflorescences were counted as they emerged.



(a) Conceptualized system scheme.

(b) Experiment system.

Figure 4.8: BudScan system.



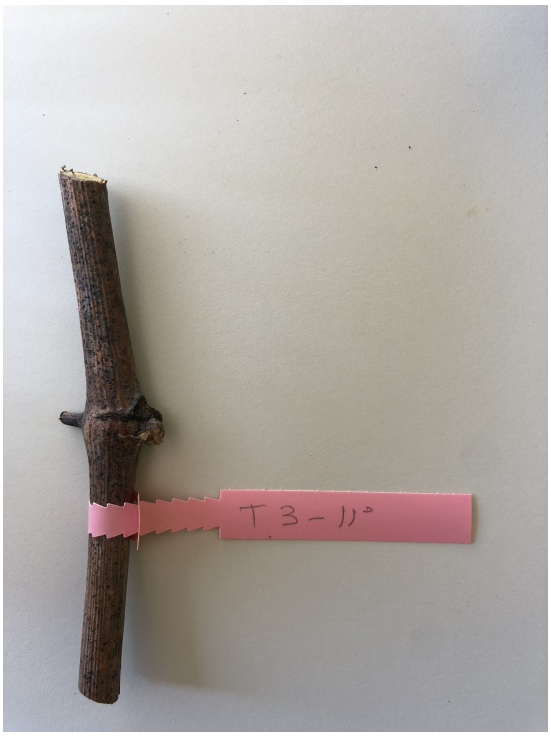
(a) Measurement in the sun.

(b) Measurement in the shadow.

Figure 4.9: In field BudScan measurements.



(a) Sample in the lower part of the cane.



(b) Sample in the higher part of the cane.

Figure 4.10: Sample labeling.



(a) Greenhouse.



(b) Canes to analyse after sprouting.

Figure 4.11: Labeled Canes in the Greenhouse.

Dataset info

- Wavelength range: 350-2500 nm
- Number of wavelength points: 2150
- Integration time: 4ms
- Total buds: 507
- Greenhouse buds: 261
- Field buds: 246
- Buds distributed by Classes:
  - Class 0: 45
  - Class 1: 168
  - Class 2: 279
  - Class 3: 15

4.4 Data analysis: BudScan

4.4.1 Initial Processing

The dataset was organized in an Excel workbook, each corresponding to a distinct orientation around the bud perimeter. To facilitate subsequent processing and analysis, each sheet was read into a pandas DataFrame and subsequently exported as an individual CSV file.

A crucial step involved the clear demarcation between the predictive features and the target variable. Columns named 'tag', 'Bud ID', 'Cane ID' were designated as identifiers and 'class' as the target. Thus, they are excluded from the feature set. Subsequently, the feature variable is created with the remaining columns.

To address potential issues arising from varying scales and units among the features, a normalization procedure was applied. The `MinMaxScaler` from the `sklearn.preprocessing` library was utilized. This preprocessing step is critical for many ML algorithms, particularly those sensitive to feature scales (e.g., KNN, SVM), ensuring that features with larger magnitudes do not dominate the learning process.

4.4.2 Data Splitting and Model Training

Data partitioning was executed using a 70/30 split, reserving roughly 70% of the data for the training phase and 30% for subsequent model evaluation. A fixed `random_state` was used to ensure the reproducibility of the split.

Two approaches will be investigated: the use of supervised classification algorithms and the use of unsupervised one-class classification algorithms.

A suite of supervised classification algorithms was selected for model development, encompassing diverse learning paradigms: LR, SVM, DT, and KNN.

Model performance was assessed using classification reports, which provided class-specific precision, recall, and F1-score values, in addition to aggregate measures including overall accuracy, macro-average, and weighted-average statistics. These metrics collectively offered a holistic view of each model's predictive capabilities across the different classes.

To further understand the predictive behaviour of the models, a confusion matrix was constructed and visualized as a heatmap. The confusion matrix provides a detailed breakdown of correct and incorrect classifications, distinguishing between true positives, true negatives, false positives, and false negatives. This visualization aided in identifying specific areas where the model performed well and where misclassifications were prevalent.

One-class classification algorithms addresses an authentication problem by learning the characteristics of the target class and identifying samples that deviate from this pattern as anomalies. The dataset was trained using both One-Class SVM model and DD-SIMCA model.

According to the Kucheryavskiy *et al.* [70], the flowchart of the DD-SIMCA pipeline can be visualized in Figure 4.12.

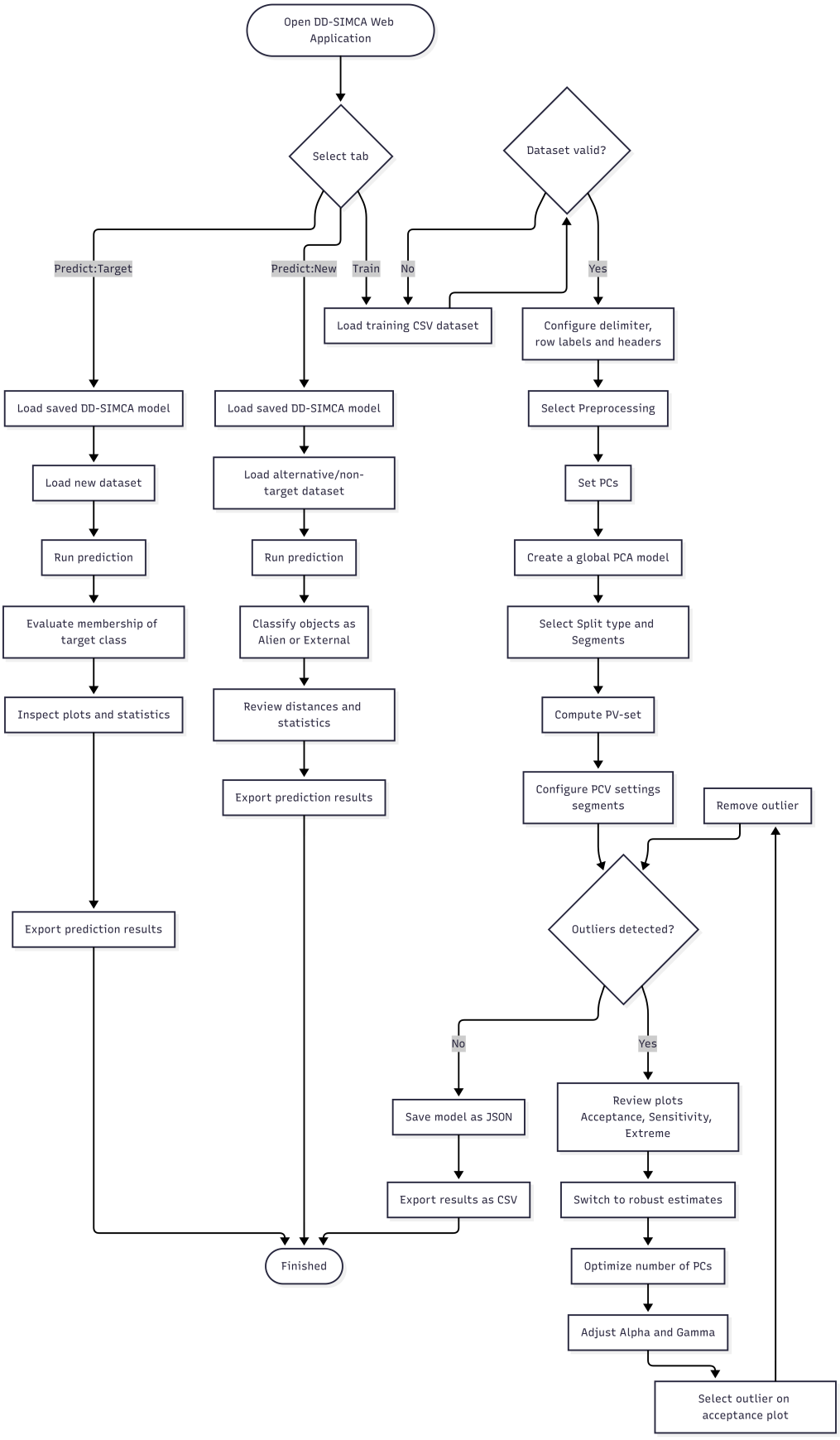


Figure 4.12: DD-SIMCA pipeline diagram.

Chapter 5

Results and Discussion

5.1 Algorithmic Benchmarking Under Wavelength Constraints

Water, chlorophyll and nitrogen content were predicted for 3 different datasets: Avocado, Olive and Vineyard. For that, classifier and regressor algorithms were used. Feature importance analysis was performed using the RF model. The objective was to identify the most informative wavelengths and compare their relative contributions to model predictions. The optimal hyperparameters obtained through GridSearchCV for the best-performing models are presented in the a table format. Both the importances graphs and the hyperparameters GridsearchCV tables can be visualized in Appendix A.

In order to access if it is possible to use data obtained from a simplified multispectral sensor instead of a hyperspectral sensor to predict plant metrics, the predictions obtained using the ranges Lite Discrete and Lite Continuous are compared in this section.
Note: In the presented tables, "Lite C" refers to the Lite Continuous configuration, while "Lite D" corresponds to the Lite Discrete configuration.

5.1.1 Classification Algorithms

In order to facilitate the results' interpretation, the complete wavelength spectra was plotted showing the mean of reflectance for each class. The five classes represent increasing levels of the target variable, namely *Very Low*, *Low*, *Medium*, *High*, and *Very High*.

To assess the impact of the spectral input representation on predictive performance, confusion matrices were generated for the best-performing models obtained using both the Lite Continuous and Lite Discrete wavelength configurations for each target metric. This comparison allows the evaluation of the differences between continuous spectral measurements and discretely sampled wavelengths. To incresase document readability, the confusion matrix corresponding to the overall best-performing model can be found in the Appendix A.

The most informative features are presented in descending order of importance, with the most significant feature listed first.

The results obtained for each spectral range are presented in the performance metrics tables. The first two columns report the performance of the baseline models without optimization. The third and fourth columns present the results obtained after applying normalization, while the fifth and sixth columns correspond to the normalized models further optimized through GridSearchCV. For each configuration, the first column reports the accuracy and the second column reports the macro-averaged F1-score.

Within each configuration, the best-performing model is highlighted in bold. Additionally, the overall best-performing method for each spectral range is also indicated in bold.

When multiple models achieved identical accuracy and overall F1-score values, the tie was resolved by comparing the class-specific F1-scores. Preference was given to the model that demonstrated superior performance across the individual classes.

To illustrate better the optimizations made with GridSearchCV, there are two tables which show the best obtained hyperparameters for each model. The first shows the LR, DT and KNN models and the second one the SVM model. The parameters of the best-performing model for this configuration are in bold. These tables can be found in the Appendix A.

To compute VIs, the feature extraction technique was applied as it can be a promising feature engineer method. The spectral indices were computed for each class and presented in a bar graph. In order to validate the simplified multispectral sensor, for each VI, the closest available wavelength was used to approximate the required spectral band. When a VI label has a number in front of it, it means that it is representing a VI that can be obtained using Lite discrete bands. A suffix "0" indicates that a single wavelength was clearly the closest match. When two neighbouring wavelengths were similarly close to the target wavelength, the corresponding approximations are denoted by the suffixes "1" and "2".

5.1.1.1 Avocado

Water Content in Avocado

The reflectance spectra of each class can be visualized in Figure 5.1. The difference between the very low and very high water content classes is highly noticeable, as their sample means are widely separated. Additionally, the standard deviations of the 5 classes almost never overlap.

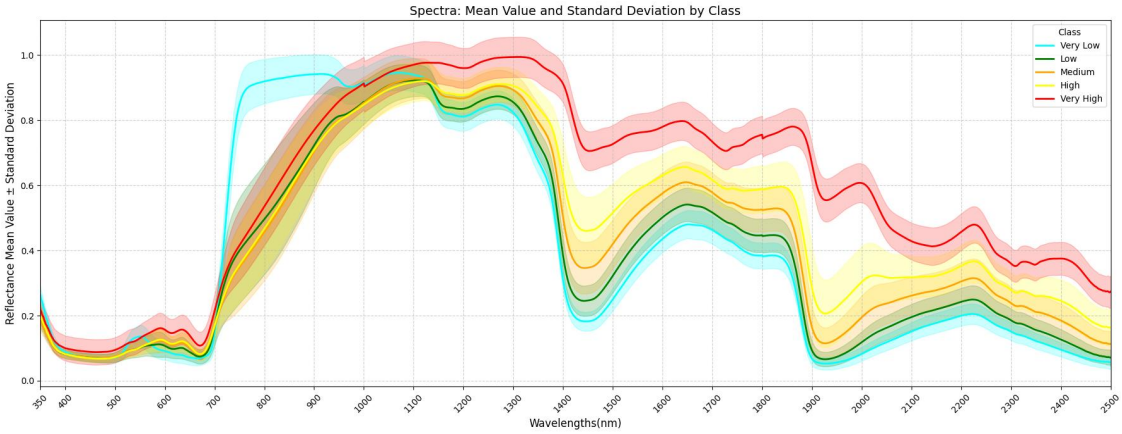


Figure 5.1: Wavelengths vs Reflectance by Class: Water Content Avocado.

The confusion matrices, in Figure 5.2, indicate that the GridSearchCV optimized SVM achieved the best classification performance for both Lite configurations. In the Lite Continuous range, almost all samples were correctly classified, whereas in the Lite Discrete range the majority of samples were assigned to their correct classes, although a higher number of misclassifications was observed.



(a) Optimized LR Confusion Matrix for Lite Continuous.

(b) Optimized SVM Confusion Matrix for Lite Discrete.

Figure 5.2: Confusion Matrices: Water Content Avocado.

The 774, 761, 764 and 755 nm are the wavelengths which give more information about the dataset, respectively, when the RF was run in the Lite Continuous range. The 630, 910, 590, 555 and 680 nm are the wavelength which give more information about the dataset, respectively, when the RF was run in the Lite Discrete.

In Table 5.1, the best result was in the Lite Continuous range, using the Optimized LR algorithm with 92% accuracy. This result was obtained with a C parameter of 1000, as shown in Table

A.1. For different ranges the strongest models were LR and SVM. If a multispectral sensor such as Lite Discrete were used, the model would achieve an accuracy of approximately 70%, suggesting that this configuration still provides satisfactory predictive performance.

Table 5.1: Performance Metrics for Wavelength Water Content Avocado.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
VIS NIR	**LR**	**0.63**	**0.64**	**0.68**	**0.69**	**0.87**	**0.88**
	SVM	0.47	0.46	0.53	0.53	0.81	0.83
	DT	0.58	0.60	0.58	0.60	0.55	0.58
	KNN	0.60	0.62	0.58	0.60	0.60	0.62
VIS	**LR**	0.56	0.56	**0.61**	**0.62**	**0.88**	**0.88**
	SVM	**0.72**	**0.74**	0.59	0.58	0.81	0.82
	DT	0.52	0.54	0.52	0.54	0.51	0.52
	KNN	0.58	0.60	0.59	0.61	0.58	0.60
NIR	LR	0.73	0.75	0.78	0.79	0.85	0.87
	SVM	**0.82**	**0.83**	**0.78**	**0.79**	**0.86**	**0.87**
	DT	0.61	0.64	0.61	0.64	0.61	0.64
	KNN	0.64	0.66	0.65	0.68	0.65	0.68
SWIR	LR	**0.78**	**0.78**	**0.79**	**0.80**	0.88	0.89
	SVM	0.77	0.77	0.68	0.68	**0.89**	**0.89**
	DT	0.75	0.75	0.76	0.76	0.78	0.77
	KNN	0.75	0.76	0.73	0.74	0.78	0.78
RE	**LR**	0.50	0.48	0.53	0.54	**0.69**	**0.70**
	SVM	0.47	0.50	0.47	0.48	0.68	0.69
	DT	0.49	0.53	0.50	0.54	0.54	0.57
	KNN	**0.56**	**0.59**	**0.58**	**0.59**	0.57	0.59
Lite C	**LR**	**0.70**	**0.71**	**0.74**	**0.75**	**0.92**	**0.92**
	SVM	0.52	0.52	0.60	0.59	0.91	0.91
	DT	0.54	0.57	0.54	0.57	0.60	0.62
	KNN	0.58	0.59	0.60	0.62	0.58	0.60
Lite D	LR	0.42	0.38	0.53	0.53	0.63	0.63
	SVM	0.46	0.45	0.51	0.51	**0.70**	**0.71**
	DT	0.55	0.57	0.56	0.58	0.56	0.58
	KNN	**0.56**	**0.59**	**0.58**	**0.60**	0.58	0.61

Chlorophyll Content in Avocado

The reflectance spectra of each class can be visualized in Figure 5.3. The Medium and High class mean are very similar. There is an overlap of the mean standard deviation of each class with the adjacent class.

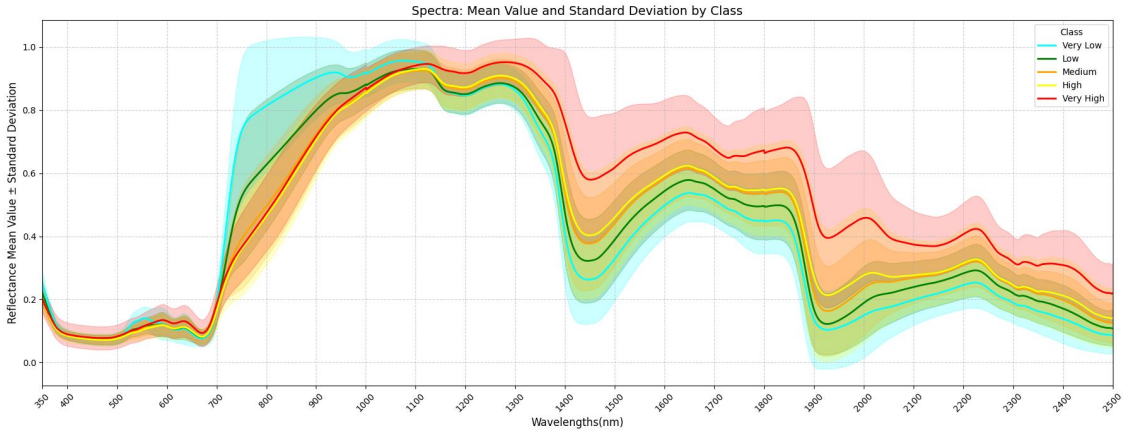
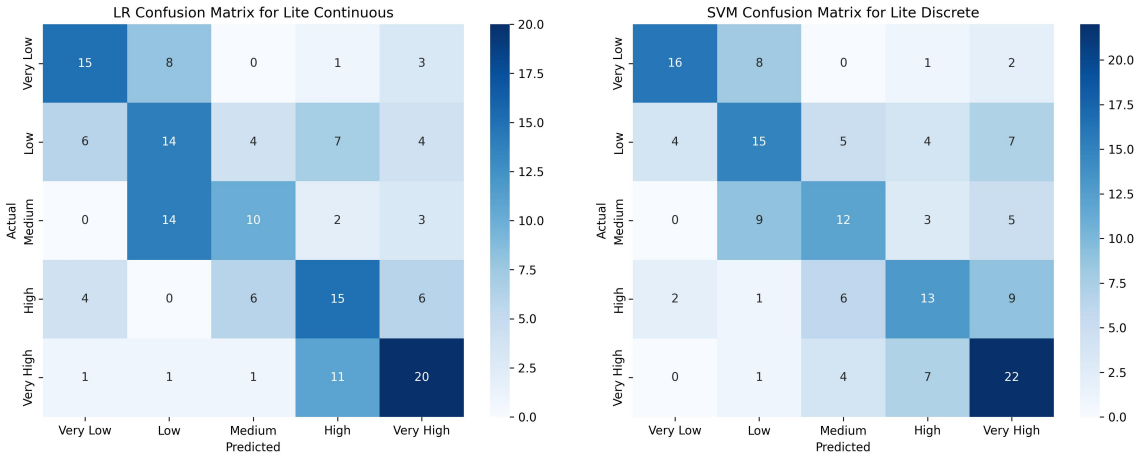


Figure 5.3: Wavelengths vs Reflectance by Class: Chlorophyll Content Avocado.

Although it is possible to identify a clear diagonal on both matrices in Figure 5.4, there are also several misclassifications.



(a) Optimized LR Confusion Matrix for Lite Continuous.

(b) Optimized SVM Confusion Matrix for Lite Discrete.

Figure 5.4: Confusion Matrices: Chlorophyll Content Avocado.

The RF importances most relevance features for NIR are the 731, 728, 729, 726 nm, respectively. They are closer to the most relevance features for Lite Continuous which are the 728, 732, 726, 725 nm, respectively. For Lite Discrete the most significant bands are 910,555,630 and 680.

The best result was in the NIR range, using the Optimized LR algorithm with 62% accuracy. For different ranges the strongest models were LR and SVM. If the Lite Discrete multispectral

sensor were considered instead of the Lite Continuous configuration, the classification accuracy would decrease to 41%. This can be seen in Table 5.2.

Table 5.2: Performance Metrics for Wavelength Chlorophyll Content Avocado.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
VIS NIR	**LR**	**0.41**	**0.40**	**0.44**	**0.44**	**0.52**	**0.52**
	SVM	0.33	0.30	0.36	0.34	0.47	0.47
	DT	0.31	0.32	0.33	0.33	0.38	0.39
	KNN	0.35	0.35	0.33	0.34	0.41	0.43
VIS	**LR**	0.36	0.31	**0.41**	**0.41**	**0.49**	**0.49**
	SVM	**0.42**	**0.43**	0.33	0.30	0.43	0.44
	DT	0.28	0.29	0.28	0.29	0.28	0.29
	KNN	0.33	0.35	0.33	0.35	0.31	0.33
NIR	LR	0.47	0.46	0.47	0.48	0.58	0.57
	SVM	**0.47**	**0.48**	**0.50**	**0.51**	**0.62**	**0.62**
	DT	0.34	0.35	0.35	0.35	0.40	0.41
	KNN	0.44	0.45	0.40	0.41	0.47	0.47
SWIR	LR	**0.38**	**0.33**	**0.38**	**0.36**	0.44	0.42
	SVM	0.35	0.30	0.35	0.28	**0.44**	**0.44**
	DT	0.22	0.23	0.22	0.23	0.31	0.31
	KNN	0.37	0.37	0.37	0.36	0.36	0.35
RE	LR	**0.38**	**0.38**	**0.40**	**0.40**	0.45	0.45
	SVM	0.35	0.32	0.34	0.32	**0.47**	**0.48**
	DT	0.28	0.29	0.28	0.29	0.29	0.31
	KNN	0.34	0.35	0.33	0.33	0.32	0.33
Lite C	**LR**	**0.41**	**0.40**	**0.46**	**0.46**	**0.54**	**0.54**
	SVM	0.29	0.23	0.37	0.35	0.51	0.52
	DT	0.35	0.35	0.35	0.35	0.32	0.33
	KNN	0.32	0.33	0.31	0.32	0.40	0.42
Lite D	LR	0.29	0.23	0.33	0.30	0.37	0.39
	SVM	0.28	0.22	0.35	0.34	**0.41**	**0.42**
	DT	**0.35**	**0.35**	**0.35**	**0.35**	0.31	0.32
	KNN	0.31	0.32	0.34	0.36	0.32	0.35

Nitrogen Content in Avocado

The reflectance spectra of each class as illustrated in Figure 5.5. Similar to Chlorophyll Content in Avocado, the Medium and High class mean are similar. There is also an overlap of the mean standard deviation of each class with the adjacent class.

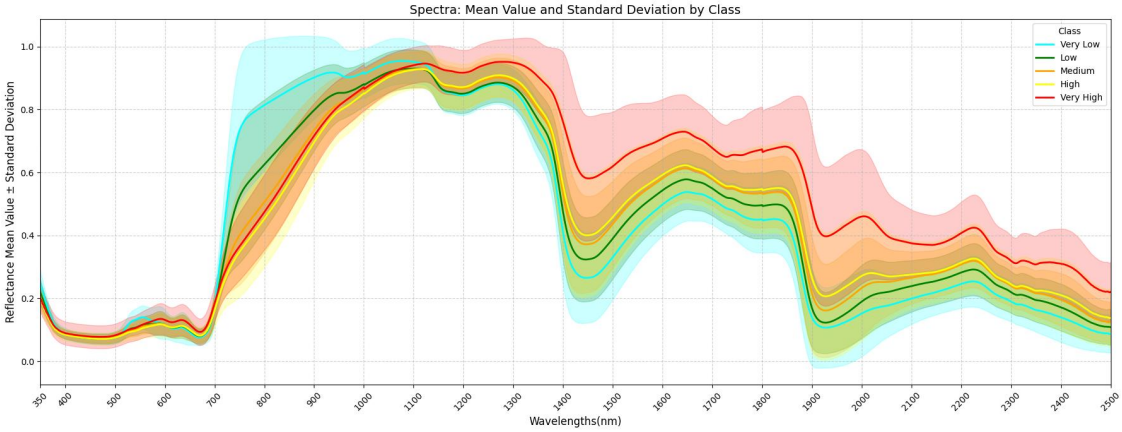
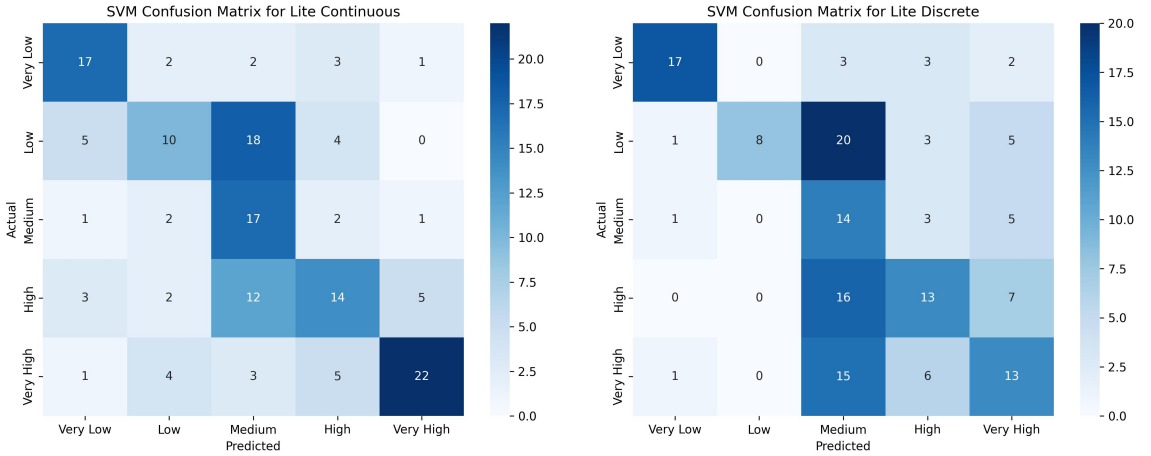


Figure 5.5: Wavelengths vs Reflectance by Class: Nitrogen Content Avocado.

A diagonal trend is still evident in Figure 5.6; however, the number of misclassifications increases, especially for the Medium class.



(a) Optimized SVM Confusion Matrix for Lite Continuous.

(b) Optimized SVM Confusion Matrix for Lite Discrete.

Figure 5.6: Confusion Matrices: Nitrogen Content Avocado.

In terms of importances, they are in the same range as Chlorophyll, most relevance features for NIR are the 724, 728, 726, 733 nm, respectively. For Lite Continuous most notorious importances are 726, 737, 728, 799 nm. Finally, the most relevance features for Lite Discrete are the 910, 555, 630, 680 nm, respectively.

In Table 5.3, the best result was in the NIR range, using the Optimized LR algorithm with 60% accuracy. For most of the different ranges the strongest model was SVM. If the Lite Discrete multispectral sensor were considered instead of the Lite Continuous configuration, the classification accuracy would decrease to 42%.

Table 5.3: Performance Metrics for Wavelength Nitrogen Content Avocado.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
VIS NIR	**LR**	**0.40**	**0.39**	**0.43**	**0.43**	**0.45**	**0.46**
	SVM	0.33	0.31	0.33	0.31	0.44	0.45
	DT	0.29	0.30	0.29	0.30	0.32	0.33
	KNN	0.32	0.33	0.35	0.36	0.41	0.43
VIS	LR	0.33	0.27	**0.41**	**0.41**	0.44	0.45
	SVM	**0.39**	**0.41**	0.30	0.28	**0.47**	**0.48**
	DT	0.32	0.33	0.33	0.34	0.33	0.34
	KNN	0.37	0.38	0.35	0.37	0.33	0.35
NIR	LR	0.44	0.42	0.45	0.44	0.56	0.55
	SVM	**0.46**	**0.46**	**0.49**	**0.49**	**0.60**	**0.59**
	DT	0.41	0.41	0.41	0.41	0.42	0.43
	KNN	0.44	0.44	0.44	0.45	0.45	0.46
SWIR	LR	**0.35**	**0.32**	**0.36**	**0.34**	0.39	0.39
	SVM	0.35	0.29	0.33	0.28	**0.40**	**0.40**
	DT	0.31	0.31	0.31	0.31	0.31	0.31
	KNN	0.34	0.34	0.35	0.35	0.36	0.34
RE	LR	**0.38**	**0.37**	**0.38**	**0.38**	0.43	0.43
	SVM	0.33	0.31	0.32	0.31	**0.46**	**0.46**
	DT	0.28	0.29	0.28	0.29	0.31	0.31
	KNN	0.33	0.34	0.34	0.34	0.32	0.33
Lite C	LR	**0.42**	**0.41**	**0.46**	**0.45**	0.50	0.50
	SVM	0.32	0.29	0.40	0.40	**0.51**	**0.52**
	DT	0.30	0.32	0.29	0.32	0.32	0.32
	KNN	0.33	0.35	0.31	0.33	0.40	0.42
Lite D	LR	0.30	0.26	0.34	0.32	0.38	0.36
	SVM	0.31	0.29	0.37	0.37	**0.42**	**0.44**
	DT	0.30	0.31	0.31	0.32	0.35	0.34
	KNN	**0.34**	**0.35**	**0.37**	**0.39**	0.38	0.41

5.1.1.2 Olive

Water Content in Olive

The reflectance spectra of each class is shown in Figure 5.7. The difference between classes is noticeable. In comparison with the avocado specie, the mean standard deviation of each class is smaller and overlaps less with the other classes.

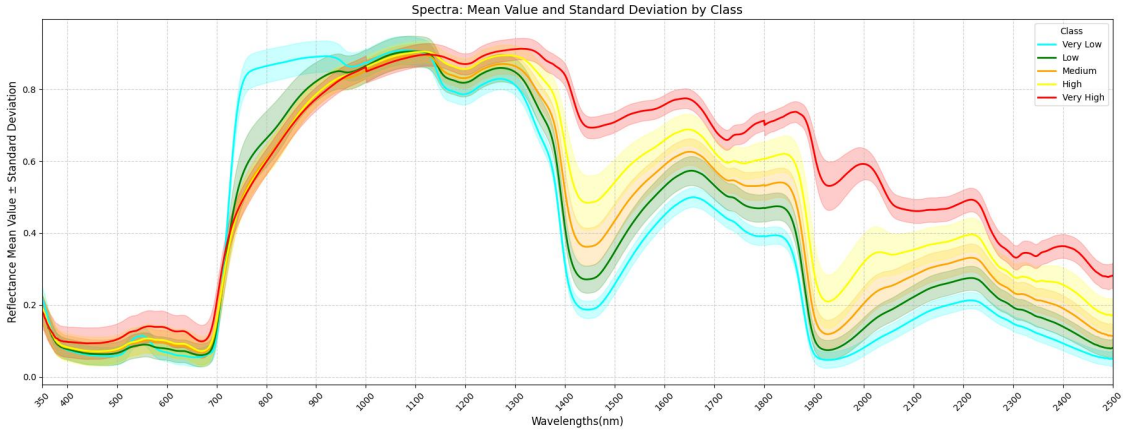
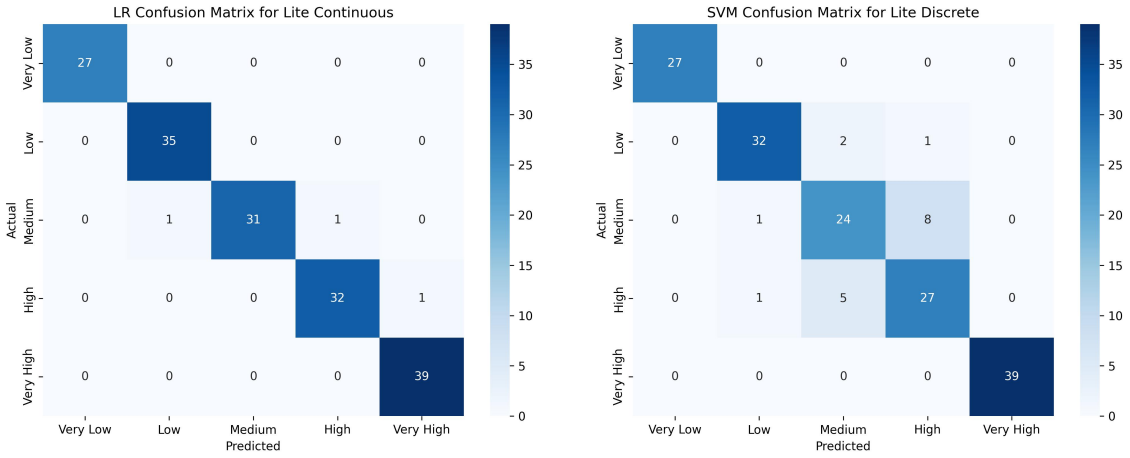


Figure 5.7: Wavelengths vs Reflectance by Class: Water Content Olive.

The clear isolated classes result in a marked main diagonal confusion matrix in Figure 5.8 and consequently higher accuracy values.



(a) Optimized LR Confusion Matrix for Lite Continuous.

(b) Optimized SVM Confusion Matrix for Lite Discrete.

Figure 5.8: Confusion Matrices: Water Content Olive.

The most relevance features for Lite Continuous are the 754, 764, 755, 761 nm, respectively and for Lite Discrete are the 910, 555, 590, 630 nm, respectively, the rest of the importances can be consulted in Appendix A.

According to Table 5.4, the best result was in the Lite Continuous range, using the Optimized LR algorithm with 98% accuracy. For different ranges the strongest models were LR and SVM. If a multispectral sensor such as Lite Discrete were used, the model would achieve an accuracy of approximately 89%, suggesting that this configuration still provides satisfactory predictive performance.

Table 5.4: Performance Metrics for Wavelength Water Content Olive.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
VIS NIR	**LR**	**0.75**	**0.75**	**0.87**	**0.87**	**0.96**	**0.96**
	SVM	0.63	0.60	0.76	0.76	0.96	0.96
	DT	0.69	0.71	0.69	0.71	0.71	0.72
	KNN	0.74	0.75	0.73	0.74	0.71	0.73
VIS	**LR**	0.65	0.65	**0.77**	**0.76**	**0.95**	**0.95**
	SVM	**0.74**	**0.71**	0.62	0.59	0.93	0.93
	DT	0.62	0.63	0.62	0.63	0.63	0.64
	KNN	0.73	0.74	0.71	0.71	0.71	0.72
NIR	**LR**	0.88	0.88	**0.92**	**0.92**	**0.96**	**0.96**
	SVM	**0.95**	**0.95**	0.93	0.93	0.95	0.95
	DT	0.76	0.77	0.76	0.77	0.76	0.77
	KNN	0.87	0.88	0.87	0.87	0.88	0.88
SWIR	LR	0.83	0.83	**0.86**	**0.86**	0.90	0.90
	SVM	**0.89**	**0.89**	0.84	0.83	**0.92**	**0.92**
	DT	0.77	0.78	0.77	0.78	0.82	0.82
	KNN	0.83	0.83	0.85	0.85	0.81	0.82
RE	**LR**	0.67	0.67	**0.80**	**0.80**	**0.95**	**0.95**
	SVM	**0.80**	**0.80**	0.76	0.77	0.93	0.93
	DT	0.68	0.68	0.68	0.68	0.68	0.68
	KNN	0.77	0.77	0.75	0.76	0.76	0.77
Lite C	**LR**	**0.78**	**0.78**	**0.93**	**0.93**	**0.98**	**0.98**
	SVM	0.64	0.64	0.87	0.87	0.96	0.96
	DT	0.72	0.73	0.71	0.72	0.71	0.72
	KNN	0.72	0.72	0.72	0.72	0.73	0.73
Lite D	LR	0.35	0.28	0.51	0.49	0.89	0.89
	SVM	0.49	0.48	**0.66**	**0.67**	**0.89**	**0.89**
	DT	**0.62**	**0.64**	0.64	0.65	0.63	0.63
	KNN	0.57	0.58	0.59	0.60	0.56	0.58

Chlorophyll Content in Olive

The reflectance spectra of each class can be seen in Figure 5.9. The chart illustrated that classes are distinct. Even thought the mean standard deviation for each class is smaller that in the Chlorophyll Content Avocado, there is still an overlap of between adjacent classes.

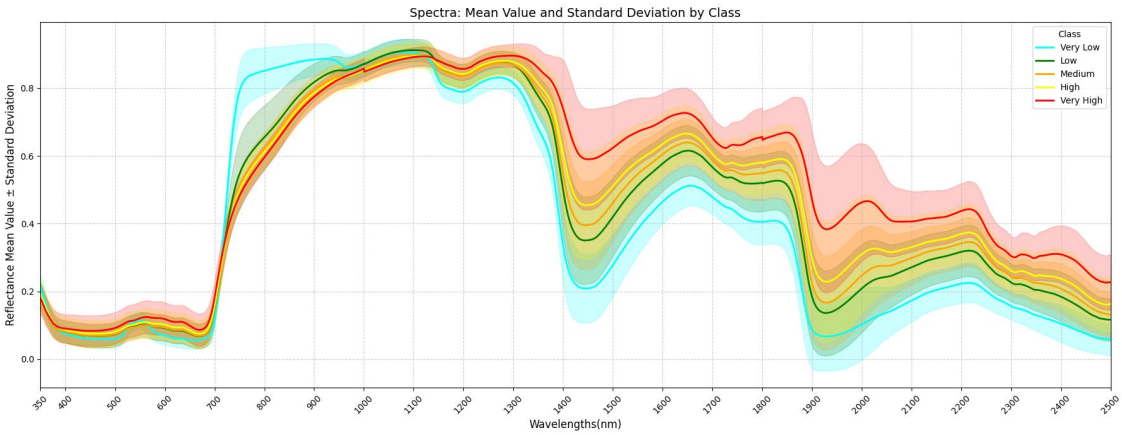
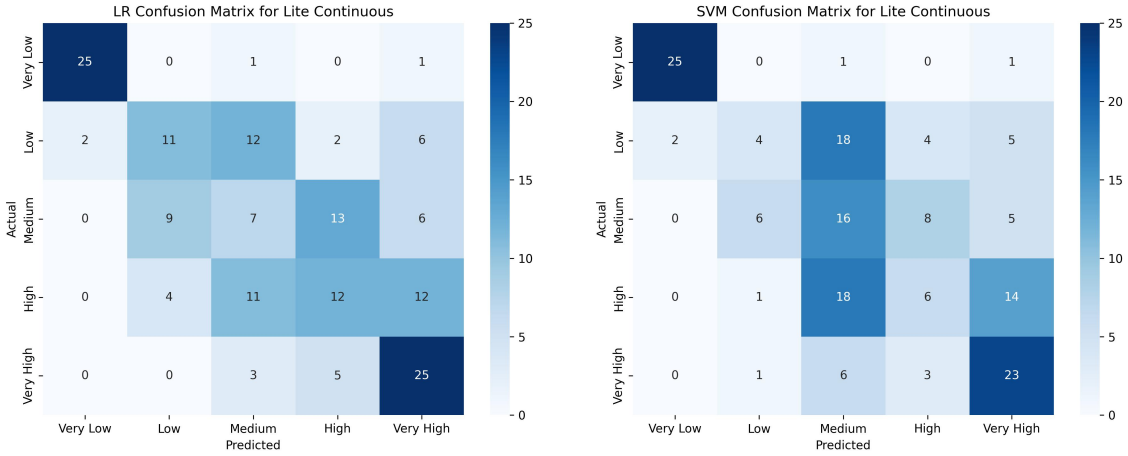


Figure 5.9: Wavelengths vs Reflectance by Class: Chlorophyll Content Olive.

Even though it is possible to identify a diagonal on both matrices in Figure 5.10, there are also several misclassifications, particularly among the three middle classes, whereas the two extreme classes are classified more accurately.



(a) Optimized LR Confusion Matrix for Lite Continuous.

(b) Optimized SVM Confusion Matrix for Lite Discrete.

Figure 5.10: Confusion Matrices: Chlorophyll Content Olive.

Most relevance features for Lite Continuous are the 761, 744, 768, 747 nm, respectively. The first most relevant importances for Lite Discrete are the same as the ones in the Avocado (910, 555, 630 nm), but the fourth one is 590 nm.

The best result was in the SWIR range, using the Optimized SVM algorithm with 51% accuracy. For different ranges the strongest models were LR and SVM. If the Lite Discrete multispectral sensor were considered instead of the Lite Continuous configuration, the classification accuracy would decrease to 49%. This can be seen in Table 5.5.

Table 5.5: Performance Metrics for Wavelength Chlorophyll Content Olive.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
VIS NIR	**LR**	**0.49**	**0.50**	**0.46**	**0.47**	**0.46**	**0.47**
	SVM	0.43	0.41	0.45	0.42	0.44	0.42
	DT	0.40	0.42	0.40	0.42	0.40	0.42
	KNN	0.40	0.42	0.43	0.46	0.45	0.47
VIS	**LR**	0.41	0.40	**0.45**	**0.43**	**0.46**	**0.44**
	SVM	0.40	0.37	0.34	0.32	0.44	0.44
	DT	0.38	0.40	0.39	0.41	0.37	0.38
	KNN	**0.44**	**0.46**	0.44	0.46	0.39	0.41
NIR	**LR**	**0.51**	**0.51**	**0.51**	**0.52**	**0.50**	**0.52**
	SVM	0.49	0.51	0.48	0.49	**0.50**	**0.51**
	DT	0.44	0.45	0.43	0.45	0.38	0.41
	KNN	0.46	0.47	0.43	0.44	0.44	0.45
SWIR	LR	**0.51**	**0.50**	**0.49**	**0.48**	0.50	0.51
	SVM	0.45	0.43	0.43	0.40	**0.51**	**0.52**
	DT	0.44	0.44	0.43	0.44	0.44	0.43
	KNN	0.49	0.49	0.48	0.49	0.50	0.50
RE	LR	**0.47**	**0.45**	**0.49**	**0.48**	0.41	0.42
	SVM	0.46	0.44	0.45	0.43	**0.49**	**0.49**
	DT	0.39	0.41	0.40	0.42	0.40	0.42
	KNN	0.41	0.43	0.41	0.43	0.42	0.44
Lite C	**LR**	**0.50**	**0.51**	**0.46**	**0.47**	**0.48**	**0.49**
	SVM	0.46	0.43	0.44	0.44	0.44	0.44
	DT	0.39	0.41	0.39	0.41	0.41	0.43
	KNN	0.42	0.44	0.44	0.47	0.46	0.46
Lite D	**LR**	0.33	0.26	0.40	0.34	**0.49**	**0.49**
	SVM	0.37	0.34	**0.47**	**0.45**	0.45	0.43
	DT	0.37	0.37	0.37	0.38	0.44	0.45
	KNN	**0.43**	**0.44**	0.41	0.43	0.42	0.43

Nitrogen Content in Olive

The reflectance spectra of each class can be seen in Figure 5.11. Similar to Chlorophyll Content in Olive, the Classes are distinct. Even thought the mean standard deviation for each class is smaller that in the Nitrogen Content Avocado, there is still an overlap of between adjacent classes.

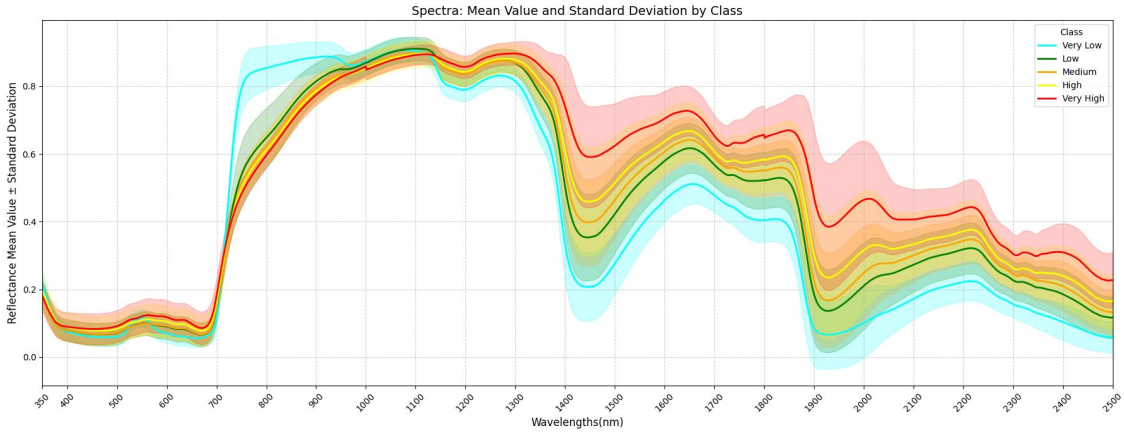
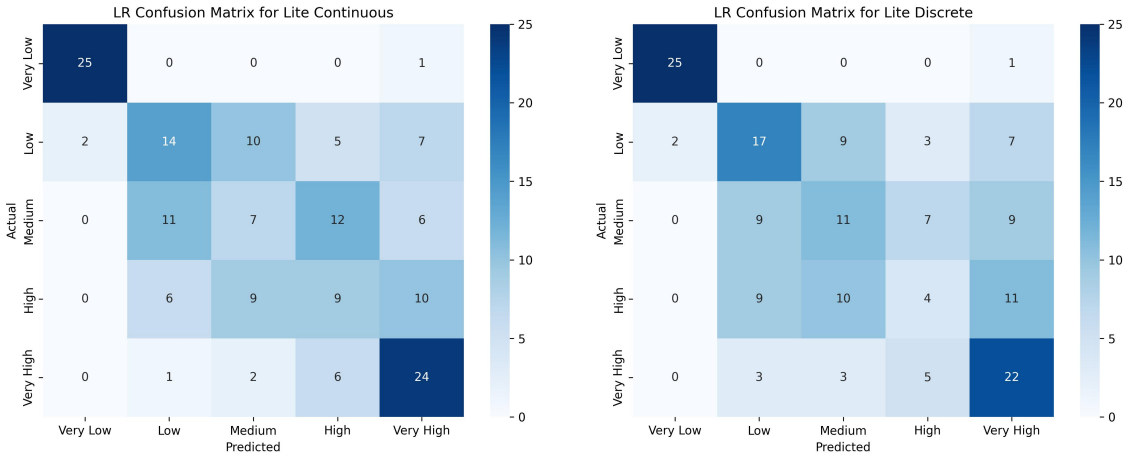


Figure 5.11: Wavelengths vs Reflectance by Class: Nitrogen Content Olive.

The main diagonal in both matrices in Figure 5.12 is not sharply defined, appearing more dispersed across adjacent classes. Misclassifications are highly concentrated among the three middle classes, whereas the two extreme classes are classified more accurately.



(a) Optimized LR Confusion Matrix for Lite Continuous.

(b) Optimized LR Confusion Matrix for Lite Discrete.

Figure 5.12: Confusion Matrices: Nitrogen Content Olive.

Most relevance features for Lite Continuous are the 764, 754, 744, 765, respectively. Most relevance features for Lite Continuous are the 910, 555, 630, 590, respectively.

The best result was in the NIR range, using the Optimized LR algorithm with 51% accuracy. For most of the different ranges the strongest model was SVM, as shown in Table 5.6 If

a multispectral sensor such as Lite Discrete were used, the model would achieve an accuracy of approximately 50%, which is almost the same value as using the NIR continuous range.

Table 5.6: Performance Metrics for Wavelength Nitrogen Content Olive.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
VIS NIR	**LR**	**0.48**	**0.48**	**0.51**	**0.52**	**0.48**	**0.49**
	SVM	0.44	0.43	0.47	0.46	0.44	0.44
	DT	0.40	0.42	0.40	0.42	0.41	0.42
	KNN	0.44	0.46	0.44	0.46	0.44	0.46
VIS	**LR**	0.43	0.42	**0.44**	**0.44**	**0.46**	**0.46**
	SVM	0.43	0.41	0.39	0.34	0.46	0.44
	DT	0.40	0.42	0.40	0.42	0.35	0.38
	KNN	**0.43**	**0.45**	0.43	0.45	0.40	0.41
NIR	LR	0.49	0.47	**0.53**	**0.54**	0.50	0.52
	SVM	**0.50**	**0.52**	0.50	0.50	**0.51**	**0.52**
	DT	0.47	0.48	0.46	0.48	0.47	0.49
	KNN	0.46	0.47	0.43	0.44	0.47	0.48
SWIR	LR	0.50	0.47	**0.49**	**0.46**	0.49	0.49
	SVM	0.46	0.43	0.44	0.41	**0.50**	**0.51**
	DT	0.43	0.45	0.43	0.45	0.43	0.45
	KNN	**0.51**	**0.51**	0.47	0.48	0.49	0.49
RE	LR	**0.48**	**0.46**	0.47	0.46	0.46	0.46
	SVM	0.48	0.46	0.45	0.43	0.46	0.45
	DT	0.47	0.49	**0.47**	**0.49**	0.40	0.41
	KNN	0.43	0.45	0.45	0.47	**0.47**	**0.48**
Lite C	LR	**0.49**	**0.49**	**0.50**	**0.50**	0.47	0.49
	SVM	0.45	0.45	0.46	0.46	**0.49**	**0.51**
	DT	0.38	0.40	0.38	0.40	0.36	0.39
	KNN	0.44	0.46	0.48	0.50	0.47	0.48
Lite D	LR	0.34	0.26	0.40	0.37	0.47	0.48
	SVM	0.38	0.35	**0.46**	**0.44**	**0.50**	**0.50**
	DT	0.41	0.42	0.40	0.41	0.40	0.42
	KNN	**0.45**	**0.46**	0.45	0.47	0.38	0.39

5.1.1.3 Vineyard

Water Content in Vineyard

The reflectance spectra of each class can be visualized in Figure 5.13. The difference between the Very High Class and the other classes is noticeable. Comparing with the avocado and olive species, the rest of the classes are closer to each other here. Hence, the classes, except for Very High, are overlapping with the adjacent classes. In comparison with the olive specie, the mean standard deviation of each class is slightly bigger, but much smaller than in the avocado specie.

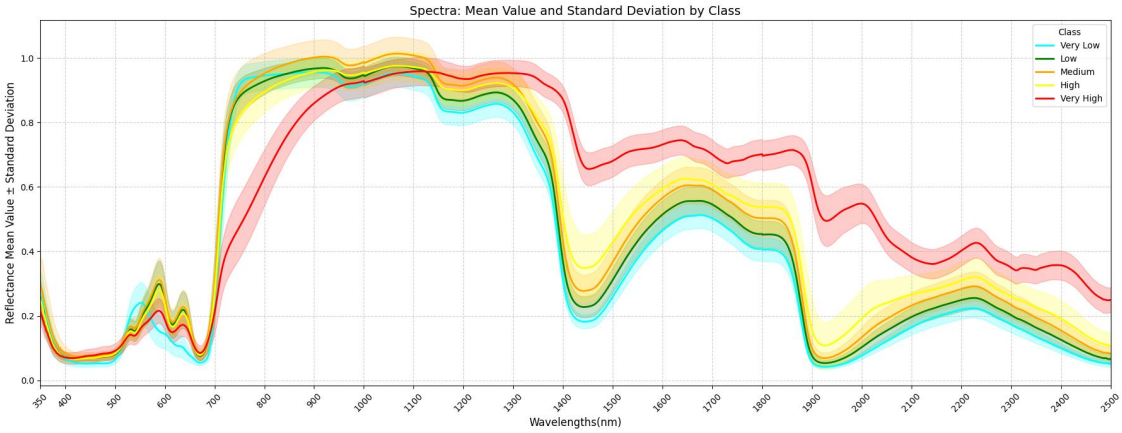
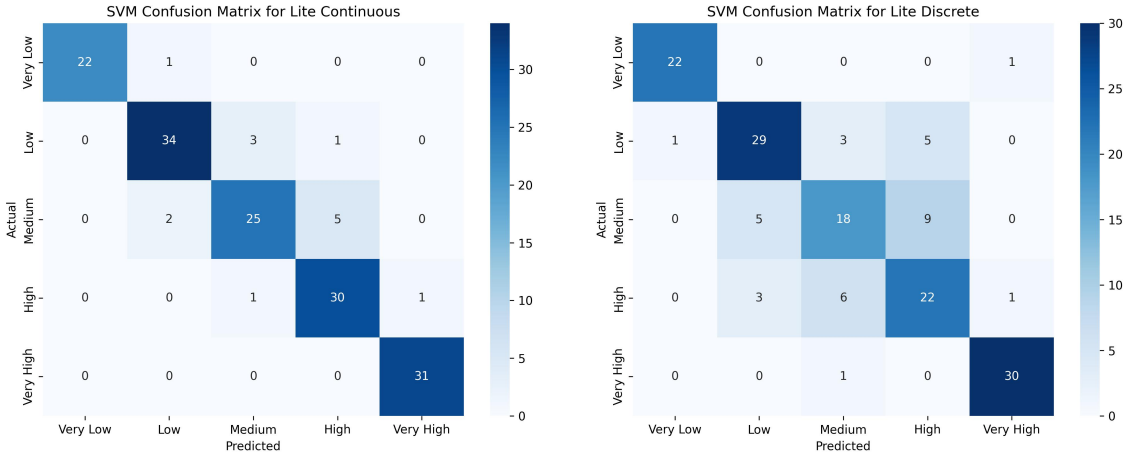


Figure 5.13: Wavelengths vs Reflectance by Class: Water Content Vineyard.

The confusion matrices, in Figure 5.14, indicate that the GridSearchCV optimized SVM achieved the best classification performance for both Lite configurations. In the Lite Continuous range, almost all samples were correctly classified, whereas in the Lite Discrete range the majority of samples were assigned to their correct classes, although a higher number of misclassifications was observed.



(a) Optimized SVM Confusion Matrix for Lite Continuous.

(b) Optimized SVM Confusion Matrix for Lite Discrete.

Figure 5.14: Confusion Matrices: Water Content Vineyard.

For this configuration, the most relevance features for Lite Continuous are the 563, 726, 727, 567, respectively. Regarding the Lite Discrete are the 555, 590, 630, 515, respectively.

According to Table 5.7, the best result was in the NIR range, using the Optimized LR algorithm with 91% accuracy. For most of the different ranges the strongest model was LR or SVM.

If a multispectral sensor such as Lite Discrete were used, the achieved accuracy would decrease to 78%, suggesting that this configuration still provides satisfactory predictive performance.

Table 5.7: Performance Metrics for Wavelength Water Content Vineyard.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
VIS NIR	LR	**0.72**	**0.74**	0.77	0.78	0.90	0.90
	SVM	0.72	0.73	**0.83**	**0.84**	**0.90**	**0.91**
	DT	0.67	0.69	0.67	0.69	0.71	0.72
	KNN	0.68	0.70	0.72	0.73	0.70	0.72
VIS	**LR**	0.60	0.61	**0.72**	**0.74**	**0.84**	**0.85**
	SVM	**0.71**	**0.73**	**0.72**	**0.74**	0.81	0.83
	DT	0.61	0.63	0.62	0.64	0.62	0.64
	KNN	0.63	0.66	0.63	0.66	0.63	0.66
NIR	**LR**	0.74	0.75	0.76	0.78	**0.91**	**0.92**
	SVM	**0.86**	**0.87**	**0.79**	**0.81**	0.85	0.86
	DT	0.67	0.69	0.67	0.70	0.62	0.65
	KNN	0.70	0.71	0.69	0.70	0.71	0.73
SWIR	**LR**	0.72	0.73	**0.76**	**0.77**	**0.85**	**0.86**
	SVM	0.72	0.73	0.64	0.65	0.81	0.82
	DT	0.60	0.61	0.60	0.61	0.62	0.63
	KNN	**0.74**	**0.75**	0.72	0.73	0.78	0.78
RE	**LR**	0.62	0.62	0.67	0.70	**0.78**	**0.80**
	SVM	**0.67**	**0.70**	**0.76**	**0.78**	0.78	0.79
	DT	0.64	0.67	0.64	0.67	0.64	0.67
	KNN	0.63	0.66	0.63	0.66	0.66	0.69
Lite C	LR	**0.78**	**0.79**	0.78	0.80	0.89	0.90
	SVM	0.72	0.73	**0.81**	**0.82**	**0.91**	**0.91**
	DT	0.68	0.69	0.69	0.70	0.71	0.71
	KNN	0.69	0.71	0.69	0.71	0.74	0.76
Lite D	LR	0.49	0.43	0.64	0.66	0.76	0.77
	SVM	0.62	0.64	**0.74**	**0.75**	**0.78**	**0.79**
	DT	0.57	0.59	0.58	0.60	0.58	0.60
	KNN	**0.66**	**0.68**	0.69	0.71	0.70	0.72

Chlorophyll Content in Vineyard

The reflectance spectra of each class can be visualized in Figure 5.15. There is a clear overlap of the mean standard deviation of each class with the adjacent class. The Very High mean is the one which is more perceptible.

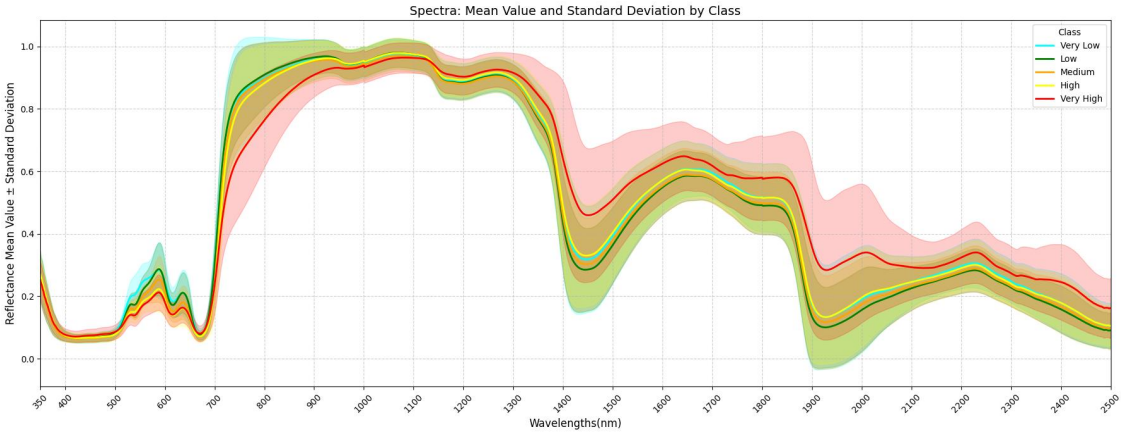
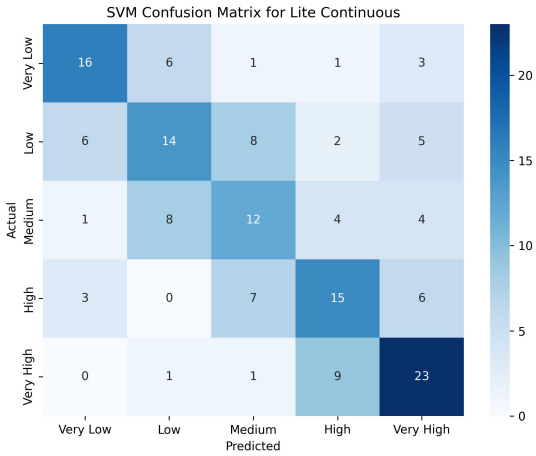


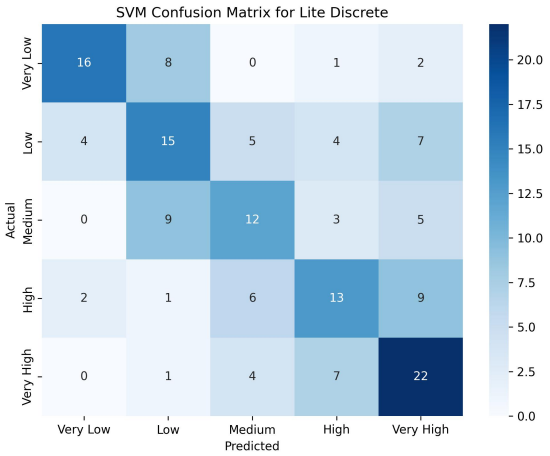
Figure 5.15: Wavelengths vs Reflectance by Class: Chlorophyll Content in Vineyard.

For Lite Continuous the most relevance features are the 563, 726, 727, 567 nm, respectively. For Lite Discrete, the most informative importances correspond to 555, 590, 630, 515 nm, respectively.

The main diagonal in both matrices in Figure 5.16 is not sharply defined, appearing more dispersed across adjacent classes. There are misclassifications across all classes.



(a) Optimized SVM Confusion Matrix for Lite Continuous.



(b) Optimized SVM Confusion Matrix for Lite Discrete.

Figure 5.16: Confusion Matrices: Chlorophyll Content Vineyard.

The best result was in the VIS range, using the Optimized SVM algorithm with 54% accuracy. For different ranges the strongest models were LR and SVM. If a multispectral sensor such as Lite Discrete were used, the model would achieve an accuracy of approximately 50%, which is almost the same value as using the VIS continuous range. This can be seen in Table 5.8.

Table 5.8: Performance Metrics for Wavelength Chlorophyll Content Vineyard.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
VIS NIR	LR	**0.50**	**0.50**	0.45	0.45	0.45	0.45
	SVM	0.48	0.49	**0.51**	**0.51**	**0.49**	**0.50**
	DT	0.43	0.43	0.43	0.43	0.46	0.46
	KNN	0.40	0.40	0.42	0.41	0.47	0.46
VIS	LR	0.47	0.46	0.49	0.48	0.49	0.49
	SVM	**0.54**	**0.54**	**0.49**	**0.49**	**0.54**	**0.53**
	DT	0.42	0.42	0.42	0.41	0.37	0.37
	KNN	0.48	0.48	0.44	0.43	0.51	0.50
NIR	LR	0.42	0.42	0.46	0.46	0.43	0.43
	SVM	**0.49**	**0.48**	**0.50**	**0.50**	**0.49**	**0.48**
	DT	0.37	0.37	0.37	0.37	0.37	0.37
	KNN	0.37	0.37	0.28	0.28	0.36	0.36
SWIR	LR	0.33	0.31	**0.35**	**0.34**	0.42	0.41
	SVM	**0.33**	**0.31**	0.29	0.25	**0.45**	**0.45**
	DT	0.22	0.22	0.22	0.22	0.27	0.27
	KNN	0.30	0.28	0.28	0.25	0.31	0.28
RE	LR	0.42	0.42	0.48	0.48	**0.51**	**0.51**
	SVM	**0.43**	**0.42**	**0.50**	**0.50**	0.50	0.50
	DT	0.38	0.38	0.38	0.39	0.45	0.45
	KNN	0.42	0.41	0.43	0.43	0.44	0.44
Lite C	LR	0.49	0.50	0.47	0.47	0.47	0.47
	SVM	**0.51**	**0.52**	**0.49**	**0.49**	**0.51**	**0.51**
	DT	0.40	0.40	0.40	0.40	0.44	0.43
	KNN	0.46	0.46	0.42	0.43	0.44	0.43
Lite D	LR	0.38	0.30	**0.49**	**0.48**	0.46	0.45
	SVM	**0.46**	**0.46**	0.47	0.47	**0.50**	**0.50**
	DT	0.37	0.37	0.37	0.36	0.40	0.38
	KNN	0.45	0.44	0.41	0.41	0.43	0.43

Nitrogen Content in Vineyard

The reflectance spectra of each class can be visualized in Figure 5.17. Similarly, to Chlorophyll Content Vineyard, there is a clear overlap of the mean standard deviation of each class with the adjacent class. The Very High mean is the one which is more perceptible.

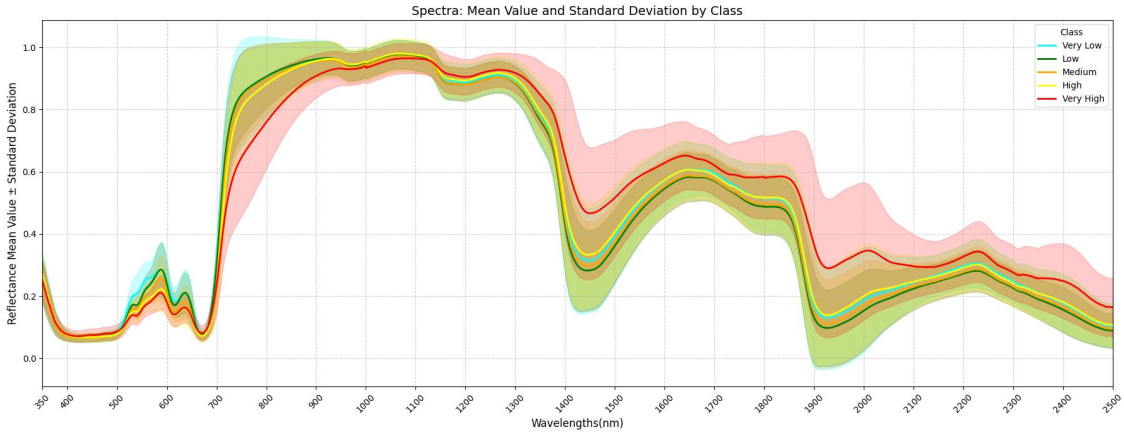
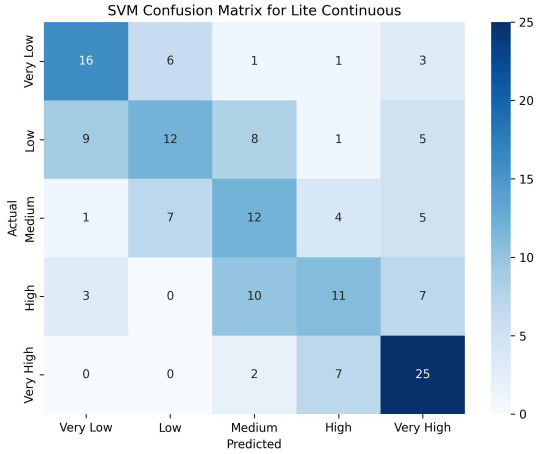
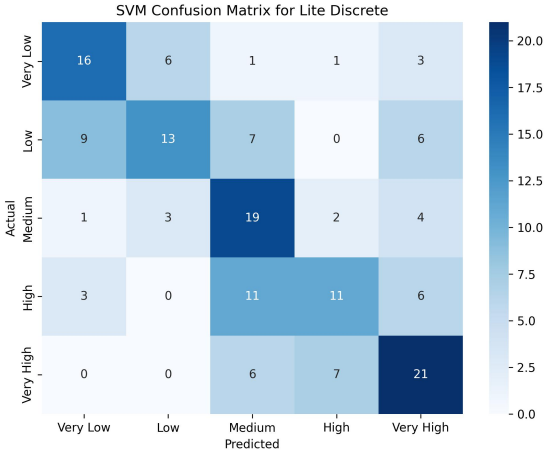


Figure 5.17: Wavelengths vs Reflectance: Nitrogen Content in Vineyard.

In Figure 5.18, the correct predictions form a broad cluster rather than a sharp diagonal, appearing dispersed across adjacent classes. There are misclassifications across all classes.



(a) Optimized SVM Confusion Matrix for Lite Continuous.



(b) Optimized SVM Confusion Matrix for Lite Discrete.

Figure 5.18: Confusion Matrices: Nitrogens Content Vineyard.

For Lite Continuous the most relevance features are the 563, 556, 557, 726, respectively. For Lite Discrete, they are the 555, 590, 630, 515, respectively.

The best result was in the VIS range, using the Optimized SVM algorithm with 56% accuracy. For different ranges the strongest models were LR and SVM. If a multispectral sensor such as Lite Discrete were used, the model would achieve an accuracy of approximately 51%, which is almost the same value as using the VIS continuous range. This can be seen in Table 5.9.

Table 5.9: Performance Metrics for Wavelength Nitrogen Content Vineyard.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
VIS NIR	**LR**	**0.51**	**0.51**	0.48	0.48	**0.50**	**0.50**
	SVM	0.50	0.50	**0.53**	**0.52**	0.50	0.49
	DT	0.38	0.38	0.38	0.38	0.40	0.40
	KNN	0.42	0.42	0.42	0.42	0.49	0.48
VIS	LR	0.51	0.49	0.50	0.49	0.51	0.51
	SVM	**0.54**	**0.54**	**0.51**	**0.51**	**0.56**	**0.56**
	DT	0.39	0.40	0.39	0.40	0.42	0.41
	KNN	0.48	0.48	0.45	0.45	0.50	0.50
NIR	**LR**	0.42	0.41	0.45	0.44	**0.44**	**0.43**
	SVM	**0.47**	**0.47**	**0.47**	**0.46**	0.43	0.42
	DT	0.28	0.28	0.28	0.28	0.32	0.32
	KNN	0.37	0.36	0.30	0.30	0.36	0.35
SWIR	**LR**	**0.31**	**0.27**	**0.38**	**0.36**	**0.44**	**0.44**
	SVM	0.31	0.25	0.29	0.21	0.44	0.44
	DT	0.22	0.23	0.22	0.23	0.22	0.22
	KNN	0.30	0.29	0.31	0.29	0.33	0.32
RE	LR	0.40	0.36	0.47	0.46	0.50	0.50
	SVM	**0.45**	**0.40**	**0.50**	**0.50**	**0.51**	**0.50**
	DT	0.39	0.39	0.38	0.38	0.37	0.36
	KNN	0.40	0.39	0.38	0.38	0.44	0.43
Lite C	**LR**	**0.51**	**0.50**	0.46	0.46	**0.50**	**0.50**
	SVM	0.48	0.44	**0.50**	**0.49**	0.49	0.48
	DT	0.40	0.40	0.40	0.40	0.36	0.35
	KNN	0.46	0.46	0.42	0.43	0.44	0.43
Lite D	LR	0.37	0.31	0.49	0.47	0.47	0.46
	SVM	0.46	0.43	**0.50**	**0.49**	**0.51**	**0.51**
	DT	0.40	0.39	0.39	0.39	0.46	0.45
	KNN	**0.46**	**0.44**	0.41	0.41	0.43	0.42

5.1.1.4 Vegetation Indices

There were plotted three VIs graphs, Figures 5.19, 5.20 and 5.21. For all three categories, the spectral indices approximations have a similar value to the value calculated with the designated wavelength. The major difference is around 0.2 in the best case for in every class. In very low

water content class that happens in the PRI550 VI and for the other in the RDVI VI. The OSAVI and NDVI are more similar across all classes.

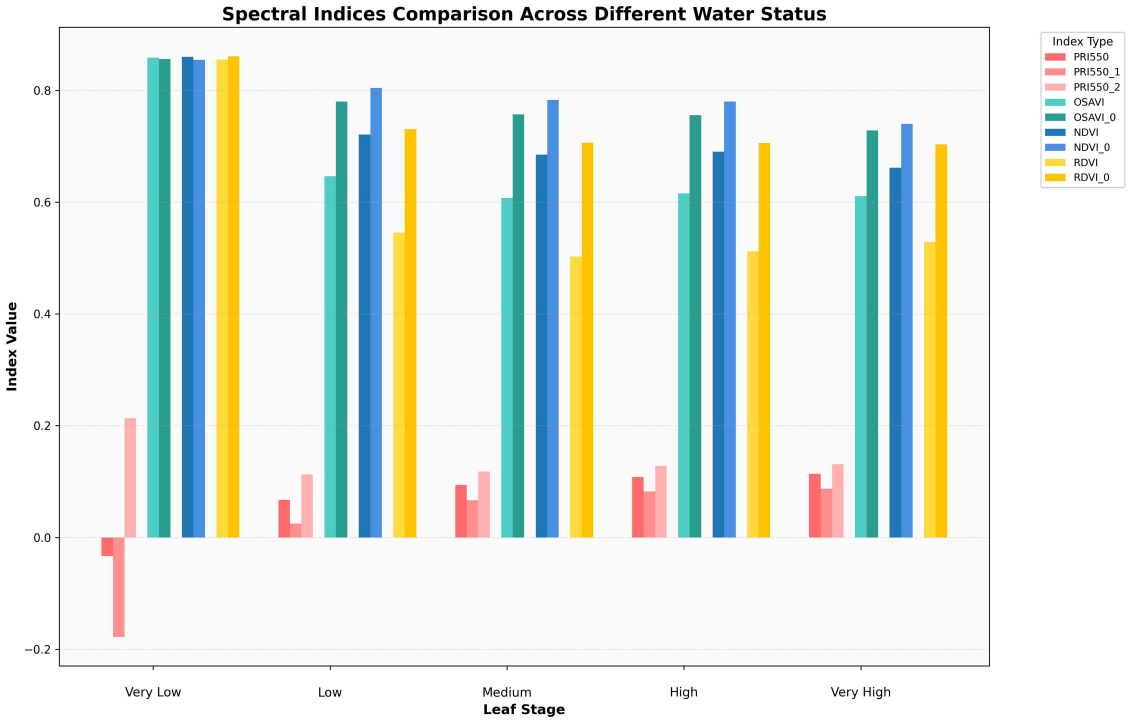


Figure 5.19: VIs Hyperspectral sensor VS VIs Multispectral sensor (Lite Discrete): Avocado.

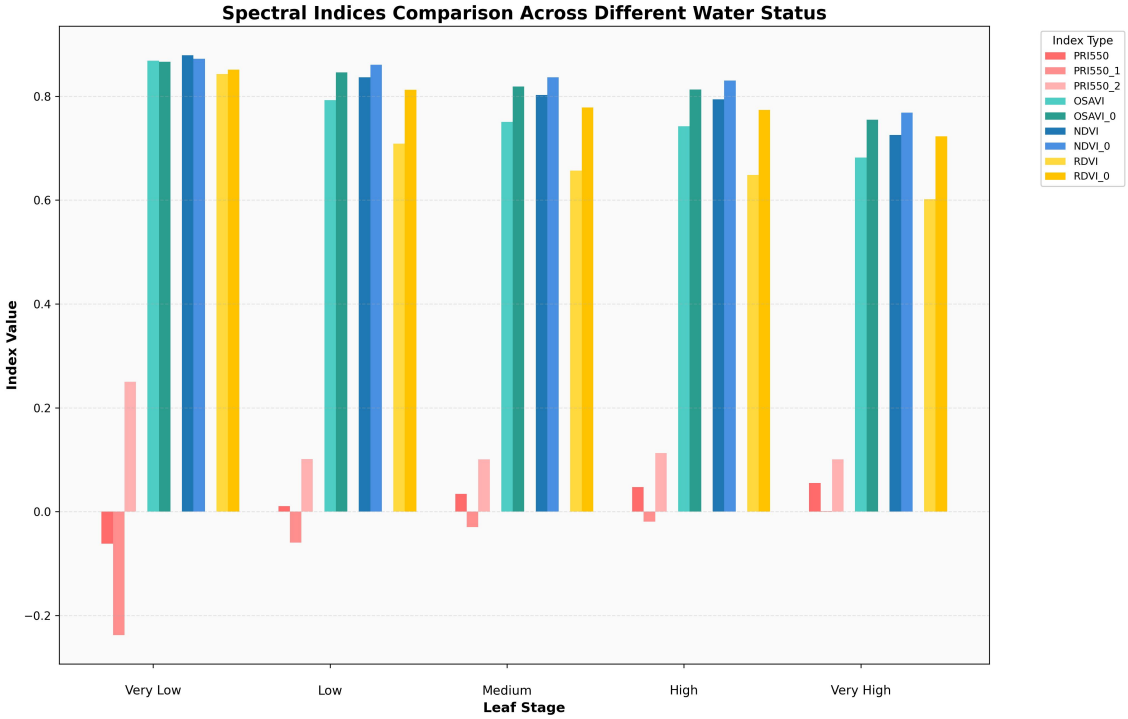


Figure 5.20: VIs Hyperspectral sensor VS VIs Multispectral sensor (Lite Discrete): Olive.

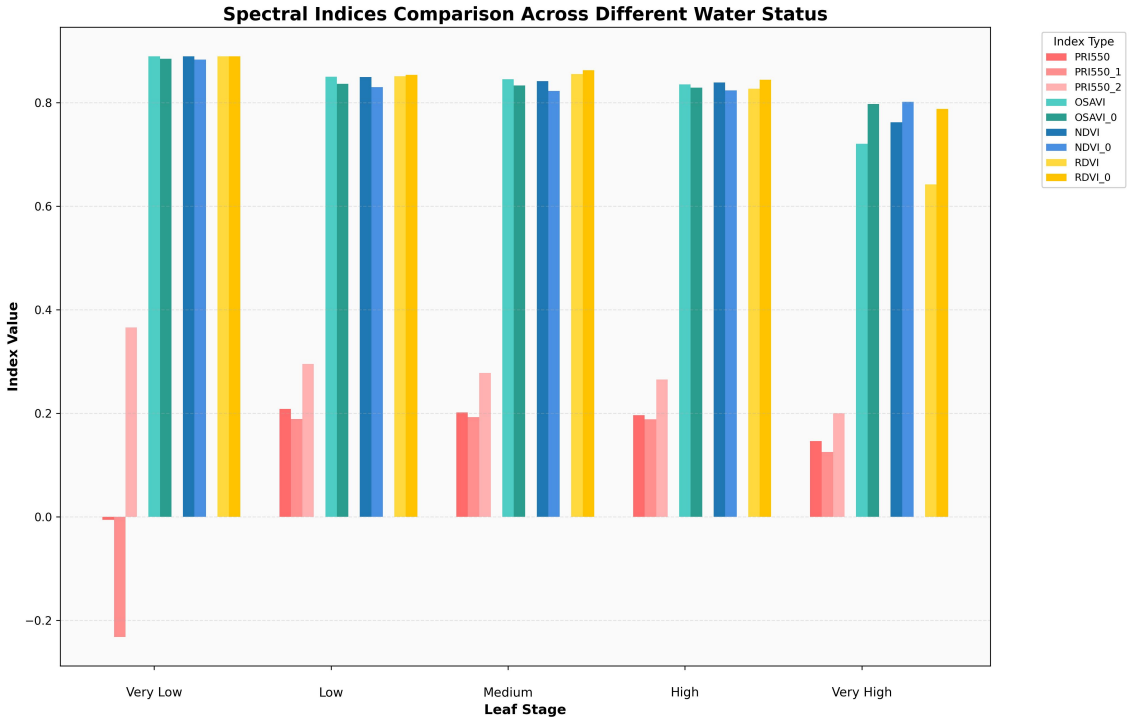


Figure 5.21: VIs Hyperspectral sensor VS VIs Multispectral sensor (Lite Discrete): Vineyard.

5.1.2 Classifier Discussion

The spectral signature varies according to both the plant species and the target plant metric. Among the evaluated variables, water content was predicted with the highest accuracy across the datasets. The reflectance curves were most distinct for olive, moderately distinct for avocado, and least distinct for vineyard. The greater class separability is reflected in the improved classification performance and the more pronounced diagonals in confusion matrices.

Across all datasets, the LR and SVM models consistently achieved the best performance after hyperparameter optimization using GridSearchCV. Furthermore, the most suitable spectral range depended on both the plant species and the target metric. For avocado and olive, which exhibited similar spectral behaviour, the Lite Continuous configuration provided the best results for water content prediction, whereas the NIR range performed best for chlorophyll and nitrogen estimation. In contrast, for vineyard data, the NIR range yielded the highest performance for water content prediction, while the VIS range was more suitable for chlorophyll and nitrogen estimation.

Regarding feature importance, avocado and olive exhibited the same three most informative wavelengths (910, 555, and 630 nm) for chlorophyll and nitrogen prediction when using the Lite Discrete configuration. For water content prediction, the most informative bands differed slightly, with the top wavelengths being 630, 910, and 590 nm for avocado, and 910, 555, and 590 nm for olive. In the Lite Continuous configuration, the most informative wavelengths varied across all metrics due to the higher spectral resolution; however, they were predominantly concentrated within the 700-800 nm range.

For the vineyard dataset, the most informative wavelengths in the Lite Continuous configuration were distributed between 555 and 727 nm. Under the Lite Discrete configuration, the chlorophyll and nitrogen models consistently identified the 555, 590, and 630 nm bands as the most informative. In contrast, water content prediction relied primarily on the 910, 680, and 480 nm bands.

5.1.3 Regressor Algorithms

In order to facilitate the results' interpretation, the wavelength spectrum was plotted showing the mean reflectance of the dataset. To assess the impact of the spectral input representation on predictive performance, ground truth vs prediction graphs were generated for the best-performing models obtained using both the Lite Continuous and Lite Discrete wavelength configurations for each target metric. This comparison allows the evaluation of the differences between continuous spectral measurements and discretely sampled wavelengths.

The results obtained for each spectral range are presented in the performance metrics tables. The first two columns report the performance of the baseline models without optimization. The third and fourth columns present the results obtained after applying normalization, while the fifth and sixth columns correspond to the normalized models further optimized through GridSearchCV. For each configuration, the first column reports the R^2score and the second column reports the RMSE.

Within each configuration, the best-performing model is highlighted in bold. Additionally, the overall best-performing method for each spectral range is also indicated in bold.

To illustrate better the optimizations made with GridSearchCV, there is a table for each model, which shows the best obtained parameters. The models are Ridge, PLS, SVM, KNN models. The parameters of the best-performing model for this configuration are in bold.

5.1.3.1 Avocado

Water Content in Avocado

The reflectance spectra of each class can be visualized in Figure 5.22. The Ground-Truth vs Predictions graphs in Figure 5.23 show a perfect regression line.

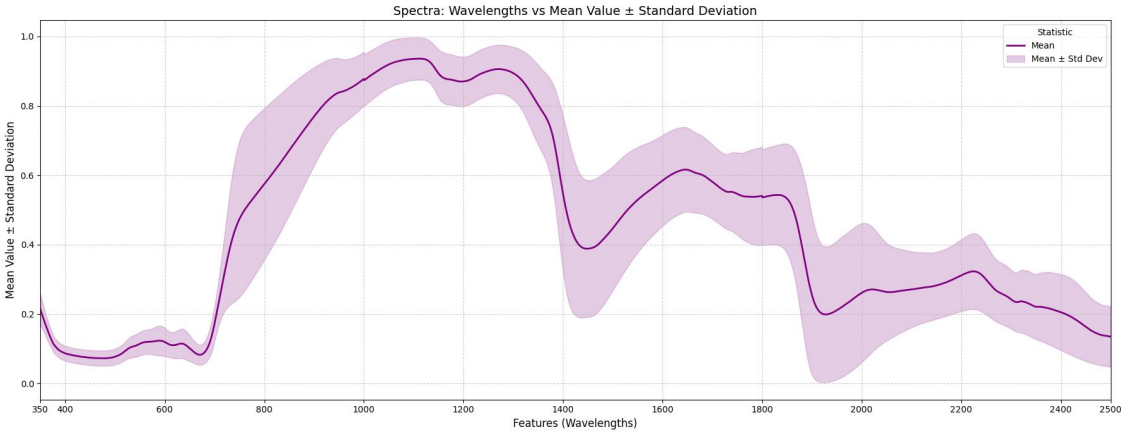
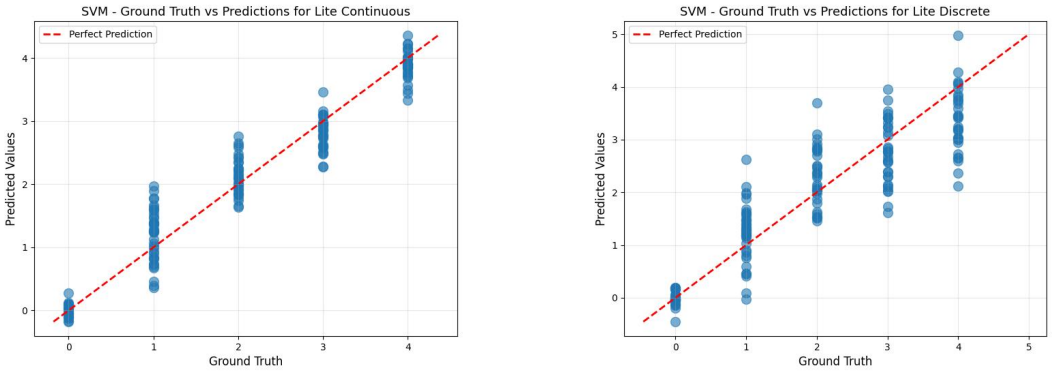


Figure 5.22: Wavelengths vs Reflectance: Water Content in Avocado.

Most relevance features for Lite Continuous are the 760, 761, 759, 754 nm, respectively. Most relevance features for Lite Discrete are the 630, 555, 910, 680 nm, respectively. As shown in Table 5.10, the best model was SVM in the SWIR range with a 0.947 R^2 score and a 0.304 RMSE value.



(a) SVM Ground Truth vs Predictions for Lite Continuous Water Content Avocado.

(b) SVM Ground Truth vs Predictions for Lite Discrete Water Content Avocado.

Figure 5.23: Ground Truth vs Predictions.

Table 5.10: Performance Metrics for Wavelength Water Content Avocado Regressor.

Range	Method	No Optim. R^2score	No Optim. RMSE	Norm. R^2score	Norm. RMSE	GridSearch R^2score	GridSearch RMSE
VIS NIR	Ridge	**0.697**	**0.744**	**0.760**	**0.663**	0.690	0.753
	PLS	0.488	0.967	0.488	0.967	0.853	0.519
	SVM	0.576	0.881	0.479	0.976	**0.889**	**0.450**
	KNN	0.657	0.792	0.634	0.819	0.634	0.819
VIS	Ridge	0.646	0.804	**0.702**	**0.739**	0.643	0.808
	PLS	0.625	0.828	0.625	0.828	0.809	0.592
	SVM	**0.710**	**0.729**	0.354	1.087	**0.848**	**0.527**
	KNN	0.698	0.743	0.688	0.755	0.687	0.756
NIR	Ridge	0.882	0.464	**0.906**	**0.415**	0.864	0.499
	PLS	0.633	0.819	0.633	0.819	**0.947**	**0.311**
	SVM	**0.923**	**0.375**	0.640	0.811	0.938	0.336
	KNN	0.799	0.607	0.808	0.592	0.808	0.592
SWIR	Ridge	0.898	0.432	**0.911**	**0.403**	0.883	0.463
	PLS	0.831	0.557	0.831	0.557	0.931	0.355
	SVM	0.852	0.521	0.718	0.718	**0.949**	**0.304**
	KNN	**0.909**	**0.409**	0.902	0.424	0.897	0.434
RE	Ridge	0.534	0.923	**0.623**	**0.831**	0.527	0.930
	PLS	0.504	0.952	0.504	0.952	0.757	0.667
	SVM	0.531	0.926	0.407	1.042	**0.796**	**0.611**
	KNN	**0.584**	**0.872**	0.590	0.866	0.590	0.866
Lite C	Ridge	**0.792**	**0.616**	**0.871**	**0.487**	0.727	0.706
	PLS	0.612	0.842	0.612	0.842	0.925	0.371
	SVM	0.627	0.825	0.485	0.971	**0.939**	**0.333**
	KNN	0.661	0.787	0.661	0.787	0.652	0.798
Lite D	Ridge	0.380	1.065	0.566	0.891	0.352	1.089
	PLS	0.533	0.924	0.533	0.924	0.749	0.678
	SVM	0.548	0.910	0.436	1.016	**0.768**	**0.652**
	KNN	**0.613**	**0.841**	**0.634**	**0.818**	0.627	0.826

Chlorophyll Content in Avocado

The reflectance spectra of each class can be visualized in Figure 5.24. The Ground-Truth vs Predictions graphs in Figure 5.25 show a that the regression line is decent, but optimal.

According to Table 5.11 the best model was SVM in the NIR range with a 0.491 R^2 score and a 6.314 RMSE value.

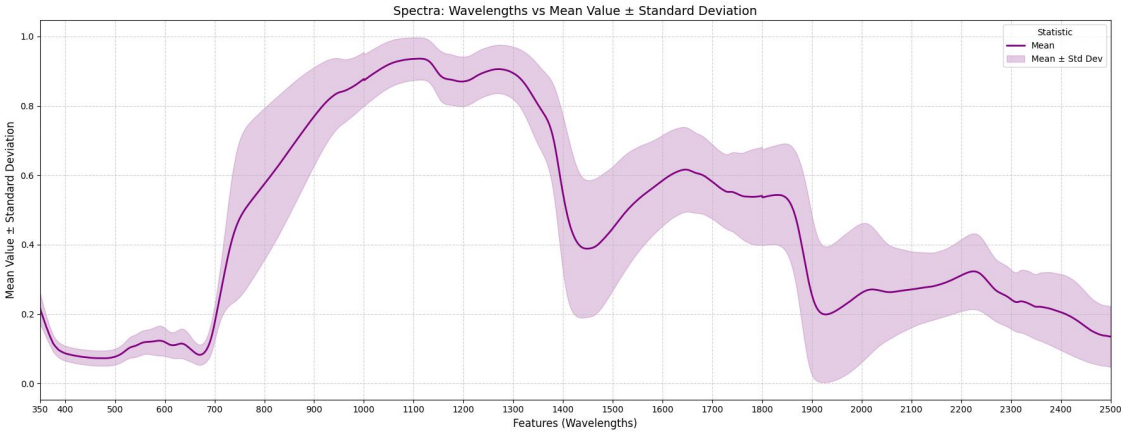
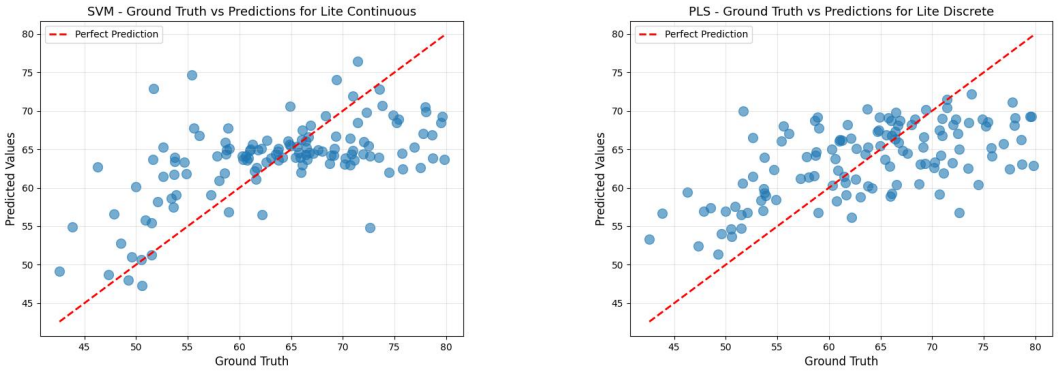


Figure 5.24: Wavelength Reflectance: Chlorophyll Content in Avocado.



(a) SVM Ground Truth vs Predictions for Lite Continuous Chlorophyll Content Avocado.

(b) PLS Ground Truth vs Predictions for Lite Discrete Water Content Avocado.

Figure 5.25: Ground Truth vs Predictions.

Table 5.11: Performance Metrics for Wavelength Chlorophyll Avocado Regressor.

Range	Method	No Optim. R^2score	No Optim. RMSE	Norm. R^2score	Norm. RMSE	GridSearch R^2score	GridSearch RMSE
VIS NIR	Ridge	0.362	7.066	0.360	7.079	0.384	6.946
	PLS	0.382	6.955	0.382	6.955	0.234	7.747
	SVM	**0.420**	**6.739**	0.345	7.163	0.375	6.999
	KNN	0.344	7.170	**0.407**	**6.813**	**0.452**	**6.549**
VIS	Ridge	0.350	7.135	**0.395**	**6.883**	0.366	7.047
	PLS	**0.358**	**7.093**	0.358	7.093	0.293	7.439
	SVM	0.293	7.439	0.254	7.642	**0.415**	**6.771**
	KNN	0.316	7.316	0.306	7.370	0.369	7.029
NIR	Ridge	0.373	7.010	0.373	7.008	0.378	6.977
	PLS	0.352	7.124	0.352	7.124	0.391	6.909
	SVM	0.391	6.908	0.352	7.125	**0.491**	**6.314**
	KNN	**0.461**	**6.499**	**0.457**	**6.524**	0.457	6.520
SWIR	Ridge	**0.185**	**7.987**	**0.201**	**7.909**	0.133	8.240
	PLS	0.023	8.747	0.023	8.747	0.260	7.612
	SVM	0.001	8.843	0.024	8.745	**0.295**	**7.429**
	KNN	0.113	8.335	0.062	8.571	0.160	8.109
RE	Ridge	0.400	6.856	**0.403**	**6.839**	0.389	6.919
	PLS	0.391	6.909	0.391	6.909	0.352	7.124
	SVM	**0.411**	**6.794**	0.366	7.047	0.386	6.935
	KNN	0.334	7.219	0.377	6.985	**0.418**	**6.749**
Lite C	Ridge	0.359	7.087	0.363	7.063	0.367	7.041
	PLS	0.351	7.128	0.351	7.128	0.184	7.993
	SVM	**0.421**	**6.735**	0.356	7.101	**0.412**	**6.787**
	KNN	0.374	7.001	**0.393**	**6.893**	0.405	6.827
Lite D	Ridge	0.284	7.489	**0.368**	**7.036**	0.278	7.520
	PLS	**0.351**	**7.132**	0.351	7.132	**0.385**	**6.939**
	SVM	0.322	7.287	0.100	8.397	0.366	7.045
	KNN	0.255	7.638	0.307	7.366	0.331	7.240

Nitrogen Content in Avocado

The reflectance spectra of each class can be visualized in Figure 5.26. The Ground-Truth vs Predictions graphs in Figure 5.27 show a that the regression line is decent, but optimal.

According to Table 5.12 the best model was SVM in the NIR range with a 0.491 R^2 score and a 6.314 RMSE value.

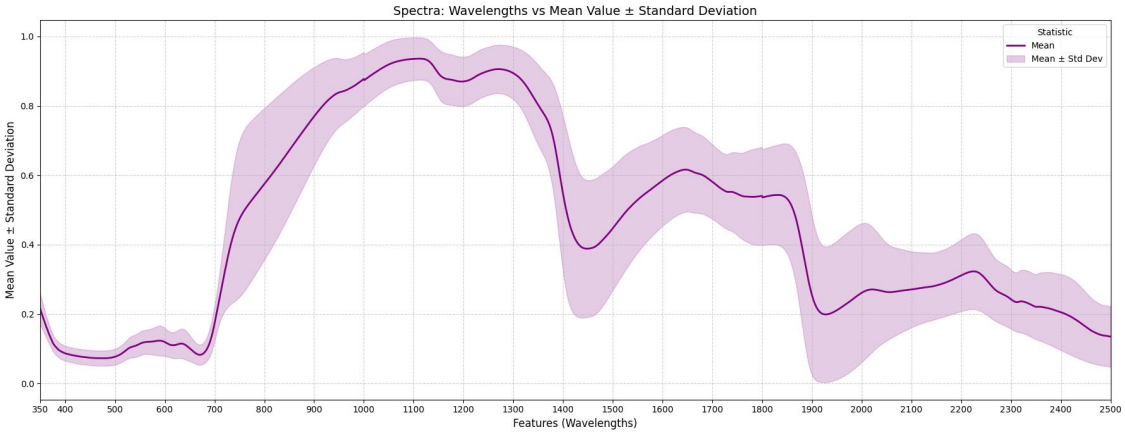
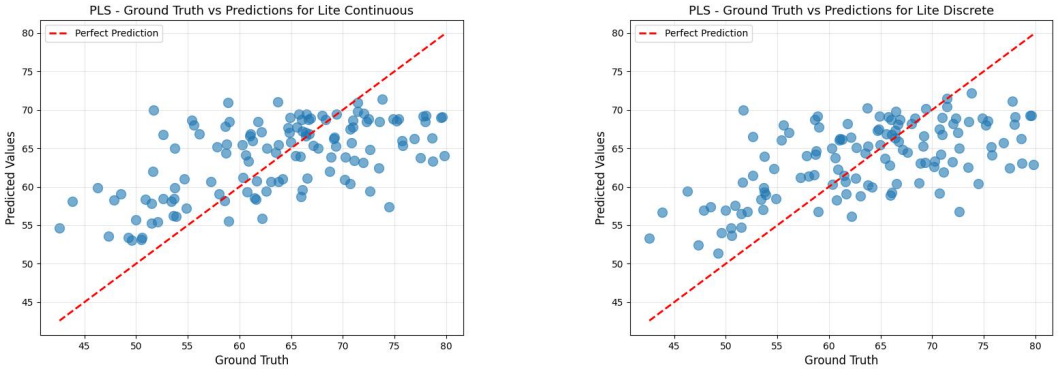


Figure 5.26: Wavelength Reflectance.



(a) PLS Ground Truth vs Predictions for Lite Continuous Water Content Avocado.

(b) PLS Ground Truth vs Predictions for Lite Discrete Water Content Avocado.

Figure 5.27: Ground Truth vs Predictions.

Table 5.12: Performance Metrics for Wavelength Nitrogen Avocado Regressor.

Range	Method	No Optim. R^2score	No Optim. RMSE	Norm. R^2score	Norm. RMSE	GridSearch R^2score	GridSearch RMSE
VIS NIR	Ridge	**0.276**	**3.404**	**0.300**	**3.347**	0.274	3.409
	PLS	0.243	3.480	0.243	3.480	**0.326**	**3.285**
	SVM	0.235	3.500	0.178	3.627	0.323	3.291
	KNN	0.265	3.430	0.278	3.400	0.256	3.451
VIS	Ridge	0.263	3.436	**0.289**	**3.374**	0.262	3.438
	PLS	0.267	3.425	0.267	3.425	**0.341**	**3.249**
	SVM	**0.300**	**3.347**	0.166	3.654	0.270	3.418
	KNN	0.260	3.441	0.275	3.407	0.275	3.407
NIR	Ridge	0.359	3.203	0.375	3.163	0.352	3.221
	PLS	0.327	3.282	0.327	3.282	**0.496**	**2.840**
	SVM	0.442	2.990	0.280	3.395	0.445	2.981
	KNN	**0.501**	**2.825**	**0.528**	**2.749**	0.474	2.901
SWIR	Ridge	**0.378**	**3.155**	**0.394**	**3.116**	0.337	3.257
	PLS	0.240	3.487	0.240	3.487	0.381	3.147
	SVM	0.319	3.301	0.327	3.283	**0.418**	**3.053**
	KNN	0.257	3.448	0.246	3.474	0.251	3.462
RE	Ridge	**0.254**	**3.457**	**0.257**	**3.449**	0.250	3.464
	PLS	0.249	3.468	0.249	3.468	0.248	3.469
	SVM	0.239	3.491	0.171	3.642	**0.290**	**3.371**
	KNN	0.163	3.661	0.174	3.638	0.234	3.503
Lite C	Ridge	0.312	3.319	**0.364**	**3.190**	0.288	3.376
	PLS	0.227	3.517	0.227	3.517	**0.408**	**3.078**
	SVM	0.244	3.479	0.192	3.597	0.370	3.176
	KNN	**0.326**	**3.285**	0.325	3.287	0.309	3.326
Lite D	Ridge	0.170	3.644	0.229	3.514	0.151	3.687
	PLS	0.220	3.533	0.220	3.533	**0.289**	**3.375**
	SVM	0.210	3.557	0.173	3.639	0.280	3.396
	KNN	**0.308**	**3.330**	**0.272**	**3.414**	0.271	3.417

5.1.4 Regressor Discussion

As in Classification Algorithms, there were better results for water content metric than in nitrogen and chlorophyll, showing a high R^2score value and low RMSE. This is also illustrated by the Ground Truth vs Predictions graphs where the samples are more aligned with the regression line in the water content that with nitrogen and chlorophyll. SVM was the algorithm with better

performance across most ranges for water content and chlorophyll content. Regarding nitrogen content, the PLS model achieve better prediction capabilities across most ranges.

5.1.5 Dataset Augmentation

In order to enhance the R^2-score of the chlorophyll and nitrogen predictions, the water content target has added to the avocado dataset.

The Table 5.13 shows an increment in the R^2-score and decrease in RMSE, which show better predictions were made.

Table 5.13: Performance Metrics for Wavelength Chlorophyll Avocado Augmentation.

Range	Method	No Optim. R^2score	No Optim. RMSE	Norm. R^2score	Norm. RMSE	GridSearch R^2score	GridSearch RMSE
VIS NIR	Ridge	**0.697**	**0.186**	**0.760**	**0.166**	0.690	0.188
	PLS	0.488	0.242	0.488	0.242	0.853	0.130
	SVM	0.635	0.204	0.570	0.222	**0.873**	**0.120**
	KNN	0.657	0.198	0.634	0.205	0.634	0.205
VIS	Ridge	0.646	0.201	**0.702**	**0.185**	0.643	0.202
	PLS	0.625	0.207	0.625	0.207	0.809	0.148
	SVM	**0.740**	**0.172**	0.427	0.256	**0.855**	**0.129**
	KNN	0.698	0.186	0.688	0.189	0.687	0.189
NIR	Ridge	0.882	0.116	**0.906**	**0.104**	0.864	0.125
	PLS	0.633	0.205	0.633	0.205	**0.947**	**0.078**
	SVM	**0.926**	**0.092**	0.665	0.196	0.934	0.087
	KNN	0.799	0.152	0.808	0.148	0.808	0.148
SWIR	Ridge	0.898	0.108	**0.911**	**0.101**	0.883	0.116
	PLS	0.831	0.139	0.831	0.139	0.931	0.089
	SVM	0.882	0.116	0.733	0.175	**0.944**	**0.080**
	KNN	**0.909**	**0.102**	0.902	0.106	0.897	0.108
RE	Ridge	0.534	0.231	**0.623**	**0.208**	0.527	0.232
	PLS	0.504	0.238	0.504	0.238	0.757	0.167
	SVM	0.552	0.226	0.426	0.256	**0.813**	**0.146**
	KNN	**0.584**	**0.218**	0.590	0.217	0.590	0.217
Lite C	Ridge	**0.792**	**0.154**	**0.871**	**0.122**	0.727	0.177
	PLS	0.612	0.211	0.612	0.211	0.925	0.093
	SVM	0.743	0.171	0.584	0.218	**0.933**	**0.088**
	KNN	0.661	0.197	0.661	0.197	0.652	0.199
Lite D	Ridge	0.380	0.266	0.566	0.223	0.352	0.272
	PLS	0.533	0.231	0.533	0.231	0.749	0.170
	SVM	0.597	0.215	0.493	0.241	**0.752**	**0.168**
	KNN	**0.613**	**0.210**	**0.634**	**0.204**	0.627	0.206

Since after optimization the results were high, it is possible to infer that the augmentation technique might be a solution to enhance the accuracy of the chlorophyll and nitrogen. However, when the dataset is unknown is not possible to use the target values as features. Hence, the added feature needs to be the water content predictions obtained with the estimator.

5.2 BudScan Results

The data configurations include all measurements (Complete), top measurements (Top), right measurements (Right), left measurements (Left), the complete measurements if they were made only using 9 bands (Lite Discrete), the measurements done in the laboratory (Lab) and the measurements done in the field (Field). Within each configuration, the best-performing model is highlighted in bold. Additionally, the overall best-performing method for each spectral range is also indicated in bold.

Making the evaluation for the best-performing models obtained using both the CropSignPro and CropSignLite (Lite Discrete) wavelength configurations, allows the evaluation of the possibility of using a sensor with discretely sampled wavelengths instead of one with approximately continuous spectral measurements.

Note: In the presented tables, "Lite D" corresponds to the Lite Discrete configuration.

In order to understand the sample results better, the samples mean and its mean standard deviation was computed. That is presented in Figure 5.28 together with the mean light and dark spectra acquired at the beginning of the experiment. Since the light spectrum exhibited a substantially higher peak intensity than the sample spectra, its values were divided by three to facilitate visualization and comparison. The dark spectrum displays small oscillations that are also present in the sample spectra. Moreover, the characteristic peak observed in the light spectrum is clearly reflected in all measured samples.

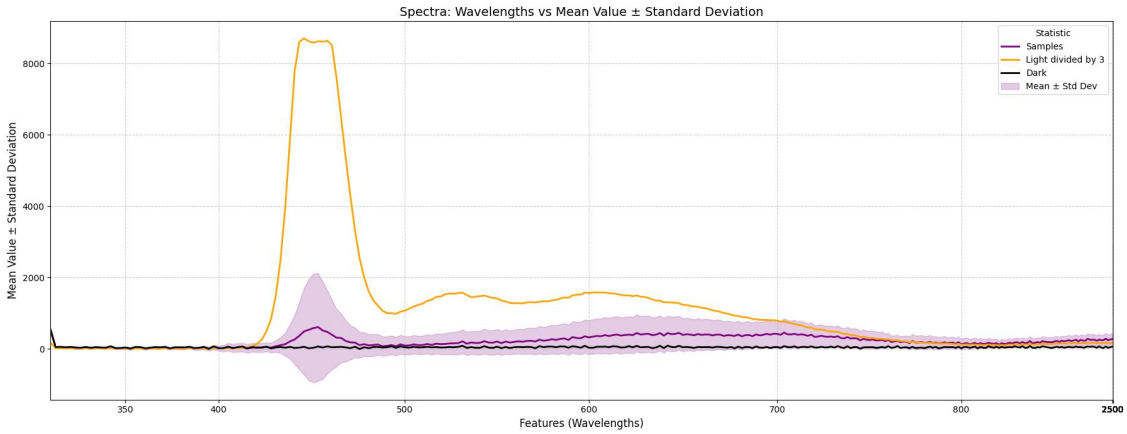


Figure 5.28: Light and Dark Mean Representations

5.2.1 Classifier Algorithm

The purpose of this experiment is to train a ML model using the collected dataset to predict the future fruitfulness of grapevine buds by classifying the number of inflorescences each bud is expected to develop.

The results obtained for each spectral range are presented in the performance metrics tables. The first two columns report the performance of the baseline models without optimization. The third and fourth columns present the results obtained after applying normalization, while the fifth

and sixth columns correspond to the normalized models further optimized through GridSearchCV. For each configuration, the first column reports the accuracy and the second column reports the macro-averaged F1-score.

5.2.1.1 Three Classes

In order to facilitate the results' interpretation, the complete wavelength spectra was plotted showing the mean of reflectance for each class. In Figure 5.29, the four classes represent increasing levels of the target variable, namely *Infertile*, *Fertile*, *Very Fertile*.

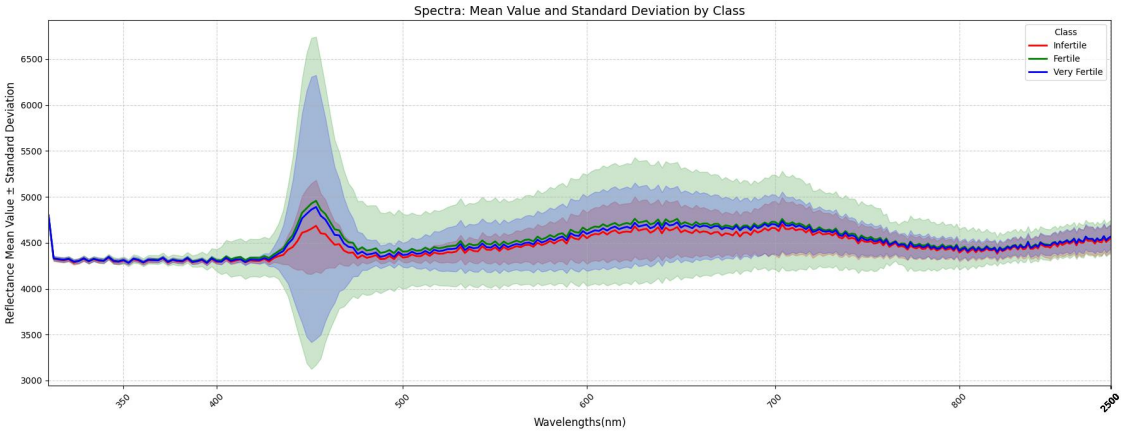


Figure 5.29: Spectra BudScan

In Table 5.14, the best result is the Field Configuration, using the Optimized SVM algorithm with 69% accuracy. This result was obtained with a C parameter of 1, gamma 1 and a RBF kernel. For different ranges the strongest models were LR and SVM. If a multispectral sensor such as Lite Discrete were used, the model would achieve an accuracy of approximately 64%, suggesting that this configuration similar performance to the Lite Continuous sensor used in the Complete configuration. The Tables 5.15 and 5.16 show the best hyperparameters using GridSearchCV BudScan.

Table 5.14: Performance Metrics for BudScan Three classes.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
Complete	LR	**0.51**	**0.33**	0.54	0.29	**0.65**	**0.29**
	SVM	0.47	0.33	**0.63**	**0.30**	0.61	0.35
	DT	0.48	0.37	0.48	0.37	0.47	0.40
	KNN	0.46	0.35	0.52	0.34	0.54	0.34
Top	LR	**0.58**	**0.39**	0.44	0.25	0.50	0.22
	SVM	0.58	0.37	0.52	0.28	0.54	0.23
	DT	**0.58**	**0.39**	**0.58**	**0.39**	**0.58**	**0.41**
	KNN	0.54	0.35	**0.58**	**0.39**	0.54	0.39
Right	LR	0.44	0.30	0.54	0.23	0.54	0.23
	SVM	0.44	0.37	0.52	0.23	0.54	0.23
	DT	**0.48**	**0.32**	0.48	0.32	0.50	0.34
	KNN	0.48	0.28	**0.54**	**0.36**	**0.60**	**0.36**
Left	LR	0.42	0.30	0.46	0.30	0.46	0.30
	SVM	0.42	0.29	**0.52**	**0.30**	**0.48**	**0.30**
	DT	0.38	0.26	0.38	0.26	0.42	0.34
	KNN	**0.42**	**0.36**	0.50	0.41	0.46	0.28
Lite D	LR	0.59	0.32	**0.64**	**0.28**	0.63	0.26
	SVM	**0.63**	**0.32**	0.63	0.27	**0.64**	**0.29**
	DT	0.51	0.40	0.51	0.40	0.51	0.40
	KNN	0.44	0.27	0.51	0.34	0.51	0.34
Lab	LR	0.46	0.34	0.48	0.29	**0.52**	**0.33**
	SVM	0.44	0.28	0.51	0.34	0.50	0.38
	DT	**0.50**	**0.44**	**0.51**	**0.44**	0.51	0.45
	KNN	0.40	0.37	0.48	0.40	0.43	0.29
Field	LR	0.53	0.31	0.64	0.37	0.62	0.26
	SVM	0.47	0.28	**0.65**	**0.33**	**0.69**	**0.37**
	DT	0.51	0.34	0.53	0.35	0.50	0.32
	KNN	**0.65**	**0.38**	**0.62**	**0.38**	0.61	0.36

Table 5.15: LR, DT and KNN Best Hyperparameters using GridSearchCV BudScan Three Classes.

Range	Method	Parameter	Value	Metric
Complete	LR	C	0.1	-
	DT	max depth	10	-
	k-NN	n neighbours	11	euclidean
Top	LR	C	0.1	-
	DT	max depth	None	-
	k-NN	n neighbours	1	euclidean
Right	LR	C	0.1	-
	DT	max depth	None	-
	k-NN	n neighbours	11	manhattan
Left	LR	C	1	-
	DT	max depth	None	-
	k-NN	n neighbours	9	manhattan
Lite Discrete	LR	C	0.1	-
	DT	max depth	None	-
	k-NN	n neighbours	9	manhattan
Lab	LR	C	0.1	-
	DT	max depth	None	-
	k-NN	n neighbours	7	euclidean
Field	LR	C	0.1	-
	DT	max depth	10	-
	k-NN	n neighbours	11	euclidean

Table 5.16: SVM Best Hyperparameters using GridSearchCV BudScan Three Classes.

Range	C	coef0	gamma	kernel
Complete	1	0	1	rbf
Top	0.1	0	1	rbf
Right	0.1	0	1	rbf
Left	10	0	0.1	rbf
Lite Discrete	1	5	1	poly
Lab	1	5	0.01	poly
Field	1	0	1	rbf

5.2.1.2 Two Classes

Since Fertile and Very Fertile have a similar reflectance signature, placing them in the same class may increase results accuracy. Not only that, but it also may facilitate the results' interpretation by the agricultor. For those reasons, the datasets were divided in two classes: Not fertile (0) and Fertile (1). Considering there is a class imbalance in the dataset, the code was also adjusted to try to mitigate class imbalance.

In Table 5.17, the best result is the Field Configuration, using the Optimized SVM algorithm with 97% accuracy. This result was obtained with a C parameter of 1, gamma 1 and a Radial basis function (RBF) kernel. For different ranges the strongest models were LR and SVM. If a multispectral sensor such as Lite Discrete were used, the model would achieve an accuracy of approximately 51%, suggesting that this configuration would have a worse performance. The Tables 5.18 and 5.19 show the best hyperparameters using GridSearchCV BudScan.

Table 5.17: Performance Metrics for BudScan Two Classes.

Range	Method	No Optim. Accuracy X	No Optim. F1-score (Macro Avg)	Norm. Accuracy $X_{normalised}$	Norm. F1-score (Macro Avg)	GridSearch Accuracy $X_{normalised}$	GridSearch F1-score (Macro Avg)
Complete	LR	0.89	0.52	**0.89**	**0.47**	0.87	0.51
	SVM	0.87	0.47	0.88	0.47	0.83	0.49
	DT	0.79	0.47	0.79	0.47	0.86	0.54
	KNN	**0.90**	**0.53**	**0.89**	**0.47**	**0.89**	**0.54**
Top	LR	0.87	0.46	0.87	0.46	0.87	0.46
	SVM	**0.88**	**0.47**	**0.88**	**0.47**	**0.88**	**0.47**
	DT	0.79	0.44	0.79	0.44	0.77	0.43
	KNN	0.87	0.46	0.87	0.46	0.87	0.46
Right	LR	0.85	0.46	**0.88**	**0.59**	**0.87**	**0.46**
	SVM	0.85	0.46	0.87	0.46	**0.87**	**0.46**
	DT	0.83	0.45	0.83	0.45	0.81	0.45
	KNN	**0.87**	**0.46**	0.87	0.46	**0.87**	**0.46**
Left	LR	0.87	0.46	0.87	0.46	0.87	0.46
	SVM	0.87	0.46	0.83	0.45	0.87	0.46
	DT	**0.79**	**0.52**	**0.79**	**0.52**	0.75	0.54
	KNN	0.87	0.46	0.87	0.46	**0.87**	**0.57**
Lite Discrete	LR	0.33	0.28	0.23	0.13	0.29	0.20
	SVM	0.38	0.32	0.28	0.21	0.40	0.29
	DT	**0.47**	**0.32**	0.47	0.32	**0.51**	**0.43**
	KNN	0.44	0.27	**0.51**	**0.34**	0.51	0.34
Lab	LR	0.80	0.45	**0.80**	**0.61**	**0.83**	**0.63**
	SVM	0.77	0.43	0.84	0.52	0.83	0.45
	DT	**0.82**	**0.59**	0.82	0.59	0.76	0.54
	KNN	0.87	0.46	0.83	0.45	0.85	0.46
Field	LR	**0.97**	**0.49**	0.93	0.48	0.88	0.47
	SVM	0.96	0.49	**0.97**	**0.49**	0.95	0.49
	DT	0.86	0.46	0.86	0.46	0.80	0.44
	KNN	**0.97**	**0.49**	**0.97**	**0.49**	**0.97**	**0.49**

Table 5.18: LR, DT and KNN Best Parameters using GridSearchCV BudScan Two Classes.

Range	Method	Parameter	Value	Metric
Complete	LR	C	1000	-
	DT	max depth	None	-
	k-NN	n neighbours	5	manhattan
Top	LR	C	1	-
	DT	max depth	None	-
	k-NN	n neighbours	5	euclidean
Right	LR	C	0.1	-
	DT	max depth	None	-
	k-NN	n neighbours	5	euclidean
Left	LR	C	0.1	-
	DT	max depth	None	-
	k-NN	n neighbours	3	euclidean
Lite Discrete	LR	C	100	-
	DT	max depth	None	-
	k-NN	n neighbours	9	manhattan
Lab	**LR**	**C**	**1**	-
	DT	max depth	None	-
	k-NN	n neighbours	9	euclidean
Field	LR	C	1000	-
	DT	max depth	None	-
	k-NN	n neighbours	5	euclidean

Table 5.19: SVM Best Parameters using GridSearchCV BudScan Two Classes.

Range	C	weights	coef0	gamma	kernel
Complete	1	0: 3, 1: 2, 2: 1	0	1	poly
Top	1	0: 3, 1: 2, 2: 1	0	0.1	poly
Right	0.1	0: 3, 1: 2, 2: 1	0	1	rbf
Left	0.1	0: 3, 1: 2, 2: 1	0	1	rbf
Lite Discrete	10	0: 5, 1: 2, 2: 1	5	1	poly
Lab	10	0: 3, 1: 2, 2: 1	0	0.1	rbf
Field	1	0: 4, 1: 2, 2: 1	5	0.01	poly

5.2.1.3 Classification BudScan Discussion

Environmental factors such as noise and light may have interfered with the measurements, complicating the classification task for the machine learning models and contributing to the sub-optimal performance. Furthermore, as detailed in the Methodology chapter, the dataset exhibited significant class imbalance. An initial evaluation using three classes produced moderate results but left room for improvement. When the problem was reduced to two classes, overall accuracy increased; however, the F1-score remained low, indicating the model still struggled to accurately identify the underrepresented class.

An initial review of the results might suggest that the *Field* dataset yielded significantly better performance, potentially implying that the field data is inherently superior or more accurate. However, this conclusion is misleading. The high accuracy is primarily an artifact of dataset imbalance, as there are significantly fewer *Non Fertile* buds present. This skew is particularly pronounced in this specific subset of data, which contains almost no instances with a target value of 0.

The change from four to two classes resulted in accuracy and F1-score enhance across all configurations, unless Lite Discrete. A higher F1-score for the minority class indicates that samples belonging to this class are being classified correctly more frequently, reflecting an improved balance between precision and recall.

5.2.2 One-Class Classification (OCC)

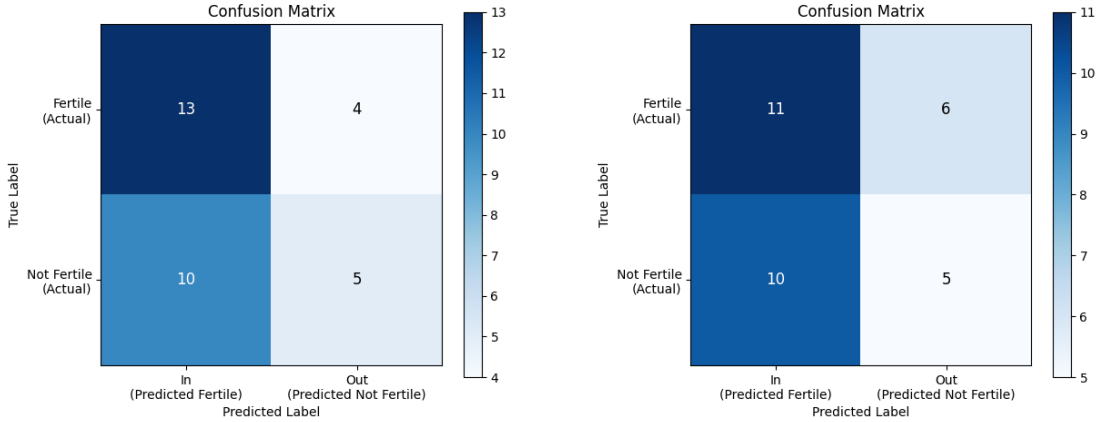
To investigate if an OCC model would be a good approach, it was decided to break the problem down into a smaller one. Therefore, one of the orientation measurements was selected. Since top measurements achieve a higher accuracy across-models, they were chosen to be the dataset for OCC models.

5.2.2.1 DD-SIMCA

The top measurement train dataset was imported and the pipeline in Figure 4.12 was followed. The preprocessing parameters used were:

- Savitzky-Golay (11, 1, 0)
- Trim tails (Left: 15 right: 288)

The following confusion matrices were plotted based on the anomaly detection report given after making predictions with the trained model (Figure 5.30).



(a) No preprocessing.

(b) With preprocessing.

Figure 5.30: DD-SIMCA BudScan Results.

DD-SIMCA model was trained in Fertile buds. If it doesn't classify a fertile bud as fertile, that doesn't mean the bud is infertile; it just means that, to the model, the spectral information of that bud doesn't match the spectral information it has learnt to associate with fertile buds.

5.2.2.2 One-Class SVM

Since one of the objectives of this study is to evaluate whether anomaly detection can be used to address the class imbalance present in the dataset, a One-Class SVM was also investigated. Unlike conventional multi-class classifiers, the One-Class SVM is trained exclusively on the majority class to learn its underlying distribution and identify samples that deviate from this pattern as anomalies. The confusion matrix in Figure 5.31 displays the results obtained.

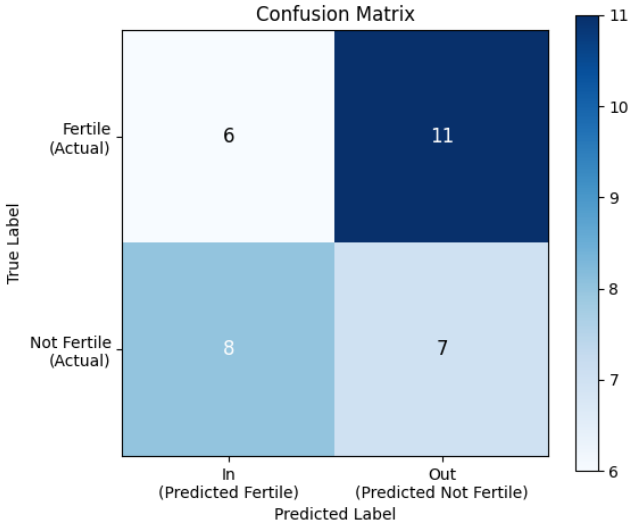


Figure 5.31: One-Class SVM.

5.2.3 One-Class Classification (OCC) Discussion

Because of the dataset imbalance, authentication models were applied for anomaly detection; however, the overall results were unsatisfactory. Interestingly, applying the DD-SIMCA model on raw data yielded better performance than using the preprocessed data. This outcome suggests that the smoothing effect of the Savitzky-Golay filter may have obscured critical features, causing the model to struggle. Furthermore, regarding the One-Class SVM, exploring alternative parameter configurations could potentially improve its anomaly detection capabilities.

Chapter 6

Conclusion

6.1 Conclusions

This dissertation contributes to the advancement of mobile-based proximal sensing in PA through the development and evaluation of ML models. To investigate which ML models are capable of making accurate predictions based on spectral analysis, two experiments were carried out: one using data collected by others and one using a spectral dataset acquired in Nelas comprising field and greenhouse data. The ML models were calibrated for a better monitoring of the plant physiological conditions.

Related to the Algorithmic Benchmarking Under Wavelength Constraints, spectral signatures vary by plant species and target metric. Across all datasets, water content was predicted with the highest accuracy. Reflectance spectra exhibited the highest distinctiveness, and subsequent classification performance, for olive, followed by avocado, and lastly vineyard. The protocol used in the dataset to collect the reflectance spectra may not be the most effective, since the SPAD measurements were taken in the veins of the leave, which is not recommended, as stated by Yuan *et al.* [71]. These might explain why the results of chlorophyll and nitrogen are not as good as the water ones. In terms of results, normalizing data may sometimes improve the model performance. LR and SVM consistently outperformed other models across all datasets after undergoing hyperparameter optimization via GridSearchCV. Optimal spectral ranges and key wavelengths are highly dependent on both the species and the target metric.

Regarding the BudScan experiment, it was not possible to access the part of the bud intended for measurement because of some tissue the grapevine forms around it for protection.. It can be concluded that the measurements with the tissue and without the tissue tend to be different. During the experiment, it was concluded that the first bud is not measurable in all canes, therefore it would be better to consider taking measurements from the second bud. Another conclusion is that the buds in the beginning of the cane are as fertile as the ones on top. In respect with the experimental setup, the probe is not as robust to the sunlight as it was assumed. Some measures seem to be affected by sunlight, while others do not. Therefore, the protocol should consider the sunlight incidence. It might be good to consider calibrating the spectra using measurements facing the sky

and facing the field, so the preprocessing can be easily done. Another thing to take in consideration is the integration time, which might have been too low, resulting in noise in the reflectance spectra.

Simplifying the classification task from a three-class system to a two-class system drastically improves model reliability, even though the Lite Discrete predictions do not follow this pattern. The use of authentication algorithms to solve the misclassification did not overcome the class imbalance challenge, since the alien and external were not always Not Fertile and some of the Not Fertile were accepted as "in". The one-class SVM also wasn't able to detect the correct anomalies. The results of the multispectral sensor were compared with a hyperspectral sensor across the datasets. Although the hyperspectral sensor is not as accurate as a multispectral sensor, it still could achieve a great accuracy. For some VI, the accuracy is similar. Therefore, it is possible to conclude that CropSign Lite is a good candidate for a low-cost plant monitoring solution. Since not all models reached higher levels of accuracy, f1-score and r2-score. It might be good to save the data collected, so when better models are discovered, the previous data is saved and can be used for training and testing.

Since it is possible to classify the plant conditions, it is possible to just show to the user the prediction made and a suggestion of how to handle the problem, therefore it is possible to presents results in an intuitive manner without exposing users to unnecessary data complexity.

The main contributions of this work are algorithm benchmarking, organization datasets, new protocol considerations when creating a bud grapevine dataset, validation of a cheaper and simplified multispectral sensor.

These findings provide a path to achieve a low-cost solution for in-field crop monitoring. An important contribution for data-driven agronomic decision-making.

6.2 Future Work

Regarding BudScan dataset acquisition, it would be worthwhile trying different ways to calibrate the light in dataset acquisition and test different integration time values to solve the measurement noise. A compelling direction for future research would be to implement data augmentation with the predictions of dependent features, in this case, use water predictions as a feature for chlorophyll and nitrogen content predictions. Another promising solution would be generating synthetic data for the minority classes to address class imbalance.

In terms of machine learning model prediction enhancement, it would be interesting to test this models in more datasets to increase robustness and reduce the risk of overfitting due to dataset imbalance. It would also be valuable to test FL models, since it allows models to be trained across multiple edge devices or servers while keeping data localized [52].

References

[1] S. Laveglia, G. Altieri, F. Genovese, A. Matera, and G. C. Di Renzo, "Advances in Sustainable Crop Management: Integrating Precision Agriculture and Proximal Sensing," en, *AgriEngineering*, vol. 6, no. 3, pp. 3084–3120, Sep. 2024, ISSN: 2624-7402. DOI: 10.3390/agriengineering6030177. Accessed: Oct. 18, 2025. [Online]. Available: https://www.mdpi.com/2624-7402/6/3/177.

[2] S. Savas and S. M. T. Gharibzahedi, "Smartphone-Integrated Electrochemical Devices for Contaminant Monitoring in Agriculture and Food: A Review," en, *Biosensors*, vol. 15, no. 9, p. 574, Sep. 2025, ISSN: 2079-6374. DOI: 10.3390/bios15090574. Accessed: Oct. 29, 2025. [Online]. Available: https://www.mdpi.com/2079-6374/15/9/574.

[3] L. Zhu, H. Gong, Z. Song, H. Liu, Z. Chen, and Y. Lu, "Smartphone-based multispectral imaging for portable spectral analysis," *Measurement*, vol. 257, p. 118 643, Jan. 2026, ISSN: 0263-2241. DOI: 10.1016/j.measurement.2025.118643. Accessed: Oct. 24, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224125020020.

[4] M. A. R. Pereira, "Early diagnosis of bacterial plant diseases based on proximal sensing from a precision agriculture perspective," eng, Mar. 2024, Accepted: 2024-05-03T23:09:51Z. Accessed: Oct. 29, 2025. [Online]. Available: https://repositorio-aberto.up.pt/handle/10216/158404.

[5] A. Soussi, E. Zero, R. Sacile, D. Trinchero, and M. Fossa, "Smart Sensors and Smart Data for Precision Agriculture: A Review," *Sensors (Basel, Switzerland)*, vol. 24, no. 8, p. 2647, Apr. 2024, ISSN: 1424-8220. DOI: 10.3390/s24082647. Accessed: Oct. 24, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11053448/.

[6] INESCTEC, *TEC4AGRO-FOOD - Innovation | INESC TEC*, en. Accessed: Jan. 10, 2026. [Online]. Available: https://www.inesctec.pt/en/innovation/tec4agro-food.

[7] E. Mamabolo *et al.*, "Application of precision agriculture technologies for crop protection and soil health," *Smart Agricultural Technology*, vol. 12, p. 101 270, Dec. 2025, ISSN: 2772-3755. DOI: 10.1016/j.atech.2025.101270. Accessed: Jan. 10, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772375525005015.

[8] K. M. Gold *et al.*, "Hyperspectral Measurements Enable Pre-Symptomatic Detection and Differentiation of Contrasting Physiological Effects of Late Blight and Early Blight in Potato," en, *Remote Sensing*, vol. 12, no. 2, p. 286, Jan. 2020, ISSN: 2072-4292. DOI: 10.3390/rs12020286. Accessed: Jan. 9, 2026. [Online]. Available: https://www.mdpi.com/2072-4292/12/2/286.

[9] S. A. D. M. Zahir, A. F. Omar, M. F. Jamlos, M. A. M. Azmi, and J. Muncan, "A review of visible and near-infrared (Vis-NIR) spectroscopy application in plant stress detection," *Sensors and Actuators A: Physical*, vol. 338, p. 113 468, May 2022, ISSN: 0924-4247. DOI: 10.1016/j.sna.2022.113468. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924424722001066.

[10] X. Du, J. Wang, D. Dong, and X. Zhao, "Development and Testing of a Portable Soil Nitrogen Detector Based on Near-infrared Spectroscopy," in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, May 2019, pp. 822–826. DOI: 10.1109/ITAIC.2019.8785499. Accessed: Oct. 24, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/8785499.

[11] L. Li *et al.*, "A strategy to improve the detection accuracy and universality of highland barley muti-quality attributes based on near infrared spectroscopy combined with model transfer method," *Food Chemistry*, vol. 480, p. 143 887, Jul. 2025, ISSN: 0308-8146. DOI: 10.1016/j.foodchem.2025.143887. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308814625011380.

[12] Rashmi and D. S. Shantala, "Evaluating Deep Learning with different feature scaling techniques for EEG-based Music Entrainment Brain Computer Interface," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 7, p. 100 448, Mar. 2024, ISSN: 2772-6711. DOI: 10.1016/j.prime.2024.100448. Accessed: Jun. 6, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772671124000305.

[13] T. Fahey *et al.*, "Active and Passive Electro-Optical Sensors for Health Assessment in Food Crops," *Sensors (Basel, Switzerland)*, vol. 21, no. 1, p. 171, Dec. 2020, ISSN: 1424-8220. DOI: 10.3390/s21010171. Accessed: Oct. 21, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7795220/.

[14] Z. Guo *et al.*, "Detection model transfer of apple soluble solids content based on NIR spectroscopy and deep learning," *Computers and Electronics in Agriculture*, vol. 212, p. 108 127, Sep. 2023, ISSN: 0168-1699. DOI: 10.1016/j.compag.2023.108127. Accessed: May 29, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016816992300515X.

[15] K. Olowe, N. Edoh, S. Zouo, and J. Olamijuwon, "Comprehensive review of logistic regression techniques in predicting health outcomes and trends," *World Journal of Advanced Pharmaceutical and Life Sciences*, vol. 7, pp. 016–026, Nov. 2024. DOI: 10.53346/wjapls.2024.7.2.0039.

[16] S. Guo, C. Jiang, S. Zhou, D. Wu, and X. Liu, "Power Oscillation Source Location Based on the Combination of Energy Function and Logistic Regression in a Fully Data-Driven Approach," en, *IET Generation, Transmission & Distribution*, vol. 19, no. 1, e70155, 2025, _eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/gtd2.70155, ISSN: 1751-8695. DOI: 10.1049/gtd2.70155. Accessed: Jun. 6, 2026. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1049/gtd2.70155.

[17] R. Anand *et al.*, "A multimodal approach for enhanced disease management in cauliflower crops: Integration of spectral sensors, machine learning models and targeted spraying technology," en, *ENGINEERING Agriculture*, vol. 12, no. 2, p. 261, Jun. 2025, ISSN: 2097-7654. DOI: 10.15302/J-FASE-2024572. Accessed: Jun. 5, 2026. [Online]. Available: https://journal.hep.com.cn/fase/EN/10.15302/J-FASE-2024572.

[18] T. Waheed, R. B. Bonnell, S. O. Prasher, and E. Paulet, "Measuring performance in precision agriculture: CART—A decision tree approach," *Agricultural Water Management*, vol. 84, no. 1, pp. 173–185, Jul. 2006, ISSN: 0378-3774. DOI: 10.1016/j.agwat.2005.12.003. Accessed: Jun. 6, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378377405003860.

[19] D. Qiao *et al.*, "Non-Destructive Monitoring of Peanut Leaf Area Index by Combing UAV Spectral and Textural Characteristics," *Remote Sensing*, vol. 16, no. 12, 2024, ISSN: 2072-4292. DOI: 10.3390/rs16122182. [Online]. Available: https://www.mdpi.com/2072-4292/16/12/2182.

[20] S. Arya *et al.*, "High-throughput chlorophyll fluorescence image-based phenotyping for water deficit stress tolerance in wheat," en, *Plant Physiology Reports*, vol. 29, no. 2, pp. 278–293, Jun. 2024, ISSN: 2662-2548. DOI: 10.1007/s40502-024-00783-7. Accessed: Jan. 9, 2026. [Online]. Available: https://doi.org/10.1007/s40502-024-00783-7.

[21] Widyaningrum, Y. A. Purwanto, S. Widodo, Supijatno, and E. S. Iriani, "Rapid assessment of vanilla (*Vanilla planifolia*) quality parameters using portable near-infrared spectroscopy combined with random forest," *Journal of Food Composition and Analysis*, vol. 133, p. 106 346, Sep. 2024, ISSN: 0889-1575. DOI: 10.1016/j.jfca.2024.106346. Accessed: Oct. 24, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0889157524003806.

[22] B. G. Reta, "Evaluating the accuracy of machine learning models in predicting soil moisture, actual evapotranspiration, and biomass: Developing normalized difference vegetation index-adjusted crop coefficient regression model for watermelon," *Journal of Agriculture and Food Research*, vol. 24, p. 102 334, Dec. 2025, ISSN: 2666-1543. DOI: 10.1016/j.jafr.2025.102334. Accessed: May 19, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666154325007057.

[23] L. Bai *et al.*, "Spectroscopic techniques combined with DD-SIMCA model and explainable artificial intelligence for rapidly and accurately identifying the quality of Astragali Radix," *Industrial Crops and Products*, vol. 231, p. 121 140, Sep. 2025, ISSN: 0926-6690. DOI: 10.1016/j.indcrop.2025.121140. Accessed: Jul. 31, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926669025006867.

[24] C. Wu, Z. Niu, Q. Tang, and W. Huang, "Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation," *Agricultural and Forest Meteorology*, vol. 148, no. 8, pp. 1230–1241, Jul. 2008, ISSN: 0168-1923. DOI: 10.1016/j.agrformet.2008.03.005. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168192308000920.

[25] M. Chen *et al.*, "Hyperspectral characteristics and chlorophyll content inversion of *Japonica* rice under elevated CO2 and temperature with different nitrogen levels," *European Journal of Agronomy*, vol. 172, p. 127 872, Jan. 2026, ISSN: 1161-0301. DOI: 10.1016/j.eja.2025.127872. Accessed: May 29, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1161030125003685.

[26] A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Boldú, "A review on the practice of big data analysis in agriculture," *Computers and Electronics in Agriculture*, vol. 143, pp. 23–37, Dec. 2017, ISSN: 0168-1699. DOI: 10.1016/j.compag.2017.09.037. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169917301230.

[27] M. Gallardo, A. Elia, and R. B. Thompson, "Decision support systems and models for aiding irrigation and nutrient management of vegetable crops," *Agricultural Water Management*, vol. 240, p. 106 209, Oct. 2020, ISSN: 0378-3774. DOI: 10.1016/j.agwat.2020.106209. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378377420303267.

[28] S. O. Araújo, R. Silva Peres, L. Bischof Pian, F. Lidon, J. Cochicho Ramalho, and J. Barata, "Smart Agricultural System Using Proximal Sensing, Artificial Intelligence, and LoRa Technology: A Case Study in Vineyard Management," en, *IEEE Access*, vol. 12, pp. 181 052–181 070, 2024, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3482179. Accessed: Jan. 8, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/10719997/.

[29] K. Vargas, "Developing Spectral Metrics as Early Indicators of Water Stress Detection at the Canopy Level," *Master's Theses and Capstones*, Sep. 2018. [Online]. Available: https://scholars.unh.edu/thesis/1215.

[30] A. Yacoob, S. Gokool, A. Clulow, M. Mahomed, V. Naiken, and T. Mabhaudhi, "A machine learning approach for quantifying crop water stress in smallholder farms using unmanned aerial vehicle multispectral imagery," *Agricultural Water Management*, vol. 324, p. 110 142, Mar. 2026, ISSN: 0378-3774. DOI: 10.1016/j.agwat.2026.110142. Accessed: Jun. 9, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378377426000235.

[31] U. Debangshi and A. Sadhukhan, "Crop water stress monitoring through precision technologies: A review," *The Pharma Innovation*, vol. 12, pp. 803–808, Nov. 2023. DOI: 10.22271/tpi.2023.v12.i11Sk.24052.

[32] A. Merot and N. Smits, "Explaining yield dynamics during vineyard conversion to organic farming," *European Journal of Agronomy*, vol. 153, p. 127 068, Feb. 2024, ISSN: 1161-0301. DOI: 10.1016/j.eja.2023.127068. Accessed: Jun. 10, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1161030123003362.

[33] F. Rançon *et al.*, "Designing a Proximal Sensing Camera Acquisition System for Vineyard Applications: Results and Feedback on 8 Years of Experiments," *Sensors (Basel, Switzerland)*, vol. 23, no. 2, p. 847, Jan. 2023, ISSN: 1424-8220. DOI: 10.3390/s23020847. Accessed: Oct. 24, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9865571/.

[34] P. Horstrand *et al.*, "A spectral imaging system for precision agriculture: From its inception till a pre-commercial prototype," in *2019 XXXIV Conference on Design of Circuits and Integrated Systems (DCIS)*, Nov. 2019, pp. 1–6. DOI: 10.1109/DCIS201949030.2019.8959891. Accessed: Oct. 24, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/8959891.

[35] H. Tømmervik *et al.*, "The northernmost hyperspectral FLoX sensor dataset for monitoring of high-Arctic tundra vegetation phenology and Sun-Induced Fluorescence (SIF)," *Data in Brief*, vol. 50, p. 109 581, Oct. 2023, ISSN: 2352-3409. DOI: 10.1016/j.dib.2023.109581. Accessed: May 29, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340923006807.

[36] R. R. Vennam, R. Bheemanahalli, K. R. Reddy, J. Dhillon, X. Zhang, and A. Adeli, "Early-season maize responses to salt stress: Morpho-physiological, leaf reflectance, and mineral composition," *Journal of Agriculture and Food Research*, vol. 15, p. 100 994, Mar. 2024, ISSN: 2666-1543. DOI: 10.1016/j.jafr.2024.100994. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666154324000310.

[37] M. Diago, A. Bellincontro, M. Scheidweiler, J. Tardaguila, S. Tittmann, and M. Stoll, "Future opportunities of proximal near infrared spectroscopy approaches to determine the variability of vineyard water status," en, *Australian Journal of Grape and Wine Research*, vol. 23, no. 3, pp. 409–414, 2017, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajgw.12283, ISSN: 1755-0238. DOI: 10.1111/ajgw.12283. Accessed: Oct. 24, 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajgw.12283.

[38] Y. Wang *et al.*, "Design and implementation of a portable snapshot multispectral imaging crop-growth sensor," *Frontiers in Plant Science*, vol. 15, p. 1 416 221, Aug. 2024, ISSN: 1664-462X. DOI: 10.3389/fpls.2024.1416221. Accessed: Oct. 18, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11381302/.

[39] A. Heiß, D. S. Paraforos, G. M. Sharipov, and H. W. Griepentrog, "Modeling and simulation of a multi-parametric fuzzy expert system for variable rate nitrogen application," *Computers and Electronics in Agriculture*, vol. 182, p. 106 008, Mar. 2021, ISSN: 0168-1699. DOI: 10.1016/j.compag.2021.106008. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169921000260.

[40] M. R. Maleki, A. M. Mouazen, B. De Ketelaere, H. Ramon, and J. De Baerdemaeker, "On-the-go variable-rate phosphorus fertilisation based on a visible and near-infrared soil sensor," *Biosystems Engineering*, vol. 99, no. 1, pp. 35–46, Jan. 2008, ISSN: 1537-5110. DOI: 10.1016/j.biosystemseng.2007.09.007. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1537511007002292.

[41] B. Rey, N. Aleixos, S. Cubero, and J. Blasco, "Xf-Rovim. A Field Robot to Detect Olive Trees Infected by Xylella Fastidiosa Using Proximal Sensing," en, *Remote Sensing*, vol. 11, no. 3, p. 221, Jan. 2019, ISSN: 2072-4292. DOI: 10.3390/rs11030221. Accessed: Jan. 8, 2026. [Online]. Available: https://www.mdpi.com/2072-4292/11/3/221.

[42] N. Schor, S. Berman, A. Dombrovsky, Y. Elad, T. Ignat, and A. Bechar, "Development of a robotic detection system for greenhouse pepper plant diseases," en, *Precision Agriculture*, vol. 18, no. 3, pp. 394–409, Jun. 2017, ISSN: 1573-1618. DOI: 10.1007/s11119-017-9503-z. Accessed: Jan. 8, 2026. [Online]. Available: https://doi.org/10.1007/s11119-017-9503-z.

[43] C. Valero *et al.*, "Single Plant Fertilization Using a Robotic Platform in an Organic Cropping Environment," en, *Agronomy*, vol. 12, no. 6, p. 1339, Jun. 2022, ISSN: 2073-4395. DOI: 10.3390/agronomy12061339. Accessed: Jan. 8, 2026. [Online]. Available: https://www.mdpi.com/2073-4395/12/6/1339.

[44] W. Woodgate, E. van Gorsel, D. Hughes, L. Suarez, J. Jimenez-Berni, and A. Held, "THEMS: An automated thermal and hyperspectral proximal sensing system for canopy reflectance, radiance and temperature," en, *Plant Methods*, vol. 16, no. 1, p. 105, Jul. 2020, ISSN: 1746-4811. DOI: 10.1186/s13007-020-00646-w. Accessed: May 29, 2026. [Online]. Available: https://doi.org/10.1186/s13007-020-00646-w.

[45] X. Zhao, J. Zhang, Y. Huang, Y. Tian, and L. Yuan, "Detection and discrimination of disease and insect stress of tea plants using hyperspectral imaging combined with wavelet analysis," *Computers and Electronics in Agriculture*, vol. 193, p. 106 717, Feb. 2022, ISSN: 0168-1699. DOI: 10.1016/j.compag.2022.106717. Accessed: Jan. 9, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169922000345.

[46] A. B. González-Fernández, J. R. Rodríguez-Pérez, V. Marcelo, and J. B. Valenciano, "Using field spectrometry and a plant probe accessory to determine leaf water content in commercial vineyards," *Agricultural Water Management*, vol. 156, pp. 43–50, Jul. 2015, ISSN: 0378-3774. DOI: 10.1016/j.agwat.2015.03.024. Accessed: Jun. 3, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378377415001079.

[47] Y. Hadiwijaya and I. E. Putri, "Spectroscopy in food and agriculture: A critical review of applications and adoption challenges," *Food and Humanity*, vol. 5, p. 100 800, Dec. 2025, ISSN: 2949-8244. DOI: 10.1016/j.foohum.2025.100800. Accessed: Oct. 24, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949824425003040.

[48] S. Chung, L. E. Breshears, and J.-Y. Yoon, "Smartphone near infrared monitoring of plant stress," *Computers and Electronics in Agriculture*, vol. 154, pp. 93–98, Nov. 2018, ISSN: 0168-1699. DOI: 10.1016/j.compag.2018.08.046. Accessed: Oct. 21, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169918305519.

[49] S. K. Gupta, B. Efrati, O. Amir, N. Francos, M. Sternberg, and E. Ben-Dor, "Smartphone-based spectroscopy as a tool to estimate soil attributes for the citizen science concept," *Smart Agricultural Technology*, vol. 5, p. 100 327, Oct. 2023, ISSN: 2772-3755. DOI: 10.1016/j.atech.2023.100327. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772375523001569.

[50] X. Wang *et al.*, "Machine learning-based in-season nitrogen status diagnosis and side-dress nitrogen recommendation for corn," *European Journal of Agronomy*, vol. 123, p. 126 193, Feb. 2021, ISSN: 1161-0301. DOI: 10.1016/j.eja.2020.126193. Accessed: Jan. 9, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1161030120302008.

[51] D. Scutelnic, R. Muradore, and C. Daffara, "Multispectral Imaging Supervised by Optical Spectrometry for Close Acquisition in Precision Agriculture," in *2024 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*, Oct. 2024, pp. 598–602. DOI: 10.1109/MetroAgriFor63043.2024.10948867. Accessed: Oct. 24, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10948867/.

[52] M. N. Mowla, N. Mowla, A. F. M. S. Shah, K. M. Rabie, and T. Shongwe, "Internet of Things and Wireless Sensor Networks for Smart Agriculture Applications: A Survey," *IEEE Access*, vol. 11, pp. 145 813–145 852, 2023, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3346299. Accessed: Oct. 24, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10371307/.

[53] G. H. Mohammed *et al.*, "Remote sensing of solar-induced chlorophyll fluorescence (SIF) in vegetation: 50years of progress," *Remote Sensing of Environment*, vol. 231, p. 111 177, Sep. 2019, ISSN: 0034-4257. DOI: 10.1016/j.rse.2019.04.030. Accessed: Oct. 21, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425719301816.

[54] V. Sharma, E. Honkavaara, M. Hayden, and S. Kant, "UAV remote sensing phenotyping of wheat collection for response to water stress and yield prediction using machine learning," *Plant Stress*, vol. 12, p. 100 464, Jun. 2024, ISSN: 2667-064X. DOI: 10.1016/j.stress.2024.100464. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667064X24001180.

[55] P. Goyal, R. Sharda, M. Saini, and M. Siag, "A deep learning approach for early detection of drought stress in maize using proximal scale digital images," en, *Neural Computing and Applications*, vol. 36, no. 4, pp. 1899–1913, Feb. 2024, ISSN: 1433-3058. DOI: 10.1007/s00521-023-09219-z. Accessed: Jan. 9, 2026. [Online]. Available: https://doi.org/10.1007/s00521-023-09219-z.

[56] A. Navarro *et al.*, "Sorting biotic and abiotic stresses on wild rocket by leaf-image hyperspectral data mining with an artificial intelligence model," en, *Plant Methods*, vol. 18, no. 1, p. 45, Apr. 2022, ISSN: 1746-4811. DOI: 10.1186/s13007-022-00880-4. Accessed: Jan. 9, 2026. [Online]. Available: https://doi.org/10.1186/s13007-022-00880-4.

[57] D. Moshou *et al.*, "Simultaneous identification of plant stresses and diseases in arable crops using proximal optical sensing and self-organising maps," en, *Precision Agriculture*, vol. 7, no. 3, pp. 149–164, Jul. 2006, ISSN: 1573-1618. DOI: 10.1007/s11119-006-9002-0. Accessed: Jan. 9, 2026. [Online]. Available: https://doi.org/10.1007/s11119-006-9002-0.

[58] M. Aggarwal, V. Khullar, N. Goyal, A. Alammari, M. Albahar, and A. Singh, "Lightweight Federated Learning for Rice Leaf Disease Classification Using Non Independent and Identically Distributed Images," *Sustainability*, vol. 15, p. 12 149, Aug. 2023. DOI: 10.3390/su151612149.

[59] O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, K.-K. R. Choo, and M. Nafaa, "FELIDS: Federated learning-based intrusion detection system for agricultural Internet of Things," *Journal of Parallel and Distributed Computing*, vol. 165, pp. 17–31, Jul. 2022, ISSN: 0743-7315. DOI: 10.1016/j.jpdc.2022.03.003. Accessed: Jan. 8, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0743731522000570.

[60] Y. Wang *et al.*, *A Comprehensive Review of Spectral Sensing Devices and Diagnostic Equipment for Field Crop Growth Monitoring*, en, 2023. DOI: 10.2139/ssrn.4635931. Accessed: Oct. 24, 2025. [Online]. Available: https://www.ssrn.com/abstract=4635931.

[61] A. N. S. Cadeado, C. C. S. Machado, and S. G. Silva, "Simultaneous reflectance and fluorescence measurements for portable formaldehyde determination in milk using a multichannel spectrometer sensor," *Food Chemistry*, vol. 464, p. 141 583, Feb. 2025, ISSN: 0308-8146. DOI: 10.1016/j.foodchem.2024.141583. Accessed: May 29, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308814624032333.

[62] ams OSRAM AG, *AS7341 11-Channel Multi-Spectral Digital Sensor*. Premstaetten, Austria: ams OSRAM AG, Jun. 2020. Accessed: Jun. 3, 2026. [Online]. Available: https://look.ams-osram.com/m/24266a3e584de4db/original/AS7341-DS000504.pdf.

[63] K. Laganovska *et al.*, "Portable low-cost open-source wireless spectrophotometer for fast and reliable measurements," *HardwareX*, vol. 7, e00108, Apr. 2020, ISSN: 2468-0672. DOI: 10.1016/j.ohx.2020.e00108. Accessed: May 29, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S246806722030016X.

[64] Hamamatsu Photonics, *Mini-spectrometers Micro Series C12880MA, C16767MA, C16767MA-01*, Published: Datasheet, Cat. No. KACC1226E25, Feb. 2026. [Online]. Available: https://www.hamamatsu.com.

[65] G. Tunens *et al.*, "Optical fiber-based open source low cost portable spectrometer system," *HardwareX*, vol. 18, e00530, Jun. 2024, ISSN: 2468-0672. DOI: 10.1016/j.ohx.2024.e00530. Accessed: May 29, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468067224000245.

[66] J. S. Estrada, R. Demarco, C. M. Johnson, M. Zañartu, A. Fuentes, and F. Auat Cheein, "A multi-spectral and hyperspectral image dataset for evaluating chemical traits and the water status of avocado, olive and grape through leaf dehydration under laboratory conditions," en, *Scientific Reports*, vol. 15, no. 1, p. 2973, Jan. 2025, ISSN: 2045-2322. DOI: 10.1038/s41598-025-85714-8. Accessed: Jul. 20, 2026. [Online]. Available: https://www.nature.com/articles/s41598-025-85714-8.

[67] E. Bridges, N. Batjes, and F. Nachtergaele, *World Reference Base for Soil Resources: Atlas*. Jan. 1998.

[68] P. Rodrigues, V. Pedroso, J. P. Gouveia, S. Martins, C. Lopes, and I. Alves, "Influence of soil water content and atmospheric conditions on leaf water potential in cv. "Touriga Nacional" deep-rooted vineyards," en, *Irrigation Science*, vol. 30, no. 5, pp. 407–417, Sep. 2012, ISSN: 1432-1319. DOI: 10.1007/s00271-012-0350-4. Accessed: May 29, 2026. [Online]. Available: https://doi.org/10.1007/s00271-012-0350-4.

[69] M. Santos, A. Fonseca, H. Fraga, G. V. Jones, and J. A. Santos, "Bioclimatic conditions of the Portuguese wine denominations of origin under changing climates," en, *International Journal of Climatology*, vol. 40, no. 2, pp. 927–941, 2020, _eprint: https://rmets.onlinelibrary.wiley.com/doi/ ISSN: 1097-0088. DOI: 10.1002/joc.6248. Accessed: May 29, 2026. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/joc.6248.

[70] S. Kucheryavskiy, O. Rodionova, and A. Pomerantsev, "A comprehensive tutorial on Data-Driven SIMCA: Theory and implementation in web," en, *Journal of Chemometrics*, vol. 38, no. 7, e3556, 2024, _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cem.3 ISSN: 1099-128X. DOI: 10.1002/cem.3556. Accessed: Jun. 27, 2026. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.3556.

[71] Z. Yuan *et al.*, "Optimal Leaf Positions for SPAD Meter Measurement in Rice," English, *Frontiers in Plant Science*, vol. 7, May 2016, ISSN: 1664-462X. DOI: 10.3389/fpls.2016.00719. Accessed: Jun. 9, 2026. [Online]. Available: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2016.00719/full.

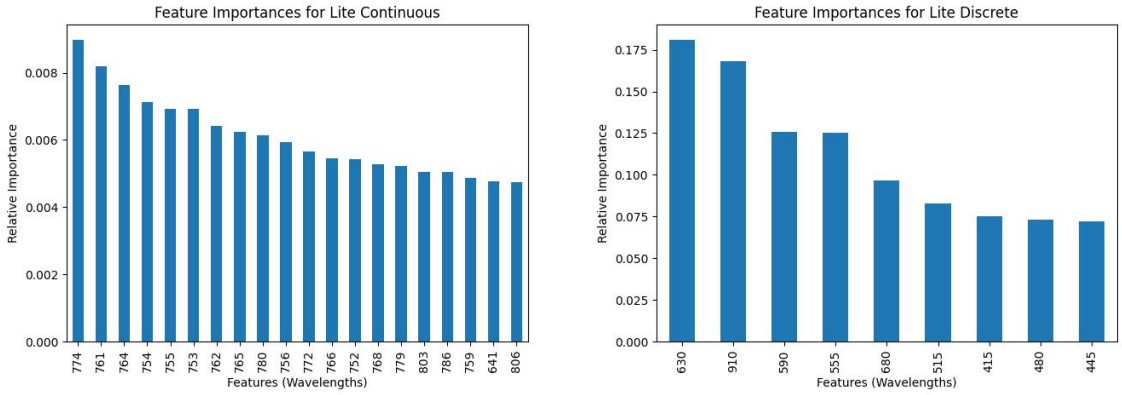
Appendix A

Algorithmic Benchmarking Under Wavelength Constraints

This appendix presents supplementary findings and detailed results about that support the primary analysis presented in Section 5.1. These graphs are provided to help contextualize the overall discussion and offer a clearer understanding of the results.

A.1 Classification algorithms

A.1.1 Water Content in Avocado



(a) Importances from RF model Lite Continuous.

(b) Importances from RF model Lite Discrete

Figure A.1: Importances of Water Content in Avocado.

Table A.1: LR, DT and KNN Best Hyperparameters using GridSearchCV Water Content Avocado.

Range	Method	Parameter	Value	Metric
VIS NIR	LR	C	10000	-
	DT	max depth	10	-
	k-NN	n neighbours	3	euclidean
VIS	LR	C	10000	-
	DT	max depth	10	-
	k-NN	n neighbours	1	manhattan
NIR	LR	C	100	-
	DT	max depth	None	-
	k-NN	n neighbours	5	euclidean
SWIR	LR	C	100	-
	DT	max depth	None	-
	k-NN	n neighbours	5	euclidean
RE	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	9	euclidean
Lite Continuous	**LR**	**C**	**1000**	-
	DT	max depth	None	-
	k-NN	n neighbours	3	euclidean
Lite Discrete	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	1	manhattan

Table A.2: SVM Best Hyperparameters using GridSearchCV Water Content Avocado.

Range	C	coef0	gamma	kernel
VIS NIR	10	1	0.1	poly
VIS	10	5	1	poly
NIR	1	0	0.1	poly
SWIR	1	5	0.1	poly
RE	10	5	1	poly
Lite Continuous	10	5	0.1	poly
Lite Discrete	10	5	1	poly

A.1.2 Chlorophyll Content in Avocado

A.1.2.1 Best result: Confusion Matrix

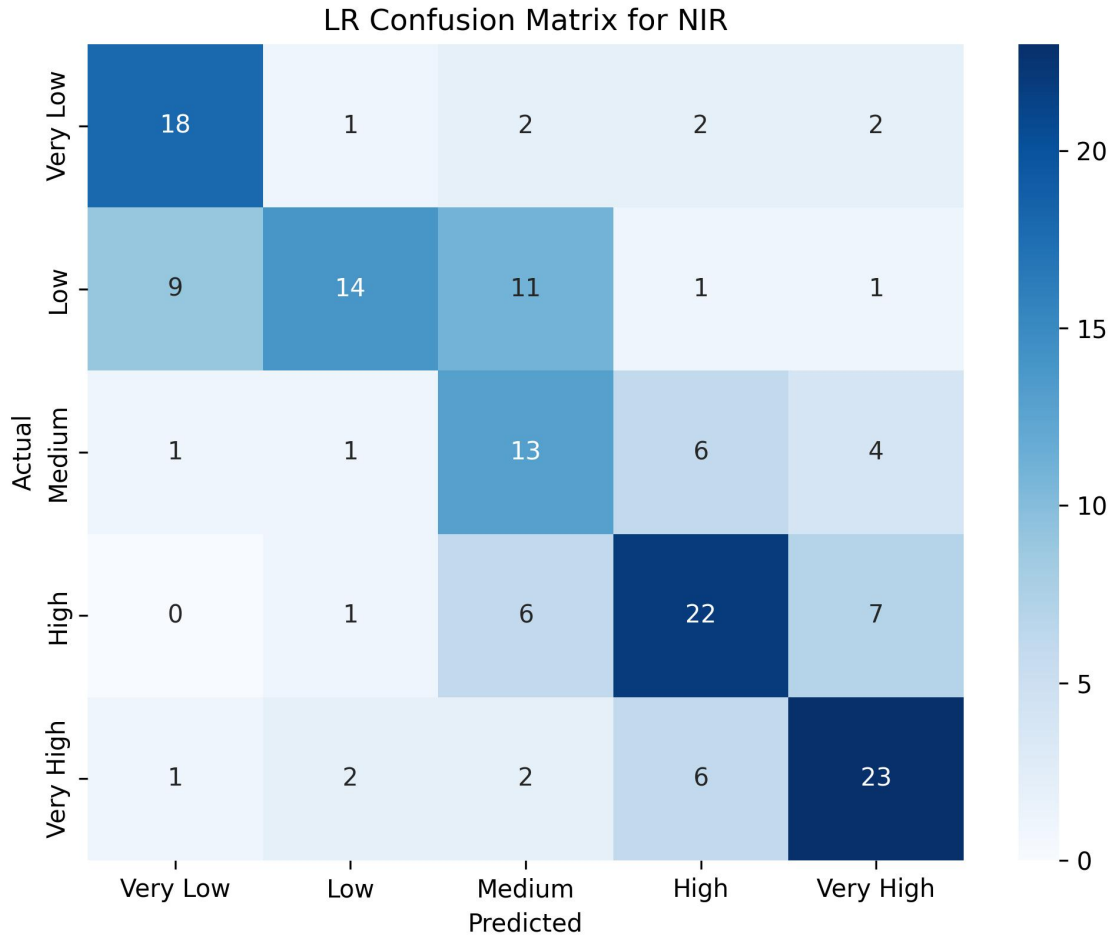


Figure A.2: Optimized LR Confusion Matrix for NIR Chlorophyll Content Avocado.

A.1.2.2 Importances

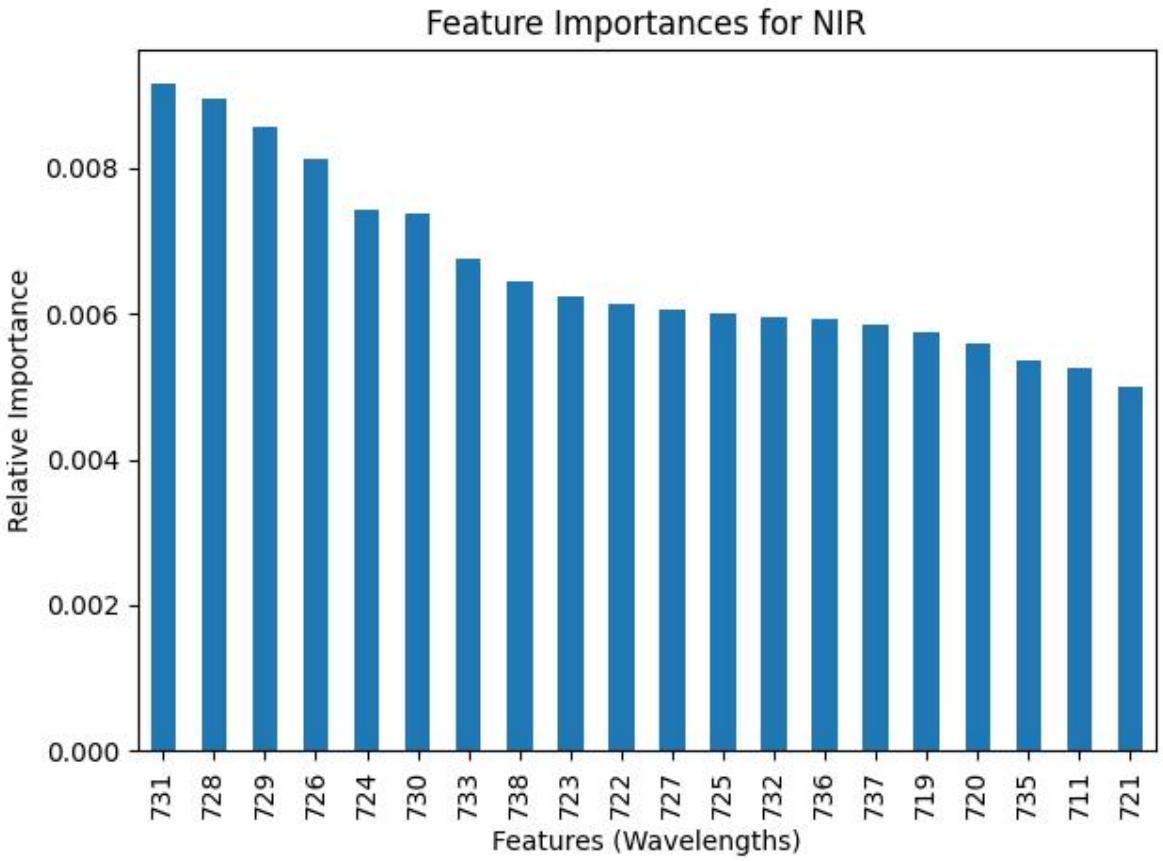
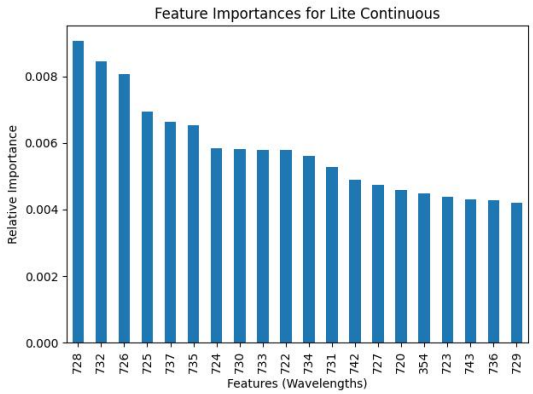
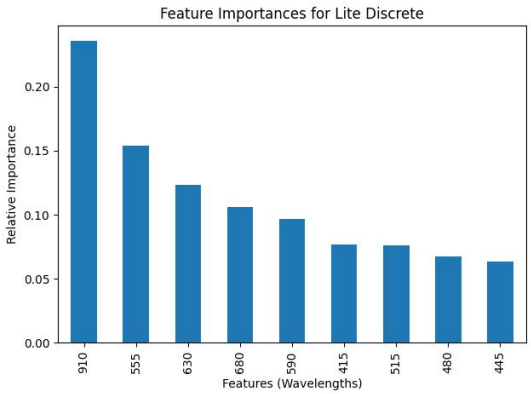


Figure A.3: Importances from RF model NIR.



(a) Importances from RF model Lite Continuous



(b) Importances from RF model Lite Discrete.

Figure A.4: Importances of Chlorophyll Content in Avocado.

Table A.3: LR, DT and KNN Best Hyperparameters using GridSearchCV Chlorophyll Content Avocado.

Range	Method	Parameter	Value	Metric
VIS NIR	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	3	euclidean
VIS	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	11	euclidean
NIR	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	11	euclidean
SWIR	LR	C	100	-
	DT	max depth	10	-
	k-NN	n neighbours	9	euclidean
RE	LR	C	100	-
	DT	max depth	10	-
	k-NN	n neighbours	7	euclidean
Lite Continuous	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	1	euclidean
Lite Discrete	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	9	euclidean

Table A.4: SVM Best Hyperparameters using GridSearchCV Chlorophyll Content Avocado.

Range	C	coef0	gamma	kernel
VIS NIR	10	1	1	poly
VIS	1	5	1	poly
NIR	**1**	**5**	**0.1**	**poly**
SWIR	10	1	0.1	poly
RE	10	5	0.1	poly
Lite Continuous	0.1	5	0.1	poly
Lite Discrete	10	0	1	rbf

A.1.3 Nitrogen Content in Avocado

A.1.3.1 Best result: Confusion Matrix

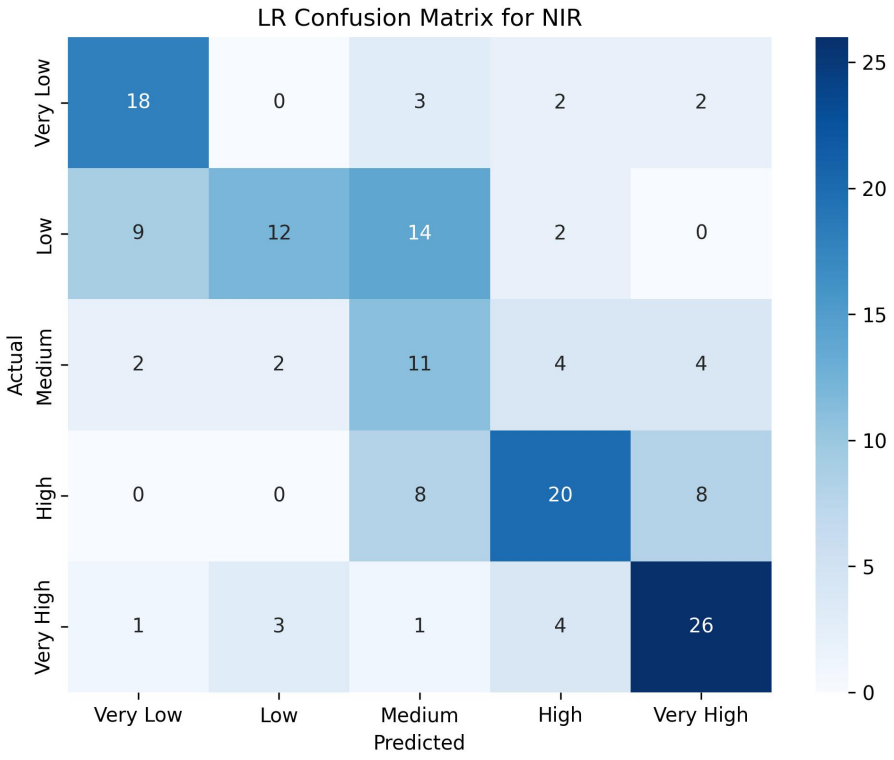


Figure A.5: Optimized SVM Confusion Matrix for NIR Nitrogen Avocado.

A.1.3.2 Importances

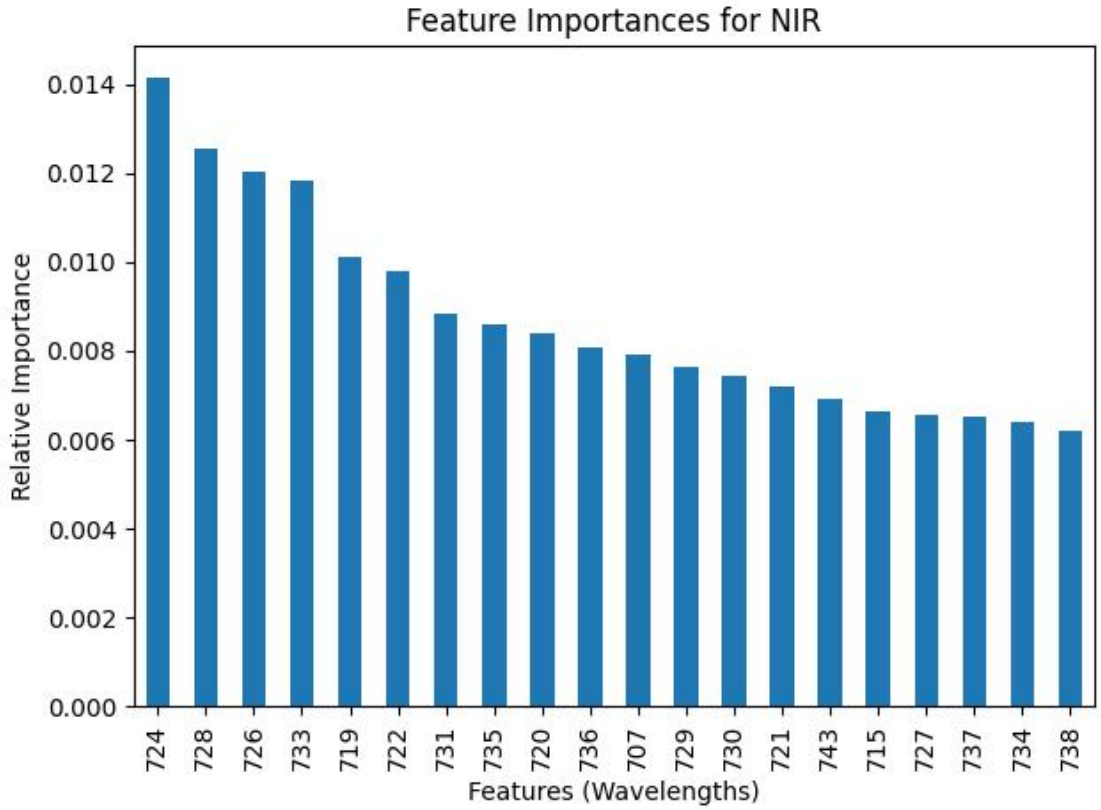
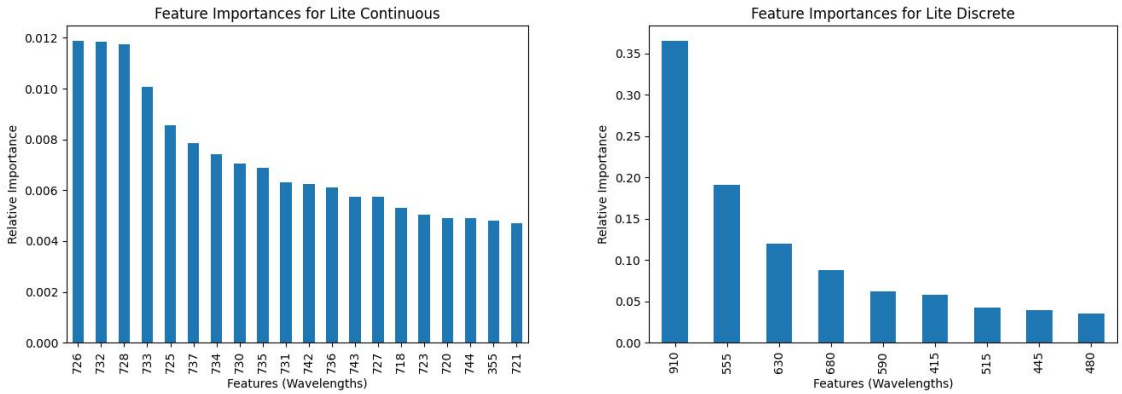


Figure A.6: Importances from RF model NIR.



(a) Importances from RF model Lite Continuous

(b) Importances from RF model Lite Discrete

Figure A.7: Importances of Nitrogen Content in Avocado.

Table A.5: LR, DT and KNN Best Hyperparameters using GridSearchCV Nitrogen Content Avocado.

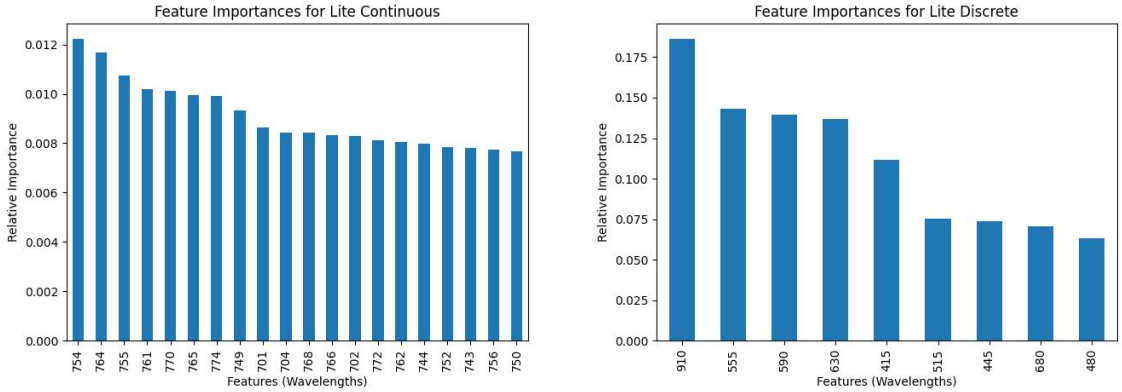
Range	Method	Parameter	Value	Metric
VIS NIR	LR	C	100	-
	DT	max depth	None	-
	k-NN	n neighbours	3	euclidean
VIS	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	3	euclidean
NIR	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	11	euclidean
SWIR	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	11	manhattan
RE	LR	C	100	-
	DT	max depth	10	-
	k-NN	n neighbours	11	euclidean
Lite Continuous	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	1	euclidean
Lite Discrete	LR	C	10	-
	DT	max depth	10	-
	k-NN	n neighbours	9	manhattan

Table A.6: SVM Best Hyperparameters using GridSearchCV Nitrogen Content Avocado.

Range	C	coef0	gamma	kernel
VIS NIR	1	5	0.1	poly
VIS	1	5	1	poly
NIR	**1**	**5**	**0.1**	poly
SWIR	10	5	0.01	poly
RE	10	5	0.1	poly
Lite Continuous	10	5	0.01	poly
Lite Discrete	10	1	1	poly

A.1.4 Water Content in Olive

A.1.4.1 Importances



(a) Importances from RF model Lite Continuous

(b) Importances from RF model Lite Discrete

Figure A.8: Importances of Water Content in Olive.

Table A.7: LR, DT and KNN Best Hyperparameters using GridSearchCV Water Content Olive.

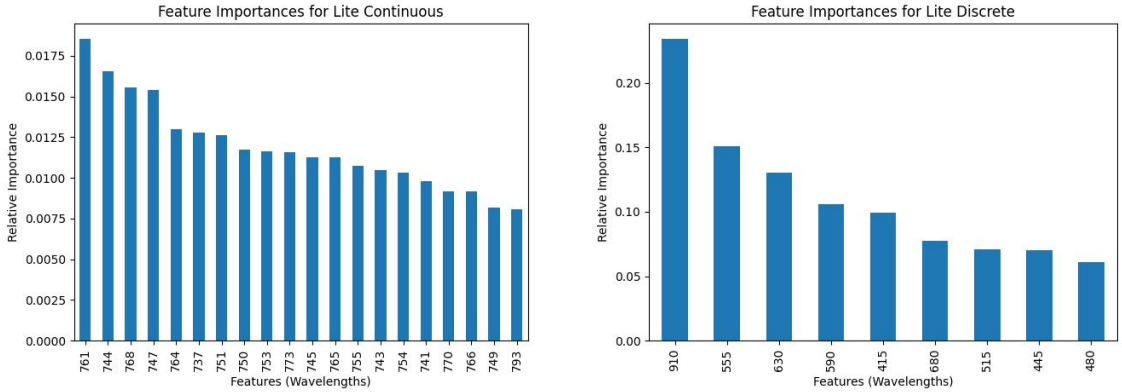
Range	Method	Parameter	Value	Metric
VIS NIR	LR	C	1000	-
	DT	max depth	None	-
	k-NN	n neighbours	1	euclidean
VIS	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	5	euclidean
NIR	LR	C	1000	-
	DT	max depth	None	-
	k-NN	n neighbours	7	euclidean
SWIR	LR	C	100	-
	DT	max depth	None	-
	k-NN	n neighbours	1	euclidean
RE	LR	C	10000	-
	DT	max depth	10	-
	k-NN	n neighbours	3	manhattan
Lite Continuous	**LR**	**C**	**1000**	-
	DT	max depth	None	-
	k-NN	n neighbours	1	euclidean
Lite Discrete	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	1	euclidean

Table A.8: SVM Best Hyperparameters using GridSearchCV Water Content Olive.

Range	C	coef0	gamma	kernel
VIS NIR	1	5	0.1	poly
VIS	10	5	1	poly
NIR	0.1	0	1	poly
SWIR	10	5	0.01	poly
RE	10	5	1	poly
Lite Continuous	10	5	0.01	poly
Lite Discrete	10	5	1	poly

A.1.5 Chlorophyll Content in Olive

A.1.5.1 Importances



(a) Importances from RF model Lite Continuous

(b) Importances from RF model Lite Discrete

Figure A.9: Importances of Chlorophyll Content in Olive.

Table A.9: LR, DT and KNN Best Hyperparameters using GridSearchCV Chlorophyll Content Olive.

Range	Method	Parameter	Value	Metric
VIS NIR	LR	C	10	-
	DT	max depth	None	-
	k-NN	n neighbours	5	euclidean
VIS	LR	C	10	-
	DT	max depth	10	-
	k-NN	n neighbours	7	euclidean
NIR	LR	C	10	-
	DT	max depth	10	-
	k-NN	n neighbours	11	manhattan
SWIR	LR	C	10	-
	DT	max depth	None	-
	k-NN	n neighbours	9	euclidean
Lite Continuous	LR	C	10	-
	DT	max depth	10	-
	k-NN	n neighbours	11	euclidean
RE	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	9	euclidean
Lite Discrete	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	7	euclidean

Table A.10: SVM Best Hyperparameters using GridSearchCV Chlorophyll Content Olive.

Range	C	coef0	gamma	kernel
VIS NIR	1	0	0.1	poly
VIS	10	5	0.1	poly
NIR	10	1	0.01	poly
SWIR	**1**	**5**	**0.01**	**poly**
RE	10	5	0.1	poly
Lite Continuous	0.1	0	0.1	poly
Lite Discrete	1	5	1	poly

A.1.6 Nitrogen Content in Olive

A.1.6.1 Best result: Confusion Matrix

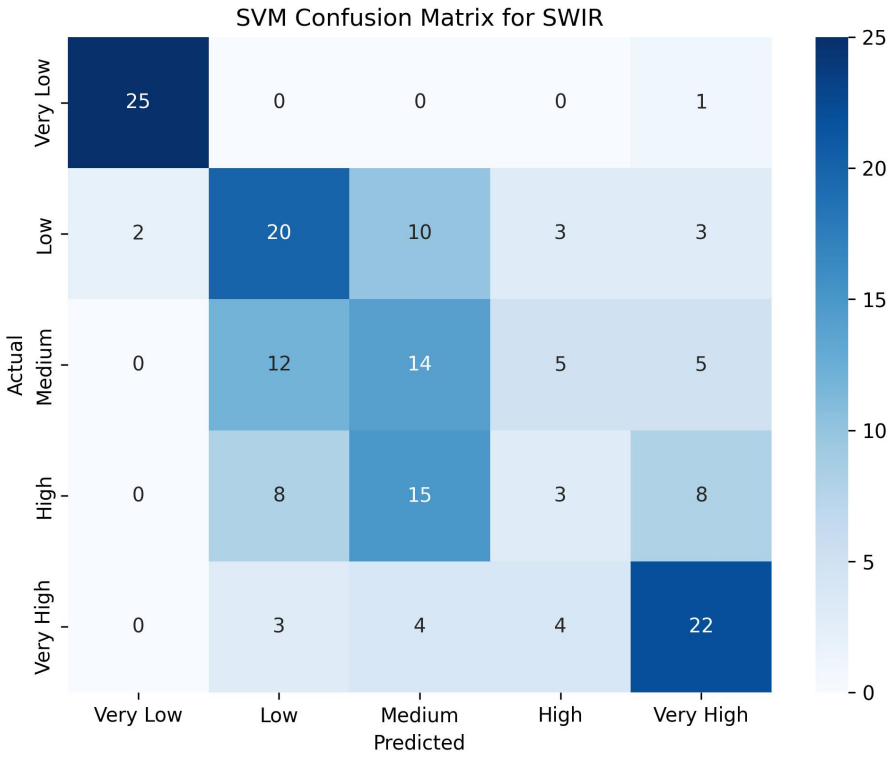
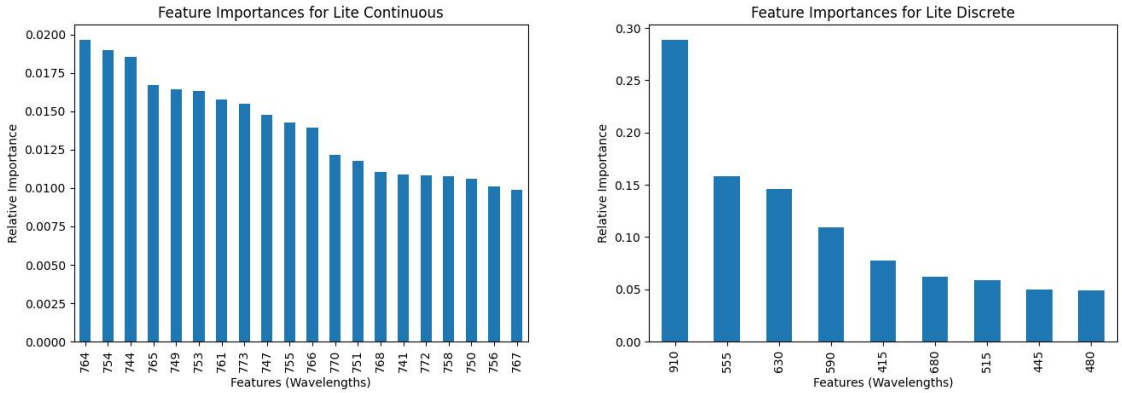


Figure A.10: Confusion Matrix: Nitrogen Content in Olive

A.1.6.2 Importances



(a) Importances from RF model Lite Continuous

(b) Importances from RF model Lite Discrete

Figure A.11: Importances of Nitrogen Content in Olive.

Table A.11: LR, DT and KNN Best Hyperparameters using GridSearchCV Nitrogen Content Olive.

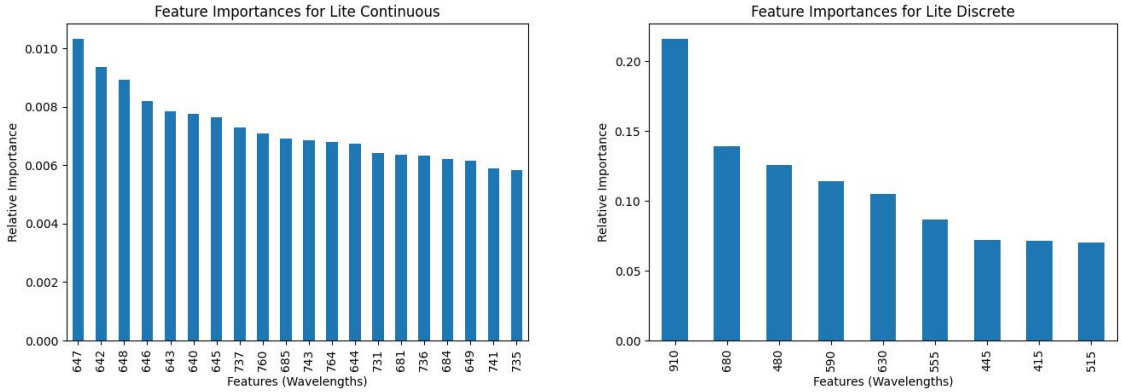
Range	Method	Parameter	Value	Metric
VIS NIR	LR	C	10	-
	DT	max depth	10	-
	k-NN	n neighbours	9	euclidean
VIS	LR	C	100	-
	DT	max depth	10	-
	k-NN	n neighbours	9	euclidean
NIR	LR	C	100	-
	DT	max depth	10	-
	k-NN	n neighbours	11	manhattan
SWIR	LR	C	10	-
	DT	max depth	None	-
	k-NN	n neighbours	9	manhattan
RE	LR	C	100	-
	DT	max depth	10	-
	k-NN	n neighbours	11	euclidean
Lite Continuous	LR	C	10	-
	DT	max depth	None	-
	k-NN	n neighbours	11	euclidean
Lite Discrete	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	11	manhattan

Table A.12: SVM Best Hyperparameters using GridSearchCV Nitrogen Content Olive.

Range	C	coef0	gamma	kernel
VIS NIR	1	0	0.1	poly
VIS	10	5	0.01	poly
NIR	**1**	**5**	**0.01**	**poly**
SWIR	1	5	0.01	poly
RE	1	0	1	poly
Lite Continuous	10	1	0.01	poly
Lite Discrete	10	5	0.1	poly

A.1.7 Water Content in Vineyard

A.1.7.1 Importances



(a) Importances from RF model Lite Continuous

(b) Importances from RF model Lite Discrete

Figure A.12: Importances of Water Content in Vineyard.

Table A.13: LR, DT and KNN Best Hyperparameters using GridSearchCV Water Content Vineyard.

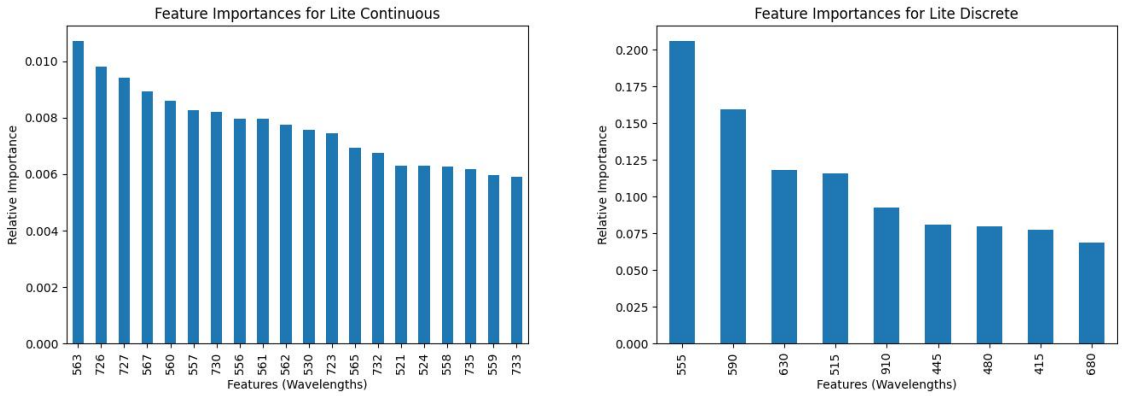
Range	Method	Parameter	Value	Metric
VIS NIR	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	11	euclidean
VIS	LR	C	1000	-
	DT	max depth	None	-
	k-NN	n neighbours	7	euclidean
NIR	**LR**	**C**	**10000**	-
	DT	max depth	10	-
	k-NN	n neighbours	1	euclidean
SWIR	LR	C	100	-
	DT	max depth	10	-
	k-NN	n neighbours	3	euclidean
RE	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	3	euclidean
Lite Continuous	LR	C	100	-
	DT	max depth	None	-
	k-NN	n neighbours	5	manhattan
Lite Discrete	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	7	manhattan

Table A.14: SVM Best Hyperparameters using GridSearchCV Water Content Vineyard.

Range	C	coef0	gamma	kernel
VIS NIR	10	5	0.1	poly
VIS	10	5	0.1	poly
NIR	0.1	1	1	poly
SWIR	10	5	0.01	poly
RE	1	1	1	poly
Lite Continuous	10	5	0.1	poly
Lite Discrete	10	5	1	poly

A.1.8 Chlorophyll Content in Vineyard

A.1.8.1 Importances



(a) Importances from RF model Lite Continuous

(b) Importances from RF model Lite Discrete

Figure A.13: Importances of Chlorophyll Content in Vineyard.

Table A.15: LR, DT and KNN Best Hyperparameters using GridSearchCV Chlorophyll Content Vineyard.

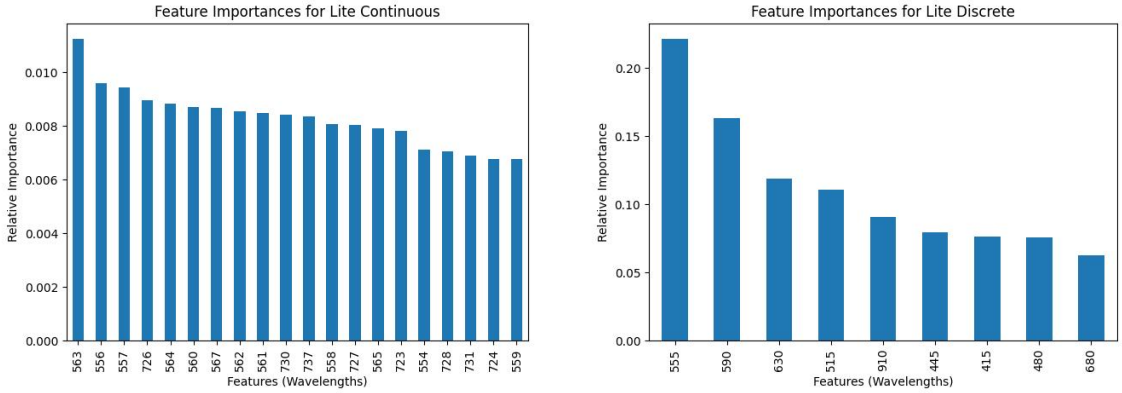
Range	Method	Parameter	Value	Metric
VIS NIR	LR	C	10	-
	DT	max depth	10	-
	k-NN	n neighbours	11	manhattan
VIS	LR	C	10	-
	DT	max depth	None	-
	k-NN	n neighbours	9	manhattan
NIR	LR	C	10000	-
	DT	max depth	None	-
	k-NN	n neighbours	9	euclidean
SWIR	LR	C	10	-
	DT	max depth	10	-
	k-NN	n neighbours	11	manhattan
RE	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	11	euclidean
Lite Continuous	LR	C	0.1	-
	DT	max depth	10	-
	k-NN	n neighbours	11	manhattan
Lite Discrete	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	11	euclidean

Table A.16: SVM Best Hyperparameters using GridSearchCV Chlorophyll Content Vineyard.

Range	C	coef0	gamma	kernel
VIS NIR	10	0	0.01	poly
VIS	**10**	**5**	**0.01**	**poly**
NIR	1	0	0.1	poly
SWIR	10	5	0.01	poly
RE	0.1	5	1	poly
Lite Continuous	1	5	0.01	poly
Lite Discrete	10	5	0.1	poly

A.1.9 Nitrogen Content in Vineyard

A.1.9.1 Importances



(a) Importances from RF model Lite Continuous

(b) Importances from RF model Lite Discrete

Figure A.14: Importances of Nitrogen Content in Vineyard.

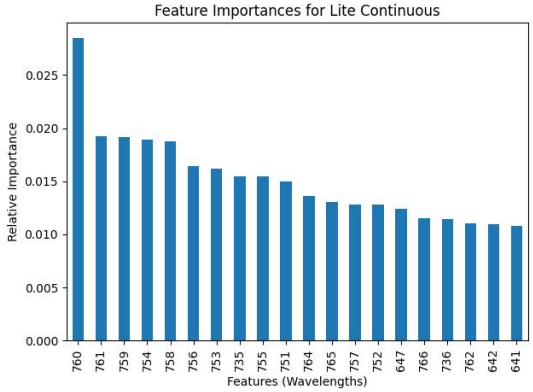
Table A.17: LR, DT and KNN Best Hyperparameters using GridSearchCV Nitrogen Content Vineyard.

Range	Method	Parameter	Value	Metric
VIS NIR	LR	C	10	-
	DT	max depth	10	-
	k-NN	n neighbours	11	manhattan
VIS	LR	C	10	-
	DT	max depth	10	-
	k-NN	n neighbours	11	manhattan
NIR	LR	C	1000	-
	DT	max depth	None	-
	k-NN	n neighbours	11	euclidean
SWIR	LR	C	100	-
	DT	max depth	None	-
	k-NN	n neighbours	11	euclidean
RE	LR	C	1000	-
	DT	max depth	10	-
	k-NN	n neighbours	11	manhattan
Lite Continuous	LR	C	10	-
	DT	max depth	None	-
	k-NN	n neighbours	9	manhattan
Lite Discrete	LR	C	100	-
	DT	max depth	10	-
	k-NN	n neighbours	11	euclidean

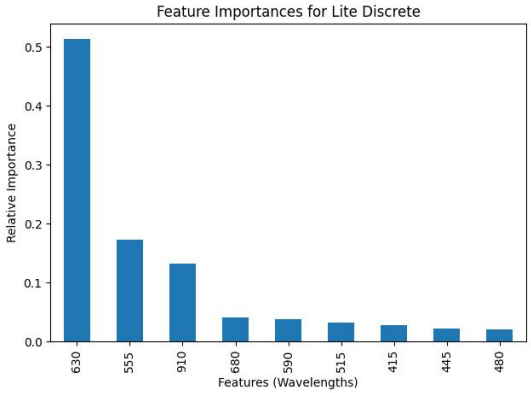
Table A.18: SVM Best Hyperparameters using GridSearchCV Nitrogen Content Vineyard.

Range	C	coef0	gamma	kernel
VIS NIR	1	5	0.01	poly
VIS	**1**	**5**	**0.01**	**poly**
NIR	1	0	0.1	poly
SWIR	10	5	0.01	poly
RE	0.1	1	1	poly
Lite Continuous	1	5	0.01	poly
Lite Discrete	10	1	1	poly

A.2 Regressor Algorithms



(a) Importances from Random Forest Lite Continuous



(b) Importances from Random Forest Lite Discrete

Figure A.15: Importances