AGRICULTURAL UNIVERSITY OF ATHENS
DEPARTMENT OF NATURAL RESOURCES MANAGEMENT & AGRICULTURAL ENGINEERING
LABORATORY OF FARM MACHINE SYSTEMS

PhD Dissertation

Investigating the application of spectral imaging and AI in precision horticulture (agriculture)

Ioannis D. Malounas

Supervisor:
Spyros Fountas | Professor, Agricultural University of Athens

Three-member committee:
Konstantinos Arvanitis| Professor, Agricultural University of Athens
Georgios Xanthopoulos | Associate Professor, Agricultural University of Athens
Spyros Fountas | Professor, Agricultural University of Athens

Athens
2024

AGRICULTURAL UNIVERSITY OF ATHENS
DEPARTMENT OF NATURAL RESOURCES MANAGEMENT & AGRICULTURAL ENGINEERING
LABORATORY OF FARM MACHINE SYSTEMS

PhD Dissertation

Investigating the application of spectral imaging and AI in precision horticulture (agriculture)

Διερεύνηση των εφαρμογών φασματοσκοπίας και τεχνητής νοημοσύνης στη λαχανοκομία (γεωργία) ακριβείας

Ioannis D. Malounas

A thesis submitted to the Agricultural University of Athens
in fulfilment of the requirements for the
degree of Doctor of Philosophy

Seven-member committee:
Spyros Fountas | Professor, Agricultural University of Athens (Supervisor)
Konstantinos Arvanitis| Professor, Agricultural University of Athens
Georgios Xanthopoulos | Associate Professor, Agricultural University of Athens
Dimitrios Savvas| Professor, Agricultural University of Athens
Aristotelis Tagarakis| Senior Researcher, Institute for Bio-economy and Agri-technology
Dimitrios Argyropoylos| Assistant Professor, University College Dublin
Manuela Zude-Sasse | Professor, ATB Leibniz-Institute of Agricultural Engineering and Bio-economy

Investigating the application of spectral imaging and AI in precision horticulture (agriculture)

Department of Natural Resources Management & Agricultural Engineering
Laboratory of Farm Machine Systems

Abstract

Spectral imaging and Artificial Intelligence in precision horticulture are commonly used for a variety of applications ranging from disease detection to quality estimation. However, most of the available solutions require deep understanding of software engineering and they mostly focus on disease detection and post-harvest applications.

This study aimed to (i)develop Artificial Intelligence models utilizing spectral data that can identify different fertilisation levels, (ii)develop Artificial Intelligence models utilizing spectral data capable of identifying plant water deficit, (iii)compare the performance of traditional machine learning algorithms with novel user-friendly Auto Machine Learning (AutoML) techniques and(iv)evaluate the feasibility of developing a generalisation-capable AI model utilizing spectral data.

Towards that end, a progressive methodology was implemented to gather data and develop the required methodologies. During the first year spectral data from broccoli plants that were submitted to different fertilization schemes were collected, while during the second year spectral data were collected from broccoli plants that were submitted to different irrigation schemes. Besides spectral data during both years, dry matter measurements were conducted not only for broccoli but also for apple, leek and mushroom. Finally, during the third year, all AI methodologies were developed, and AI experiments were conducted.

Throughout these three years, this study evaluated and compared traditional Artificial Intelligence approaches with AutoML systems towards water/ acclimation and nutrient deficiency stress identification using spectral imaging. For both types of stress, AutoML was compared to a traditional machine learning approach (Partial Least Squares – Discriminant Analysis) used for classification of spectral data. On both occasions, data were captured with the use of the IMEC snapscan Visible Near Infrared hyperspectral camera (400-900nm). Moreover, the study aimed to investigate generalisation capabilities of spectral imaging and how each step of the “traditional” pre-processing pipeline followed for spectral data modelling affects its generalisation capabilities and performance. The pipeline, followed by both stress experiments and tested for its generalisation capabilities, consisted of the following steps: Outlier removal→Data smoothing→Data Scaling→Feature selection→Feature Extraction→Modelling. Techniques used for various steps across the pipeline included Savitzky Golay smoothing, Standard and Min Max scaling, f and mutual info regression for feature

selection, umap, autoencoeder and PCA feature extraction and various machine learning models ranging from linear to quadratic models and reaching the complexity of neural networks.

For identifying nutrient stress, the AutoML system achieved results that were superior to those achieved by the Partial Least Squares – Discriminant Analysis (PLS-DA) algorithm. Namely, an accuracy of 0.72 was achieved when using the CIELAB colour space and 0.94 when combining the CIELAB colour space with the hyperspectral data.

When using the hyperspectral data standalone, the results improved (accuracy 1.00), this performance was achieved using all 150 bands, however, it is worth mentioning that the same performance was maintained even when using the single statistically most important wavelength (874 nm). On the other hand, for the identification of water/acclimation stressed plants, both the Automated Machine Learning system and the PLSDA algorithm achieved an accuracy of 1.00 across all stress levels. Finally, hyperspectral imaging has proven capable of generalizing across different fruits and vegetables, achieving an (RMSEP) = 0.0137 using the Partial Least Squares Regression algorithm on a 10x5-fold cross-validation protocol.

Overall, the results suggest that Automated Machine Learning can achieve and even outperform traditional spectral imaging machine learning approaches for detecting water/ acclimation and nutrient deficiency stress. Moreover, the use of the CIELAB colour space for training the models failed to match the performance of using the spectral data, while combining the two did not lead to a performance increase compared to just using the spectral data. The evaluated techniques used for preprocessing affected the two regression algorithms, Automatic Relevance Determination Regression (ARD) and Partial Least Squares Regression (PLSR) in a different way, with the best performance achieved when the complete pipeline was used. Furthermore, feature selection appeared to be the preprocessing technique that had the most negative impact on the linear regression performance when used standalone. However, its use to fit a quadratic transformation of the features was found to be a good compromise. Overall, the pipeline using either ARD algorithm or PLSR algorithm showed strong generalization and performance in the Visible Near Infrared wavelength based dry matter content estimation across diverse crops.

To conclude, the use of Spectral imaging with AutoML solutions may provide a user friendly and cost-effective method for detecting plant stress, while at the same time, spectral imaging model generalisation can be achieved provided that a universal data acquisition protocol is followed, with promising results even without following complex data preprocessing pipelines. Finally, it should be noted that the present study did not examine the lag factor. It is assumed that with the prevalence of water or nutrient scarcity, the change in spectral data will not be automatic.

Scientific area: Agricultural Engineering

Keywords: Precision agriculture; Artificial Intelligence; Spectral Imaging; Remote sensing; AutoML; Fertilisation; Irrigation, Abiotic stress, Acclimation stress, Generalisation, Dry Matter Content, Polynomial Regression, Feature Extraction

Διερεύνηση των εφαρμογών φασματοσκοπίας και τεχνητής νοημοσύνης στην λαχανοκομία (γεωργία) ακριβείας

Τμήμα Αξιοποίησης Φυσικών Πόρων &ι Γεωργικής Μηχανικής
Εργαστήριο Γεωργικής Μηχανολογίας

Περίληψη

Η φασματική απεικόνιση και η Τεχνητή Νοημοσύνη στη λαχανοκομία (γεωργία) ακριβείας χρησιμοποιούνται συνήθως για μια ποικιλία εφαρμογών που κυμαίνονται από την ανίχνευση ασθενειών έως την εκτίμηση της ποιότητας. Ωστόσο, οι περισσότερες από τις διαθέσιμες λύσεις απαιτούν βαθιά κατανόηση της μηχανικής μάθησης και επικεντρώνονται κυρίως στην ανίχνευση ασθενειών και σε εφαρμογές μετά τη συγκομιδή.

Αυτή η μελέτη είχε ως στόχο: (i) την ανάπτυξη μοντέλων τεχνητής νοημοσύνης που χρησιμοποιούν φασματικά δεδομένα και μπορούν να προσδιορίσουν διαφορετικά επίπεδα λίπανσης, (ii) την ανάπτυξη μοντέλων τεχνητής νοημοσύνης που χρησιμοποιούν φασματικά δεδομένα και είναι ικανά να προσδιορίσουν το έλλειμμα νερού των φυτών, (iii) τη σύγκριση των επιδόσεων των παραδοσιακών αλγορίθμων μηχανικής μάθησης με νέες φιλικές προς τον χρήστη τεχνικές Αυτόματης Μηχανικής Μάθησης, και (iv) την αξιολόγηση της δυνατότητας ανάπτυξης ενός μοντέλου τεχνητής νοημοσύνης με δυνατότητα γενίκευσης που χρησιμοποιεί φασματικά δεδομένα.

Προς το σκοπό αυτό, εφαρμόστηκε μια προοδευτική μεθοδολογία για τη συλλογή δεδομένων και την ανάπτυξη των απαιτούμενων μεθοδολογιών. Κατά το πρώτο έτος συλλέχθηκαν φασματικά δεδομένα από φυτά μπρόκολου που υποβλήθηκαν σε διαφορετικές μεταχειρίσεις λίπανσης, ενώ κατά το δεύτερο έτος συλλέχθηκαν φασματικά δεδομένα από φυτά μπρόκολου που υποβλήθηκαν σε διαφορετικές μεταχειρίσεις άρδευσης. Εκτός από τα φασματικά δεδομένα κατά τη διάρκεια και των δύο ετών πραγματοποιήθηκαν μετρήσεις ξηρής ουσίας όχι μόνο για το μπρόκολο αλλά και για το μήλο, το πράσο και το μανιτάρι. Τέλος, κατά τη διάρκεια του τρίτου έτους αναπτύχθηκαν όλες οι μεθοδολογίες τεχνητής νοημοσύνης και πραγματοποιήθηκαν και τα αντίστοιχα πειράματα.

Κατά τη διάρκεια αυτών των τριών ετών η παρούσα μελέτη αξιολόγησε και συνέκρινε τις παραδοσιακές προσεγγίσεις τεχνητής νοημοσύνης με τα συστήματα Αυτόματης Μηχανικής Μάθησης για τον εντοπισμό της καταπόνησης λόγω έλλειψης νερού/εγκλιματισμού και θρεπτικών στοιχείων με τη χρήση φασματικής απεικόνισης. Και για τους δύο τύπους καταπόνησης η Αυτόματη Μηχανική Μάθηση συγκρίθηκε με μια παραδοσιακή προσέγγιση μηχανικής μάθησης,Partial Least Squares – Discriminant Analysis (PLSDA), που χρησιμοποιείται για την ταξινόμηση φασματικών δεδομένων. Και στις δύο περιπτώσεις τα δεδομένα λήφθηκαν με τη χρήση της υπερφασματικής κάμερας IMEC snapscan Visible Near Infrared (400-900nm). Επιπλέον, η μελέτη αποσκόπησε στη

διερεύνηση των δυνατοτήτων γενίκευσης της φασματικής απεικόνισης και του τρόπου με τον οποίο κάθε βήμα της "τυπικής" διαδικασίας προεπεξεργασίας που ακολουθείται για τη μοντελοποίηση φασματικών δεδομένων επηρεάζει τις δυνατότητες και τις επιδόσεις της γενίκευσης. Η διαδικασία που ακολουθήθηκε τόσο από τα πειράματα καταπόνησης όσο και από τον έλεγχο των δυνατοτήτων γενίκευσης αποτελείται από τα ακόλουθα βήματα: Απομάκρυνση εκτόπων τιμών→Εξομάλυνση δεδομένων→Κλιμάκωση δεδομένων→Επιλογή χαρακτηριστικών→Εξαγωγή χαρακτηριστικών→Μοντελοποίηση. Οι τεχνικές που χρησιμοποιήθηκαν για τα διάφορα βήματα της διαδικασίας περιλαμβάνουν εξομάλυνση SavitzkyGolay, κλιμάκωση Standard και MinMax, παλινδρόμηση f και mutualinfo για επιλογή χαρακτηριστικών, εξαγωγή χαρακτηριστικών umap, autoencoeder και pca και διάφορα μοντέλα μηχανικής μάθησης που κυμαίνονται από γραμμικά έως τετραγωνικά μοντέλα και φτάνουν στην πολυπλοκότητα των νευρωνικών δικτύων.

Για τον εντοπισμό της θρεπτικής καταπόνησης το σύστημα αυτόματης μηχανικής μάθησης πέτυχε αποτελέσματα που είναι ανώτερα από εκείνα που πέτυχε η ανάλυση Partial Least Squares – Discriminant Analysis. Συγκεκριμένα, επιτεύχθηκε ακρίβεια (accuracy) 0,72 όταν χρησιμοποιήθηκε ο χρωματικός χώρος CIELAB και 0,94 όταν συνδυάστηκε ο χρωματικός χώρος CIELAB με τα υπερφασματικά δεδομένα. Κατά τη χρήση των υπερφασματικών δεδομένων αυτοτελώς, τα αποτελέσματα βελτιώθηκαν (ακρίβεια 1,00), η επίδοση αυτή επιτεύχθηκε με τη χρήση και των 150 φασμάτων, ωστόσο αξίζει να αναφερθεί ότι η ίδια επίδοση διατηρήθηκε ακόμη και όταν χρησιμοποιήθηκε το μοναδικό στατιστικά σημαντικότερο φάσμα (874 nm, near infrared).

Από την άλλη πλευρά, για την ταυτοποίηση των φυτών που έχουν υποστεί στρες από νερό/κλιματισμό, τόσο το σύστημα αυτόματης μηχανικής μάθησης όσο και ο αλγόριθμος PLSDA. Επέτυχαν ακρίβεια 1,00 σε όλα τα επίπεδα στρες. Τέλος, η υπερφασματική απεικόνιση αποδείχθηκε ικανή να γενικεύει σε διάφορα φρούτα και λαχανικά, επιτυγχάνοντας Μέση Τετραγωνική απόκλιση (RMSEP) = 0.0137 χρησιμοποιώντας παλινδρόμηση Partial Least Squares Regression σε πρωτόκολλο διασταυρούμενης επικύρωσης 10x5 φορές.

Συνολικά, τα αποτελέσματα υποδηλώνουν ότι η αυτόματη μηχανική μάθηση μπορεί να επιτύχει και ακόμη και να ξεπεράσει τις παραδοσιακές προσεγγίσεις μηχανικής μάθησης φασματικής απεικόνισης για την ανίχνευση του στρες του νερού/εγκλιματισμού και της θρεπτικής ανεπάρκειας. Επιπλέον, η χρήση του χρωματικού χώρου CIELAB για την εκπαίδευση των μοντέλων απέτυχε να φτάσει την απόδοση της χρήσης των φασματικών δεδομένων, ενώ ο συνδυασμός των δύο δεν οδήγησε σε αύξηση της απόδοσης σε σύγκριση με τη χρήση μόνο των φασματικών δεδομένων. Τέλος, οι τεχνικές προ επεξεργασίας που αξιολογήθηκαν επηρέασαν διαφορετικά τους δύο αλγορίθμους παλινδρόμησης (Automatic Relevance Determination και Partial Least Squares), με τα καλύτερα αποτελέσματα να επιτυγχάνονται όταν χρησιμοποιήθηκε η πλήρης διαδικασία. Επιπλέον, η επιλογή χαρακτηριστικών φάνηκε να είναι η τεχνική προ επεξεργασίας που

έχει τον πιο αρνητικό αντίκτυπο στην απόδοση της γραμμικής παλινδρόμησης όταν χρησιμοποιείται μεμονωμένα. Ωστόσο, η χρήση της για την προσαρμογή ενός τετραγωνικού μετασχηματισμού των χαρακτηριστικών διαπιστώθηκε ότι αποτελεί έναν καλό συμβιβασμό. Συνολικά, η διαδικασία που χρησιμοποίησε είτε την Automatic Relevance Determination παλινδρόμηση είτε την Partial Least Squares Regression παλινδρόμηση παρουσίασε ισχυρή απόδοση και γενίκευση για την εκτίμηση της ξηρής ύλης με βάση το ορατό και κοντινό υπέρυθρο σε διάφορα φρούτα και λαχανικά.

Συμπερασματικά, η χρήση της φασματικής απεικόνισης με λύσεις Αυτόματης Μηχανικής Μάθησης μπορεί να παρέχει μια φιλική προς τον χρήστη και οικονομικά αποδοτική μέθοδο για την ανίχνευση της καταπόνησης των φυτών, ενώ ταυτόχρονα μπορεί να επιτευχθεί γενίκευση του μοντέλου φασματικής απεικόνισης, εφόσον ακολουθείται ένα καθολικό πρωτόκολλο απόκτησης δεδομένων, με πολλά υποσχόμενα αποτελέσματα ακόμη και χωρίς να ακολουθούνται πολύπλοκες σωληνώσεις προ επεξεργασίας δεδομένων. Τέλος, πρέπει να σημειωθεί ότι η παρούσα μελέτη δεν εξέτασε τον παράγοντα υστέρησης. Εκτιμάται πως με την επικράτηση έλλειψης νερού και θρεπτικών συστατικών η μεταβολή των φασματικών δεδομένων δεν θα είναι αυτόματη.

Επιστημονική περιοχή: Γεωργική Μηχανική

Λέξεις κλειδιά: Γεωργία ακριβείας; Τεχνητή Νοημοσύνη; Φασματοσκοπία, Αυτόματη Μηχανική Μάθηση; Τηλεπισκόπηση; Λίπανση; Άρδευση; Αβιοτικό στρες, Προσαρμοστικό στρες, Γενίκευση, Περιεκτικότητα σε ξηρή ύλη, Πολυωνυμική παλινδρόμηση, Εξαγωγή χαρακτηριστικών

Certificate of Originality

I hereby certify that the text of this thesis does not contain any material that has been accepted as part of the requirements for a degree or diploma at any university, nor does it contain any material that has been previously published or written, unless such material is referenced.

In addition, with my permission, this thesis has been checked and verified for validity and originality by the examination board using plagiarism detection software available from the Agricultural University of Athens.

Ioannis Malounas

List of publications

International scientific journal publications

1. Malounas, Ioannis, Diamanto Lentzou, Georgios Xanthopoulos, and Spyros Fountas. "Testing the Suitability of Automated Machine Learning, hyperspectral imaging and CIELABcolor space for proximal in situ fertilization level classification." *Smart Agricultural Technology* (2024): 100437.

2. Malounas, Ioannis, Georgios Paliouras, Dimosthenis Nikolopoulos, Georgios Liakopoulos, Panagiota Bresta, Paraskevi Londra, Anastasios Katsileros, and Spyros Fountas. "Early detection of broccoli drought acclimation/stress in agricultural environments utilising proximal hyperspectral imaging and AutoML." Smart Agricultural Technology (2024): 100463.

3. Malounas, Ioannis, Wout Vierbergen, Sezer Kutluk, Manuela Zude-Sasse, Kai Yang, Ming Zhao, Dimitrios Argyropoulos, Jonathan Van Beek, Eva Ampe, and Spyros Fountas. "SpectroFood dataset: A comprehensive fruit and vegetable hyperspectral meta-dataset for dry matter estimation." *Data in Brief* 52 (2024): 110040.

4. Malounas, Ioannis, Borja Espejo-Garcia, Konstantinos Arvanitis, and Spyros Fountas. "Evaluation of a hyperspectral image pipeline toward building a generalisation capable crop dry matter content prediction model." Biosystems Engineering 247 (2024): 153-161.

Participation in other publications

International scientific journal publications

1. Wieme, Jana, Kaveh Mollazade, Ioannis Malounas, Manuela Zude-Sasse, Ming Zhao, Aoife Gowen, Dimitrios Argyropoulos, Spyros Fountas, and Jonathan Van Beek. "Application of hyperspectral imaging systems and Artificial Intelligence for quality assessment of fruit, vegetables and mushrooms: A review." *biosystems engineering* 222 (2022): 156-176.

2. Espejo-Garcia, Borja, Ioannis Malounas, Nikos Mylonas, Aikaterini Kasimati, and Spyros Fountas. "Using EfficientNet and transfer learning for image-based

diagnosis of nutrient deficiencies." *Computers and Electronics in Agriculture* 196 (2022): 106868.

3. Espejo-Garcia, Borja, Ioannis Malounas, Eleanna Vali, and Spyros Fountas. "Testing the Suitability of Automated Machine Learning for Weeds Identification." *Ai* 2, no. 1 (2021): 34-47.

Acknowledgments

Completing a significant research project like this requires the collective efforts of more than one person, each contributing differently. Therefore, I would like to express my gratitude to the following people:

Firstly, I would like to express my sincere gratitude to my supervisor, Professor Spyros Fountas, for giving me the opportunity, inspiring me, and guiding me throughout my PhD journey. With his vibrant, enthusiastic character, Spyros encouraged me to chase my goals while offering his unwavering support and allowing me the freedom to pursue diverse projects. Thank you, Spyros, for being such and exceptional mentor! Throughout my Ph.D. years, your guidance has been invaluable in nurturing my research and professional growth.

I want to thank the other two members of my three-member committee: Professor Georgios Xanthopoulos, whose expertise, encouragement, and guidance significantly enriched my doctoral experience, enabling me to develop my skills further and succeed in achieving my academic goals, as well as Professor Konstantinos Arvanitis for his invaluable experience. It was a pleasure to have you on my PhD committee!

Additionally, I extend my gratitude to Professors Dimitrios Argyropoylos, Manuela Zude-Sasse and Dimitrios Savvas as well as to Senior Researcher Aristotelis Tagarakis for graciously accepting the invitation to serve as members of my committee. Your involvement is greatly appreciated.

To my colleagues in the Smart Farming Technology Group Borja Espejo García, Katerina Kassimati, Nicoleta Darra, Loukas Athanasakos, and Kalliopi Kounani for the support they have given me at all levels of my research project. Special thanks to Georgios Paliouras for his catalytic role in setting up the field experiments and for his assistance in the data collection process.

At this point, I would like to mention that funding for this research has been provided by the European Union’s ICT AGRI research and innovation programme SpectroFood. I also want to sincerely thank all research partners for sharing their knowledge, expertise, and data.

Finally, I would like to express my gratitude to my family, who provided me with their support and a firm foundation to build my life.

Executive Summary

Precision agriculture aims to optimize and improve primary production through the use of modern technological solutions. The majority of those solutions require big amounts of data and as a result, precision agriculture heavily relies on a variety of sensors for data collection, such as spectral cameras. Moreover, Artificial Intelligence (AI) is crucial as it enables data analyses of these large amounts of data in an efficient and accurate way, enabling data driven decision making.

Chapter 1 begins with an introduction to precision agriculture and broccoli production. It continues by providing the fundamental principles and knowledge around Spectral Imaging and Artificial Intelligence and ends by introducing the main sources of variability in agriculture and how spectral imaging and AI have contributed towards improving primary production in terms of resource efficiency, yield and quality improvements. The chapter also highlights the synergies and trade-offs between all technologies described.

Chapter 2 gives an overview of the materials and methods with information on the selected experimental decision and equipment used. It then focuses on the data collection protocols and the various techniques used for data preprocessing, analyses and sample classification. In this study, data from water acclimated/stressed plants and nutrient deficient plants were collected using spectral imaging. Finally, a hyperspectral dataset for dry matter estimation comprising of a variety of crops was constructed and a plethora of preprocessing methods were evaluated towards improving spectral model generalisation performance.

Chapter 3 presents the research fundings of this Ph.D. dissertation. An exploratory analysis and evaluation of the spectra collected, and the various techniques used are presented. Spectral imaging was found to be superior to just using the CIELAB colour space for identification of stressed plants, while at the same time AutoML reached excellent performance comparable to the use of traditional machine learning techniques (PLS-DA) that require in depth knowledge of software engineering. Moreover, this study concluded that just normalizing spectral data can improve the generalisation capabilities of machine learning models that make use of Spectral data, while at the same time adding more data can allow for algorithms to uncover previously hidden patterns. Finally, the synergistic effect of various spectral preprocessing techniques was proven towards improving the performance of generalized spectral data models as well as the effect of sample size on improving model performance even when data are heterogeneous.

Chapter 4 discusses the contributions of the three research papers produced as part of this PhD thesis: 1. Testing the Suitability of Automated Machine Learning, hyperspectral imaging and CIELAB colour space for proximal in situ fertilisation level classification, 2. Early detection of broccoli drought acclimation/stress in agricultural environments utilising proximal hyperspectral imaging and AutoML and 3. Evaluation of

a hyperspectral image pipeline toward building a generalization capable crop dry matter content prediction model.

Chapters 5, conclusions and 6, future work, are the final segments of this dissertation. They are interrelated, built on each other and collectively draw conclusions regarding the aforementioned objectives. They also delve into potential avenues for future investigation concerning the application of precision agriculture, spectral imaging, and Artificial Intelligence.

Contents

Table of Figures

Table of Tables

Table of Equations

Chapter 1 – Introduction

1.1 Problem statement

The agricultural sector is facing significant challenges and will undergo substantial transformations in the near future. The consequences of climate change, such as rising global temperatures, increases in heavy precipitation, and widespread water shortage, directly impact food production and threaten the future of farming. On the other hand, agriculture is a primary driver of climate change. Current primary production practices contribute to air, water and soil pollution, with agriculture being responsible for 10.3% of the EU's Green House Gas emissions(European Environment Agency and European Commission, 2022), while consuming excessive amounts of natural resources and energy. Moreover, pesticides and fertilisers overuse severely affects humans and other life forms as well as the environment. Two examples of fertiliser overuse are the nitrogen and phosphorus cycles, which exceed their safe operating space in Europe by a factor of 3.3 and 2 respectively(European Environment Agency, 2020),

Furthermore, by 2050, agriculture will have to produce 70% to 100% more food. Food production will need to be accompanied by sustainable management of agricultural lands to stop or slow down the negative impacts on the quality and quantity of water and soil resources, land degradation, greenhouse gas emissions and biodiversity (Gomiero et al., 2011). However, this shift will not be easy as it will need to take place in a highly uncertain, variable, and constantly changing agricultural landscape. To address these challenges technological disruption is required.

The technological advancements in the field of precision agriculture technologies and agri-environmental monitoring over the last years have been staggering in terms of hardware (variety of available sensors and platforms, edge devices) and software processing power, resulting in an unprecedented collection of daily observations of crop status and environmental conditions(Glass and Gonzalez, 2022).

These technologies empower farmers to optimize management practices such as fertilisation, irrigation, and plant protection product application, enabling significant cost reduction, improved crop quality and yield, and increased competitiveness(Sharma et al., 2020). With precision agriculture, data are collected to assist farmers in making data-guided sub-field decisions, including applications of fertilisers and pesticides, distribution densities for seeds, irrigation application rates, and tillage regimes(Taylor, 2023).

Summing up, precision agriculture technologies are considered one of the most promising ways to deal with agriculture uncertainty and variability, improve its performance sustainably, reduce its environmental impact, and help it achieve sustainable food production.

1.2 Uncertainty and variability in agriculture

Food security, being able to provide all people, at all times, with physical and economic access to sufficient, safe, and nutritious food that meets their dietary needs and food preferences for an active and healthy life (Shaw, 2007), is one of the most significant problems the world is facing. Ensuring food security has become crucial to numerous countries with different degrees of economic development, with the agricultural sector playing a strategic role in improving food availability(Pawlak and Kołodziejczak, 2020). However, achieving food security under climate change is a complex public policy issue or a so-called "wicked problem." (Vermeulen et al., 2013). The main reason behind all previous statements is modern agriculture's high uncertainty and variability. Agriculture does not suffer from a single source of uncertainty and risk; instead, it has to face multiple and diverse ones ranging from climate and weather-related events to fluctuations in the prices of agriculture inputs such as fertilisers.

Financial uncertainties and policy and regulatory changes also pose severe threats to modern production systems. Diving deeper into each one, farmers first have to deal with the natural uncertainty and risks that directly impact production and are uncontrollable; examples are diseases and weather. Secondly, they have to deal with market uncertainty as the majority of decisions in agriculture are made in advance when the market price for the output is usually unknown. Thirdly, farmers have to face policy uncertainty with economic and environmental policies having a direct impact, such as the mandated reduction of fertiliser use or indirect impact with their effect on taxes and provision of public goods (Aimin, 2010).

Over the past couple of years, the agricultural landscape structure has been shifting towards a simpler one via changes in management, land use, agricultural development, modernisation, and intensification (Benton et al., 2003). However, it remains far more complex compared to the environment of other industries, such as warehouses and factories, where the majority of the variables, such as illumination, obstacles, and landmarks, remain unchanged, with this environment complexity not foreseen to be simplified in the near future. On the contrary, on some occasions, it is predicted to become more complex, with policies aimed at enhancing landscape complexity to increase biodiversity being introduced (Commission and Environment, 2017). It becomes, therefore, apparent that modern agriculture solutions will have to work in such environments. The following paragraphs present a breakdown of the primary sources of agricultural variability, namely, i) soil, ii) climate, iii) illumination, and iv) plant growth.

According to the USDA soil taxonomy, there are 12 major soil types, each with its own taxonomy (Great Group, Subgroup, and Family)(Natural Resources Conservation

Service. U.S. Department of Agriculture, 1999) and its characteristics such as colour and texture, thus making it more difficult for precision agriculture solutions and machinery to work universally. On top of that, specific soil parameters such as soil organic matter and soil total nitrogen are also affected by the farming practices used (Huang et al., 2007), introducing another variable that precision agriculture solutions have to consider. To conclude, spatial variability plays a crucial role in advancing precision agriculture, as site-specific management is currently treated on an average basis(López-Granados et al., 2002).

Climate change is one of the most critical problems the modern world faces, with agriculture being extremely vulnerable. This also leads to new challenges for the agricultural technology industry, as machinery needs to operate in more unpredictable and harsher environments. One of the climate change factors that affect agriculture is higher temperatures, which, besides decreasing yield, they promote weed and pest expansion, rapidly changing the environment where machinery have to work. The second factor is precipitation patterns, primarily affecting irrigated crops(Nelson et al., 2009), which once again challenge disease and plant management. Despite that, climate change will also affect irrigation demands as the physiology and phenology of the plant change (Shahid, 2011). Therefore, precision irrigation management will become more critical to achieving stable yields in constantly changing conditions.

Moreover, precision agriculture solutions have to overcome specific challenges closely related to the technologies being used. For vision-based applications, the constantly changing illumination conditions outdoors represent a significant factor contributing to variability in image quality. In open fields, illumination can vary from direct solar light to diffuse light caused by clouds, from sunrise to sunset, and from sloping winter to straight summer sunlight(Ruiz et al., 2009). Such variations can potentially modify the appearance of objects or the overall content captured in the image(Silwal et al., 2021).

Finally, plant growth is an additional source of uncertainty and variability as it is not restricted to strict guidelines, with the predominance of branch and leaf shade in agricultural environments posing an additional challenge(Sun et al., 2023).

1.3 Agricultural inputs and their effects

Agriculture relies on various inputs to sustain production and growth in order to meet the increasing population's needs. These inputs include fertilisers, irrigation water, seeds, pesticides, and energy for farm machinery and equipment use, with each input category showcasing constant technological developments (Sheahan and Barrett, 2017). This dissertationhas focused on fertilisation and irrigation, and as a result, these two inputs have been further analysed in the next sub-sections.

1.3.1 Fertilisation

Fertilisers are maybe the most critical input to increase yield, with studies reporting a coefficient of 7.85, which means that a 1 kg/ha increase in fertiliser is associated with higher yields of nearly 8 kg/ha; this coefficient is the highest among other agricultural inputs(McArthur and McCord, 2017). Besides directly increasing yield, fertilisers are linked to improved quality (Siavoshi et al., 2011) and enhanced growth (Nkaa et al., 2014). However, the most important fertiliser pollution concerns are associated with nitrogen-based fertilisers. This type of pollution stands out as a significant environmental concern in the 21st century, playing a role in air and water pollution, climate change, and stratospheric ozone depletion with agriculture being one of the predominant sources of it (Kanter et al., 2015). Additionally, the per capita nitrogen is not going to decrease in the near future. On the contrary, conservative projections estimate that it will remain unchanged, while high projections estimate an increase of more than 33% by 2050 (Lim et al., 2021). It, therefore, becomes crucial to optimize fertiliser use and increase efficiency in order to maintain and even increase yields while at the same time making sure to minimize the environmental footprint of their use. Studies suggest that in some applications fertiliser use efficiency can be as low as 0.60, indicating that, on average, half of the fertiliser utilized is excessive(HU et al., 2019). On the bright side, knowledge of potential crop Nitrogen demand could reduce fertilisation rates by 3 to 10%. Meanwhile, site-specific management could lead to substantial reductions without yield loss in various cropping systems, thus increasing profitability and environmental quality.

1.3.2 Irrigation

As mentioned earlier, the population increase and the improvement of living standards will result in a sharp increase in food demand during the following decades, raising questions about food security. The majority of this increase in food production will be covered by irrigated agriculture (Playán and Mateos, 2006). Irrigation provides water, which is vital for plant growth, with water stress being linked to yield reduction by

diminishing crop growth of canopy and biomass (Marutani and Cruz, 1989). As a result, irrigation helps agriculture achieve higher production, which is linked to lower crop failure risk, while improving quality. Moreover, irrigation has (i)allowed farmers to switch from low-value subsistence production to high-value market-oriented production, (ii)enabled smallholders to adopt diversified cropping patterns, and (iii) made food more available and affordable to people experiencing poverty (Hussain and Hanjra, 2004). However, over-irrigation could have the opposite effects, making irrigation planning crucial (Yuan et al., 2003).

Despite the importance of irrigation, special attention should be paid to its optimization, as water resources are finite and there is competition between agricultural, industrial, and urban consumers, making it an expensive input (Sarwar et al., 2010).Despite the cost, water scarcity and environmental concerns also necessitate the reduction of water input per irrigated area unit. Modern agricultural solutions promise to help achieve that goal by optimizing irrigation and increasing water productivity (Playán and Mateos, 2006).

1.4 Precision Agriculture

"Precision agriculture" or "smart farming" is a farming management strategy that makes use of data, communication technologies (ICTs) and equipment such as sensors, drones and GPS to increase agriculture's productivity and efficiency(Linaza et al., 2021).

Precision agriculture methodologies focus on gathering information on crops and their surroundings through the use of proximal and remote sensors, global positioning systems (GPS), and other technologies. Upon data collection, optimum resource and crop management practices are defined. Examples of such practices are the determination of the best time to irrigate fertilise and harvest.

Advancements of GPS technology in the 1980s led to the conception of precision agriculture, however, farmer adoption did not begin until the late 1990s and early 2000s. Nowadays, precision agriculture software is becoming more and more available, while at the same time GPS and IoT technologies are becoming less expensive. This has resulted in GPS guidance being used by 82% of agricultural retailers, and in GPS-enabled sprayer booms adoption increasing from 39% in 2011 to 53% in 2013(Franzen and Mulla, 2015). Moreover, resource optimization through precision agriculture is currently offering many advantages, the most popular of which are: i) higher crop yields; ii) decreased environmental impact; iii) enhanced profitability; and iv) increased sustainability.

Precision agriculture relies heavily on sensors, GPS, and ICTs. There is a variety of sensors on the market suitable for precision agriculture applications. These might be as simple as soil humidity sensors and colour cameras, or they can be more complex like spectral cameras and internal microchip implants for plants. Such sensors are used to gather data on crops and their surroundings, in a proximal or remote way. These data include information about crop health, yield, nutrient levels, and soil moisture. Moreover, innovations and novel technologies such as miniaturized computer components, GIS, mobile computing and automatic control have expanded the precision agriculture applications leading to a new era of increased agricultural productivity(Pathak et al., 2019). Lastly, ICTs are used for communication purposes, as well as for collection, storing, and analysing data from sensors and GPS. As previously said, ICTs range in complexity from simple networking devices to Artificial Intelligence and 5G technology, depending on the needs of the farmer and the use case.

Up until now, precision agriculture has focused on applications with a high Return On Investment (ROI), as since its early years its economic feasibility has been questioned by farmers and researchers alike (Mulla and Khosla, 2016). Some of the applications that have received the majority of attention are the following: i) crop monitoring for growth and development as well as for pest and disease detection; ii) irrigation by directing water to the plants and areas where it is most needed to improve water use efficiency; iii) fertilisation by directing nutrients to the plants and areas where they are most needed to

improve fertiliser use efficiency; and finally iv) harvesting by determining which parts of the field are ready to be harvested in terms of maturity.

Precision agriculture has gained popularity over the last year, but there are still many obstacles to overcome. The first is high costs as precision agriculture equipment can be costly to both install and operate, leading to lengthy depreciation periods and high acquisition prices. The second is that most solutions are tough to use and maintain, and to make matters worse, there is a lack of knowledge regarding such novel technologies in the agricultural sector, which makes it challenging for non-experts like farmers to accept and comprehend them. Lastly, because of the fact that precision agriculture produces vast volumes of data that are challenging to manually process and interpret, it heavily relies on the use of Artificial Intelligence (AI) systems to perform analysis quickly and effectively. AI technologies with a focus on agriculture are still lacking compared to other industries while at the same time farmers are not familiar with them, thus leading to further adoption difficulties and mistrust. On the bright side, AI breakthroughs are already speeding up development of such solutions (Redhu et al., 2022). A general overview of the precision agriculture cycle is presented in the figure below.

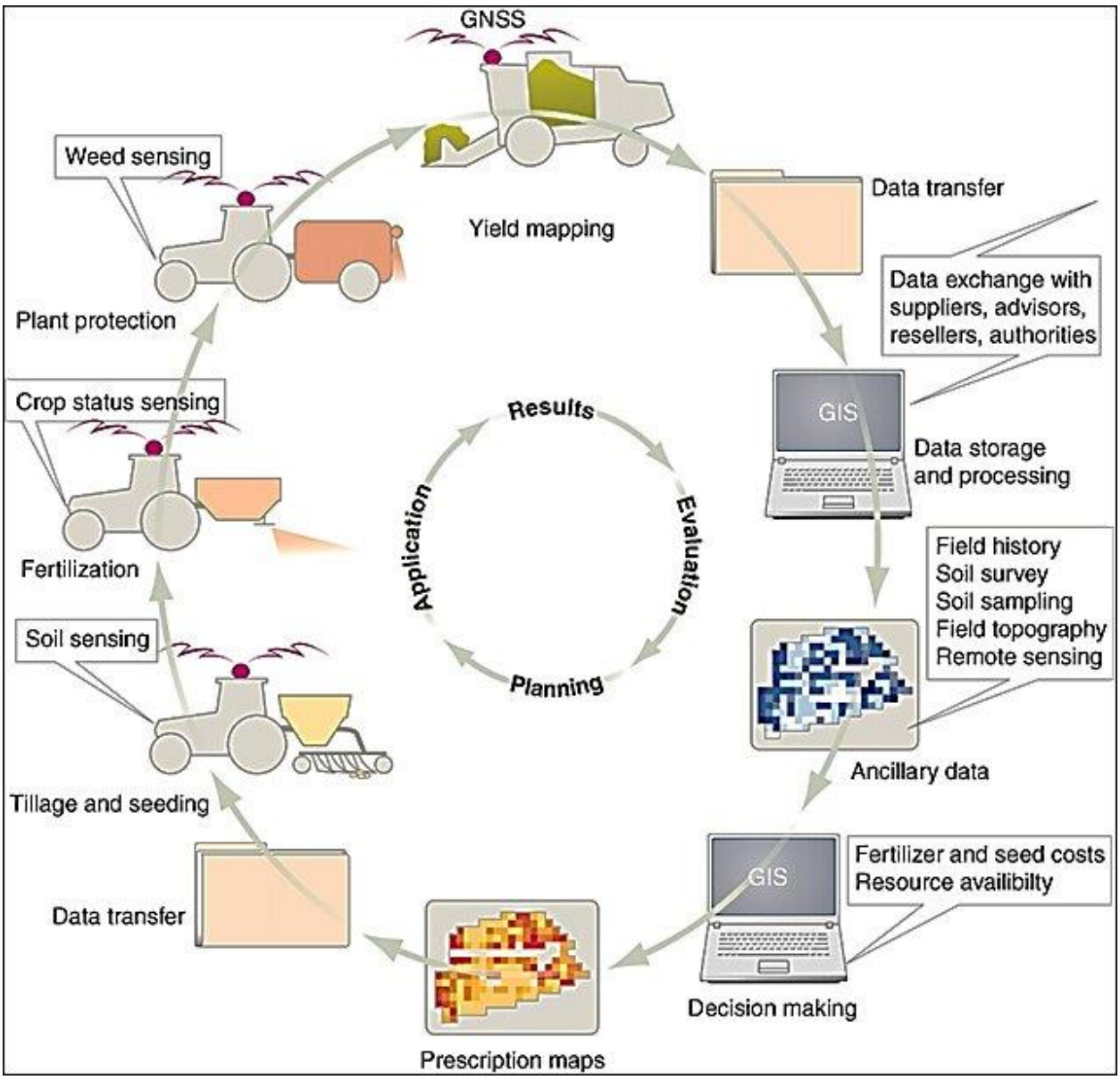

Figure 1. Precision agriculture cycle, Source:(Gebbers and Adamchuk, 2010)

1.5 Artificial Intelligence

Intelligence is the ability to learn, understand, solve problems, and make decisions. Artificial Intelligence (AI) aims to enable machines to perform tasks requiring intelligence as if performed by humans(Boden, 1980). Artificial Intelligence can, in turn, be divided into smaller subsets, namely Machine Learning and Deep Learning (Figure 2).

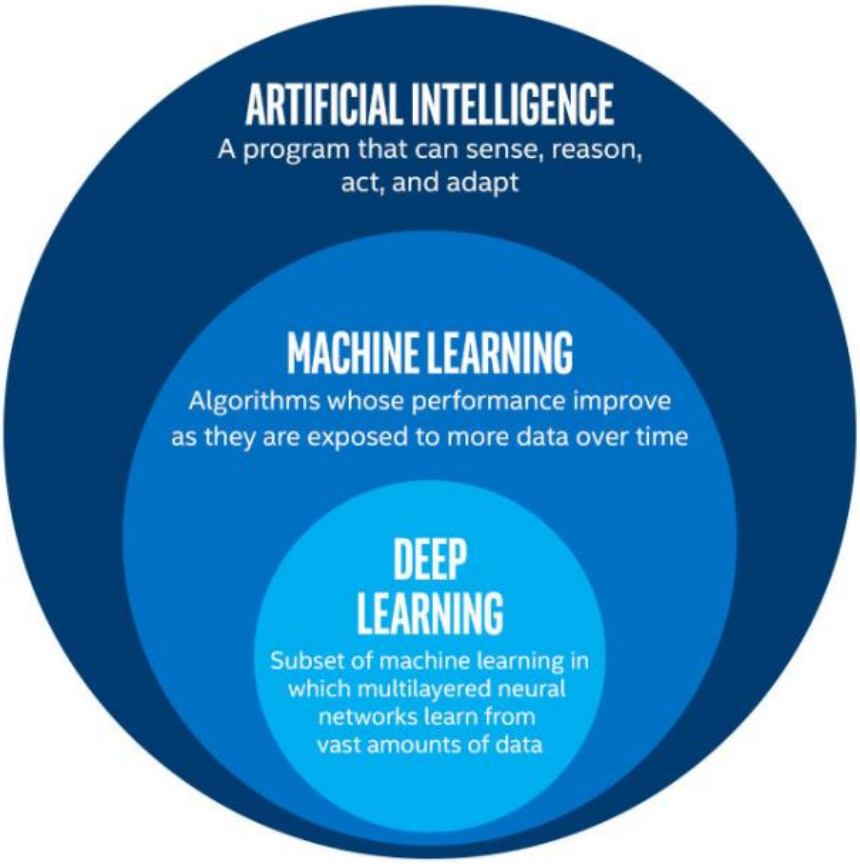

Figure 2. Subsets of AI. Source: towardsdatascience

The term "machine learning" refers to a broad category of approaches and strategies that systems use to learn from data and become more efficient at a task. The main concept of machine learning is to facilitate computers to discover patterns and come to conclusions or forecasts without having to be specifically trained to do so. AutoML, a subset of machine learning, is the application of automated tools and methods to automate the process of learning with the aim of increasing accessibility to machine learning for people with little to no background in software engineering. It aims to do so by helping them choose and configure the best algorithms for a given use case, as well as automate a number of steps in the machine learning pipeline, including feature selection, model selection, hyperparameter tweaking, and deployment.

Deep learning, another branch of machine learning, focuses specifically on neural networks with numerous layers, also referred to as deep neural networks. By attempting to mimic the structure of the human brain, deep learning algorithms enable machines to automatically learn and represent data in hierarchical levels. Deep learning can automatically learn complex features and representations from raw data, and therefore can replace laborious feature engineering, which has led to its rise in popularity. Common deep learning designs include Convolutional Neural Networks (CNNs) for image processing and Recurrent Neural Networks (RNNs) for sequential data. In conclusion, deep learning is a subset of machine learning that uses deep neural networks, machine learning is the broad area that encompasses many learning methodologies, and AutoML is a collection of tools and methods intended to automate and streamline the machine learning process.

At this point it is crucial to note that AI like all other technologies can be categorized into several eras, each marked by significant advancements and changes in AI research and technology. The commonly recognized eras of AI include(Council of Europe, 2024):

- Birth of AI (1940s-1960s): The period between 1940 and 1960 was strongly marked by rapid technological developments and the ambition to merge the functioning of machines and humans. However, hardware limitations at the time made the use of computer language difficult. Despite these limitations some foundations still present today were developed. Examples are LTM (logic theorist machine) which was developed as early as 1956.
- Expert systems (1980s-1990s): The development of the first microprocessors at the end of 1970 led AI to take off, thus leading to the golden age of expert systems. Examples of such systems are the DENDRAL (expert system specialized in molecular chemistry) developed by MIT in 1965 and the MYCIN (system specialized in the diagnosis of blood diseases and prescription drugs) developed at Stanford University in 1972.
- Data and computing power boom (2010-Present): The main factors leading this era are the easy access to massive volumes of data and the vast improvements in the efficiency of computer graphics card processors which accelerated the calculation of learning algorithms.

These eras represent a broad overview, and it is essential to note that AI is continually evolving, with ongoing research and developments shaping its trajectory.

1.5.1 Machine learning

Machine learning (ML) is one of the largest subsets of AI with three main areas of focus: 1) task-oriented studies aimed towards analysing learning systems to increase their performance in a predetermined set of tasks, 2) cognitive simulation, which involves the investigation and computer simulation of human learning processes and 3) theoretical analysis which is the theoretical investigation of possible learning methods and algorithms independent of the application domain(Michalski et al., 2013). Machine learning has many applications in modern life; examples are social network content filtering, object detection by autonomous vehicles, and speech to text transcription(LeCun et al., 2015). Moreover, its application is not limited to a single domain; on the contrary, ML algorithms are used in multiple scientific fields, for example, genetics and genomics(Libbrecht and Noble, 2015),medicine(Kourou et al., 2015), remote sensing (Belgiu and Drăguţ, 2016) and agriculture(Gao et al., 2019). A short definition of machine learning would be the use of algorithms to learn from existing data and make predictions about unseen data (Figure 3).

Figure 3. Machine learning representation. Source: dltlabs

The two main machine learning configurations are supervised and unsupervised learning (Figure 4). In supervised learning, the algorithm is presented with the input variables(x) and an output variable(y) and is asked to learn the mapping function from the input to the output y=f(x). In supervised machine learning, the algorithm is provided with known quantities to support future judgments and is usually used for classification problems where the association between input and output labels is sought or for regression problems where the aim is to map an input to a continuous output. For the classification problem, the goal is to create a mapping function (f) from input variables (x) to discrete output variables (y) such as 'apple' or 'banana,' 'green' or 'red.' In regression problems, the algorithm needs to create a function(f) that maps input variables(x) to a continuous output variable(y), such as the 'salary' or 'weight' of a person.

On the other hand, unsupervised learning is a technique used when only the input data (x) are available with the aim of finding patterns in them. The algorithm tries to model the structure or distribution in the data in order to learn more about them and to infer patterns from a dataset without labelled outcomes. An example of unsupervised learning could be a market survey. The responses are gathered, and the market manager can choose whether to cluster the customers using their demographic variables (age, sex, education level, income level) or to cluster the responses according to changes in price.

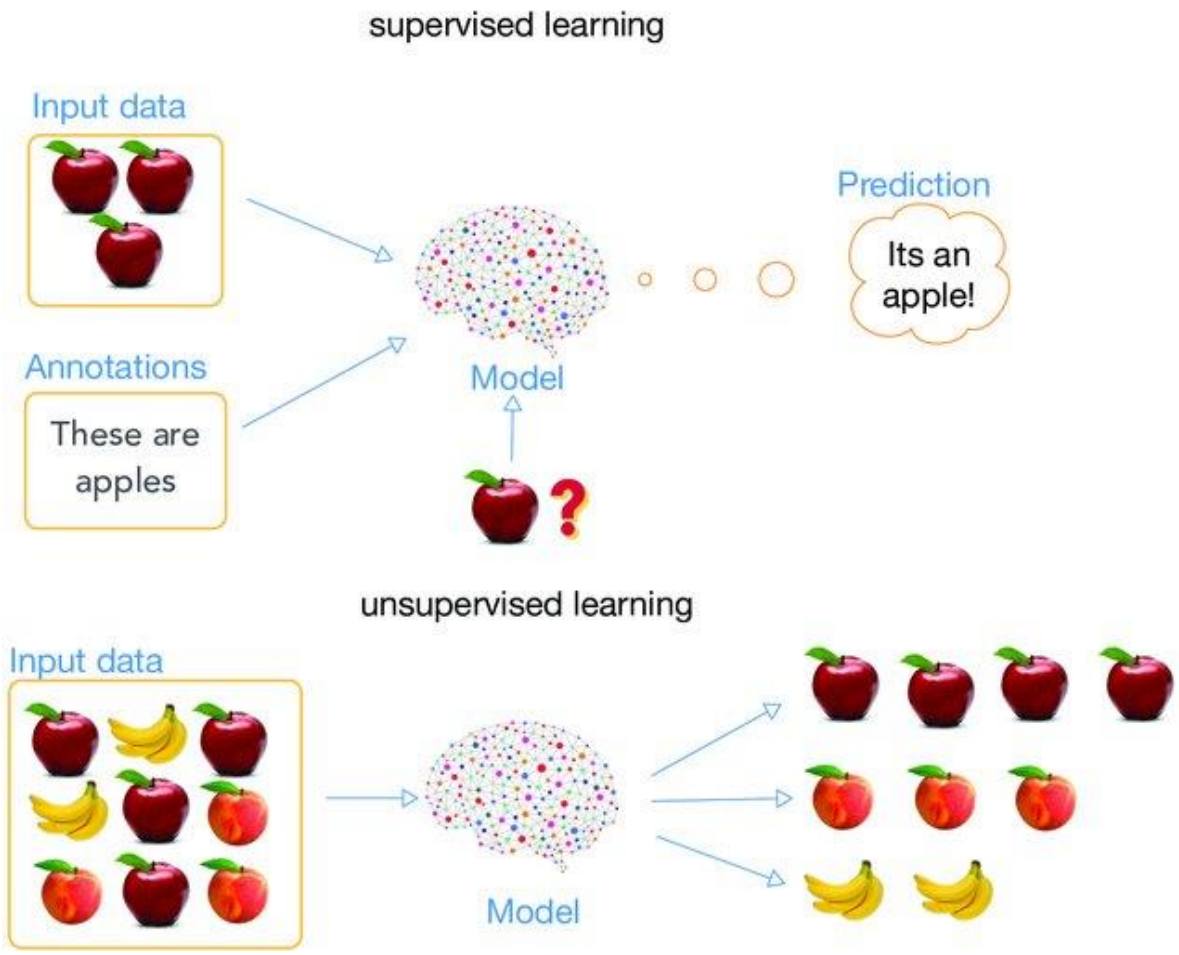

Figure 4. The two types of machine learning techniques. Source: (Ma et al., 2018)

1.5.2 Automated Machine Learning (AutoML)

Automated Machine Learning (AutoML) is a paradigm-shifting technique in the field of Artificial Intelligence that primarily focuses on supervised learning tasks, such as regression and classification and addresses the challenges and complications associated with putting machine learning models into real-world applications. As the need for machine learning solutions grows across a variety of industries, AutoML emerged as a

critical tool that democratizes the application of advanced analytical techniques through model creation automation. It essentially makes use of automated tools and procedures to streamline and optimize the entire machine-learning pipeline. Without AutoML, every step in a typical data science pipeline, such as data preprocessing, feature engineering, and hyperparameter optimization, is executed manually by machine learning experts. On the other hand, using AutoML allows a simpler development process where a few lines of code can generate the code necessary to begin developing a machine learning model(IBM, 2024). Thus, making previously labour-intensive procedures less complicated and ultimately making machine learning more accessible to a larger audience, including those with little to no prior experience in the field.

Every AutoML solution consists of several key components (Figure 5):

1. Feature Preprocessing: AutoML tools automate transformation, and normalization of raw data, ensuring it is appropriately prepared for the modelling phase.
2. Feature Selection: Leveraging advanced algorithms, AutoML assists in the automatic extraction of relevant features from raw data, reducing the need for manual intervention and domain expertise.
3. Model Selection: AutoML algorithms intelligently explore a range of machine learning models, selecting the most suitable architecture for a given dataset and problem.
4. Feature construction: Creating new features from existing data, which designed to be more informative helping the model learn better and ultimately being more accurate.
5. Parameter optimization: The optimization of model hyperparameters, a crucial and often intricate task, is automated through AutoML, enhancing the performance of the selected model.

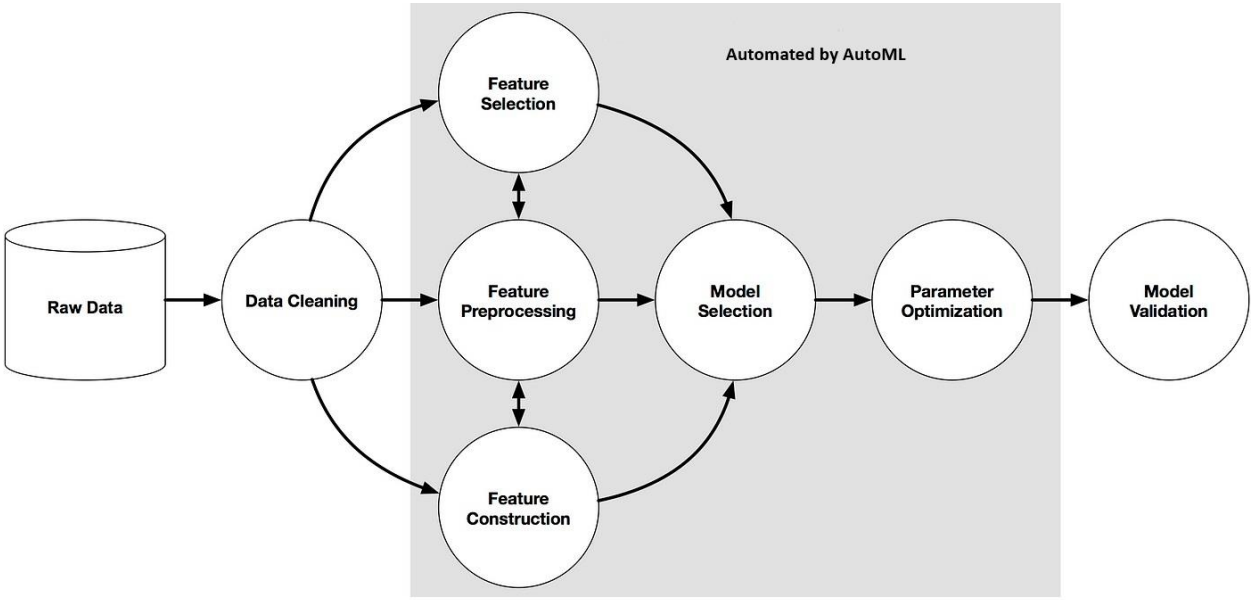

Figure 5. AutoML pipeline, Source: https://towardsdatascience.com/automated-machine-learning-d8568857bda1

As a result, AutoML holds a significant promise in democratizing machine learning by lowering the barriers to entry for practitioners. Its automated nature reduces the dependency on domain-specific expertise, allowing stakeholders to use machine learning more efficiently and cost-effectively. Moreover, by accelerating the model development lifecycle, AutoML enables rapid prototyping and iteration, which is crucial in dynamic environments where timely decision-making is imperative.

To sum up, AutoML stands at the forefront of advancing the application of machine learning techniques, offering a comprehensive and accessible solution to practitioners across diverse domains, including agriculture. As the field continues to evolve, the integration of AutoML into standard data science workflows promises to revolutionize the landscape of predictive analytics, empowering experts and non-experts alike to unlock insights from their data with unprecedented efficiency.

1.6 Artificial Intelligence in agriculture

Artificial Intelligence (AI) is rapidly transforming various industries, and agriculture is no exception. With the global population expected to reach 9.7 billion by 2050 and the demand for food increasing at an unprecedented rate, AI-powered solutions hold the promise to increase food quality and production by enhancing agricultural productivity, sustainability, and resource efficiency(Javaid et al., 2023). As mentioned earlier, precision agriculture is a farming management approach that uses various sensors and different data types to improve the efficiency and effectiveness of agricultural production. However, as the number of data points and the complexity of the data gathered increases, humans cannot cope with data processing. One of the most promising solutions is the use of AI algorithms that can quickly analyse vast amounts of data from sensors, drones, and satellites to identify areas that require specific attention. The use of AI, therefore, causes agriculture to shift from empirical decision-making to data-driven decision-making, allowing farmers to make informed decisions that maximize resource use and minimize environmental impact. The major subdomains of agriculture that AI techniques have found application are the following (Bannerjee et al., 2018):

- General crop management
- Pest management
- Disease management
- Agricultural product monitoring and storage control
- Soil and irrigation management
- Weed management
- Yield prediction

AI applications have increased in popularity, with numerous academic and commercial solutions having been presented over the years, with a systematic review identifying more than 150 papers based on the existing automation applications in agriculture from 1960 to 2021(Wakchaure et al., 2023).

1.6.1 Artificial Intelligence for water stress detection

As mentioned in the previous chapters, irrigation is one of the most critical agricultural inputs for achieving high-quality products and high yields; as a result, it has drawn the attention of many AI researchers. Numerous publications can be found using a variety of sensors and cameras as well as various AI algorithms.

Estimating water stress using satellite imagery is one of the most popular approaches, as it allows for large spatial coverage with minimal manual labour. AI is crucial in this approach as it allows for fast data processing. Various ML and AI algorithms

have been used, such as Genetic Algorithms (Hassan-Esfahani et al., 2015), and Gaussian mixture models (Huang et al., 2007)(Sun et al., 2017), all with promising results.

Another source of data is UAV, where multispectral imagery is used to produce various vegetation indices that, in turn, are used for water status estimation using AI algorithms such as Artificial Neural Networks (ANN) (Romero et al., 2018)(Poblete et al., 2017). Besides multispectral data, colour (RGB) data captured from UAV systems have also been used to characterize water stress in combination with ANN(Chandel et al., 2022). However, as AI algorithms become more sophisticated and computational power cheaper, more complex approaches, such as fusing thermal and RGB UAV-captured images, are being investigated (Aversano et al., 2022).

A completely different approach that has received much consideration is using AI algorithms to estimate Evapotranspiration (ET) (Virnodkar et al., 2020). Once again, ANN and SVM are among the most commonly used algorithms (Dou and Yang, 2018), with additional state-of-the-art algorithms such as extreme learning machine (ELM) and adaptive neuro-fuzzy inference system (ANFIS) being tested (Dou and Yang, 2018).

Finally, a different approach towards identifying water stress was to use canopy temperature calculations (Andrade et al., 2018) or Crop Water Stress Indices (CWSI) together with ML algorithms such as Bayesian regularized neural network (BRNN), SVM with radial basis function (RBF) kernel, least absolute shrinkage and selection operator (LASSO), ridge regression, generalized linear model (GLM), multivariate adaptive regression splines (MARS), conditional inference tree (CIT), RF, eXtreme gradient boosting and cubist (Xu et al., 2018).

1.6.2 Artificial Intelligence for fertilisation

Fertilisation is another crucial input to achieve high quality and yield. However, overuse of fertilisers can lead to environmental pollution. Therefore, researchers have focused on the problem of fertilisation determination and quantification with nitrogen((Cilia et al., 2014)(Quemada et al., 2014)(Argento et al., 2021)(Bagheri et al., 2013)(Link et al., 2004)(Yi et al., 2007)(Lammel et al., 2001)(Basso et al., 2016)) and phosphorus (Siedliska et al., 2021) being the most commonly studied nutrients. Moreover, studies focusing on fertilisation intensity and not on specific nutrients have also been conducted (Hollberg and Schellberg, 2017)(Papadopoulos et al., 2023).

UAV and satellite imagery are widely used as data acquisition platforms for fertilisation applications, with data processing approaches including various AI agents. An example of this is the use of spectral and vegetation indices in conjunction with the Spectral angle mapper classifier (SAM)(Bagheri et al., 2013), quadratic linear regression

(Argento et al., 2021), Linear regression(Quemada et al., 2014) to estimate nitrogen status and create fertilisation maps (Cilia et al., 2014).

Despite the sizeable spatial resolution of satellite and UAV imagery, proximal sensing appears to be the most common approach when determining fertilisation levels. Various sensors have been used depending on the wavelengths of interest and the environment in which the data acquisition occurred. One approach suggests the use of spectroradiometers in lab conditions combined with MLR (multiple linear regression) and ANN (artificial neural network) modelling(Yi et al., 2007). Another one, is the use of spectral proximal sensing and supervised classification (Backpropagation Neural Network, Random Forest, Naive Bayes, and Support Vector Machine) (Siedliska et al., 2021). Moreover, spectral sensors have also been mounted to tractors as an alternative to UAVs or satellites to cover large areas (Link et al., 2004)(Lammel et al., 2001), while the Vegetation indices approach has also been tested for proximal remote sensing applications in combination with a random forests classifier(Hollberg and Schellberg, 2017).

1.6.3 Most common AI algorithms in agriculture

Building upon the previous chapters, it becomes apparent that AI is a crucial part of modern precision agriculture. As a result of the rapid growth of precision agriculture solutions, numerous systematic reviews have emerged in recent years, trying to encapsulate and present the latest trends in the use of AI and ML in agriculture. Based on those reviews, some of the most important and widely used algorithms are showcased below.

Table 1. Summary of AI algorithms used in agriculture

Reference (Review)	Title	Algorithms presented	Conclusions
(Gupta et al., 2022)	Analysis of Some Popular AI & ML Algorithms Used in Agriculture	• ANN • GA • Fuzzy logic • SVM • KNN	• AI promotes agricultural growth • SVM provides more accurate results
(Wakchaure et al., 2023)	Application of AI techniques and robotics in	• Fuzzy logic • Genetic algorithm	• FL, ANN, and GA are widely

	agriculture: A review	• ANN • Particle swarm optimization • Ant colony optimization • Firefly algorithm • Bat algorithm • Artificial potential field approach • Artificial bee colony algorithm • Harmony search algorithm • Cell decomposition • Simulated annealing.	accepted in the field of agriculture • Most robot applications are developed using FL, GA, and ANN
(Megeto et al., 2021)	Artificial Intelligence applications in the agriculture 4.0	• CNN • Vector quantization • Gaussian mixture models (GMMs) • SVM • Random Forest • Hidden Markov Models • Multilayer Perceptron	• SVMs and small NN are very popular

(Eli-Chukwu, 2019)	Applications of Artificial Intelligence in Agriculture: A Review	• Median of Medians • Fuzzy logic • ANN • Genetic algorithm • SVM	AI improved the agricultural sector Below the average impact compared to its impact in other sectors .
(Bannerjee et al., 2018)	Artificial Intelligence in Agriculture: A Literature Survey	• Expert systems • ANN • Fuzzy logic • K-means	• Since 1990, ANNs and fuzzy inference systems have been the most popular. • Recently, hybrid systems, neuro-fuzzy or image processing coupled with ANN are gaining popularity.
(Bhat and Huang, 2021)	Big Data and AI Revolution in Precision Agriculture: Survey and Challenges	• ANN • Random Forest • Decision Trees • SVM • CNN	• CNNs have been gaining popularity

Based on the above findings, the most popular AI algorithms in agriculture are Artificial Neural Networks (ANN) and Support Vector Machines (SVM). There are several reasons why these algorithms are chosen on many occasions.

ANNs offer the following advantages and characteristics:

- Fault Tolerance: Even if one or more of their cells are faulty, ANNs can still operate. The spread nature of information storage throughout the network is the cause of this fault tolerance.

- Parallel Processing: ANNs can process information in parallel, they can handle numerous tasks at once. This enables them to effectively manage challenging jobs.
- Training and Learning: ANNs are capable of learning from events and inferring patterns from them to make judgments. They can adjust to new data and withstand extended training periods.
- Gradual Corruption: ANNs do not corrode or malfunction right away. Rather, individuals experience a slow deterioration over time that is controllable.
- Speed: When it is important to quickly evaluate the taught target function, artificial neural networks (ANNs) come in handy. They are able to make decisions instantly and assimilate information quickly.
- Effective Visual Analysis: Similar to how humans interpret images, ANNs can conduct efficient visual analysis. They can therefore be applied to applications such as picture recognition and classification.
- Processing Unorganized Data: An important advantage in today's data-driven environment is that ANNs can handle and organize disorganized data. They are fast and effective at organizing and classifying data.
- Adaptive Structure: Because of their adaptable structure, artificial neural networks (ANNs) can change how they behave depending on the task they are assigned. They are adaptable and helpful in a variety of applications because of this.

While SVMs, in turn, come with their advantages:

- Effective in High-Dimensional Spaces: SVMs perform well in high-dimensional feature spaces, making them suitable for tasks where the data may have many features.
- Kernel Trick: SVMs can use the kernel trick to transform the input data into a higher-dimensional space, allowing them to handle non-linear relationships between features.
- Global Optimization: SVMs aim to find the hyperplane that maximally separates different classes, leading to a global optimization objective. This contributes to robust and well-generalized models.
- Margin Maximization: SVMs focus on maximizing the margin between different classes, which helps achieve better generalisation and resilience to noise in the data.
- Versatility: SVMs can be adapted for classification and regression tasks, making them versatile for various machine-learning applications.

In summary, ANNs are popular for their ability to process information in parallel and handle unorganized data, while SVMs are valued for their effectiveness in high-dimensional spaces, global optimization objectives, and versatility in handling different

types of tasks. However, both algorithms require deep knowledge and extensive experience to optimize them and achieve good performance results.

1.7 Spectral imaging

Spectral imaging is an imaging technique employed that involves acquiring and analysing spectral data across a range of wavelengths. This method provides a detailed and comprehensive understanding of the electromagnetic spectrum's interaction with a given object or scene. By capturing information beyond what is visible to the human eye, spectral imaging enables researchers to discern subtle differences in material composition, identify specific chemical components, and gain insights into biological specimens' structural and functional characteristics. The combination of such sensors with AI algorithms facilitates the extraction of valuable spectral signatures, contributing to advancements in various fields, including agriculture. The precision and versatility spectral imaging provides a significant promise for better understanding complex phenomena and fostering innovative solutions across various scientific domains.

When referring to spectral imaging, both hyperspectral and multispectral sensors are included. The exact thresholds for what is considered "multispectral" or "hyperspectral" can vary, but the key distinction lies in the density and granularity of the spectral information captured. Multispectral sensors can generally capture a limited number of discrete bands, often ranging from a few (e.g., 3 to 10) to several dozen. On the other hand, hyperspectral sensors can usually capture a much larger number of spectral bands, ranging from tens to hundreds or more.

Finally, it is worth mentioning that although the modern practice is to use aggressive designations such as multi and hyper added to “spectral imaging” to characterize the number of wavelength bands, it is proposed to avoid using such vague adjectives and use scientifically sound terminology instead, such as “imaging spectroscopy” or “spectral imaging”(Polder and Gowen, 2021).

1.7.1 Multispectral Imaging

As mentioned above, there is a common distinction between multi- and hyper-spectral imaging, with a multispectral image consisting of limited specific wavelength ranges. Multispectral imaging aspires to allow the fast acquisition of spatial and spectral information, which can be processed by simple image processing and decision-making algorithms. The increased efficiency compared to other spectral imaging methods results from the reduction of the total size of the data, achieved with relatively low spatial resolution (capturing selected wavelengths). Three are the main capturing/scan methods: 1) The point scan, 2) The line scan, and 3) The area scan. Because of the fast image capture constraint, the point-scan method is not used in practice, as scanning along two dimensions is time-consuming. The line-scan and area-scan methods (Figure 6) are used with minor adjustments, while they can be both tuned to capture images at selected wavelengths. For the line-scan method, specifying the positions of all the useful tracks

along the spectral dimension of the detector allows for the collection of fewer wavelengths. As a result, only the information from the specified tracks is collected, reducing the capturing time and the size of every line-scan image (y, λ). On the other hand, the area-scan method reduces capturing time by simultaneously allowing single-band image capturing at multiple selected wavelengths (Qin et al., 2013).

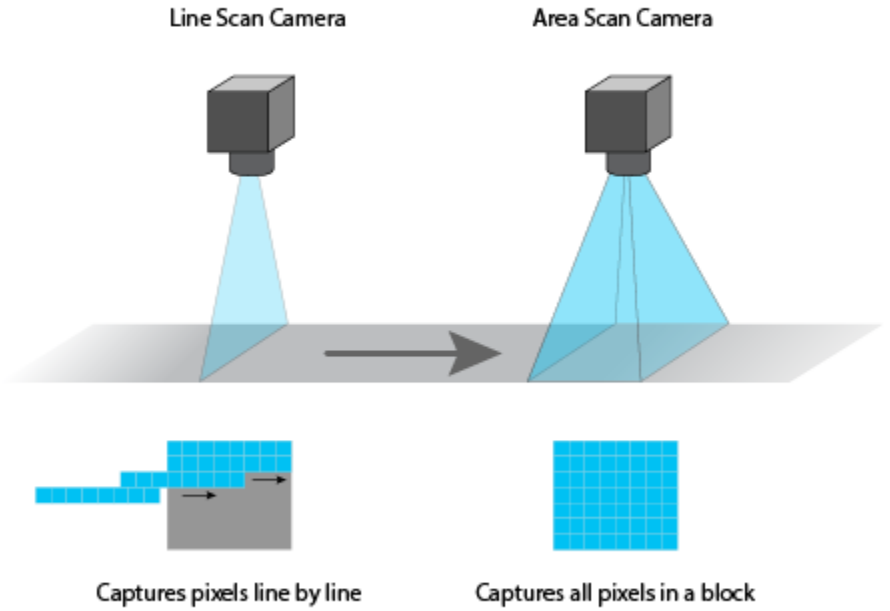

Figure 6. Line scan vs Area scan cameras. Source: Fainstec

1.7.2 Hyperspectral Imaging

Compared to multispectral imaging, hyperspectral imaging allows the capturing of more extensive spatial and spectral information(Table 2) and is one the most promising areas in remote sensing, proximal sensing, close-range sensing, etc. (Gowen et al., 2007). This promise enabled the development of improved optics and sensor technologies that have not only improved the spatial and spectral resolution of those cameras but also reduced their size and cost(Monteiro et al., 2007).

Table 2. Comparison of RGB imaging near-infrared spectroscopy (NIRS), multispectral imaging (MSI), and hyperspectral imaging (HSI).

Feature	RGB imaging	NIRS	MSI	HSI
Spatial information	✓		✓	✓
Spectral Information		✓	Limited	✓
Multi-constituent information	Limited	✓	Limited	✓
Sensitivity to minor components			Limited	✓

A hyperspectral image can be described as a cube consisting of two spatial dimensions and one wavelength dimension(Lu and Chen, 1999). The wavelength dimension is specified by hundreds of contiguous wavebands (Figure 7). This results in each image pixel being a column vector with dimensions equal to the number of wavebands. More spectral bands, therefore, result in richer spectral information. Continuing the multi- and hyper-spectral comparison, as the size of the third dimension is considerably larger in the hyperspectral images, it can be argued that each pixel vector contains more spectral information, an attribute crucial for data analysis(Gowen et al., 2007). However, using more information-dense images comes with a trade-off: computational time increase. Primarily, this is counteracted by the continuous advancements in computer science, both hardware, and software, that have enabled the use of information-rich images without sacrificing much time. Despite the advancements, the accuracy speed trade-off will always exist, with increased accuracy coming from a sacrifice of speed(Huang et al., 2017).

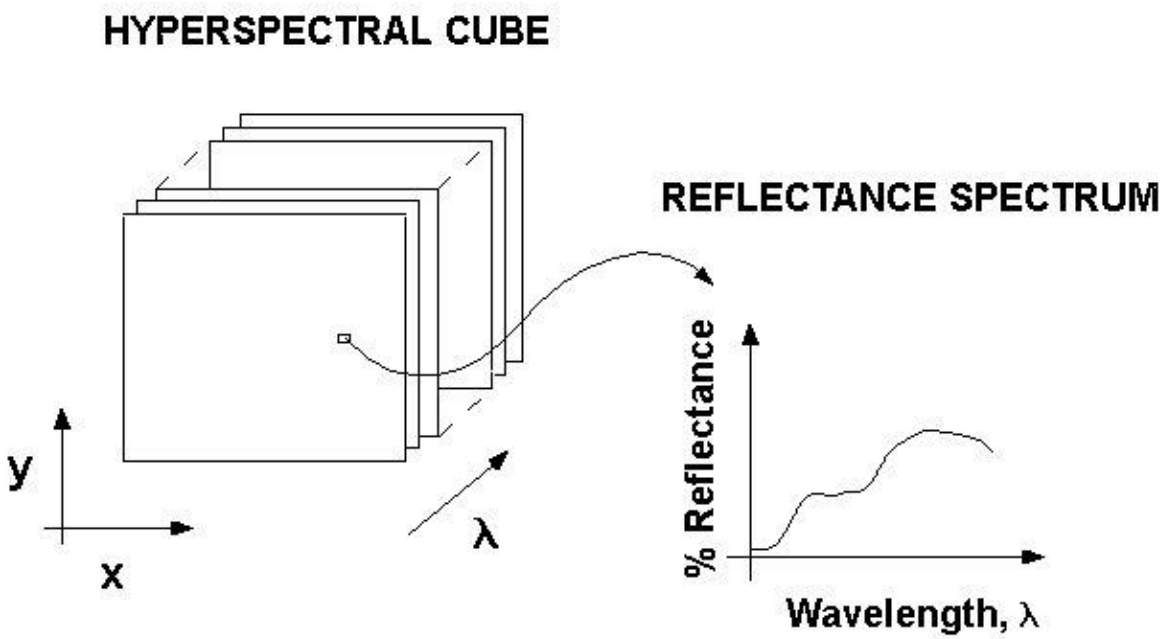

Figure 7. Schematic representation of hyperspectral image cube Source:(Stamatas et al., 2003)

Similarly to multispectral imaging, three are the main approaches for obtaining hyperspectral images: two spatial scan methods (point-scan and line-scan) and one spectral-scan method(area-scan) (Figure 8). In the point-scan or whiskbroom method, a single point is scanned along the two spatial dimensions (x, y) by moving the object or the sensor. The single spectrum of each pixel is then captured using a point detector-equipped spectrophotometer. Using this method, the hyperspectral image is assembled pixel by pixel. The second approach is the line scan or pushbroom method, an extension of the point-scan method. Using this method for every spatial point in the linear field of view (FOV), a slit of spatial information as well as complete spectral information is acquired simultaneously. This results in a two-dimensional image with one spatial dimension (y) and a spectral dimension (λ) captured each time. The hypercube is progressively completed as the slit is scanned in the direction of motion (x). Both scan methods provide good quality results but have very low efficiency, with long integration times being a prerequisite to obtaining a full hyperspectral image. The last method is the area-scan or band sequential method, which captures a two-dimensional single-band greyscale image (x, y) with full spatial information at once. The hypercube is completed by stacking single-band images as the scan is performed. One of the main benefits of this method is that it does not require movement between the object and the sensor(Qin et al., 2013).

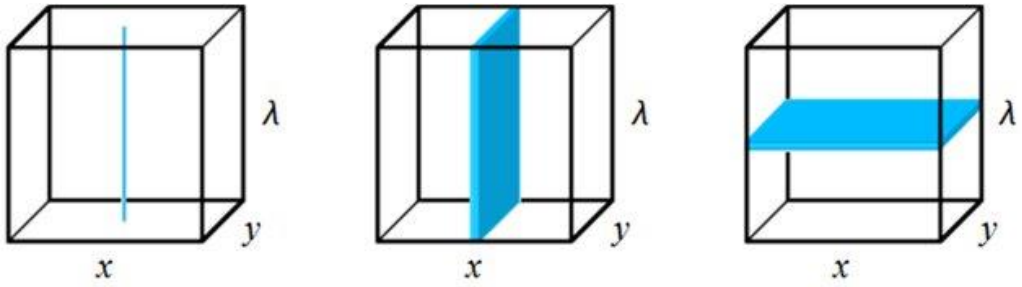

Figure 8. Methods of spectral image acquisitions with left image: point scan, middle image line scan, and right image area scan. Source: (Qin et al. 2013)

Despite the three different methods used to build a hyperspectral cube, a hyperspectral device usually consists of the following parts: a monochrome charged couple device (CCD) or complementary metal oxide semiconductor (CMOS) image sensor which is used to acquire the spectral image, and a dispersive means (prism or grating) which is integrated into the optical system. The drawback of dispersive means is that the image is analysed per line, and the mechanics are integrated into the optical train. Optical band-pass filters, tunable or fixed, e.g., rotary filter wheels, liquid crystal tunable filters, and acousto-optic tunable filters, are a solid alternative (Schelkanova et al., 2015). Using optical band-pass filters to acquire hyperspectral images means the spectrum must be scanned in steps. Finally, a third method to obtain the spectrum of a light source is the Fourier transform spectroscopy. In this measurement technique, spectra are collected based on measurements of the coherence of a radiative source, using time-domain or space-domain measurements of electromagnetic radiation or another type of radiation(Pisani and Zucco, 2009).

1.7.3 Spectral imaging in agriculture

As mentioned earlier, spectral imaging in agriculture has emerged as a powerful tool, leveraging advanced technology to enhance various aspects of crop management and monitoring. Recapping: this imaging technique involves capturing and processing a broad spectrum of wavelengths, providing detailed information about the composition and health of crops.

One of the primary applications of spectral imaging in agriculture is precision farming. By analysing the reflected light from crops across different spectral bands, farmers can gain insights into the nutritional status, moisture levels, and overall health of plants. This enables precise and targeted interventions, such as optimized irrigation and fertilisation, ultimately improving crop yield and resource efficiency. On this topic (Ruett et al., 2022) investigated the applicability of spectral imaging for determining the vitality of shoots and roots in ornamental plant production, while (Kim et al., 2010) used spectral

imaging for the detection of water stress in apple trees and (Williams et al., 2023) for differentiating between biotic and abiotic stress in raspberry plants.

Moreover, spectral imaging plays a crucial role in disease detection and pest management. The unique spectral signatures of diseased plants or infestations can be identified, allowing for early detection and timely intervention. Disease detection has been among the most researched topics as each disease requires a different treatment and, if left untreated, can cause a reduction in yield and quality and, on some occasions, put at risk human health (e.g., Aflatoxin produced by the fungi *Aspergillus flavus* and *Aspergillus parasiticus* in crops such as corn). According to the Scopus statistics, there are 412 relevant papers from 2005 to 2020 where 'plant disease' and 'hyperspectral' are used as keywords for the search (Cheshkova, 2022). The focus has not been on a single crop but instead spread on multiple ones depending on each researcher's topic of interest. The following table presents a short list of publications on various crops and disease combinations to showcase the variability in crops and diseases.

Table 3. Diseases and crops where spectral imaging has been used.

Crop	Disease	Reference
Tomato	Gray mold	(Xie et al., 2017)
Wheat	Powdery mildew	(Khan et al., 2021)
Apple	Apple scab	(Gorretta et al., 2019)
Strawberry	Gray mold and anthracnose	(Zhang et al., 2023)
Tea plants	Anthracnose	(Yuan et al., 2019)
Grapevine	Ochratoxin A	(Templalexis et al., 2023)
Squash	Powdery mildew	(Abdulridha et al., 2020)
Pistachios	Aspergilus Flavus	(Mastrodimos et al., 2022)
Palm trees	BSR disease	(Lee et al., 2022)
Rice	Sheath blight	(J. Zhang et al., 2021)
Banana	Black Sigatoka	(Ugarte Fajardo et al., 2020)

Another significant benefit of spectral imaging is its capacity to assess soil health. By analysing the spectral information reflected from the soil surface, farmers can gather data on soil composition, moisture content, and nutrient levels. This information guides informed decisions on soil management practices, helping to optimize crop growth conditions and reduce environmental impact. (Jia et al., 2017) used spectral imaging to classify soil types and determine soil total nitrogen, (Nanni et al., 2021) used it to map organic matter and soil particle size, while (Haijun et al. 2017) used it to predict soil moisture.

Furthermore, spectral imaging supports crop phenotyping, which involves the comprehensive analysis of plant traits. Researchers and agronomists can utilize this technology to study and understand the genetic characteristics and variations in different

crop varieties. This information is invaluable for crop breeding programs, enabling the development of more resilient and productive plant varieties.(Bodner et al., 2018) used spectral imaging for characterizing the root system architecture, (Banerjee et al., 2020) used it to derive biomarkers for genotypic nitrogen response, while (Pandey et al., 2017) focused on the analysis of plant leaf chemical properties.

To sum up, spectral imaging in agriculture had found uses in a variety of tasks ranging from crop phenotyping and soil health assessment to precision farming and disease detection. By incorporating spectral imaging into farming methods output and resource efficiency improvements are expected as well as resilience and sustainability strengthening of modern farming systems.

1.8 CIELAB Colour space

Colour is defined as the aspect of things caused by differing qualities of light being reflected or emitted by them. It is explicitly associated with electromagnetic radiation of a specific range of wavelengths visible to the human eye. Colour is a perception of energy and specific wavelengths of light that reach our eyes. Perceived colour can vary based on a person's biology and how their brain receives signals, so two people may not see an object as the same colour(Nassau, 2024).

Colour is perceived differently by each person, like smells or sounds, so it is a subjective stimulus to which many factors contribute, such as the light source, the mood, the angle of observation, and the colour sensitivity of the observer. Despite their subjectivity, however, colours can be compared objectively as long as the conditions in which they are viewed are stable and independent of external factors. Three basic properties characterize colour:

- The hue refers to an object's primary colour and is the first criterion for distinguishing colour.

- The chromatic saturation (chroma) describes the colour's purity.

- The brightness when the hue is held constant, and the three components of the colour (red, green, blue) are reduced simultaneously while maintaining their proportion, reducing the brightness of the colour.

Various colour coordinate systems can describe the colour of an object. Some of the most widely used systems are Munsel, RGB (red-R, green-G, blue-B), and the Comission Internationale de l'Eclairage's-CIE): CIEL*a*b* (Figure 9), CIEXYZ, CIEL*u*v*, CIEYxy and CIELCh. According to the CIE, the human eye has three receptors: red, green, and blue, and all colours result from the combination of these three primary colours. The amounts of red, green, and blue required to form any colour are called tristimulus factors and are denoted respectively by the letters X, Y, and Z (Pathare et al., 2013)

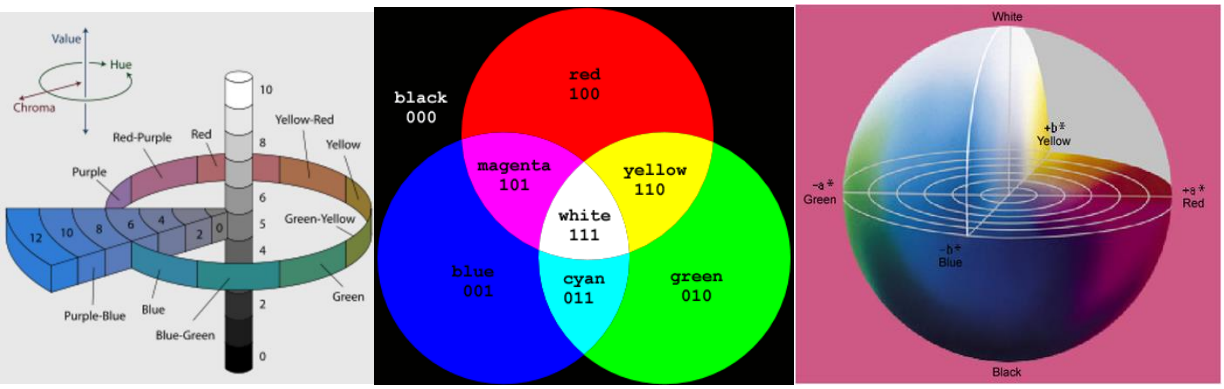

*Figure 9. Left: Munsell colour space, Centre: RGB colour space, Right: CIEL*a*b* colour space Sources: Munsell colour, 2013, Centre, Hernandez 2007, OPI 2013*

The CIE introduced the CIELAB or L*a*b* colour model in 1976, derived from the CIE XYZ colour space. It is a visually uniform colour space that best approximates the human perception of colour differences of all colour systems. Each colour is described by three factors as in the RGB colour model and is influenced by the Munsell colour system.

In CIELAB, the colour factors are called L*, a*, and b* and are represented in a 3D Cartesian coordinate system. The L* (Lightness) factor carries the information of the brightness of the image and takes values from 0 (black) to 100 (white), while the a* and b* factors, respectively, carry the following colour information without any numerical boundaries for them. Positive values of a* represent shades of red, and negative values represent shades of green. Positive values of b* represent shades of yellow, and negative values represent shades of blue(Schanda, 2007). These values can be placed in the three-dimensional CIE colour coordinate space, so that each colour-hue is characterized by a distinct point in it.

In other words, CIEL*a*b* compares a sample with a standard colour sample and performs a numerical determination based on their colour differences. The difference in luminance L* when positive means the sample is brighter than the standard, and when negative, darker. The difference a* when positive is redder than the standard and when negative is greener than the standard. Similarly, when the difference b* is positive, the sample is more yellow, and when negative, it is bluer than the standard.

The CIEL*a*b* colour model can also be represented in a cylindrical polar coordinate system with the CIEL*C*h* model (Figure 10). Where L* is the luminance, C* is the chromatic saturation, which defines the intensity or purity of the colour, and h* is the hue, which is measured in degrees and defines the hue by taking values of 0º for red, 90º for yellow, 180º for blue-green and 270º for blue.

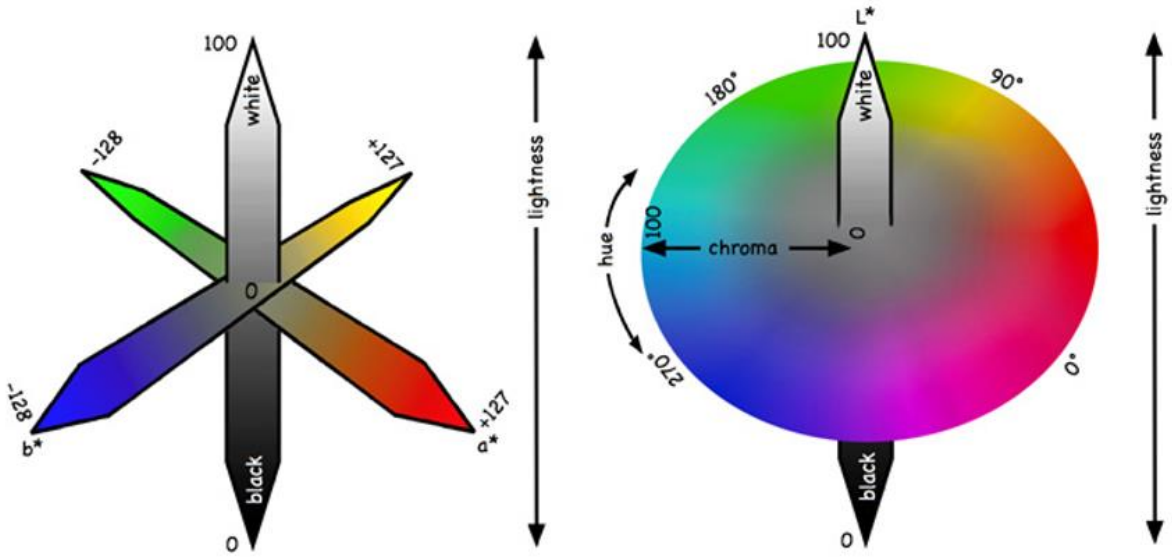

*Figure 10. CIE L*a*b* (Left) and CIE L*C*h* (Right) colour spaces, Source: OHWEB, 2013.*

When the C difference is positive, the colour has a higher density, while when negative it has a lower density than the standard. Finally, the difference h when positive will be closer to the opposite colour than the standard, for example, for a red sample, the colour will be bluer than the standard.

Each of the above-mentioned values is calculated using one of the following equations:

$$C* = \sqrt{(a*)^2 + (b*)} \quad [1]$$

$$h = tan^{-1}(\frac{b*}{a*}) \quad [2]$$

$$h = 0 \quad [3]$$

$h = 90$ when a* = 0 and b* > 0 [4]

$$h = 180^o + \tan^{-1}(\frac{b*}{a*}) \quad [5]$$

$h = 270$ when a* = 0 and b*<0 [6]

$h = 360^o + \tan^{-1}(\frac{b*}{a*})$ when a*>0 abd b* <0 [7]

The following equation can determine the total change in colour between two points in space:

$$\Delta E = \sqrt{(\Delta L*)^2 + (\Delta a*)^2 + (\Delta b*)^2} \quad [8]$$

Where ΔL, Δα και Δb are, respectively, the differences from an original colour point or reference point. However, as the human eye cannot distinguish such colours, a* and b*, the data are converted into colour functions of hue and chroma (Bartz and Brecht, 2002).

1.8.1 CIELAB in agriculture

The CIELAB colour space has valuable applications in various fields, including agriculture. This colour space is based on human perception of colour, making it more perceptually uniform compared to other colour models like RGB or CMYK. In agriculture, where colour can be indicative of crop health, quality, and ripeness, utilizing CIELAB offers several advantages.

One significant application of the CIELAB colour space in agriculture is crop monitoring and management. Farmers and researchers can assess crop health and detect potential issues such as nutrient deficiencies, diseases, or pest infestations by analysing colour variations in plant leaves, fruits, and other agricultural products. The perceptual uniformity of CIELAB allows for more accurate and consistent colour measurements across different lighting conditions and environments, enhancing the reliability of such assessments. Examples of such applications are growth monitoring of onion and garlic (Kim et al., 2023), generic plant disease detection (El Sghair et al., 2017), detection of unhealthy citrus leaves(Goyal et al., 2022) and leaf blight (Fayyaz et al., 2022)

Moreover, CIELAB facilitates the development of colour-based sorting and grading systems for agricultural products. By establishing standardized colour thresholds based on Lab values, sorting machines can efficiently categorize produce according to quality parameters such as ripeness, size, and blemishes. This not only improves the efficiency of pre and post-harvest processing but also ensures uniformity in product quality, benefiting both producers and consumers. More precisely, CIELAB has been used for determining the maturity of pomegranate(Pérez, 2021), lemons (Conesa Martinez et al., 2019) and peaches (Ferrer et al., 2005), for evaluation of tobacco leaves towards automatic harvesting (Guru and Mallikarjuna, 2010), for quality grading (Pandey et al., 2014) and defect detection such as browning in mango (Zheng and Lu, 2012).

In addition to crop monitoring and sorting, the CIELAB colour space is instrumental in agricultural research and development. Researchers leverage Lab values to quantify and compare colour attributes of different plant varieties, helping identify traits associated with desirable characteristics such as flavour, nutritional content, and shelf-life. This knowledge informs breeding programs and agronomic practices aimed at enhancing crop yield, resilience, and marketability. One commonly studied colour pigment is anthocyanin, found in the berry skin of grapes (Liang et al., 2011) and plums (Rampáčková et al., 2021)

Furthermore, CIELAB-based colour analysis supports precision agriculture techniques by enabling targeted interventions at the plant level. By precisely identifying areas of concern based on colour deviations, farmers can implement localized treatments such as fertilisation or pesticide application, optimizing resource utilization and minimizing environmental impact. Applications such as weed recognition (Dyrmann and Jørgensen, 2015), identifying and segmentation of vegetation (Setyawan et al., 2018)(Concepcion II

et al., 2021), and evaluation of nitrogen status in wheat (Yakushev and Kanash, 2011) are just some examples.

Finally, CIELAB has also found more niche applications such as land use and land cover classification (Vignesh et al., n.d.), soil classification and determination of its physical, chemical, and biological properties based on colour(Baek et al., 2022)(Roy et al., 2006)(Ibáñez-Asensio et al., 2013) and even i*n situ* detection of glyphosate on plant tissues in combination with cysteamine-modified gold nanoparticles(Tu et al., 2019).

Overall, the adoption of the CIELAB colour space in agriculture is wide and underscores its significance in advancing crop management practices, quality control measures, and scientific research within the agricultural sector. Its perceptual uniformity, coupled with advancements in technology and data analysis, has led to novel opportunities for improving productivity, sustainability, and profitability in farming systems worldwide.

1.9 Spectral Imaging vs. CIELAB: Unveiling the differences

Understanding colour goes beyond perceiving its basic hue. Both spectral imaging and the CIELAB colour space delve into the world of colour but through vastly different approaches. Choosing the proper technique depends on the specific needs and desired information.

Spectral imaging does not capture just an image, but an entire spectrum of light reflected from each pixel, gathering light information across hundreds of narrow wavelength bands and providing detailed spectral fingerprints for every point in the image. This "spectral data" unlocks a wealth of information beyond simple colour perception.

Spectral imaging can thus facilitate applications such as:

- **Material differentiation:** Identify subtle differences in visually identical materials based on their unique spectral signatures.
- **Chemical analysis:** Analyse the presence and concentration of specific chemicals based on their absorption patterns in the spectrum.
- **Non-destructive testing:** Analyse objects without physically altering them.

However, there is a trade-off to the information-dense images it can capture. To name a few:

- **Data complexity:** Analysing large datasets with hundreds of spectral bands requires specialized software and expertise.
- **Cost:** Hyperspectral cameras and associated equipment can be expensive investments.
- **Real-time limitations:** Processing such vast data volumes often limits real-time applications.

On the other hand, CIELAB colour space**,** as mentioned earlier, aims to quantify colour perception in a way that closely resembles human vision, providing some unique characteristics:

- **Simplicity:** Easy to understand and use, requires only three values to represent colour. However, it is worth noting that there also many complex indices such as BI and YI which integrate the L,a,b in a more complex way.
- **Standardization:** Widely adopted across various industries, enabling meaningful comparisons across different measurements.
- **Computational efficiency:** Data analysis is relatively straightforward, making it suitable for real-time applications.

On the contrary, to spectral imaging, the simplicity of the images comes with its weaknesses:

- **Limited information:** Only perceived colour is represented, lacking spectral details captured by spectral imaging.
- **Material differentiation limitations:** Unable to distinguish materials based on subtle spectral differences invisible to the human eye.

Thus, the choice between hyperspectral imaging and CIELAB depends on the specific use case and goals. Spectral imaging is most suitable if detailed material analysis, chemical identification, or non-destructive testing is required. However, if quantifying perceived colour is only needed, then CIELAB is a suitable option, providing ease of use and efficiency. Finally, both techniques are often combined, with CIELAB providing initial colour classification and hyperspectral imaging offering a more profound analysis of specific areas of interest.

1.10 Spectral imaging and Artificial Intelligence: a perfect fit

Spectral imaging, a technology capturing a wide range of electromagnetic wavelengths, seamlessly aligns with Artificial Intelligence's (AI) capabilities due to its inherent ability to provide rich and detailed spectral information. This synergy between spectral imaging and AI is particularly advantageous for various applications.

Firstly, the vast amount of spectral data acquired by spectral sensors serves as a robust foundation for AI algorithms. The high spectral resolution enables the identification and discrimination of subtle differences, enhancing the precision of AI models. This comprehensive spectral information provides a nuanced understanding of the target, allowing for more accurate classification and analysis.

Secondly, spectral imaging complements AI by addressing the challenges posed by complex and dynamic environments. Spectral imaging excels in capturing detailed information about scenes with diverse and overlapping objects, a scenario where conventional imaging may fall short. When integrated with spectral data, AI algorithms can navigate through intricate datasets, discerning patterns and features that might be imperceptible to the human eye.

Moreover, the synergy between spectral imaging and AI facilitates automation and efficiency. AI algorithms can be trained to process and interpret spectral data rapidly, expediting crop monitoring, disease detection, and environmental assessment. This streamlined automation not only reduces manual effort but also enhances the scalability of applications across diverse domains.

Furthermore, combining spectral imaging and AI contributes to improved decision-making processes. The detailed spectral signatures captured by spectral sensors enable AI models to generate insightful analyses, aiding in informed decision-making across

agriculture and environmental monitoring. This amalgamation of technologies empowers users with actionable insights derived from a deeper understanding of the data.

In conclusion, spectral imaging and AI compatibility lie in the former's ability to capture rich spectral information, which serves as a valuable input for AI algorithms. This fusion enhances precision, allows for complex environment applications, automates processes, and facilitates data driven decision-making. Spectral imaging and AI integration emerge as a powerful combination, promising advancements in diverse fields through high resolution data and efficient data analysis.

1.11 Common problems with spectral imaging and AI

1.11.1Big data

One of the most common problems of spectral sensors is the large amounts of data generated due to their large spectral resolutions (hundreds of bands) and considerable spatial resolution (Adão et al., 2017). From an AI perspective, this would be considered a benefit; however, these amounts of data are linked to relatively few samples, with one image/sample containing lots of information but being used only once by the AI models. Another side effect is the significant increase in the resources and computational time required for extracting main features from the spectral images (Mahesh et al., 2015), which makes handling it cumbersome. Both the resources required as well as the complexity of data acquisition and analysis hinder the use of spectral imaging(Adão et al., 2017).

All the previous can be summarised to what is known as the dimensionality problem of spectral imaging, which has been identified by numerous researchers (Khan et al., 2022)(Liu et al., 2015) and that guides the research community to develop cost-effective and efficient algorithms to speed up spectral data processing while increasing model performance. There are two ways towards that goal: band selection and feature selection. Bands that contain more information, show less data correlation, and present good separability are preserved. Feature extraction indexes and methods that are easy to implement and with high extraction accuracy are then used .(Yu et al., 2022)

1.11.2 Model generalisation

Finally, another major problem is model generalisation, which refers to the challenge of applying spectral imaging techniques across different conditions, environments, or contexts. Most of the results published that utilise hyperspectral imaging follow a common methodology with minor adjustments to achieve good results. The common methodology used for data processing once the hyperspectral images have been acquired can be summarized in a specific number of steps, as presented by (Wang et al., 2015):

1. Pretreatment (e.g., Derivative correction, Smoothing)
2. Methods for Variable selection
3. Discriminant methods
4. Calibration of the model

5. Evaluation of the model

The algorithms selected and what extent they are used is use case specific and can vary. (Medic, 2023) made use Partial Least Square (PLS) regression algorithm and spectral smoothing, namely Savitzky–Golay, to achieve a coefficient of determinationR^2 of 0.91 for estimation of DM in apples. Apples were also been the focus of studies by (Zhang et al., 2019) that achieved similar results using feature extraction (Principal Component Analysis) and PLS algorithm and (Kavuncuoğlu et al., 2023), that made use of feature selection (Bootstrap Random Forest) and Artificial Neural Networks. Furthermore, (Taghizadeh et al., 2009) used PLS regression, spectrum smoothing (Savitsky–Golay) and normalization (Standard Normal Variate) for moisture content prediction in white button mushrooms, with a performance of R^2 of 0.8. Lastly, (Muruganantham et al., 2022) by focusing on unpeeled whole potato tubers achieved an R^2 of 0.53 using PLS and feature selection (β-coefficient and VIP).

Despite the excellent results, all previously mentioned research focuses on single fruit/ vegetable/crop. Currently, models capable of generalizing additional crops have not been at the center of research.

The most important factors that impede generalizability are the following:

1. **Variability**: Different crops, varieties, regions, high intra-class spectral-spatial variability, atmospheric and daylight conditions make spectral models lack universality (Signoroni et al., 2019).
2. **Calibration**: Spectral imaging and chroma meter systems require precise calibration to ensure accurate and reliable measurements. However, variations in sensor response, illumination sources, and environmental factors can introduce calibration errors, leading to inaccuracies in spectral data interpretation.
3. **Limited Training Data**: Training spectral imaging algorithms typically requires large volumes of labelled data to learn complex patterns and relationships. However, such datasets are limited, and collecting and annotating such data can be labour-intensive and costly, particularly for niche or specialized applications, leading to a scarcity of training data and potential overfitting issues.
4. **Transferability**: Spectral imaging models trained on data from specific environments or conditions may struggle to generalize to new, unseen scenarios. This lack of transferability can limit the practical utility of spectral imaging technologies, particularly in applications where adaptability to diverse operating conditions is essential.

As a result of the above, the generalisation and applicability of spectral imaging developed methods have yet to be explored(Shuai et al., 2024).

1.12 Broccoli

The crop of selection for this dissertation is broccoli. The selection of broccoli as the focal point of this dissertation stems from a multifaceted rationale ranging from its importance to modern agriculture and human nutrition to the challenges its morphology poses to precision agriculture solutions. In essence, the selection of broccoli as the subject of this dissertation is driven by a commitment to advancing agricultural knowledge and addressing the pressing challenges precision agriculture is facing.

1.12.1 General Information

Broccoli has been of considerable interest worldwide in recent decades, and its consumption has increased significantly during the winter months due to the publicity it has received for its dietary qualities as well as its medicinal properties for the prevention of various forms of cancer in humans. Large quantities of broccoli are produced in the USA, Italy, northern European countries, and cold regions of the Far East (Olympios, 2015).

Broccoli is considered a native plant of southern Europe and the eastern Mediterranean, a popular vegetable of the Italians since Roman times, who consumed it raw or cooked, but mainly for medicinal purposes. Today, there is still confusion about the origin of broccoli and cauliflower. The most widely accepted view is that broccoli is the ancestor of the early cauliflowers grown today. It is one of the few vegetables that have become very popular worldwide in recent years. Until 1920 it was not popular in the U.S. until Italian immigrants brought it to California and began growing it (Olympios, 2015).

In Greece today, mainly varieties with green flower heads and much less varieties with violet flower heads are cultivated. The head or edible part of the broccoli consists of densely arranged flower buds in an inflorescence and tender parts of the stem end, dark green or violet in colour, depending on the variety, with the stem being about 15 cm long (Olympios, 2015).

1.12.2Botanical Characteristics

Broccoli is an annual or biennial dicotyledonous, herbaceous plant and belongs to the family *Cruciferae* in the Brassica oleracea var italica species. The plant grows to a size of 50-90 cm and forms larger spaces between shoots compared to cabbage and cauliflower. During its growth, a short stem is formed, the top of which is divided into a number of secondary shoots that enlarge, become fleshy, form closed flowers, and

together form the marketable head. The leaves first appear in a rosette, and later, the space between leaves spaces become elongated. The leaves have a strong central nerve which is colourless-greyish-green. A central flower head develops in the centre of the plant on the unbranched central stem. The flower head appears branched, and a compact hemispherical head is formed. The colour of the head is green or violet, depending on the variety, and is surrounded by leaves without being entirely covered by them. In broccoli, after the central head has been removed, lateral floral inflorescences of smaller size develop at the bases of the lower leaves. The dominance of the top flower head influences the development of the secondary flower heads. After the top flower head is harvested, the secondary develops and is harvested later (Olympios, 2015).

1.12.3 Varieties

The cultivated varieties of broccoli are divided into five categories:

i. Depending on their earliness, i.e., the time required from sowing to harvesting the final product. They are divided into early, medium and late.
ii. Depending on the growing season, they are divided into autumn, winter and spring varieties.
iii. Depending on the flower head size, large heads are preferred for the fresh market and small heads for the frozen market.
iv. Depending on their ability to form only one flower head, the central one, or to form lateral (second-order) flower heads on the axils (bases) of the leaves (parapillar bracts).
v. Depending on the colour of the flower head (dark or light green or red-violet).

The essential quality characteristics sought in broccoli cultivars and hybrids are the shape, color, consistency of the flower head, the size of the individual flower buds of the florets, the extent of branching, the length of the intercalary spaces, the production or not of lateral flower heads, the ability of the flower heads to flower and after harvest and disease resistance.

The main varieties and hybrids cultivated in Greece today are the following:

- Marathon: a popular variety that grows mainly in August or October. It is a popular crop grown in most parts of the world, mainly in autumn, and it shows resistance to powdery mildew.
- Parthenon: with a biological cycle of 105 days and grown in autumn (September-October). This variety is resistant to low temperatures.
- Fidel F1: Hybrid with very consistent, heavy-weight heads and high quality. It has high cold hardiness and is resistance to powdery mildew. Its Biological

cycle duration is 95-105 days. Recommended growing season: summer and autumn (August-October).

- Milady F1: Hybrid with dark green small flowers. It shows resistance to stem formation. It has a long biological cycle of 65 days. Harvested in April-June and September-November.
- Mon Top F1: Hybrid with a biological cycle of 65-70 days.

(Olympios, 2015)

1.12.4 Climatic requirements

Broccoli is a cool season plant, and for a good quality product, the average monthly temperature should not exceed 16 ºC. At higher temperatures, the plant generally stops growing. Also, low temperatures during the early stages of the plant's growth cause the formation of early immature flower heads, and the plant grows very slowly when the temperature is below 5 ºC. The plant is sensitive and is damaged by freezing temperatures after forming flower heads. Finally, there are variations between varieties in terms of the need for exposure to low temperatures to form flower stems (Olympios, 2015).

1.12.5 Agricultural inputs

Broccoli (*Brassica oleracea var. Italica*) is a crop that requires irrigation and fertilisation to reach high yields. Both play a vital role as they determine productivity and quality (Wien and Wurr, 1997)(Vågen et al., 2004)(Thompson et al., 2002). However, despite the crop being highly responsive to N fertilisation, excessive amounts can cause quality degradation(Doerge, 1991)(Stivers et al., 1993). Moreover, the nutritional demands are not constant, and they change depending on the broccoli phenology, with fewer nutrients required in the first two weeks after transplanting and demand increasing as the plant grows(Carranza et al., 2008)(Cecílio Filho et al., 2017). It therefore, becomes clear that to increase production, a well-scheduled nitrogen and irrigation plan is required throughout the growing season to provide the plant with the required nutrients, water and soil moisture (Erdem et al., 2010).

1.12.6 Physiological disorders

Yellowing of flower spikes may occur in over-ripe broccoli when stored at higher than optimum temperatures or in response to exposure to ethylene. The presence of yellow florets reduces the marketability of broccoli. There is sometimes confusion between the yellow florets associated with aging and the peripheral head florets, which are yellow to light green and are also affected by shading by overlying flower tissue. Something typical for tissues not exposed to light during head development. Also, a disorder called "black spot" on stems occurs in stored broccoli, with some varieties being more susceptible than others. Finally, broccoli is very susceptible to bruising (Vasilakakis, 2006).

1.12.7 Harvest

The time from sowing or transplanting to harvest is influenced by the variety, season, prevailing climatic conditions, soil nutrient availability, and moisture availability during cultivation and usually ranges from 90-150 days after transplanting. The central flower head (top) is harvested when it has reached the marketable size (desired size, small and closed flowers, the head is immature, compact, and cohesive (tight)). Removing the central flower head allows the development of the lateral flower heads, which are small and short-stemmed and harvested when they have reached the right size (same stage of maturity as the central flower head). At harvest, the flower heads are cut with a knife or small sickle (pruning shears) with part of the stem (shoot) about 15-25 cm long (Figure 11), with mechanical and robotic harvesting solutions being in the research and development stage. Manual harvesting from a plantation takes 1-2 months and is completed in 5-10 rounds. The harvesting period starts in September and continues until April. The central flower heads vary in weight from 100-1000 g and in diameter from 10-20 cm (Olympios, 2015).

Figure 11. Broccoli harvesting.

Commercial maturity is based on the diameter of the head; super-mature heads are characterized by open florets or enlarged florets on the verge of opening, resulting in a loose flower head (Vasilakakis, 2006). High-quality broccoli is either dark or bright green with closed florets. The head should be firm and compact when pressed by hand, and the stem should be cleanly cut to the appropriate length for a given quality standard (USDA, 2016). Broccoli sold as a "whole flower head" should be tight and well-developed. Upon harvest leaves are removed, and heads are sold by the piece or weight (Vasilakakis, 2006).

1.12.8 Nutritional value

Broccoli is one of the most affluent vegetables in vitamin A (Table 4). The dark green colour is an indication of high carotenoid content. Although they have a slightly bitter taste, broccoli leaves are edible and contain a high concentration of vitamin A. They also contain vitamins B1, B2, B3, B6, iron, magnesium, potassium, and zinc. Frozen broccoli contains more beta-carotene than fresh broccoli because it consists mainly of flowers. However, the stems also contain considerable amounts of calcium, iron, thiamine, riboflavin, niacin, and vitamin C. The darker the colour of the inflorescences, the more vitamins A and C they contain. Broccoli contains sulforaphane, which helps antioxidants such as vitamins C and E (Vasilakakis, 2006).

Table 4. Broccoli nutritional value (per 100 g fresh weight) (Vasilakakis 2006)

Nutritional Value	Amount
Energy (Kcal)	24
Carbohydrates (%)	5.8
Protein (%)	3
Fat (%)	0
Vitamin A (mg)	874
Folic acid (mg)	40
Calcium (mg)	27
Vitamin C (mg)	113

Folk traditional medicine and pharmacology, but especially the latest research, show the vital contribution of the cultivated vegetables of the cruciferous vegetable group in preventing various forms of cancer, i.e., they act against carcinogenesis and mutations. It is also noted that broccoli helps to reduce cholesterol in the blood (Olympios, 2015).

1.13 Precision agriculture applications in broccoli production

Precision agriculture has found applications in a variety of horticultural crops. Most of those solutions are suitable and can be applied to most crops (e.g., NDVI mapping) without significant modifications, while others are tailored for specific crops (e.g., robotic harvesting, yield prediction models). Despite the popularity of broccoli in recent years, the precision agriculture applications tailored to its cultivation are still limited compared to other crops such as maize, wheat, and strawberries.

However, solutions have been developed for the whole primary broccoli production process. Starting with phenotyping (Chengquan et al., 2020), moving to growth monitoring (Psiroukis et al., 2022)(Lee et al., 2023),quality monitoring (Zhou et al., 2020),yield estimation (Noé et al., 2002)(Zhou et al., 2022),weeding (Pallottino et al., 2018), irrigation (Kumar et al., 2021), pest damage evaluation (Zou et al., 2021), stress monitoring (El-Shikha et al., 2007)(Tremblay et al., 2008)(Graeff et al., 2008) and ending with selective harvesting (Garcia-Manso et al., 2021)(Montes et al., 2020).

1.13.1 AI and broccoli

As mentioned earlier, broccoli has not been at the centre of precision agriculture research. The same can be said for AI solutions, with numerous solutions having been developed for a variety of crops ranging from model plants such as lettuce (Rahimikhoob et al., 2023) to crops farmed intensively such as wheat (Mehta et al., 2023) and from apple orchards (Mazzia et al., 2020) to vineyards (Fraiwan et al., 2022).

The lack of AI/ machine learning research on broccoli is further proven by conducting a quick search in the Scopus database using the keywords “Machine learning” plus the” Crop of interest.” At the moment the search was conducted (January 2024), broccoli yielded the least number of results among the crops investigated (Table 5) accounting for only 0.4% of the documents among 10 crops.

Table 5. Results per crop searching the Scopus database ranked in ascending order.

Crop	Documents found
Broccoli	22
Cucumber	98
Lettuce	117
Strawberry	170

Grape	396
Potato	422
Tomato	690
Maize	948
Apple	954
Wheat	1,257

The lack of research interest could be attributed to a combination of factors such as:

1. **Complexity of Plant Biology**: Broccoli has complex biological processes that govern its growth, development, and response to environmental factors. Moreover, its morphology and geometry pose significant challenges.
2. **Data Availability and Quality**: AI algorithms rely heavily on data for training and validation. However, comprehensive and high-quality data specific to broccoli cultivation are scarce and not readily accessible.
3. **Resource Allocation**: Research focuses on major staple crops like rice, wheat, or corn with high economic significance and broader impact.
4. **Industry Priorities**: The direction of AI research in agriculture is influenced by industry priorities and market demands. If broccoli production does not represent a significant market opportunity or there is limited demand for AI-driven solutions in this sector, research efforts may be directed toward other crops or agricultural applications.

From the AI solutions investigated for broccoli, autonomous harvesting is the most popular with multiple publications, such as (Ramirez, 2006)that developed a computer vision system to locate and classify mature and immature broccoli heads for selective harvesting, (Montes and Cielniak, 2022) that used 3D point cloud for multiple broccoli head detection and tracking, (Kusumam et al., 2017) that besides detection went a step further as also to assess the broccoli size and (Blok et al., 2021)who suggested a novel image-based size estimation to overcome occlusion problems.

Besides tasks related to harvesting, disease detection has also been investigated. (Ferdinand and Al Maki, 2022) used AI to classify broccoli leaf diseases.(Pineda et al., 2022) introduced novel vegetation indices to identify broccoli plants infected with Xanthomonas campestris. Finally, (Zou et al., 2019) used machine vision to discriminate between broccoli seedlings and weeds and to estimate pest damage (Zou et al., 2021), while (Makino and Amino, 2020) focused on the post-harvest evaluating broccoli freshness. However, besides their importance for broccoli cultivation, AI solutions focusing on fertilisation and irrigation have yet to be investigated.

1.13.2 Spectral imaging and broccoli

The same that applied to AI research and broccoli production applies and for spectral imaging research. It is, however, worth noting that from the limited studies available for broccoli production, the majority focuses on the post-harvest stage, and only a few on the pre-harvest stage, more specifically only 29% of the studies found focuses on the pre-harvest. The limited studies available for the pre-harvest stage focus on disease and pest damage detection. Table 6 presents an overview of the available research on combining broccoli and spectral imaging at the post-harvest and pre-harvest stages to allow for direct comparisons.

Table 6. Pre- and post-harvest broccoli characteristics investigated with the use of spectral imaging.

Growth stage	Investigated characteristic	Reference
Post-harvest	1. Senescence	(Guo et al., 2022)
	2. Total glucosinolates	(Hernández-Hierro et al., 2014)
	3. Pesticide residue	(Gui et al., 2019)
	4. Degradation rate	(Hosaka et al., 2012)
	5. Degreening velocity	(Makino and Kousaka, 2020)
Pre-harvest	1. Seedling pest damage	(Zou et al., 2021)
	2. Disease detection	(Pineda et al., 2022)

Aim and Objectives

This research aims to offer a meaningful addition to the domain of precision agriculture by investigating the capabilities of spectral imaging and Artificial Intelligence (AI) for optimizing fertilisation and irrigation. To achieve this, the research aimed at accomplishing the following objectives:

(i) Develop AI models utilizing spectral data that can identify different fertilisation levels.
(ii) Develop AI models utilizing spectral data capable of identifying plant water deficit.
(iii) Compare the performance of traditional machine learning algorithms with novel user-friendly AutoML techniques.
(iv) Evaluate the feasibility of developing a generalisation-capable AI model utilizing spectral data.

The results of this research provide advantages that reach beyond academia, providing valuable assistance to diverse groups, such as researchers, farmers, and stakeholders engaged in decisions regarding agricultural inputs distribution, food security, and sustainable agricultural practices.

Chapter 2 – Materials and Methods

2.1 Workflow overview

This study followed a parallel exploratory methodology (Figure 12). One line of experiments was focused on the pre-harvest stage while the second on model generalisation. The primary focus of this research which expanded over three years was to investigate i) the potential of spectral imaging and Artificial Intelligence in optimizing pre-harvest stage primary production and ii) the generalisation capability of spectral models.

During the first year, the emphasis was on fertilisation and whether spectral imaging can identify different fertilisation levels among plants grown under the exact same conditions. Moreover, the same year, dry matter measurements were conducted across a variety of fruits and vegetables, namely apple, broccoli, leek, and mushrooms in cooperation with partners from abroad.

Through these joint measurements, we aimed to understand Spectral Imaging and Artificial Intelligence better and share novel ideas, opinions, and thoughts with fellow researchers and PhD students. Moreover, it allowed us to investigate and test the generalisation capabilities of AI models developed using spectral imaging datasets. It is worth mentioning that the different treatments used for studying the preharvest stage proved invaluable for dry matter estimation as they resulted in variations among the grown broccoli plants, providing the necessary variability in the collected data to develop robust Artificial Intelligence algorithms.

During the second year the focus was on irrigation. Namely, the spectral response of plants being exposed to water stress was investigated. During the third year, the generalisation capabilities of models trained with spectral data were evaluated using the large joint effort spectral dataset (Malounas et al., 2024) collected in year one. More precisely, the steps followed throughout the typical development of an Artificial Intelligence model used for spectral imaging data processing were evaluated. This approach yielded a better understanding of how data size affects model performance and how different data-preprocessing techniques influence the generalisation capabilities of AI models.

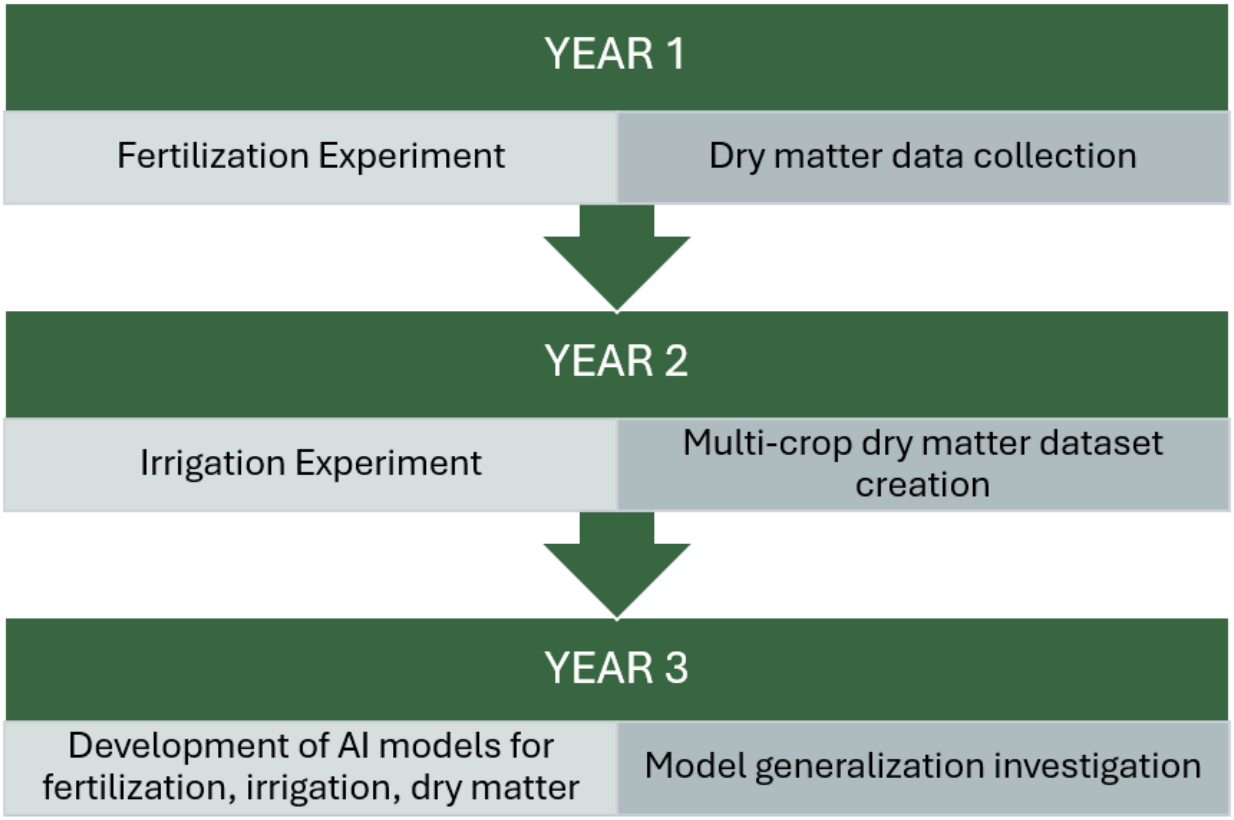

Figure 12. Workflow followed by this study. With light grey the pre-harvest stage, while with dark grey the experiments related to model generalisation.

2.2 Description of the study area

The study was conducted during 2021 - 2023 in a glasshouse (Figure14) located at the Agricultural University of Athens premises, Athens, Greece (37.986039570505596, 23.706417286906994) with a surface area of 100 m^2(Figure 13).

Figure 13. The glasshouse position (red dot) at the Agricultural University of Athens (Google Earth, 2024)

2.3 Growing conditions

The study covered two growing seasons: winter 2021-2022 and winter 2022-2023. Some growing conditions were kept the same during both growing seasons, while others were adapted based on the agricultural input under investigation.

Figure 14. Broccoli glasshouse

The unchanged growing conditions were the broccoli variety, the substrate, the fertiliser, the irrigation water, and the temperature. More precisely, the Nerone variety, a 60-day variety suitable for the climate conditions of Greece, was selected. The variety selection was made for the short time it requires to reach maturity, as most broccoli varieties require between 70 and 100 days. The brief time to reach maturity was a prerequisite as the ideal growing condition for broccoli is between 18-24 ^{o}C and not higher than 27 ^{o}C, which can be maintained only during the winter months in a glasshouse in Greece with no active cooling only. The fertiliser used was a balanced fertiliser (20-20-20) containing approximately equal proportions of the three primary nutrients essential for plant growth: nitrogen (N), phosphorus (P), and potassium (K). The tap water supply system was used for irrigation purposes. Finally, to ensure that the temperature was kept within the desired threshold, the automated window system was set to open when the temperature inside the glasshouse reached 22 degrees ^{o}C and close again once it dropped down to 18 degrees.

During the first growing season, 90 plants were grown, and fertilisation deficiency was investigated. The fertiliser used was a typical 20/20/20 N/P/K fertiliser as described above. However, three different fertilisation schemes were followed. The first one followed the typical dosage followed by commercial broccoli farms (15 g of granular fertiliser/ plant), the second one used the typical dosage cut in half (7.5 g of granular fertiliser/ plant), while the plants under the third fertilisation scheme did not receive any fertiliser (0 g / plant), with the plant relying entirely on the soil substrate. To that end, the soil substrate used was a typical soil used for growing vegetables that also covered the needs of broccoli growth in order to simulate real-life growing conditions as closely as possible. More precisely, it was loamy soil (a balanced mix of sand, silt, and clay) with a pH of around 7. Finally, regarding irrigation, a drip irrigation system supplied water daily until soil moisture saturation was reached.

During the second growing season, 60 plants were grown, and irrigation and water acclimation/stress were investigated. To that end, the irrigation scheme had to allow for precision measurements. Irrigation was carried out manually, maintaining the soil at 40% of field capacity through daily weighing. Finally, the recommended dosage (15 g granular fertiliser 20-20-20 / plant) was applied to all plants.

2.4 Data Collection

2.4.1 Remote sensing equipment

During both growing seasons spectral images were captured in-situ inside the glasshouse. The setup used consisted of the following components. A hyperspectral camera (IMEC snapscan VNIR) (Figure 15) and

Figure 15. Imec snapscan VNIR hyperspectral camera

a three-wheel platform (Figure 16), which provided the necessary mobility to the system, allowing it to move on rough terrain and narrow rows while simultaneously allowing all individual components to be mounted.

Figure 16. Three-wheeled platform

The components fitted to the platform besides the camera were the following: a three-joint arm where the camera was mounted, which allowed the adjustment of the height of the camera as well as the angle at which the images are captured, a 12-volt battery, a power inverter used to provide the needed power to the spectral camera and finally the laptop used to control the camera functions (Figure 17). The system did not involve an illumination system; instead, it relied on the sun's presence for the necessary illumination.

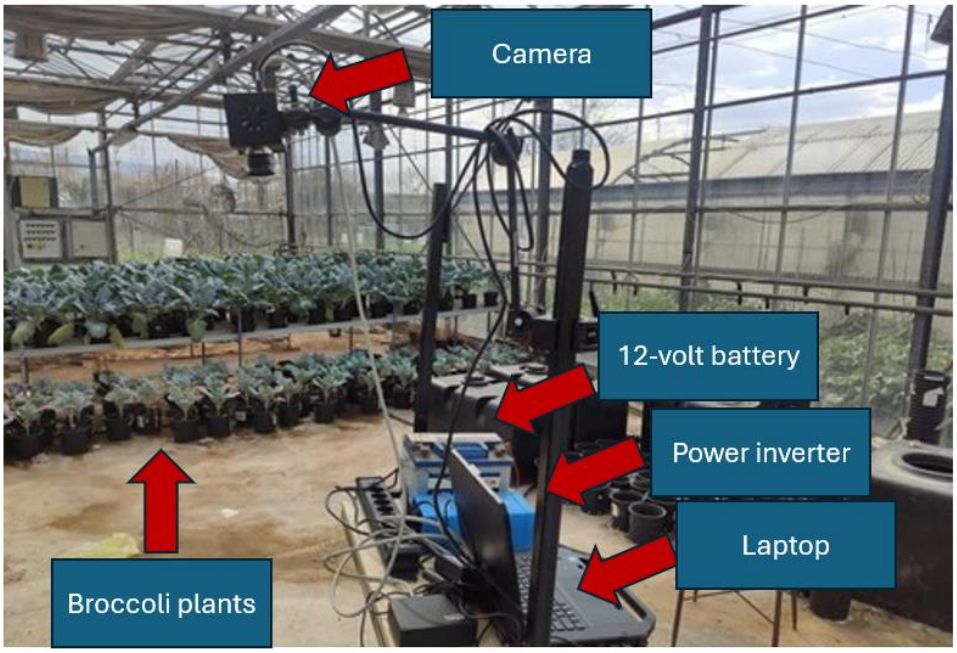

Figure 17. Three wheeled platform with all components mounted.

The camera specifications are presented in the table below.

Table 7. Imec Snapscan product specifications

Spatial resolution	up to 3650 x 2048 pixels (7Mpx RAW per band)
Spectral resolution	150+ bands
Spectral range	470-900nm
FWHM	~10-15nm (collimated)
Acquisition speed	~200ms – 20 seconds, depending on acquisition parameters, lighting and object (without including pre- and post-processing time)
SNR	>100-200, flat SNR over spectral range
Dynamic range	8/10 bit
Optics	Schneider Kreuznach Apo-Xenoplan lens, f2.0, Focal length: 35 mm lens
Dimensions	10x7x7 cm (WxDxH)
Weight	580g (camera without optics)
Input voltage	24V DC 2.7A (external controller)

Additionally, once the spectral imaging took place, CIELAB colour measurements were conducted using the Lovibond RT300 spectrophotometer (Figure 18).

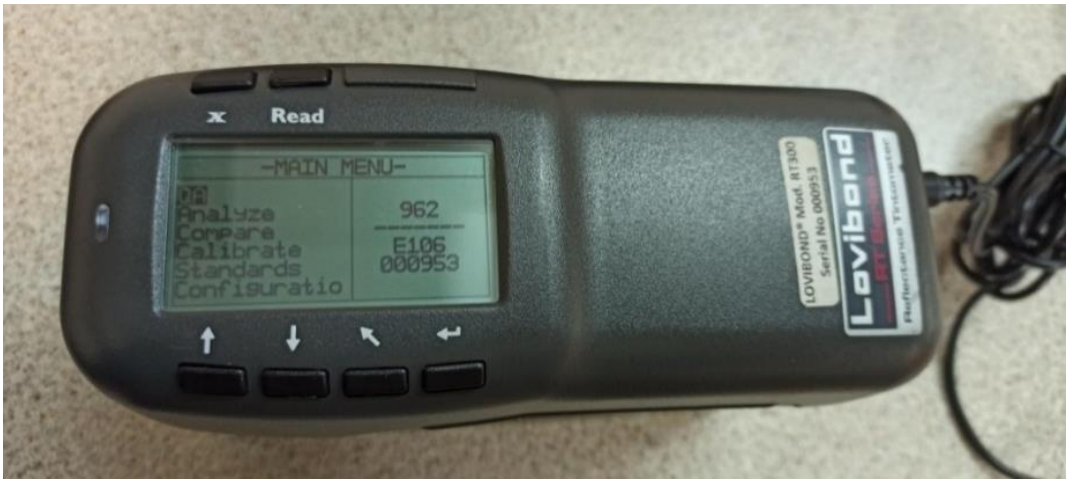

Figure 18. Lovibond RT300 spectrophotometer

The spectrophotometer specifications can be found in the table below.

Table 8 Lovibond RT300 product specifications

Spectral Interval	10 nm - measured; 10nm - output
Measurement Range	0 to 200 % reflectance
Spectral Range	400 - 700 nm
CIE L* a* b Scale resolution	0.01

For both growing seasons during the spectral imaging sessions no artificial illumination was used; instead, sunlight was utilised for the necessary illumination. Imaging sessions took place during midday to ensure the best possible illumination conditions. Before each imaging session, integration time and gain were set to optimal, and a white reference was acquired to estimate the incident radiant flux density and a dark reference to minimize the camera sensor's inherent imperfections. The white reference was a Zenith Lite™ diffuse target (Spectralon, Labsphere) (Figure 19), which reflects 95% of the incident radiation. The dark reference was captured by completely closing the camera's mechanical shutter.

Figure 19. Spectralon diffuse target

Upon image and reference acquisition, spectral images were radiometrically corrected using the calibration function below:

$$Rc = \frac{Ro-D}{W-D} x\ 100 \text{ [9]}$$

where Ro is the raw spectral image, W is the image of a white reference object of uniform, stable, and high reflectance standard (~100% reflectance), D is the dark image/reference (~0% reflectance), and Rc is the corrected spectral image. Moreover, to obtain the extracted spectra, spikes and dead pixels, were excluded using thresholding and more precisely by using fixed values. Finally, the background removal was conducted using background removal techniques (e.g. Otsu algorithm) or manually.

2.4.2 Plant physiology measurements

During the second growing season, in order to validate that plants were under drought stress, gas exchange parameters were measured prior to imaging with the focus being onthe photosynthetic rate of each plant. The measurements were conducted using the LC pro+ gas analyser under ambient environmental conditions (25.5 °C and 425 ppm CO2) and saturating light levels.

Figure 20. LC pro+ gas analyser

2.4.3 Dry matter measurements

Once the spectral measurements were completed, the samples were immediately transported for dry matter measurements. All samples were weighed to measure their fresh weight and then placed in an oven, convective air dryer, up to the moment that their mass reached a constant value following the dry matter estimation protocol established by (Cunniff and Washington, 1997)

Based on fresh weight of each sample and their final constant mass (final weight), their moisture content was calculated using the following equation.

$$Moisture\ content\ (\%) = \frac{\text{Fresh weight} - \text{Dry matter}}{\text{Fresh weight}} x\ 100 \quad [10]$$

One moisture content value (% wet basis (w.b.)) was recorded for each sample, and then the dry matter was calculated by subtracting moisture percentage (%) from 100%.

2.4.5 Datasets used

Four datasets were compiled during this 3-year study: the first focused on fertilisation, the second on irrigation, the third on broccoli dry matter and the last one on dry matter estimation of multiple crops. At this point it should be pointed out that the multiple crop dry matter dataset includes measurements of apple, broccoli, leek and mushroom. The decision to create this dataset was made in order to investigate the generalisation capabilities of spectral datasets across a variety of crops. Therefore, crops belonging to different families, and with different color and shape characteristics were chosen.

Each of the four datasets consists of more than one type of measurements besides spectral measurements. Additional measurements were gathered either to validate findings or to investigate novel applications and approaches. A detailed overview of the datasets is presented in the table and in the following sections.

Table 9. Used datasets overview

Experiment	Dataset	Type of data	Number of samples	Number of features (bands)
Fertilisation	Spectral	• Spectral images	49	150
Fertilisation	CIELAB	• CIELAB colour measurements	49	3
Fertilisation	Spectral + CIELAB	• Spectral Images • CIELAB colour measurements	49	153
Irrigation	Drought onset	• Spectral images • Physiological measurements (validation)	60	150
Irrigation	Drought acclimated	• Spectral images • Physiological measurements (validation)	60	150
Irrigation	Drought onset + Drought acclimated	• Spectral images	120	150

		• Physiological measurements (validation)		
Dry matter model generalisation	Dry matter (Broccoli)	• Spectral images • Dry matter measurements	250	109
Dry matter model generalisation	Dry matter (various fruit and vegetables)	• Spectral images • Dry matter measurements	779	109

Fertilisation dataset

For the fertilisation experiment three datasets were used. The first consisted only of spectral measurements and the second one only of CIELAB measurements. The reason for acquiring the two datasets is that hyperspectral imaging relies on chlorophyll absorption while CIELAB on the phenomenological background of chlorophyll (green colour). A third dataset was produced by merging the two. By combining both types of measurements, machine vision systems can achieve more accurate colour perception, leading to improved classification accuracy, and quality control. Finally, as measurements took place *in situ,* using the sun as an illumination source, the integration of spectral and CIELAB measurements allowed for colour consistency and accuracy across the varying illumination conditions, leading to more robust models.

Plants were cultivated until the harvestable vegetative plant parts reached 60-70% of their final head diameter (BBCH 46-47) and then imaging took place. The zero fertilisation plants failed to reach this stage as they did not develop the harvestable vegetative plant parts and were discarded. Regarding the half and full fertilisation plants, broccoli plants that showed defects (e.g. yellowing, flowering) were rejected as well. The harvested broccoli from both fertilizer applications did not differ in appearance (figure 22). The only difference was the average weight with the full fertilisation broccoli weighing 20% more on average. The distance between the camera and the highest point of the sample (broccoli) was kept constant (60cm) while avoiding movement to prevent motion blur. Moreover, to facilitate a top-down view the camera was constantly perpendicular to the broccoli head.

Figure 21. Broccoli with full fertilization dosage left and with half fertilization dosage right. No visible differences.

After spectral imaging, CIELAB measurements were carried out. For each sample, 5 points, selected to cover as much of the broccoli head color and shape variation, were measured and averaged to achieve a sample colour representation as complete as possible.

Each of the three datasets consisted of 49 samples (21 samples from the half deficiency class and 28 from the recommended fertilisation class. A 70/30 train test split was used resulting in 34 samples/images for training and 15 samples/images used for testing. The hyperspectral dataset had 150 features (bands) per sample, the CIELAB dataset had 3 features (L*, a*, b) while the merged dataset 153.

Irrigation dataset

For the irrigation experiment, plants were divided into two treatments (drought and control). Two datasets consisting of 60 images were constructed. For the first dataset imaging took place at the phenological stage were 70% of the expected head diameter was reached (BBCH-scale 47) and for the second when the typical size and form had been reached and the head remained tightly closed (BBCH-scale 49). The first dataset of images was captured after the broccoli were not irrigated for 4 days and while the substrate water content reached 40% of its pot capacity, while the second dataset was captured 12 days later, maintaining (with daily weighing) the substrate water content at 40% of its pot capacity over the duration of these days. The images were acquired in situ (inside the greenhouse) as top views of each plant. Prior to imaging, gas exchange parameters were measured using the LC pro+ gas analyser under ambient environmental conditions (25.5 °C, and 425 ppm CO_2) and saturating light levels. Upon image acquisition outliers (single images) were detected and removed, Table 10 shows the final datasets used for the experiments. From the first dataset of images (drought onset), a total of forty-

two (42) out of sixty (60) images were kept, and from the second dataset (drought short-term acclimation), forty-eight (48) out of sixty (60) images. Finally, the two datasets were integrated into a third one including both drought onset, drought acclimated and control broccoli plants.

Table 10. Irrigation experiment Image dataset before and after outliers' removal.

Dataset	No. of images with outliers	No. of images without outliers
drought onset	60	42
drought acclimated	60	48
mixed	120	90

Both the drought onset dataset and the drought acclimated dataset were unbalanced. The drought onset dataset contained more drought samples while the drought acclimated dataset contained more control samples. More precisely, the drought onset dataset contained seventeen (17) control and twenty-five (25) drought broccoli images, and the drought acclimated dataset, twenty-seven (27) control and twenty-one (21) drought acclimated broccoli images. Lastly, the combined dataset contained forty-four (44) control, twenty-five (25) drought onset and twenty-one (21) drought acclimated plant images. Table 11 shows the distribution between the different classes for each dataset.

Table 11. Data distribution among the datasets described in %.

Dataset	control	drought onset	drought acclimated
drought onset	40%	60%	-
drought acclimated	56%	-	44%
mixed	49%	28%	23%

Multiple crop and vegetable dry matter dataset

The dataset consists of three different crops, Apple, Broccoli and Leek, captured with different sensors covering the VIS-NIR range of 700-900nm. To acquire reflectance spectra the samples (edible parts of the crops) were placed in an environment that was illumination-controlled in order to maximize the dynamic range of all the sensors used for each sample. Across the three crops, the same acquisition protocol was followed to ensure the consistency of all measurements. The protocol consisted of the following actions: i) imaging mode set to Reflectance, ii) use halogen lamps (Apple: 150W from Illumination Technologies; Broccoli and Leek: 50W from Osram;) with excellent performance at VIS-NIR range of 400-900 nm, together with a stabilized DC power supply, iii) capture dark and white reference images using a high reflectance and stable standard (~100% reflectance) and a ~0% reflectance standard, respectively iv) maintain a constant distance between sample and sensor throughout the imaging campaign. The distance was kept the same for each imaging experiment; however, it differed for each crop and camera setup to optimize data acquisition quality based on the specific characteristics of each camera (e.g., linescan, snapscan) and crop (e.g., shape). To minimize sample exposure to the heat produced by the halogen lamps they were placed in the image acquisition stage only once the setup was ready for capturing. The sensors, crops and specific details of all use cases can be found in the table below. Broccoli image acquisition was conducted by Agricultural University of Athens, while the apple and leek image acquisition were conducted by Leibniz Institute of Agricultural Engineering and Bio-economy and the Flanders Research Institute for Agriculture, Fisheries and Food respectively.

For the broccoli dataset the previously described Imec Snapscan hyperspectral camera was used, while for the apple dataset the Cubert ULTRIS S20 hyperspectral camera. The specific hyperspectral camera comes with a global shutter, a spectral resolution of 141 bands and a FWHM of 12nm. Finally for the leek measurements the Specim FX10 hyperspectral camera was used. The FX10 operates in the region of 400-100nm with a FWHM of 2.62-2.82nm capturing in total 224 concrete spectral bands. The specifications of all hyperspectral cameras can be found in the table below.

Table 12. Technical specifications of the hyperspectral cameras used for the dry matter content dataset

Crop	**Apple**	**Broccoli**	**Leek**
Camera	Cubert ULTRIS S20	Imec Snapscan	Specim FX10

Distance between sample and camera	50cm	30cm	60cm
Spectral range	430-990nm	470-900nm	398-931 nm
Spectral resolution/No. of bands	141	150	224
Full width half maximum (FWHM).	12nm	10-15nm	2.62-2.82 nm
Number of measurements	240	250	288

However, the use of various hyperspectral sensors, lead to a difference in the centre wavelength value, number of bands and available wavelength for each of the selected crop.

Namely the spectral resolution/ No. of wavelength bands was the following for each of the crops: apple/ 141, Broccoli/ 150, and Leek/ 421. Moreover, spectral range for apple/ 430-990nm, broccoli/ 470-900nm, and Leek/ 398-1717nm. As a result, if a band differed more than 2nm was discarded, taking into account the Full Width at Half Maximum and the centre wavelength value of the sensors used. Once those bands were discarded each consisted of 109 discrete bands whose wavelength ranged from 469 to 900 nm. The spectral signature of all three (3) crops in the VIS-NIR region can be seen in the figure below with their standard deviation in brackets.

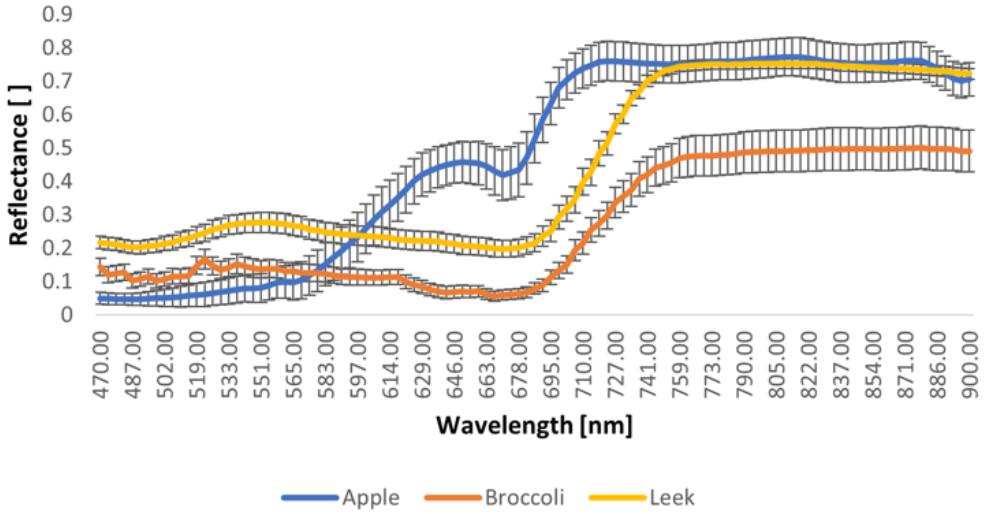

Figure 22. Multi crop dataset averaged spectral signatures.

The spectral measurements were accompanied with dry matter measurements (%). The Min, Max, and Average dry matter content (DMC) of each crop, following proper irrigation and fertilisation, can be found in the table below.

Table 13. Dry matter content (DMC, %) per crop

DMC in %	Apple	Broccoli	Leek
Min	14 %	12%	8.1%
Max	17 %	20%	19%
Average	15 %	15%	12%

2.5 Spectral Data Pre-Processing

For the purposes of this dissertation Python programming language code was developed to conduct the spectral data preprocessing. The pipeline used in this study which is also the typical pre-processing pipeline (Wieme et al., 2022) used to handle spectral data, is presented in the figure below.

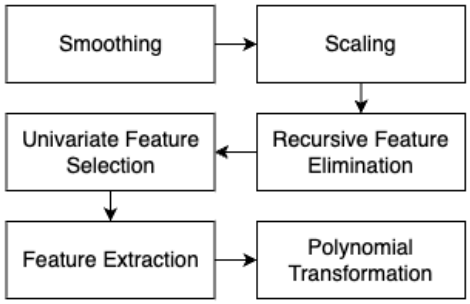

Figure 23. Common spectral data pre-processing pipeline

All the components shown in Figure 23 are explained in detail in the upcoming subsections. The Recursive feature elimination and polynomial transformation steps were skipped for the irrigation and fertilisation experiments as excellent results were achieved without them, thus they would be just adding complexity to the machine learning pipeline.

2.5.1 Smoothing

The first pipeline component was smoothing. For the fertilisation and dry matter experiment the Savitzky-Golay filter (Press and Teukolsky, 1990) was used for data smoothing. Savitzky-Golay is a smoothing algorithm used in signal processing and data analysis that applies a convolution operation with a polynomial window to the input data, intending to smooth noisy signals through successive subsets fitting of data points that are adjacent together with a polynomial of low-degree using the method of linear least squares. This filter is commonly used in signal processing and data analysis (Vivó-Truyols and Schoenmakers, 2006). The reason for that is it diminishes high-frequency signals, such as noise, while simultaneously maintaining essential characteristics of the signals, such as relative peaks, troughs, height, and width(Zimmermann and Kohler, 2013). Moreover, it is computationally efficient. For the irrigation experiment the LOWESS method was used (Locally Weighted Scatterplot Smoothing), a non-parametric regression technique used for smoothing data points, which works by fitting

a weighted polynomial regression to localized subsets of the data and adjusting the weights based on proximity.

The aim of this stage was (i) smoothing of different bands signal and (ii) to reduce noise. This was essential as overfitting can be caused by noise damaging, in the end, the ability of the model to generalize new previously unseen data. Moreover, the filter aids highlighting the underlying trends and patterns in the signal data. The Savitzky-Golay filter can be adjusted by changing the window size and the polynomial order according to the dataset characteristics and the machine-learning task requirements.

2.5.2 Scaling

An important preprocessing step in machine learning is data scaling. This step adjusts the data attributes range to make sure that the contribution of each is equal to the learning process. Various machine learning algorithms, and more specifically the ones that involve distance calculations or gradient descent-based algorithms such as linear regression and neural networks, presume that all features have the same scale. In case of the opposite, it may lead to some features dominating the calculations for the distance or the steps of gradient descent, therefore leading to models that are biased and that prioritize these features. Moreover, scaling makes sure that all variables are affected by the regularization term equally and that the kernel functions calculate similarities based on a standardized feature space. Selecting the correct scaling method depends on the specific requirements of the machine learning model being used and the nature of the data and. For the generalisation experiment three (3) different strategies for scaling were tested: (i) no scaling (ii) Standardization and (iii) 0-1 Scaling (Normalization). Normalization transforms each feature individually so that it falls in the zero- one range while Standardization adjusts the features so that they have standard normal distribution properties with a mean of zero (0) and a standard deviation of one (1). Combining Normalization and Standardization was not evaluated to avoid importing redundancy in the pipeline. Furthermore, these two strategies may cause an increased sensitivity to outliers, and as result distort performance metrics, or they may reduce how comparable the performance of each algorithm is as each algorithm has lower or greater compatibility with the scaling techniques listed. Finally, for the irrigation and fertilisation experiments scaling was conducted using normalization.

2.5.3 Recursive Feature Elimination

Recursive Feature Elimination (RFE) technique is used in regression problems to select a small set of relevant features to be used for model construction through the repeated elimination of features according to the scores of a fitted model. This dissertation selected a Bayesia-based regressor to obtain the scores. The process of selection was repeated till half the original features were eliminated or until the model's performance stopped improving significantly.

2.5.4 Univariate Feature Selection

Univariate feature selection was used in this dissertation to refine processing of data by removing redundant or irrelevant attributes. Techniques such as this prevent overfitting by selection of the most important wavelengths and increase the efficiency of the selected ML models. Univariate approaches evaluate the relationship of each pair of individual feature and its target variable. Two univariate filter methods were used for the dry matter content generalization study: (i) Mutual information and (ii) F-test for regression. The first one assesses the linear association between a feature and its target variable and reports the F-statistic and its corresponding p-values which show the linear relationship strength. Features with higher F-statistics and lower p-values are considered significant features and are selected. Mutual information quantifies the level to which the knowledge of one variable lowers the uncertainty regarding another variable by calculating the mutual information of each feature and its target. A benefit of Mutual information is that it can identify nonlinear associations. Finally, both irrigation and fertilisation experiments made use of only the F-test.

2.5.5 Feature Extraction

In the next step of the data pre-processing pipeline followed in this dissertation, a feature extraction step was used to reduce data dimensions. Firstly, as the number of dimensions increases, the data required to adequately represent and generalize patterns grows exponentially (Hughes, 1968). As a result, ML models might struggle to find meaningful patterns due to sparsity, leading to overfitting or increased computational complexity. Secondly, multidimensional data often contain redundant or irrelevant features. This can confuse models, impacting their ability to distinguish between relevant and noisy information, potentially reducing predictive accuracy (Loggenberg et al., 2018). Lastly, high-dimensional data can make models more complex, leading to longer training times, increased computational resource requirements, and challenges in model

interpretability(Cozzolino et al., 2010). This step apart from feature selection, could also reduce computational load and noise. For the fertilisation and irrigation experiments Principal Component Analysis (PCA) was used. In the generalisation experiment three (3) methods were evaluated for feature extraction: (i) Uniform Manifold Approximation and Projection (UMAP), (ii) Autoencoder, and (iii) Principal Component Analysis (PCA).

PCA reduces complexity of the dataset by retaining the most important principal components, with their ranking being based on dataset variance they account for. The principal components, derived from PCA, are intentionally uncorrelated. This lack of correlation ensures that each component provides unique information, leading to a more concise representation of the data. This concise representation simplifies the dataset, enhancing the efficiency and clarity of subsequent analyses. UMAP is another technique for reducing dimensions in a nonlinear way, whose aim is to maintain both the global and local structures of data that is high-dimensional. UMAP highlights the local proximity of data points, thus preserving nonlinear correlations and intricate patterns that are ignored by other methods such as PCA. Finally, Autoencoders, are neural networks designed for unsupervised learning. They consist of two parts, the encoder which compresses data into a condensed latent-space representation, and the decoder which reconstructs the data from this latent representation. During the training process, the network ensures that the latent space depicts the most important data characteristics by minimizing the difference between the original data and its reconstruction. Due to the nonlinearity of its transformations, more insightful embeddings could be achieved compared to PCA. The selected architecture for the generalisation experiment consisted of two (2) layers in the encoder. The first layer duplicated the original component number, and the second layer projected the components into a feature space of 8, 16, 24, or 32. The decoder was used to reconstruct the original features and had symmetrical architecture.

2.5.6 Polynomial Transformation

A way to significantly enhance the model's ability to capture complex relationships within the data is through the use of polynomial feature transformation. Datasets often contain non-linear relationships which a simple linear model cannot capture effectively. Moreover, polynomial features include the interaction terms between different features as well as the higher degree terms of those individual features. These terms can provide valuable information about the combined effect of two or more variables on the target variable. This dissertation evaluated only the quadratic transformation since, in early experiments, higher-degree polynomial features led to overfitting, and the performance was poor on the test set. Furthermore, adding polynomial features can rapidly increase features number, especially with higher degrees and with datasets that contain many original features. This was the case in this dissertation. As a result, the use of all original

features with the polynomial regression was producing an unstable experimental pipeline whose reproducibility was heavily affected by the computational costs and the reiterative crashes after hours of execution without reporting a single result.

An overview of all the pre-processing methods used across all three experiments is shown below:

Table 14. Pre-processing methods used for the experiments (part 1)

Experiment	**Smoothing**	**Scaling**
Fertilisation	• Savitzky-Golay filter	• Normalization
Irrigation	• LOWESS	• Normalization
Dry matter/Generalisation	• Savitzky-Golay filter	• No scaling • Normalization • Standardization

Table 15. Pre-processing methods used for the experiments (part 2)

Experiment	**Recursive Feature Elimination**	**Univariate Feature selection**
Fertilisation	• N/A	• F-test
Irrigation	• N/A	• F-test
Dry mater/Generalisation	• Yes	• F-test • Mutual Information

Table 16. Pre-processing methods used for the experiments (part 3)

Experiment	**Feature extraction**	**Polynomial Transformation**
Fertilisation	• N/A	• N/A
Irrigation	• PCA	• N/A
Dry mater/Generalisation	• PCA • UMAP • Autoencoder	• Quadratic transformation

2.6 Model generalisation pipeline configuration developed for dry matter estimation

To investigate whether the generalisation performance of models trained using a multi class dataset could be further improved by introducing data pre-processing steps. The different stages of data processing were evaluated towards reporting which components were empirically the most best performing ones. The pre-processing steps (Figure 24) include data smoothing, data scaling, feature selection and extraction, and finally feature polynomial transformation. The target of the integration of the previously mentioned components is boosting performance in an iterative way while at the same time allowing the understanding of how each of the elements negatively or positively affects the regression problem. It is crucial to note that despite the potential of each component to improve performance, the interaction among each of them and the nature of the dataset used in this particular research could result in negative performance that could open the discussion for its use or not.

Different preprocessing method orders were used during the experiments. For instance, the polynomial transformation was also evaluated before applying feature extraction. However, as they failed to achieve high performance, they are not shown in Figure 24, which depicts a synthetic version of the most reliable pipeline. Additional details are presented in the Discussion section.

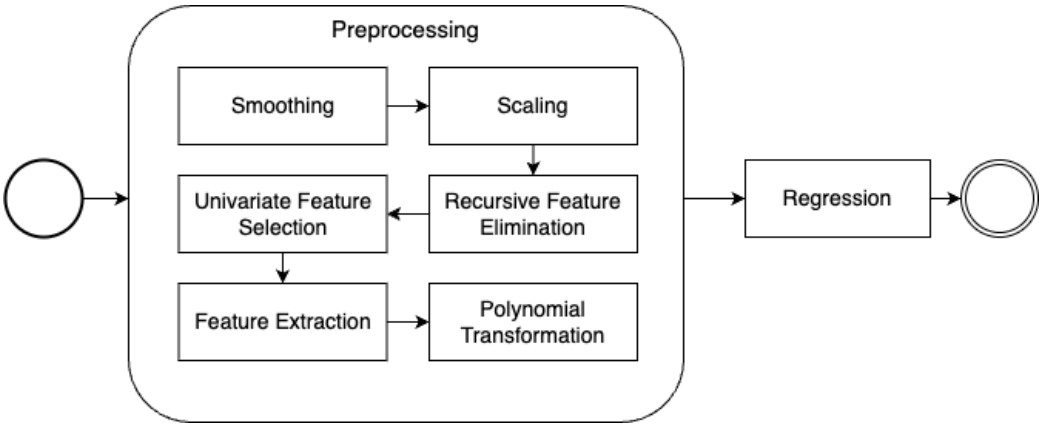

Figure 24. Proposed dry-matter analysis pipeline procedure set up during this study.

Different values and algorithms could be used for each of the preprocessing steps. The configurations used and tested in this particular study are listed in the table below. Finally, it is essential to emphasize that each time a processing step was added to the pipeline, the whole pipeline was executed from the beginning. As a result, the selection of the hyper-parameters and algorithms achieving the best results could be different.

Table 17. Evaluated configurations for each pre processing stage for the dry matter content generalisation experiment.

Preprocessing Stage	Value
Smoothing [Window sizes]	[0, 4, 10]
Scaling techniques	[Standardization, 0-1 Normalization]
Univariate Feature Selection [Criteria]	[Mutual Information, F-Test]
Univariate Feature Selection [Output Features]	[8, 16, 24, 32]
Feature Extraction [Algorithms]	[UMAP, PCA, Autoencoder]
Feature Extraction [Output Features]	[Selected Features / 4, Selected Features / 2, 2 * Selected Features / 3]
Polynomial Transformations	[Quadratic / No Transformation]
Regression Algorithms	[ARD / PLS]

2.7 Machine learning experimentation framework

AutoML was evaluated for the irrigation and fertilisation experiments. In recent years, various Automated Machine Learning (AutoML) frameworks have emerged that enable computers to autonomously discover the most appropriate machine learning pipeline tailored to a particular task and dataset. AutoML solutions can either be cloud-based, such as Microsoft Azure Machine Learning(Barnes, 2015), Google Cloud AutoML Vision(Bisong and Bisong, 2019), , and Apple's Create ML, or open-source such as PyCaret AutoKeras(Jin et al., 2023), , Auto-WEKA 2.0 (Kotthoff et al., 2017), H2O AutoML(LeDell and Poirier, 2020), AutoSklearn (Feurer et al., 2015), TPOT(Le et al., 2020), autoxgboost(Thomas et al., 2018), and OBOE(Yang et al., 2019). For this dissertation, PyCaret (Ali, 2020) was the AutoML framework of choice.

PyCaret is a Python-based, low-code machine learning library designed to streamline the experimental model building and deployment process. It offers a wide array of algorithms and automated processes for feature engineering, model selection, hyperparameter tuning, and model evaluation, covering the entire machine learning workflow. PyCaret offers significant advantages to both beginners and experts, providing a range of features for quick experimentation and comparative analysis across different models and datasets. This allows users to concentrate on model conceptualization and analysis rather than coding intricacies.

For the Classification task, PyCaret searched for the best machine learning algorithm from a list of 14 algorithms, which ranged from simple to more complex ones. Specifically, PyCaret searched through. Linear classifiers (Logistic Regression and Ridge Classifier with L2 regularization). Tree-based models such as Random Forest, Extra Trees, Gradient Boosting, and Decision Tree classifiers, using them independently or in an ensemble to make predictions Instance-based classifiers (K Neighbours Classifier) classifying samples based on the majority class among their nearest neighbours. Boosting algorithms (Ada Boost and Light Gradient Boosting Machine) to sequentially build weak classifiers to form a robust model. Hyperplane based classifiers (Support Vector Machine with a linear kernel) to separate classes. Dimensionality reduction and linear discrimination classifiers (Linear Discriminant Analysis and Quadratic Discriminant Analysis) and the Naive Bayes classifier which utilizes probabilistic methods based on Bayes' theorem with the "naive" independence assumption.

Finally, for the AutoML experiments, a StratifiedKFold with 10 folds was used to evaluate the models based on accuracy, recall, precision, and F1-score. The model evaluation metrics were recorded, and then the AutoML system searched for a possibly better-performing solution by executing automated hyperparameter tuning. However, while AutoML can operate without specific configurations, setting experimental constraints becomes necessary to extract insights into this process and understand both

its limitations and advantages. The AutoML framework used is presented in the figure below.

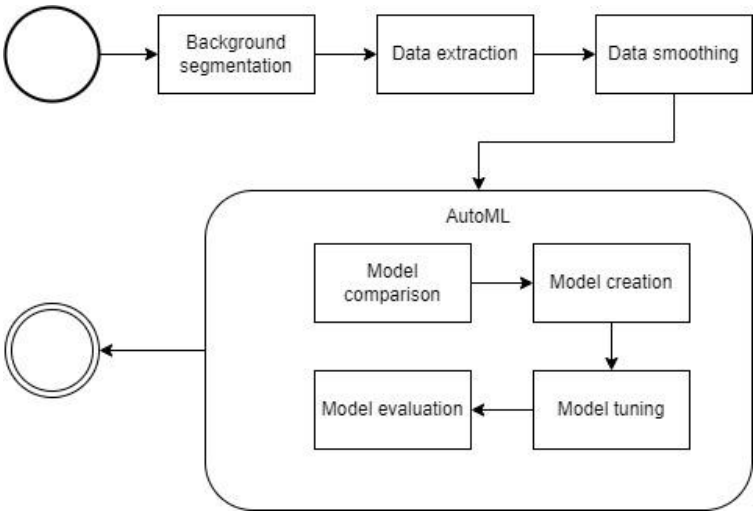

Figure 25. AutoML framework pipeline used in the fertilization experiment

For classification experiments (fertilisation and water stress/ acclimation experiments), the results of the AutoML system were compared with Partial Least Squares Discriminant Analysis (PLS-DA). PLS-DA is a multivariate statistical technique used to analyse high-dimensional data to discriminate or classify between categories/ groups and is a widely used machine learning algorithm for hyperspectral imaging applications.

PLS-DA operates by establishing components/latent variables, which are linear combinations of the original predictors. These components/ latent variables are derived in a way that maximizes the covariance between predictor variables (independent) and categorical class variables (dependent). This process enables PLS-DA to identify underlying structures that differentiate between classes while taking into consideration multicollinearity issues commonly encountered in high-dimensional data.

As a result, PLS-DA achieves dimensionality reduction of the dataset while retaining essential information crucial to the classification task. Moreover, it facilitates the prediction of categorical outcomes for new observations using the learned associations between predictors and class labels.

Two PLS-DA algorithm versions have been developed, PLS1-DA(Cozzolino et al., 2010)(Liu et al., 2008)(Lee and Jemain, 2019)(Pan et al., 2015)(Xia et al., 2019) and PLS2-DA(Vieira et al., 2020)(Marquetti et al., 2016)(Bronzi et al., 2020)(Manheim et al., 2016). PLS2-DA is applied when the objective is to discriminate among multiple groups

simultaneously and when the dataset contains multiple sets of response variables associated with the same set of predictors. In this dissertation, PLS1-DA was selected to discriminate between a set of response variables (control – test).

Regarding the regression experiment (dry matter estimation), two (2) regression methods were tested: Automatic Relevance Determination (ARD) (Wipf and Nagarajan, 2007) and Partial Least Squares Regression (PLS) (Wold et al., 2001). ARD regression is a type of linear regression that differentiates itself from typical linear regression by incorporating Bayesian inference. ARD Regression introduces another Gaussian prior on each regression model weight. These priors' variance enables the model to automatically adjust each feature relevance. Features that show nearly zero variances have their weights reduced toward zero, thus eliminating them from the model.

PLS Regression is an algorithm that projects the input features and the target to a different space to find a linear regression model. It achieves better results on occasions where the predictor matrix consists of more variables than observations and when input values show multicollinearity. The main advantage of PLS Regression is its ability to handle cases with numerous correlated predictors and cases with more predictors than observations.

For the dry matter regression problem, the experiments were executed ten (10) times using a 5-fold cross-validation setting (10x5-fold cv) to further enhance precision of the final performance assessments.

Finally, for all experiments, data processing was carried out by programming in Python 3.10, while data preprocessing was conducted using scikit-learn 1.3.2 and SciPy v.14.1. For each experiment different code was written and used. The irrigation and fertilization experiments further used PyCaret 3.0 for the data analysis, while the dry matter generalization experiments made additional use of the UMAP library. For background removal and scikit-image 0.21.0 was used.

2.8 Evaluation metrics

The performance of all tested classification models was evaluated using Accuracy, the most common way to evaluate a classification model, which works well when the dataset is balanced.

$$Accuracy = 100 \text{ x } \frac{\text{tp+tn}}{tp+tn+fp+fn} \quad [11]$$

However, since the datasets for the water stress and the fertilisation were unbalanced, it was crucial to compute the micro-averaged F1 score for comparative purposes. This study favours this specific aggregation technique over the macro-average, especially when dealing with class imbalances, as observed in all datasets.

$$F1Score = \frac{2 \times precision \times recall}{precision+recall} \quad [12]$$

Recall measures the proportion of accurately identified categories from the original dataset,

$$Recall = 100 \text{ x } \frac{\text{tp}}{tp+fn} \quad [13]$$

while precision gauges the accuracy of labels in the classifier's output

$$Precision = 100 \text{ x } \frac{\text{tp}}{tp+fp} \quad [14]$$

where t_p = true positives, f_p = false positives, t_n = true negatives, and f_n = false negatives

Finally, the dry matter content estimation study assessed the different preprocessing components towards enhancing the modelling abilities of the two selected and tested regression methods. Root Mean Squared Error on Prediction (RMSEP) was used to measure prediction accuracy. Moreover, to improve the accuracy of the assessments, experiments were conducted ten (10) times under a 5-fold cross-validation setting (10x5-fold cv). Additionally, R^2 ,the adjusted coefficient of determination, was utilized for removing the pipelines that did not manage to report in a constant way performances above the mean of the dry matter output. Lastly, it's important to note that the whole pipeline run from the start whenever a new processing step was introduced. This meant that the optimal combination of algorithms and hyperparameters could vary at each stage, depending on the adjustments made.

2.9 Statistical analysis

Pearson's correlation coefficient (r) calculations were conducted to explore the relationships between CIELAB and Spectral wavelengths. The objective was to assess whether there is relationship between the two types of optical measurements. Descriptive statistics, including average, min, max and standard deviation, were computed for the broccoli weight during first year to provide a comprehensive overview of broccoli production under different fertilisation schemes.

Chapter 3 Results

3.1 Testing the Suitability of Automated Machine Learning, hyperspectral imaging and CIELAB color space for proximal in situ fertilisation level classification

Over the first year, two different fertilization levels were investigated, full fertilisation (control) and half fertilisation with the first class consisting of 28 plants/samples and the second one of 21 plants/ samples summing up to a total of 49 samples. Hyperspectral and CIELAB measurements were conducted for each of the broccoli plants, and three datasets were constructed. The first comprised only of hyperspectral data, the second only of CIELAB data and the third one was the result of merging the two. Each of those datasets was then used to train artificial intelligence classification models either using traditional algorithms and namely PLS-DA or AutoML solutions and the PyCaret library.

In Figure 26, each fertilisation class's extracted average spectra (spectral signature) are visualized. In contrast to the color image, differences between the two classes in the red-edge (660-760 nm) and NIR wavelengths (>770 nm) are observable.

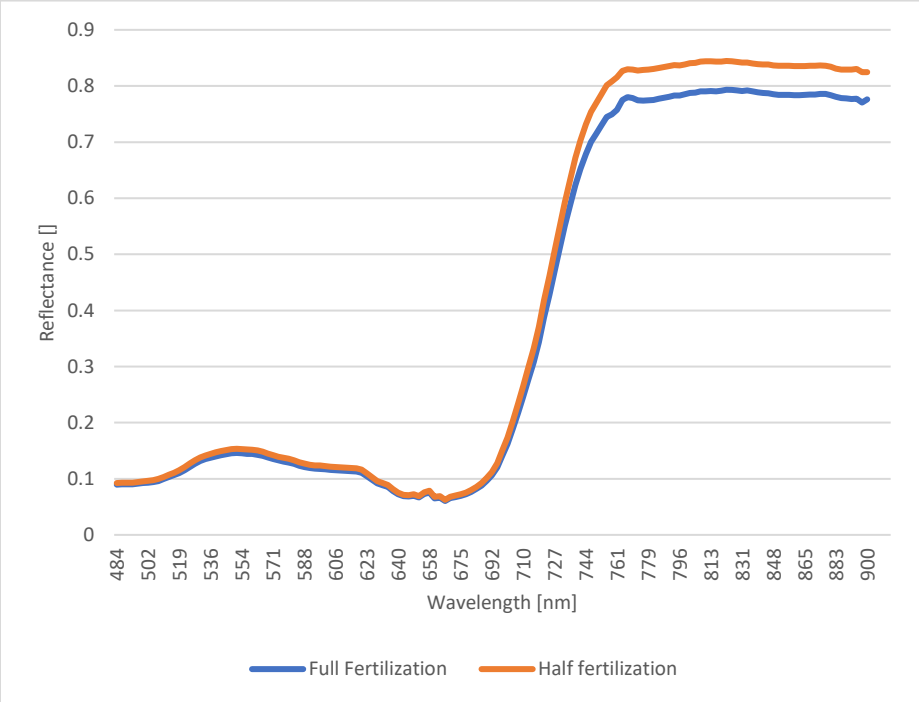

Figure 26. Spectral signature (average spectra) for the two fertilization classes

3.1.1 Training using AutoML

For the classification task the AutoML system, achieved an accuracy and F1-score of 1.00 using the hyperspectral dataset, outperforming the CIELAB dataset, which achieved an accuracy of 0.72 and F1-score of 0.68. Table 18 lists the best-performing machine learning algorithms compared by PyCaret.

Table 18. Best-performing PyCaret algorithms

Dataset	ML algorithm	Accuracy	Recall	Precision	F1-score
CIELAB	MLP	**0.72**	0.65	0.75	0.68
CIELAB	Gaussian Process Classifier	0.67	0.70	0.63	0.64
CIELAB	Logistic Regression	0.66	0.55	0.60	0.57
CIELAB	SVM - Radial Kernel	0.64	0.60	0.67	0.60
Hyperspectral	Ada Boost Classifier	**1.00**	1.00	1.00	1.00
Hyperspectral	Decision Trees	**1.00**	1.00	1.00	1.00
Hyperspectral	Extra Trees	**1.00**	1.00	1.00	1.00
Hyperspectral	Random Forest	**1.00**	1.00	1.00	1.00
Hyperspectral	Extreme Gradient Boosting	**1.00**	1.00	1.00	1.00
Hyperspectral	CatBoost	**1.00**	1.00	1.00	1.00
Hyperspectral	Gradient Boosting	**1.00**	1.00	1.00	1.00

The hyperspectral dataset yielded the best results with the Decision Trees, Extra Trees, Random Forest, Ada Boost Classifier, CatBoost, Extreme Gradient Boosting and Gradient Boosting algorithms. Meanwhile, the MLP algorithm performed the best on the CIELAB dataset, followed by the Gaussian Process Classifier, Logistic Regression, and SVM - Radial Kernel.

The study also explored the interaction between hyperspectral and CIELAB color data, aiming to provide a more comprehensive analysis of the subject. The results demonstrated that the combined dataset offered better performance than the CIELAB data alone, though it did not surpass the hyperspectral dataset in accuracy. The three top-performing algorithms were Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and the MLP Classifier. Among these, LDA performed the best, achieving an accuracy of 0.94 and an F1-score of 0.87. Table 19 presents the performance metrics of all three algorithms for the combined dataset.

Table 19. Combined dataset performance

Dataset	ML algorithm	Accuracy	Recall	Precision	F1-score
Hyperspectral + CIELAB	LDA	**0.94**	0.85	0.90	0.87
Hyperspectral + CIELAB	QDA	0.77	0.70	0.58	0.63
Hyperspectral + CIELAB	MLP	0.71	0.55	0.65	0.55

3.1.2 Training using PLS-DA

Upon finishing the training and evaluation of both datasets using the AutoML pipeline, the hyperspectral dataset, the best-performing dataset, was used to train a PLS-DA model. PLS-DA is a machine learning algorithm commonly used in hyperspectral imaging applications and can serve as a benchmark for comparing AutoML classification performance. Figure 27 presents the PLS latent variable cross-decomposition plot. The cross-decomposition graph of latent variables serves as a visual tool that illustrates the relationships between the latent variables and the variability present in the original dataset. This type of graph helps to uncover patterns and connections that may not be immediately evident, offering deeper insights into how the latent variables capture the underlying structure of the data.

Figure 27 shows that the two classes can be effectively distinguished by using latent variables 1 and 2. In the lower left quadrant the half fertilisation samples are gathered, with the full fertilisation samples being scattered in the upper left and lower right quadrants.

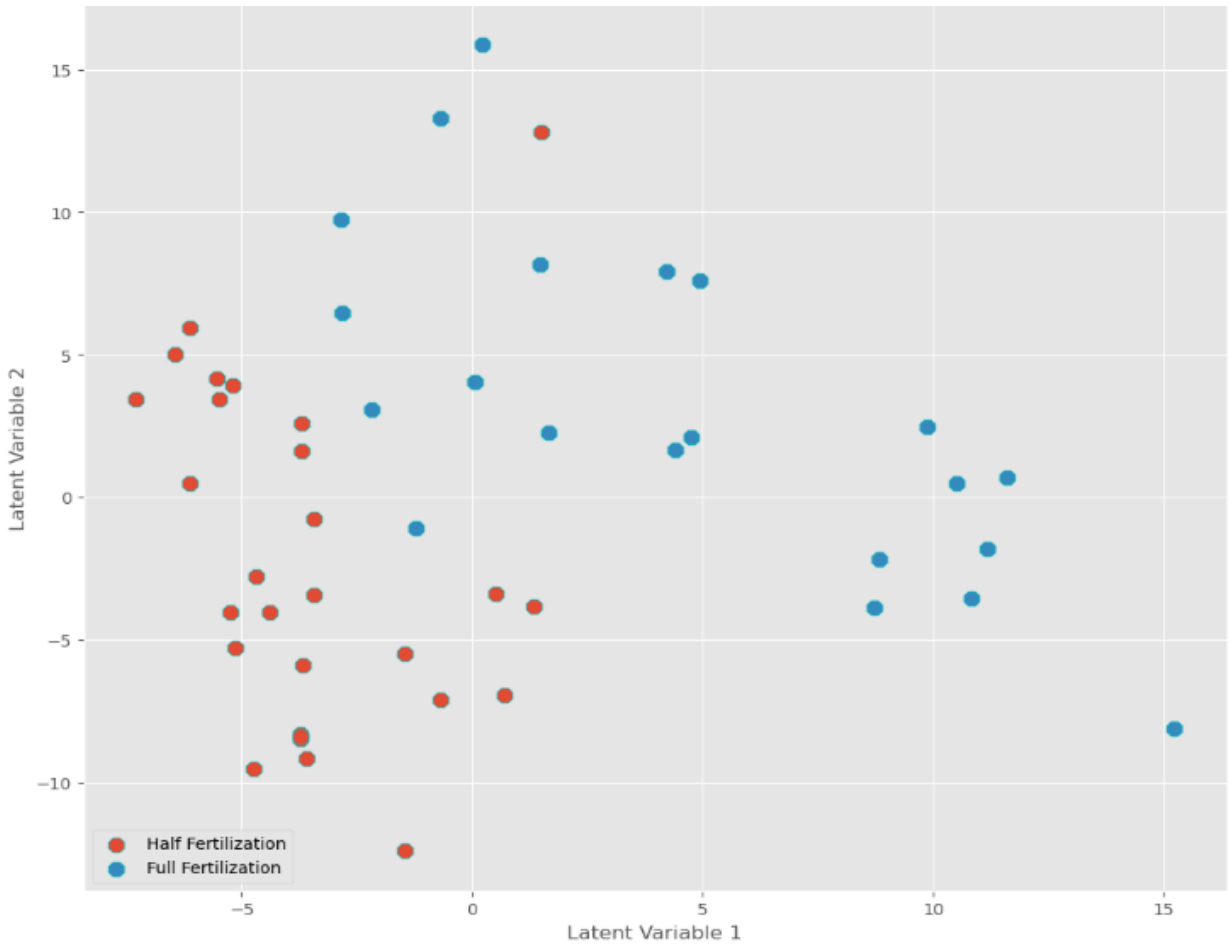

Figure 27. PLS cross-decomposition Score plot. The red dots represent half fertilization samples, and the blue dots represent the full fertilization samples.

However, the performance of the PLS-DA algorithm does not match that of the AutoML system, achieving an accuracy and F1-score of 0.91, compared to the perfect score of 1.00 achieved by the AutoML system for both performance metrics. Table20 contains the detailed performance evaluation for the PLS-DA.

Table 20. PLS-DA algorithm performance

Dataset	ML algorithm	Accuracy	Recall	Precision	F1-score
Hyperspectral	PLS-DA	**0.91**	0.88	0.95	0.91

3.1.3 Training using AutoML and a single-feature dataset

The performance of the AutoML system was further assessed using a dataset with a single feature, building on the strong results previously obtained with the full hyperspectral dataset. Feature selection was conducted with the scikit-learn library, utilizing the ANOVA F-statistic as the scoring function to narrow down the dimensions of the hyperspectral data.

The possibility of achieving good classification performance with just one wavelength was explored by setting the number of desired wavelengths to 1. The chosen wavelength based on the Anova F-statistic was in the near-infrared (NIR) region, specifically at 874 nm.

The AutoML system accurately classified all samples, achieving an accuracy and F1-score of 1.00. The top-performing algorithms were the Extra Trees Classifier, Decision Tree Classifier, Ada Boost Classifier, CatBoost Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Extreme Gradient Boosting. These classifiers also performed the best when using the full hyperspectral dataset (Table 21).

Table 21. Single feature dataset performance

Dataset	ML algorithm	Accuracy	Recall	Precision	F1-score
Hyperspectral single wavelength	Decision Tree Classifier	**1.00**	**1.00**	**1.00**	**1.00**
Hyperspectral single wavelength	Ada Boost Classifier	**1.00**	**1.00**	**1.00**	**1.00**
Hyperspectral single wavelength	Random Forest Classifier	**1.00**	**1.00**	**1.00**	**1.00**
Hyperspectral single wavelength	CatBoost Classifier	**1.00**	**1.00**	**1.00**	**1.00**
Hyperspectral single wavelength	Extra Trees Classifier	**1.00**	**1.00**	**1.00**	**1.00**
Hyperspectral single wavelength	Gradient Boosting Classifier	**1.00**	**1.00**	**1.00**	**1.00**
Hyperspectral single wavelength	Extreme Gradient Boosting	**1.00**	**1.00**	**1.00**	**1.00**

3.2 Early detection of broccoli drought acclimation/stress in agricultural environments utilising proximal hyperspectral imaging and AutoML

During the second year the focus was on detecting drought acclimation/stress. Two different irrigation schemes were used, which in turn led to the creation of two datasets, drought onset dataset and drought acclimated dataset, each consisting of two classes control and drought onset/ drought acclimated respectively. The drought onset dataset consisted of 42 images/samples while the drought acclimated of 48. Finally, a third dataset was created by merging the two datasets. The merged dataset contained 90 images/samples of all three classes control, drought onset and drought acclimated. All datasets were used to train artificial intelligence classification models. The models were trained either traditional algorithms and namely PLS-DA or AutoML and the PyCaret library.

This section presents the classification results using PLS1-DA and the AutoML framework on all three datasets. Various pre-processing techniques (smoothing combined with either dimensionality reduction or feature selection) were evaluated separately to determine how they affect both the PLS1-DA and AutoML classification metrics.

The extracted spectral signatures are visualized in the following figures to provide insight into the data used for classification purposes. Namely, Figure 28 displays the average spectral signatures of the plants imaged at the drought onset, and Figure 29 the average spectral signatures of the plants imaged at the drought acclimation stage.

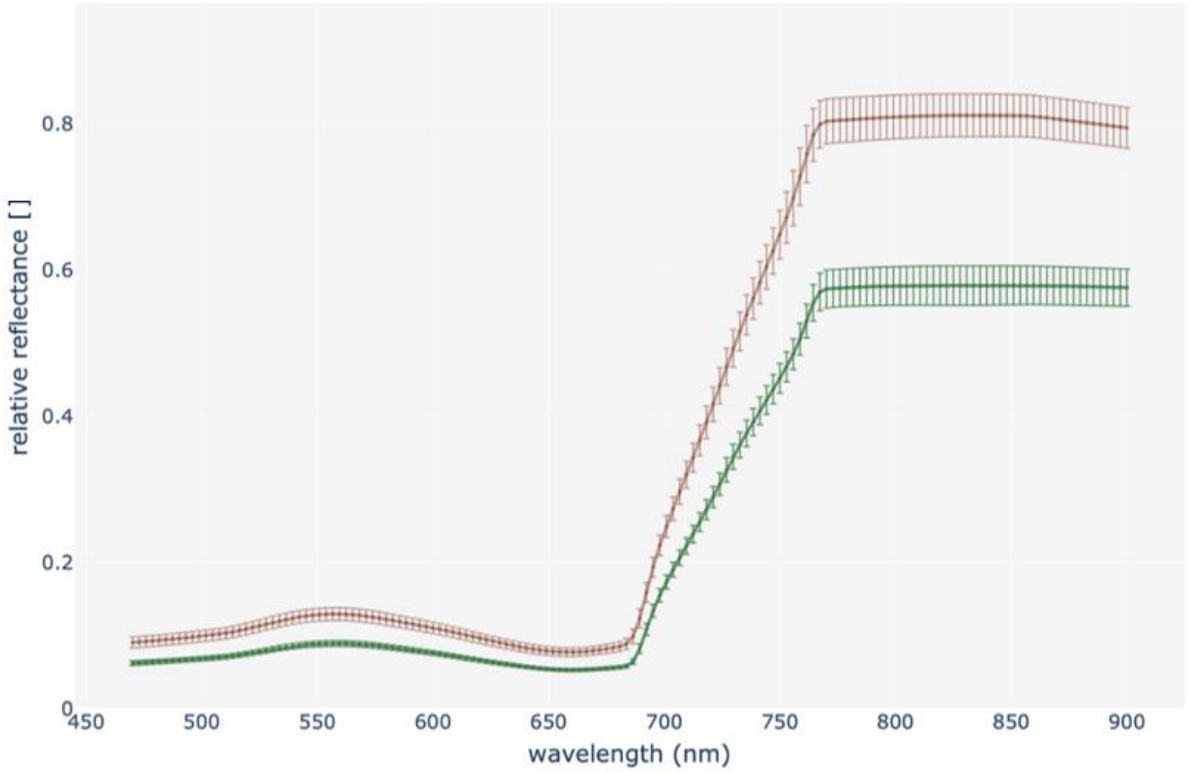

Figure 28. Mean spectral signature of broccoli canopy at the drought onset. With red is depicted the control group, and with green the drought. 95% CI are also presented.

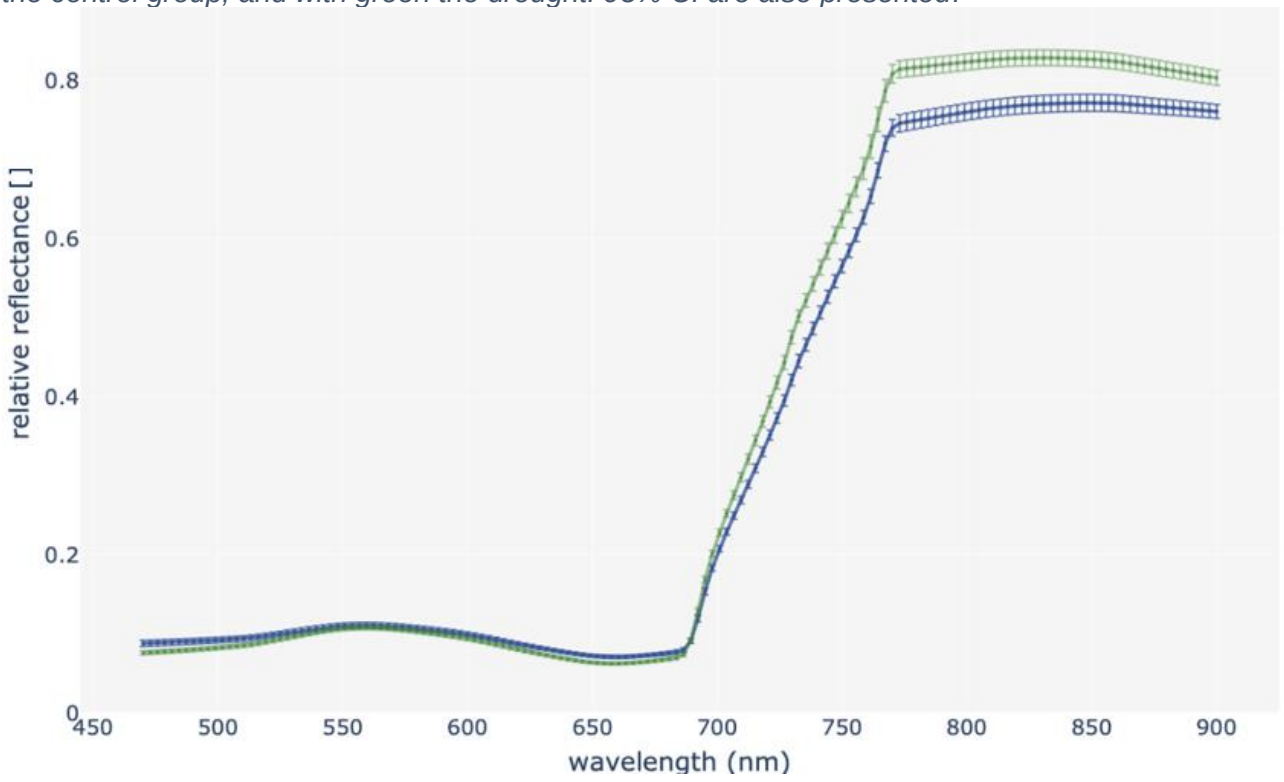

Figure 29. Mean spectral signature of broccoli canopy at the drought acclimation. With green is depicted the control group, and with blue the drought. 95% CI are also presented.

Statistically significant differences (independent t-test at $\alpha = 0.05$) are observed within both datasets in the NIR region, as opposed to the visible spectrum (including red-edge), in which statistically significant differences are found only in the onset of drought dataset. Between the different datasets, statistically significant differences (paired t-test at $\alpha = 0.05$) are observed for the control group only in the visible spectrum and for the drought group in the full measured spectrum. At the onset of drought, the maximum photosynthetic rate (p-value = 0.06 at $\alpha = 0.05$) and the stomatal conductance (p-value = 0.055 at $\alpha = 0.05$) were nearly significant, while the transpiration rate was statistically significant (p-value = 0.03 at $\alpha = 0.05$) between the treatments, indicating a drought stress onset. This could be attributed to the timing of hyper-spectral imaging and gas exchange sampling. The drought group did not receive any irrigation for four days, from which the initial three were cloudy (low evapotranspiration), in contrast to the fourth day which was particularly hot (high evapotranspiration, limited available substrate water content). Physiological parameters were measured early in the morning of the fourth day to validate the onset of the short-term drought acclimation that progressed due to the environmental conditions, in an early loss of turgor late in the afternoon, when imaging took place. Based on the laboratory measurements, drought stress did not occur on any date. Though, statistically significant differences were observed in the stomatal conductance, maximum photosynthetic rate, and the transpiration rate between the treatments 16 days after drought initiation. From the previous can be concluded that the drought group was acclimated due to the drought conditions. On that day, hyper-spectral imaging and physiological measurements were conducted within a two-hour difference in the morning.

Although this dissertation provides preliminary insights into drought acclimation level, the observed difference in spectral signatures requires further investigation in future research to draw solid conclusions. Prior to training the model PCA was used due to the high collinearity of the spectral data collected (figures 30 and 31).

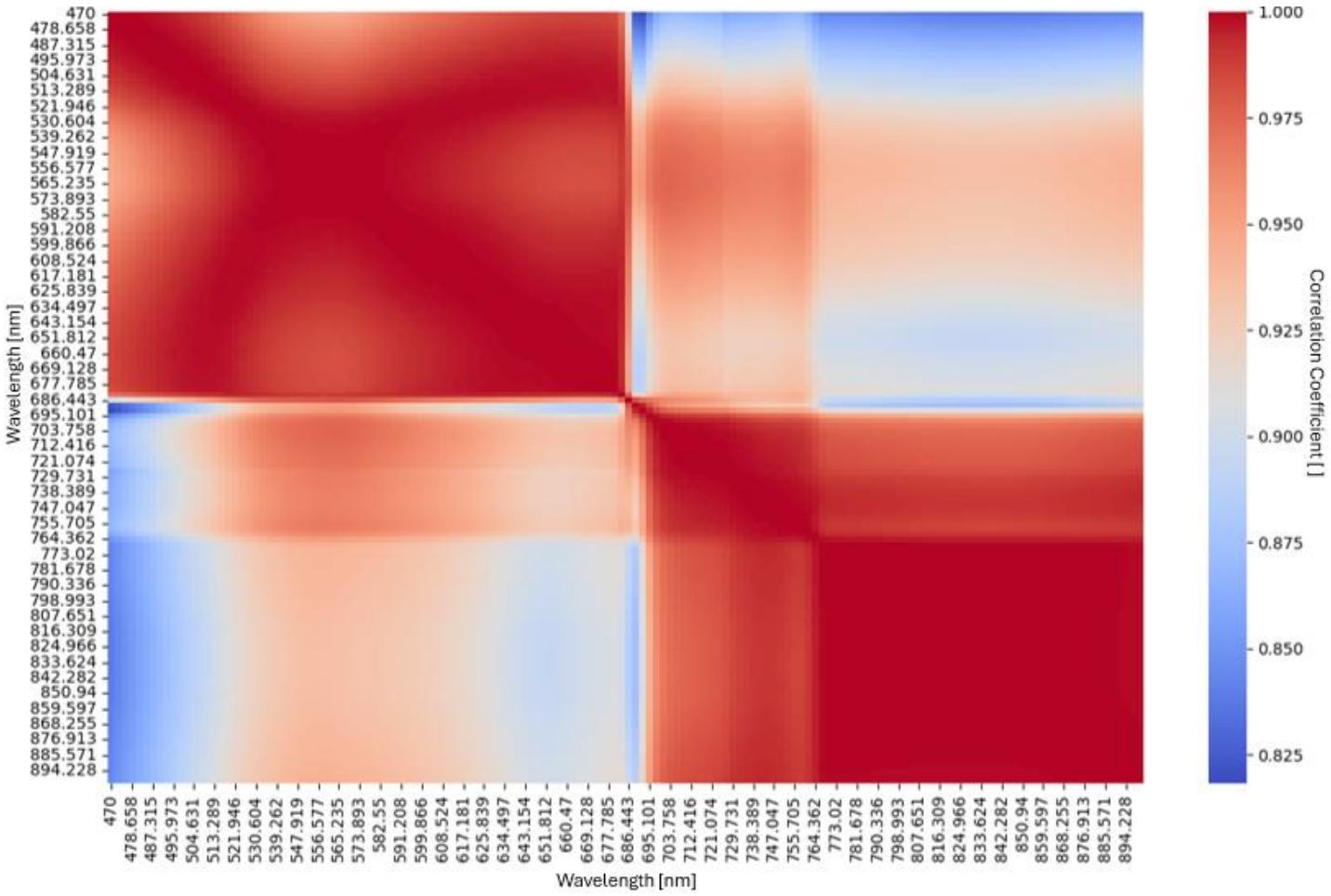

Figure 30. Correlation matrix of drought onset dataset. Highly correlated data appear in red. Should be noted that the baseline of the correlation coefficient for this dataset is 0.825.

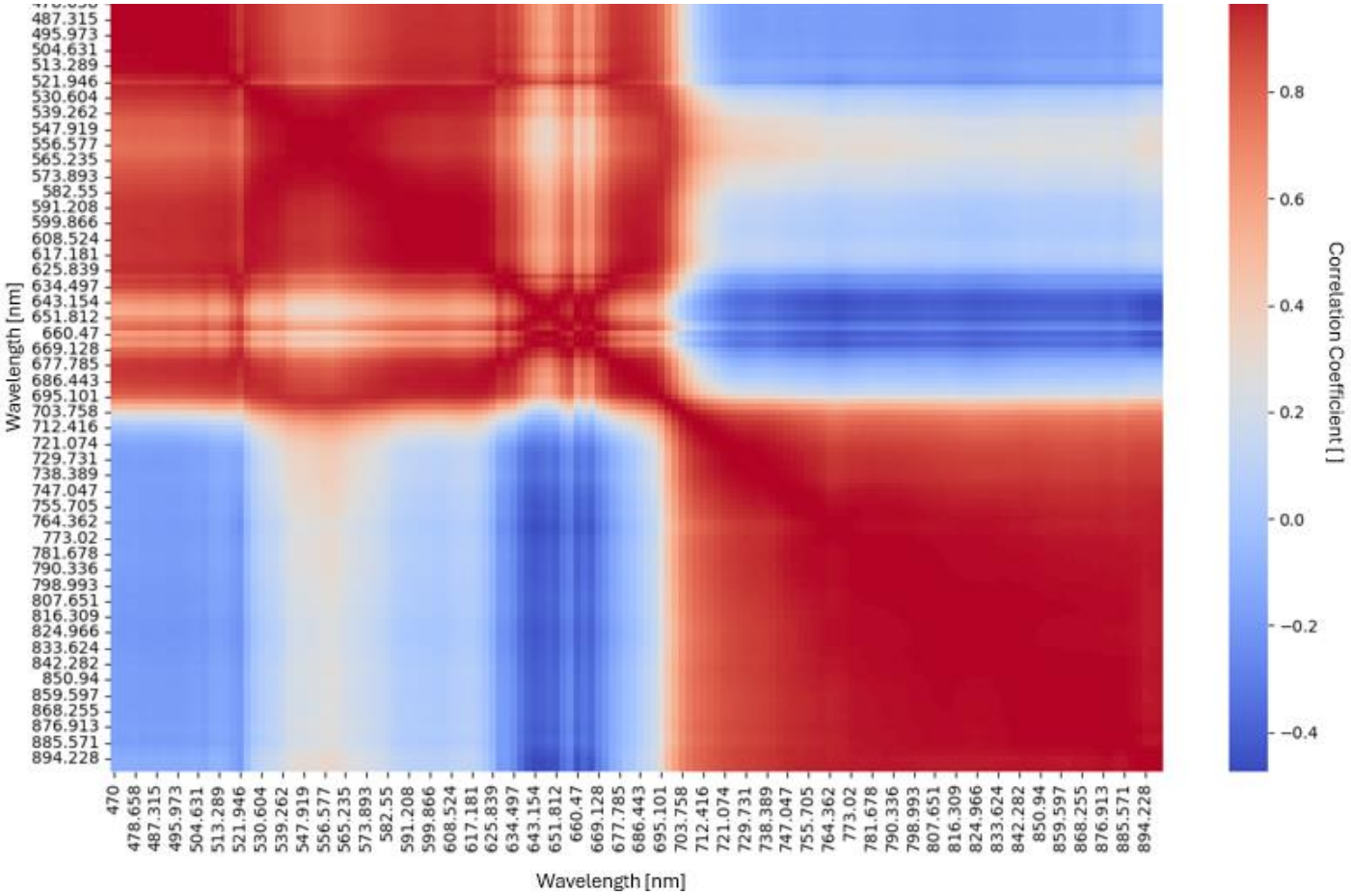

Figure 31. Correlation matrix of drought acclimated dataset. Highly correlated (1.0) data appear in red, while the least correlated in whitish blue (0.0).

3.2.1 Training using AutoML

The AutoML framework evaluated 14 different classifiers. Each classifier was cross-validated and fine-tuned using PyCaret, which automatically searches and applies the best hyperparameter configurations.

The AutoML framework managed to accurately classify between both drought onset/ control and drought acclimated/control plants, accuracy and F1-score of 1.00 on the hold-out subset regardless of the drought level. All the pre-processing techniques and combinations achieved excellent performance (accuracy > 0.90). The use of only the five least correlated wavelengths (~478nm, ~530nm, ~672nm, ~770nm, ~850nm) instead of all 150, provided similar perfomance for most of the classifiers evaluated by the AutoML framework, underlining the impact of multicollinearity in hyperspectral data.

Table 22 presents the highest performing classifiers within the 5-fold cross validation (CV-val) of the train subset using AutoML across both acclimation levels and using the various pre-processing techniques. Because of the imbalance in the two classes both F1-score and accuracy are presented. CV-train metrics (mean validation results within the 5-fold cross validation by using the training folds) were calculated as well but not presented for simplicity. Both CV-train and CV-val performance was quite similar, indicating non-overfitted data.

Table 22. AutoML CV-val performance across both drought levels and pre-processing techniques. The standard deviation (SD) is provided in parentheses

Dataset	**Pre-processing technique**	**Architecture**	**Accuracy**	**F1-Score**
drought onset	LOWESS	qda/ridge/ svm/lda	**1.00 (0.00)**	1.00 (0.00)
drought onset	LOWESS & PCA	ridge/et	0.9667 (0.067)	0.971 (0.057)
drought onset	LOWESS (5 features)	qda/nb/ridge	0.9600 (0.0800)	0.9333 (0.1333)
drought acclimated	LOWESS	lr/knn/nb qda/ridge/et/lda	**1.00 (0.00)**	1.00 (0.00)

drought acclimated	LOWESS & PCA	all classifiers	**1.00 (0.00)**	1.00 (0.00)
drought acclimated	LOWESS (5 features)	lr/knn/qda/nb/ridge/et/svm/lda	**1.00 (0.00)**	1.00 (0.00)

The trained classifiers were ultimately validated on a subset of data that the model had never encountered before. This step provided an unbiased evaluation of the model's ability to perform on unseen data, helping to identify potential overfitting and offering a more accurate estimate of the model's real-world performance prior to deployment. The performance on the never seen data (Table 23) was comparable to the performance of the mean hold-out fold (CV-val) metrics thus model overfitting probability is minimal.

Table 23. AutoML hold-out subset performance across both drought levels and pre-processing techniques.

Dataset	**Pre-processing technique**	**Architecture**	**Accuracy**	**F1-Score**
drought onset	LOWESS	ada/nb/et/svm/dt/rf/gbc	**1.00**	1.00
drought onset	LOWESS & PCA	all classifiers except et/knn	**1.00**	1.00
drought onset	LOWESS (5 features)	ada/ridge/et/dt/rf/gbc	**1.00**	1.00
drought acclimated	LOWESS	lr/knn/ada/qda/ridge/et/lda/svm	**1.00**	1.00
drought acclimated	LOWESS & PCA	all classifiers	**1.00**	1.00
drought acclimated	LOWESS (5 features)	lr/knn/qda/nb/ridge/et/lda	**1.00**	1.00

Only the best performing classifiers are presented in the tables in terms of accuracy and F1-score with classifiers performing slightly worse being omitted.

Finally, the mixed dataset comprising of both drought onset and drought acclimated plants was used to evaluate the discrimination capabilities of the AutoML classifiers between the three classes (control - drought onset- drought acclimated). The performance achieved was slightly lower compared to the binary classification problems but still excellent achieving an F1-score and accuracy of 1.00 on the hold-out subset (Table 24).

Table 24. AutoML performance for the mixed dataset. In total three classes were used for classification. The standard deviation (SD) is provided within parentheses.

Dataset	**Pre-processing technique**	**Architecture**	**Accuracy**	**F1-Score**
mixed (CV-val)	LOWESS	lr/knn/qda/ svm/lda	**0.9679 (0.0393)**	0.9673 (0.0402)
mixed (CV-val)	LOWESS & PCA	lr/knn/qda/rf/gbc	**0.9679 (0.0393)**	0.966 (0.0402)
mixed (CV-val)	LOWESS (5 features)	rf/gbc	**0.9679 (0.0393)**	0.9673 (0.0402)
mixed (hold-out)	LOWESS	ridge/lda	**1.00**	1.00
mixed (hold-out)	LOWESS & PCA	et/lda/dt	**1.00**	1.00
mixed (hold-out)	LOWESS (5 features)	svm	0.9643	0.9647

3.2.2 Training using PLS1-DA

The second algorithm used was PLS-DA. Table 25 presents the accuracy, and F1-score achieved using PLS1-DA for both the CV-val and hold-out subsets. Cross validation (CV-val) achieved an accuracy and F1-score of 0.966 on the drought onset dataset, and 1.00 on the drought end dataset. Finally, for the mixed dataset, cross validation accuracy and F1-score were 0.922 and 0.934, respectively. The results on the hold-out subset were similar (slightly better), suggesting that the PLS1-DA model does not overfit the data.

Table 25. PLS1-DA performance for both acclimation levels and pre-processing techniques. The standard deviation is provided within parentheses.

Dataset	**Pre-processing technique**	**Architecture**	**Accuracy**	**F1-Score**
drought onset (CV-val)	LOWESS	PLS1-DA	0.966 (0.076)	0.966 (0.076)
drought onset (hold-out)			1.00	1.00
drought acclimated (CV-val)	LOWESS	PLS1-DA	1.00 (0.00)	1.00 (0.00)
drought acclimated (hold-out)			1.00	1.00
mixed (CV-val)	LOWESS	PLS1-DA	0.922 (0.053)	0.934 (0.037)
mixed (hold-out)			1.00	1.00

3.3 Evaluation of a hyperspectral image pipeline toward building a generalisation capable crop dry matter content prediction model

The third year further capabilities of the artificial intelligence regression models and spectral data were investigated. Namely their ability to generalize across heterogeneous data. As a result, a multi-crop and vegetable dry matter dataset was constructed. The dataset consisted of three different crops (apple, broccoli, and leek) with hyperspectral data collected using different hyperspectral cameras in the VIS-NIR range. The dataset consisted of 779 pairs of dry matter and hyperspectral measurements. Additionally, besides investigating the generalisation capabilities of the model, the effect of various preprocessing techniques as well as the effect of the dataset size were investigated.

In the upcoming tables, an extra preprocessing step is added as a column transitioning from a single step pipeline to the full six step pipeline: 1) Smoothing, 2) Scaling, 3) Recursive Feature Elimination, 4) Univariate Feature Selection, 5) Feature Extraction, 6) Polynomial Transformation. The first table (Table 26) consists of three (3) columns. The first column contains the first preprocessing step configuration, column two (2) and three (3) contain the algorithm performance for the pipeline evaluation. The last table (Table 31) consists of eight (8) columns, one column for each of the preprocessing steps, six in total) and two that contain the model performance. The following tables are sorted based on the ARD algorithm performance in descending order. The order of each table may be different, as the best performing combination and value of the preprocessing steps may differ.

Table 26 contains the performance of using all the available (wavelengths) features in order to predict dry matter content without applying any smoothing filter and with the application of one (window size 8 and 16). Various window sizes were evaluated, however, the no smoothing configuration reported higher performances for both the ARD and PLS algorithms compared to any smoothing. ARD regression achieved a better performance than PLS with RMSEP=0.0162. This served as the baseline for all following subsequent processing stages.

Table 26. RMSEP without and with smoothing. Dry Matter Min. Content Value = 0.0811; Max. Value = 0.2019

Smoothing Window	ARD	PLS
-	**0.0162**	**0.0163**
8	0.0165	0.0167

16	0.0168	0.0172

The use of scaling improved the performance of the ARD regression (Table 27). A minimum RMSEP=0.0151 was achieved with standardization, and an improvement of RMSEP=0.0153 was achieved when using normalization. PLS showcased the same pattern in the best performances: Standardization (0.0152), normalization (0.0153). Once again, smoothing caused a decrease in performance.

Table 27. RMSEP upon the addition of the scaling step. Dry Matter Content Min. Value = 0.0811; Max. Value = 0.2019

Smoothing Window	Scaling	ARD	PLS
-	Standardization	**0.0151**	**0.0152**
-	Normalization	0.0154	0.0153
8	Standardization	0.0155	0.0154
8	Normalization	0.0156	0.0157
16	Normalization	0.0161	0.0159
16	Standardization	0.0162	0.0161

The addition of the RFE preprocessing stage further improved performance (Table 28). ARD achieved RMSEP = 0.0147, while PLS RMSEP = 0.0149. For one more time, smoothing failed to improve performance. On the contrary, integrating feature scaling with RFE achieved better performance compared to tests conducted without any feature scaling.

Table 28. RMSEP upon the addition of the recursive feature elimination step,. Dry Matter Content Min. Value = 0.0811; Max. Value = 0.2019

Smoothing Window	Scaling	RFE	ARD	PLS
-	Standardization	Yes	**0.0147**	**0.0149**
-	Normalization	Yes	0.0150	0.0150
-	Standardization	No	0.0151	0.0152
-	-	Yes	0.0152	0.0150
-	Normalization	No	0.0154	0.0153
8	Normalization	Yes	0.0155	0.0151
8	Standardization	Yes	0.0157	0.0154
-	-	No	0.0161	0.0155
16	-	Yes	0.0162	0.0157

As shown in Table 29, including a feature selection step improved the overall performance and PLSR achieved RMSEP = 0.0137. The inclusion of a feature selection step also managed to improve the ARD regression performance (RMSEP = 0.0140) for the first time. The best feature selection algorithm was permutation which showed a better performance over its two competitors (Mutual Information and F-statistic). Pipeline performance increased only when using permutation as both the F-statistic and Mutual Information criteria led to a decrease in performance.

Table 29. Top-10 RMSEP upon the addition of the univariate feature selection step. Dry Matter Content Min. Value = 0.0811; Max. Value = 0.2019

Smoothing Window	Scaling	RFE	No. of features selected by the Feature selection	Feature selection criteria	ARD	PLS
-	Standardization	Yes	32	Permutation	**0.0140**	**0.0137**
-	Standardization	No	32	Permutation	0.0141	0.0139
-	Standardization	Yes	24	Permutation	0.0142	0.0143
-	Normalization	Yes	32	F-statistic	0.0143	0.0148
8	Standardization	No	32	Permutation	0.0144	0.0150
-	Normalization	Yes	32	Permutation	0.0145	0.0145
-	-	Yes	32	Permutation	0.0148	0.0147
-	-	Yes	24	permutation	0.0149	0.0143
-	Standardization	Yes	32	F-statistic	0.0150	0.0148
8	Standardization	Yes	32	Mutual Information	0.0152	0.0145

Including a feature extraction stage decreased performance pattern, in contrast to the previous pre-processing stages (Table 30). Additionally, both Autoencoder and the UMAP algorithms failed to achieve one of the best performances, making PCA the best performing feature extraction algorithm. The best performance when including feature extraction was achieved with PLS regression (RMSEP = 0.0148), however it fell back to the performance of only using the feature selection preprocessing stage.

Table 30. Top-10 RMSEP upon the addition of the feature extraction step. Dry Matter Content Min. Value = 0.0811; Max. Value = 0.2019.

Smoothing Window	Scaling	RFE	No. of features selected by the feature selection	Feature selection criteria	Feature Extraction Algorithm	ARD	PLS
8	-	Yes	32	Mutual Information	PCA	**0.0149**	**0.0148**
0	-	No	32	Permutation	PCA	0.0152	0.0152
0	-	Yes	32	Permutation	PCA	0.0153	0.0151
0	Normalization	Yes	24	Permutation	PCA	0.0155	0.0154
0	Normalization	Yes	24	Mutual Information	PCA	0.0156	0.0153
8	-	No	32	Permutation	PCA	0.0158	0.0151
8	-	Yes	32	Permutation	PCA	0.0160	0.0150
8	Normalization	Yes	24	Mutual Information	PCA	0.0161	0.0153
0	-	Yes	32	Mutual Information	PCA	0.0163	0.0155
0	-	Yes	24	Permutation	PCA	0.0164	0.0157

Table 31 summarises the ten (10) preprocessing pipelines that achieved the worst performances. They all showcased a noticeable pattern regarding the selection of the minimum number of features. Namely eight (8) features were selected by the F-statistic as the criteria. Finally, using PCA worsened the performance by reducing the number of selected features to four (4).

Table 31. The worst 10 RMSEP upon the addition of all the steps. Dry Matter Content Min. Value = 0.0811; Max. Value = 0.2019.

Smoothing Window	Scaling	RFE	No. of features selected by the feature selection	Feature selection criteria	Feature Extraction Algorithm	ARD	PLS
16	Standardization	Yes	8	F-statistic	PCA	0.0235	0.0222
16	Standardization	Yes	8	F-statistic	PCA	0.0217	0.0205

-	Standardization	Yes	8	F-statistic	PCA	0.0219	0.0204
8	Standardization	Yes	8	F-statistic	PCA	0.0219	0.0204
0	Standardization	Yes	8	F-statistic	PCA	0.0216	0.0201
8	Standardization	Yes	8	F-statistic	PCA	0.0215	0.02
-	-	No	8	F-statistic	PCA	0.0194	0.0194
16	-	No	8	F-statistic	PCA	0.0193	0.0193
0	Normalization	No	8	F-statistic	PCA	0.02	0.0193
16	Normalization	No	8	F-statistic	PCA	0.02	0.0193

In addition to examining the preprocessing steps that could result in the most effective pipeline for dry matter content prediction, analysing the wavelengths most commonly chosen by the top-performing models could help identify the regions of the electromagnetic spectrum that contain the most valuable information. This insight could be crucial for refining the prediction process and enhancing model accuracy. In Fig. 32 the most commonly used wavelengths which reported a RMSEP lower than 0.0140 can be seen. Notably, all 19 highest score wavelengths are located in the visible (VIS) region of the spectrum. This indicates that the VIS region contains the most informative features for the analysis. This pattern is rather stable with the exception of wavelength 538 nm, which was selected less compared to the 535nm wavelength which is larger. It is worth mentioning that wavelengths longer than 543nm were not selected frequently.

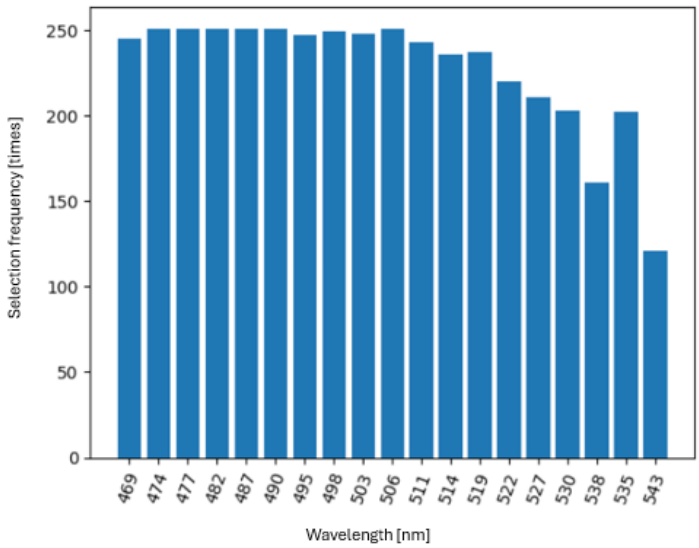

Figure 32. Most commonly selected wavelengths by the best-performing pipelines

3.3 Individual crop dry matter prediction

The same methodology used for the generalization experiments was also applied to single crops. In Table 32, the results for leek are shown. Permutation for the feature selection stage and not feature extraction was most important for achieving the lowest RMSEP. Again, neither Autoencoder nor UMAP was able to provide high performances as observed within the whole dataset. On the other hand, a key difference compared to using the entire dataset was that selecting 8 features, rather than 32, consistently resulted in a lower RMSEP, with the exception of when PCA was applied without RFE and scaling. This suggests that a more focused selection of features can improve model performance, though specific preprocessing techniques may influence the outcome. It is also worth noting that the best performance was obtained by PLS RMSEP = 0.0154.

Table 32. Top-10 RMSEP for leek upon the addition of the feature extraction step. Dry Matter Content Min. Value = 0.0811; Max. Value = 0.1910.

Smoothing Window	Scaling	RFE	No. of features selected by the univariate feature selection	Feature selection criteria	Feature Extraction Algorithm Output	ARD	PLS
-	Normalization	Yes	8	Permutation	-	**0.0162**	0.0163
-	-	No	24	Permutation	PCA	0.0163	0.0166
16	Normalization	No	8	Permutation	-	0.0164	0.0164

16	Standardization	Yes	8	Mutual Information	-	0.0165	**0.0154**
8	Normalization	No	8	Permutation	-	0.0166	0.0165
-	Standardization	No	8	Permutation	-	0.0168	0.0167
8	Standardization	Yes	8	Permutation	-	0.0170	0.0169
-	Normalization	Yes	8	Mutual Information	-	0.0171	0.0163
-	Standardization	Yes	8	Permutation	-	0.0173	0.0167
-	Normalization	Yes	8	Permutation	-	0.0174	0.0169

Broccoli showed a similar pattern to leek for permutation and low number selected bands eight (8) (Table 33). Use of large smoothing windows failed to achieve good performances. Finally, feature extraction algorithms failed to place themselves among the best performing pipelines. PLS performed the best with average RMSEP = 0.0103.

Table 33. Top-10 RMSEP for broccoli upon the addition of the feature extraction step. Dry Matter Content Min. Value = 0.1187; Max. Value = 0.2019

Smoothing Window	Scaling	RFE	No. of features selected by the univariate feature selection	Feature selection criteria	Feature Extraction Algorithm Output	ARD	PLS
8	Standardization	No	8	Permutation	-	**0.0104**	**0.0103**
0	Standardization	No	8	Permutation	-	0.0105	0.0105
0	Normalization	Yes	8	Permutation	-	0.0106	0.0107
0	-	Yes	8	Permutation	-	0.0107	0.0108
0	Standardization	Yes	8	F-statistic	-	0.0108	0.0109
0	Standardization	Yes	8	Permutation	-	0.0109	0.0107
0	Normalization	Yes	8	F-statistic	-	0.0111	0.0108
8	Standardization	Yes	8	Permutation	-	0.0113	0.0111
8	Normalization	Yes	8	Permutation	-	0.0114	0.0110
0	-	No	8	Permutation	-	0.0115	0.0109

Apple showcased the same patterns for feature selection. Low number of selected spectral bands and permutation as the selection criteria (Table 34). Moreover, using feature extraction did not improve the RMSEP obtained. As it was the case for broccoli, applying large window sizes for smoothing did not improve performance and using feature scaling and RFE did not show a clear pattern. The lower RMSEP values compared to the other crops and the whole dataset are caused by the different dry matter contents of each crop.

Table 34. Top-10 RMSEP for apple upon the addition of feature extraction step. Dry Matter Content Min. Value = 0.1349; Max. Value = 0.1743.

Smoothing Window	Scaling	RFE	No. of features selected by the univariate feature selection	Feature selection criteria	Feature Extraction Algorithm Output	ARD	PLS
0	Normalization	Yes	8	Permutation	-	**0.0075**	**0.0073**
0	-	No	8	Permutation	-	0.0077	0.0074
8	Standardization	Yes	8	Permutation	-	0.0078	0.0076
8	-	No	8	Permutation	-	0.0079	0.0075
8	Standardization	No	8	Permutation	-	0.0080	0.0077
0	Standardization	No	8	Permutation	-	0.0081	0.0076
0	Standardization	Yes	8	Permutation	-	0.0082	0.0080
0	-	Yes	8	Permutation	-	0.0084	0.0081
8	-	Yes	8	Permutation	-	0.0086	0.0079
8	-	Yes	8	Mutual Information	-	0.0089	0.0082

3.4 External validation

Real generalization (external validation) was evaluated by training the algorithms on pairs of crops and testing on the remaining crop. Table 35 depicts the performances achieved using all possible configurations. It is crucial to highlight that all crop pairs were evaluated as training data: apple-leek, apple-broccoli and broccoli-leek. However, only training on the pair of apples and broccoli and using the leek as the test set consistently reported performances with a positive R^2 (adjusted coefficient of determination). A performance of

RMSEP = 0.0224 was achieved using the F-statistic selection criteria and twenty-four (24) wavelengths as input variables.

Table 35. RMSEP on holdout dataset upon the addition of the feature extraction step.

Smoothing Window	Scaling	RFE	No. of features selected by the univariate feature selection	Feature selection criteria	Feature Extraction Algorithm Output	ARD	PLS
0	-	No	24	F-statistic	-	**0.0224**	**0.0226**
0	-	Yes	24	F-statistic	-	0.0226	0.0227
16	Normalization	No	24	F-statistic	PCA	0.0232	0.0234
16	Normalization	Yes	24	F-statistic	PCA	0.0235	0.0236

3.5 Statistical analysis results

For the fertilisation experiment agronomic data were also collected besides the spectral. This diverse collection of data allowed for more traditional statistical analysis. Pearson correlation was calculated for the CIELAB and Spectral Bands. This was conducted to explore the possibility of reducing spectral data complexity through replacing selected bands with CIELAB values. Possible correlations would mean that spectral data could be compressed and simplified while preserving the most relevant information to human vision, ultimately leading to faster analysis and reduced storage requirements.

No strong correlations were found. However, it is worth pointing out that for the wavelengths between 470-480nm the correlation for the a*, and b* value was the same but opposite, 0.2 and -0.2. At the same time there was a negative correlation of -0.72 among the a* and b* values.

Moreover, a detailed comparison for the broccoli weights among the different fertilisation treatments is provided in the table below.

Table 36. Broccoli weight statistical analysis

Treatment	Average Weight	Min weight	Max weight	Weight St.dev
Full Fertilisation	125.3	97.3	183.9	15.2
Half Fertilisation	95.1	69.6	134.2	22.6

Chapter 4 - Discussion and contributions

This research aimed to offer a meaningful addition to Precision Agriculture by investigating the capabilities of spectral imaging and Artificial Intelligence for model generalisation as well as fertilisation and irrigation optimization. Towards that end, this study compared novel user-friendly machine learning tools with traditional techniques, investigated whether spectral imaging and AI can identify different fertilisation and irrigation treatments and finally dived deeper into one of the most prominent problems spectral imaging faces: model generalisation.

4.1 Testing the Suitability of Automated Machine Learning, hyperspectral imaging and CIELAB colour space for proximal in situ fertilisation level classification

The study compared the performance of PyCaret, an open-source AutoML framework to PLS-DA algorithm, which is a norm for spectra classification, using fertiliser level classification as a potential use case. Both approaches performed well, with PyCaret showing a slight performance advantage over PLS-DA. This study provides evidence of the effectiveness and efficiency of modern ML architectures for classification tasks. Moreover, the potential of combining CIELAB colour space with hyperspectral data for fertiliser level classification was tested and compared to using only hyperspectral imaging. Both cases showed promising results but slightly lower when using the combined approach.

The CIELAB colour space achieved the best classification performance when used alone with PyCaret, with an accuracy of 0.72 using a fine-tuned MLP algorithm. When combined with the hyperspectral data, the accuracy improved to 0.94 using the LDA algorithm. These results are consistent with previous studies that have found a correlation between the CIELAB colour space and nitrogen status in barley (Christ et al., 2021), wheat (Yakushev and Kanash, 2016), and broccoli (Graeff et al., 2008). None of these studies reports classification metrics, making direct comparisons impossible. Instead, they focus on establishing a correlation between the two variables. However, colour and most precisely HSV values have been used to determine fertilisation levels using rice leaves (Sari and Alkaff, 2020). This particular study achieved an accuracy of 0.825. It is worth noting that in this occasion imaging took place under controlled laboratory conditions. From the above mentioned discussion it could be argued that colour information could yield satisfactory results when classifying different fertilisation dosages. After all the visual identification of either nutrient deficiencies or phytopathological cases is the current tool of an agronomist.

Nonetheless, the classification results obtained from the CIELAB dataset are inferior to those derived from the hyperspectral imaging dataset. The PLS-DA algorithm applied to the hyperspectral data produced excellent results with an accuracy of 0.91. Combining PLS-DA with hyperspectral data is known to provide excellent performance(Lee et al., 2024)(Tunny et al., 2023). The technology has been utilized in agriculture for various tasks. For instance, it has been used to detect decay lesions in citrus fruit with a classification rate of 0.91 (Folch-Fortuny et al., 2016), identify rice seed cultivars with classification rates over 0.8 (Kong et al., 2013), and predict viability and vigour in muskmelon seeds with a classification accuracy of 0.95 (Kandpal et al., 2016).

However, the most accurate results were achieved by combining the hyperspectral dataset and PyCaret analysis tool, resulting in an accuracy of 1.00, with minimal user intervention required, limited to data preprocessing. It is important to note that due to the relatively small sample size, the models may be overfitting the data(Hawkins, 2004), despite the measures taken to prevent this from affecting the performance metrics such as using a Stratified KFold cross validation to provide a more representative and unbiased evaluation of the model's performance across different folds. However, it is not surprising that perfect classification rates are achieved when combining machine learning and hyperspectral imaging for fertilisation estimation tasks, as other studies have reported similar results. For instance, in tea plants fertilisation experiments, perfect classification rates (100% accuracy) have been achieved (Wang et al., 2018), while accuracies of 75% and 80% have been reported for bok choy and spinach fertilisation studies respectively (Nguyen et al., 2020). Moreover, experiments on fertilisation in corn and cucumbers have reported accuracies of 99.46% (Goel et al., 2003) and 96.14%(Sabzi et al., 2021) respectively. Although potentially more powerful machine learning approaches, such as CNNs, were not investigated due to an insufficiently large dataset, evidence suggests that they could outperform the PLS(Mishra and Passos, 2022).

It is important to note that combining hyperspectral data with CIELAB data had an adverse effect on the classification performance in this particular use case, despite the contrary being reported for grape samples(Rodríguez-Pulido et al., 2021). The lack of a strong enough synergistic effect may explain why the potential noise introduced by the CIELAB measurements could not be overcome.

Moreover, the dataset using a single wavelength (874 nm) matched the performance of using the whole hyperspectral dataset. This signifies that the AutoML system is strong enough to classify samples even when fewer wavelengths are available. Additionally, the NIR region is the most promising in detecting fertilisation deficiencies. This could be attributed to the fact that water is a strong absorber of infrared (IR) energy (Tsenkova, 2010). When a plant is under stress such as nutrient deficiency, its water content can change, leading to alterations in its NIR reflectance spectrum. This, also, explains the poor performance of the CIELAB dataset, which captures information in the

visible region. The importance of the NIR region is also supported by additional research that focused on fertilisation (Gómez-Casero et al., 2007)

Finally, the weaker performance of the CIELAB dataset compared to the hyperspectral one could be attributed to the fact that hyperspectral imaging focuses on chlorophyll absorption, which refers to the process by which chlorophyll molecules absorb light energy during photosynthesis and it involves the absorption of specific wavelengths of light by chlorophyll pigments, which allows them to convert light energy into chemical energy. More specifically chlorophyll content is calculated using wavelengths 663 and 645 according to the Lambert-Beer law (Liu et al., n.d.). On the other hand, CIELAB focuses on the phenomenological background of chlorophyll, which refers to the overall influence of various factors, such as the surrounding environment and the presence of other pigments. In more detail, the increase in a^* has been linked exclusively to chlorophyll in the absence of anthocyanin pigments (Ferrer et al., 2005). Ultimately, these factors can affect the overall absorption spectrum of chlorophyll and contribute to the background noise or interference in the measurement of chlorophyll absorption. Additionally, CIELAB and the phenomenological background of chlorophyll (green colour) exhibit a hysteresis effect (Peng et al., 2017)which can manifest itself in various ways, such as non-linearities in colour transitions or differences in perceived colour changes at different points in the colour space. This effect is important and can add complexity to image processing applications, thus lowering the model performance.

All of the aforementioned studies employ complex ML and DL techniques that necessitate a thorough comprehension of ML concepts to create and refine algorithms. However, this study has achieved comparable outcomes using an AutoML framework and PyCaret analysis tool. AutoML systems, have demonstrated their capabilities in agriculture tasks, as shown in studies using AutoML for weed identification (Espejo-Garcia et al., 2021)(Jiang et al., 2020), pest identification (Hayashi et al., 2019), stress detection (Karthickmanoj et al., 2021), and yield prediction (Duan et al., 2022). Therefore, AutoML holds the potential to replace labour-intensive manual tasks.

4.2 Early detection of broccoli drought acclimation/stress in agricultural environments utilising proximal hyperspectral imaging and AutoML

In the context of short-term drought effect classification, this study explored the efficacy of the open-source AutoML framework PyCaret, along with the Partial Least Squares Discriminant Analysis (PLS1-DA) algorithm, a typical method for spectra classification. Both PyCaret and PLS1-DA exhibited commendable outcomes, demonstrating that the classification of drought acclimated broccoli is feasible even at the beginning of drought. PyCaret classifiers showcased very similar results with PLS1-DA, achieving accuracy and F1-score of 1.00 for every evaluated dataset. It is worth pointing out that longer drought periods (further developed water stress/acclimation) lead to better classification performance. This could be explained by the higher SD of the drought onset dataset. All tested methods performed excellent in the binary classifications (drought onset - control, drought acclimated - control), while for the mixed dataset containing drought onset, drought acclimated and control samples, results were slightly worse. These findings are further supported by additional research that has shown that is possible to determine water stress using spectral reflectance on sweet corn (Genc et al., 2013). While (W. Zhang et al., 2021) reported accuracy of around 0.9 for classifying greenhouse tomato plants under water stress using visible near-infrared and (Nampally et al., 2023) accuracy of 0.91 for water stress classification in maize once again looking at visible near-infrared (VIS-NIR) region.

These results encourage research for identifying water stress using spectral data. However, they also showcase the complexity of the interpretation of water dynamics on plant material and the potential of hyperspectral imaging. As in the current work, non-stressed short-term acclimated broccoli plants showcased outstanding classification results, emphasizing the high sensitivity of this method. Moreover, further research is required on how the relative or gravimetric leaf water content and/or the dynamic cuticle of broccoli plants influence the reflectance.

It is worth pointing out that similar results have also been achieved using colour imaging, with accuracies higher than 0.95 being reported for maize (An et al., 2019) and sunagoke moss (Ondimu and Murase, 2008). These results further support the use of the VIS-NIR region for water stress identification, thus enabling possible future low-cost solutions that do not rely upon expensive high sensitivity cameras.

Finally, the excellent results achieved using the AutoML framework, which requires minimal user intervention and not prior knowledge regarding ML algorithm come as no surprise as they align with prior studies leveraging AutoML and hyperspectral imaging which achieved excellent results (accuracy >90%) for plant phenotyping(Koh et al., 2021) and crop yield and mass estimation (Ondimu and Murase, 2008). Additionally, the

successful application of AutoML solutions in achieving comparable outcomes to widely used algorithms such as PLS-DA and complex machine learning methods, underscores the versatility and potential of AutoML systems in streamlining ML algorithm development for agricultural tasks. The validity of the results and the possible problem of overfitting were set to the test using a holdout dataset, in which the proposed model performed equally well. Consequently, AutoML framework emerges as a potential candidate for replacing or become valuable aid to labour-intensive manual plant monitoring in agriculture.

However, it should be noted that due to the nature of spectral imaging, solutions that have been developed for a specific crop-problem pair cannot be expanded and generalized to other crops. Moreover, to the best of our knowledge, no generalisation-capable hyperspectral models have been developed yet due to various reasons such as limited data availability. Despite the developed solution's limitations, its value to the primary production industry is substantial as it provides a proof of concept for developing water stress detection software, thus facilitating irrigation optimization while protecting crops from yield and quality losses related to water stress.

4.3 Evaluation of a hyperspectral image pipeline toward building a generalisation capable crop dry matter content prediction model

The performance of using an incremental pipeline towards establishing a baseline methodology for developing a global spectral model for various crops was presented in this research. The study focuses on predicting dry matter content in various crops and serves as a baseline for generalization-capable regression models. PLSR reported the best performance of RMSEP = 0.0137 against the RMSEP = 0.0140 of ARD. However, the external validation was worse, RMSEP = 0.024, as the model tries to model unseen data. It is worth pointing out that the dataset containing all three (3) crops achieved better performance compared to the leek model, RMSEP = 0.0154.

Another valuable remark is that RMSEP is highly influenced by the dry matter content range of each crop, making a fair comparison more challenging. For example, it is not easy to say if the larger size of the combined dataset allows better performances since the dry matter content range is also larger and more challenging. However, both the leek subset and the combined dataset exhibited similar ranges, though the combined dataset had a much lower RMSEP. This reduction in RMSEP could be attributed to two factors: the increased diversity of information in the combined dataset, and the complementary nature of the different data sources, which likely enhanced the model's ability to capture relevant patterns more effectively: Namely: the lower values of the dry matter content in apple and broccoli or the emergence of better statistical properties in specific wavelengths allowed the creation of better regressor at the cost of using more features.

The caveat with agricultural produce, which was also validated in this study, is that each crop has its own particularities in colour, size and shape making it hard for AI solutions to identify patterns. Moreover, when diving deeper, using the higher spectral resolution hyperspectral imaging offers can make matters even more complicated. Despite the three (3) main pigment groups: i) carotenoids, ii) chlorophylls, and iii) flavonoids, scientists calculate that there could be up to 4,000 different phytonutrients. The model developed in this study had to cope with various pigments. In apples, the main flavonoids present are carotenoids and anthocyanins; in leek, chlorophylls and flavonoids and in broccoli, carotenoids and chlorophylls.

One of the goals of this paper was to understand how different preprocessing stages influence the regression problem. It could be discussed that the use of feature extraction with this dataset caused a performance drop. This behaviour could be

attributed to various reasons such as (i) nonlinear relationships, (ii) removal of variables that are not highly correlated but show a synergistic effect with other variables, (iii) loss of valuable information, and (iv) sampling variability. However, feature extraction could be necessary in other situations.

Another important insight was that a stage (e.g., using RFE) reporting the best performance under one specific scenario/setting could be beaten by the opposite approach upon including another post-processing step (e.g., feature selection). This can be observed after the integration of the feature extraction stage, when the Mutual Information criteria for feature selection reported the best performance. This was the opposite behaviour compared to previous results, where mutual information never was among the best pipelines. In other words, discarding non-working preprocessing stages can damage the final performance if new processing stages are added afterwards. Similar behaviour can arise when using specific preprocessing stages that reduce the relevance of previous stages.

Contrary to (Medic, 2023) the results of this study showed that the Savitzky-Golay filter was not a good match for this problem. Although using the smallest window size eight (8) reported the highest performance, overall, avoiding the use of smoothing was a better solution. Furthermore, this pattern was made more apparent because of the low performance achieved by large window sizes. There were two (2) reasons for using this filter: (i) to smooth too steep changes (wavelengths between 678 nm to 741 nm), possible meaningless peaks (seen in the initial spectrum of broccoli), or random noise in the input, and (ii) to reveal potential unexpected performances upon integration of this pre-processing stage with next processing stages. However, the spectral data used in this study are relatively smooth in their raw nature, and therefore, using an extra filter could delete critical information towards building the dry matter content regression. Regarding the appearance of unexpected behaviors, while integrating new processing stages to the image smoothing, it could be argued that its performance for single crops is more variable with a tendency to be higher, but this could be a consequence of the inability to find strong statistical patterns by the machine learning pipeline and of the smaller dataset size.

One of the main problems of using the proposed methodology is that the number of combinations while integrating the different processing components, their hyper-parameters, and their specific order may give a combinatorial explosion. Therefore, some a priori decisions were made to accomplish realistic research that could provide robust insights. Some of these decisions were based on early exploration experiments. Despite that, the complete evaluation of all components should be conducted since their performances, although initially less promising, could converge into the best or most reliable performance. For example, many regressors were not covered (e.g., Support Vector Regressor, tree-based regressors). Additionally, more powerful machine learning

approaches, such as Convolutional Neural Networks (CNNs), were not investigated despite evidence that they could perform better than PLS (Mishra and Passos, 2022). The insufficiently large dataset required for training complex regression models was the driving force of the decision. Early experiments utilising polynomial transformation were used to explore its potential improvement in performance. However, although the RMSEP slightly decreased, the large number of created features compared to the dataset size deserved a deeper investigation, which was out of the scope of this study. It is essential to highlight that polynomial features include the higher-degree terms of individual features and the interaction terms between different features. On the other hand, the goal of this research was not to achieve state-of-the-art performance through exhaustive grid search optimization of components and hyperparameters. Instead, the focus was on gaining an empirical understanding of how each component in the pipeline influences the others, either positively or negatively, and to establish a baseline for future studies.

Finally, it is worth noting that the most significant wavelengths were in the VIS region, which appears counterintuitive considering that the presence of water, as absorption bands of the O-H group, is observed at the NIR region, namely 740, 840, 960, and 1,440 nm(Sun et al., 2020). This behaviour could be attributed to the heterogeneity of pigments among the different crops, thus enabling the model to focus on a spectral region where data shows a higher variability. Furthermore, on some occasions, changes in dry matter content might be accompanied by changes in pigment concentration. This procedure takes place in olives, where during ripening dry matter increases and colour changes (Conde et al., 2008)(Conde et al., 2008).

4.5 Statistical analysis agronomical insights

As mentioned previously the main difference between the two fertilisation levels was the size of the broccoli head when measurements were conducted. From the data collected it can be concluded that optimal fertilisation does not only provide larger broccoli heads but also more homogeneous plants. Despite the above, weight alone cannot be used for determining fertilisation levels as broccoli size is plant density and cultivar besides fertilization dependent (Schellenberg et al., 2009). During the irrigation experiment, no agronomical data were recorded.

Chapter 5 - Conclusions

Two of the main objectives of this study were to **Develop Artificial Intelligence models utilizing spectral data capable of identifying different fertilisation levels** and to **Compare the performance of traditional machine learning algorithms with novel user friendly AutoML** techniques. Both were achieved by *Testing the Suitability of Automated Machine Learning, hyperspectral imaging and CIELAB color space for proximal in situ fertilisation level classification.* The research findings demonstrate that AutoML outperforms PLS-DA, a traditional machine learning approach. Additionally, the results indicate that hyperspectral data are capable of identifying different fertilisation levels. Diving deeper, the use of hyperspectral data outperformed both the use of CIELAB data and the combination of them. Hyperspectral imaging datasets analysed with the Partial Least Squares Discriminant Analysis (PLS-DA) are often used as a robust starting point for computer vision tasks, yielding promising results. However, effective implementation of PLS-DA requires machine learning and statistics proficiency. AutoML, an upcoming field in machine learning, offers a promising alternative, automating many of the intricate processes involved and thus making it accessible to both experienced practitioners and those new to the domain. The proliferation of portable multi- and hyper-spectral sensors across ground-based and aerial platforms is driving a surge in the agricultural application of spectral imaging. This promising field presents an opportunity for the deployment of AutoML. Moreover, as AutoML solutions continue to evolve, incorporating increasingly sophisticated machine learning and deep learning algorithms, their potential applications within agriculture and beyond are poised to expand significantly." Additionally, the simplicity of training these models provides an unprecedented opportunity to create bespoke machine learning models that are tailored to specific sites and problems, overcoming one of the most persistent challenges in machine learning and hyperspectral imaging: model generalisation. The study’s results provide a foundation for using hyperspectral imaging and AutoML in precision agriculture tasks related to fertilisation.

Another objective was to **develop Artificial Intelligence models utilizing spectral data that are capable of identifying plant water deficit**. This objective was achieved by investigating the early detection of broccoli drought acclimation/stress in agricultural environments utilising proximal hyperspectral imaging and AutoML. Moreover, this publication supports the previous claims that novel user-friendly AutoML techniques can match the performance of traditional machine learning algorithms. To further elaborate, the study's findings demonstrate that the AutoML framework matched the performance of the PLS1-DA and that hyperspectral data can facilitate drought stress/acclimation identification in broccoli plants. Within the domain of computer vision, hyperspectral imaging datasets coupled with Partial Least Squares Discriminant Analysis

(PLS1-DA) often serve as a reliable starting point, yielding excellent results. However, effective implementation of PLS1-DA requires a degree of technical proficiency. Additionally, AutoML demonstrated exceptional performance in discriminating between control and acclimated plants. Therefore, AutoML frameworks offer a promising alternative to traditional manual machine learning approaches, catering to users of varying skill levels. As the concerns of water scarcity are continuously increasing together with the imperative for efficient irrigation management, the integration of hyperspectral imaging with machine learning and deep learning algorithms presents a promising solution. The widespread adoption of spectral sensors in proximal and remote agricultural applications further underscores the potential of spectral data. As mentioned earlier, AutoML, by automating many of the intricate processes involved in model development, offers a valuable tool for both experienced practitioners and those new to the field. By doing so it enables the rapid experimentation with diverse architectures and hyperparameters. To conclude, the findings of this study establish a foundation for advocating and implementing AutoML and hyperspectral imaging in precision agriculture and irrigation management. Additionally, it underscores the importance of caution when interpreting plant material water dynamics through spectral data.

Finally, the last objective of this study was to **evaluate the feasibility of developing a generalisation capable AI model utilizing spectral data**. This was achieved by *evaluating of a hyperspectral image pipeline toward building a generalization capable crop dry matter content prediction model*. Pre-processing and modelling components were incrementally evaluated, resulting in a performance of RMSEP=0.0140 using ARD and RMSEP=0.0137 using PLSR.

Testing the proposed pipelines and algorithms on an open-access dataset highlighted the limitations of machine learning pipelines in effectively generalizing complex, non-uniform data. This revealed challenges in adapting models to diverse data patterns, underscoring the need for further refinement in pipeline design for better generalization. On the other hand, the results on a dataset with multiple crops showed a lower RMSEP than the single crop leek subset, where the dry matter range was similar. This could be attributed to the emergence of statistical properties only available for multiple crops. Testing with additional datasets should be conducted to validate further the proposed pipeline's ability to generalise. Moreover, there is more space for improvement since other algorithms like Support Vector Regressor, tree-based methods, or deep learning techniques could complement each proposed component and regressor.

However, as the complexity of the algorithms increases, so do the time and computational resources needed to determine the optimal architecture for a given problem. More sophisticated models demand greater processing power and longer training times, significantly raising the cost of model development and optimization. Additionally, the complexity added by those algorithms is expected to further hinder the identification of the effects of each pipeline component on the model performance and

the model explainability. However, training and implementation of both selected algorithms (ARD and PLSR) are straightforward once the pipeline and optimal architecture are found.

Chapter 6 - Future work

Precision agriculture is an evolving field that employs technology and data-driven methodologies to optimize crop production and enhance quality. The integration of various technologies, including geospatial technologies, Internet of Things (IoT), Big Data analysis, and Artificial Intelligence (AI), offers opportunities for informed decision-making to enhance crop production. Precision agriculture encompasses the utilization of these technologies to optimize agricultural inputs, thereby increasing production and minimizing losses. In recent decades, Remote sensing technologies have seen significant growth in precision agriculture, driven by the widespread availability of high-resolution satellite images. These images have facilitated numerous precision agriculture applications, including crop monitoring, irrigation management, nutrient application, disease and pest management, and yield prediction. Commercial agriculture has already integrated remote-sensing-based PA technologies like variable fertiliser rate application systems such as Green Seeker and Crop Circle. The use of unmanned aerial vehicles (UAVs) has surged due to their cost-effectiveness and ability to capture high-resolution images, which are essential for precision agriculture. Additionally, the abundance of satellite and spectral data has spurred researchers to explore advanced data storage and processing methods such as cloud computing and machine learning.

Precision agriculture, spectral imaging and AI have the potential to enhance the efficiency, sustainability, and profitability of small and medium-sized farms by enabling informed decision-making regarding irrigation, fertilisation, and other management practices. This can result in cost savings and increased yields. Moreover, it can replace manual inspection of crops with automated solutions, thus relieving workers from a tedious and strenuous activity while at the same time making enabling the monitoring and investigation of large sample sizes and even whole batch instead of the random sampling currently performed to assess the quality of fruit and vegetable.

Despite the extensive research on spectral imaging applications in precision agriculture, there is a notable absence of established techniques or frameworks that are both accurate and reproducible across various climatic, soil, crop, and management conditions. The accuracy of those methods depends on several factors, including image resolution (spatial, spectral, and temporal), atmospheric conditions, weather patterns, crop growth stages, land cover, and the analysis technique employed (e.g., regression-based, machine learning, physically based modelling). Further research is required to comprehend the spatio-temporal patterns of uncertainty in estimating biotic and abiotic stress and other crop parameters. Further elaborating on that, an irrigation or fertilisation deficiency detection method may perform well under controlled experimental conditions but may not exhibit similar performance in real-world scenarios where various stressors influence crop response.

Moreover, AI solutions need to cope with several challenges to facilitate their adoption and implementation. Firstly, as mentioned earlier, the diversity and variability of agricultural environments, including different crops, soil types, weather patterns, and management practices, pose significant hurdles to developing AI models that are universally applicable and capable of generalisation. Additionally, the interpretability and trustworthiness of AI models in making critical decisions regarding crop management and resource allocation remain essential concerns for farmers and stakeholders. Moreover, the need for continuous model adaptation and validation to accommodate evolving agricultural conditions further complicates AI adoption in precision agriculture.

Therefore, future research in spectral imaging and AI for precision agriculture should focus on developing robust, generalisation capable and interpretable models that can effectively handle the complexity and variability of agricultural systems. As well as developing solutions that are easy to use by non-experts, while trying to maintain acquisition costs of those future solutions low enough to allow adoption by middle sized farmers. Ultimately, increasing their impact and transforming modern primary production systems.

References

Abdulridha, J., Ampatzidis, Y., Roberts, P., Kakarla, S.C., 2020. Detecting powdery mildew disease in squash at different stages using UAV-based hyperspectral imaging and artificial intelligence. Biosyst. Eng. 197, 135–148.

Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., Sousa, J.J., 2017. Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry. Remote Sens. 9, 1110.

Aimin, H., 2010. Uncertainty, risk aversion and risk management in agriculture. Agric. Agric. Sci. procedia 1, 152–156.

Ali, M., 2020. PyCaret: An open source, low-code machine learning library in Python.

An, J., Li, W., Li, M., Cui, S., Yue, H., 2019. Identification and classification of maize drought stress using deep convolutional neural network. Symmetry (Basel). 11, 256.

Andrade, M.A., Evett, S.R., O'Shaughnessy, S.A., 2018. Machine learning algorithms applied to the forecasting of crop water stress indicators. Proc. Tech. Irrig. Show.

Argento, F., Anken, T., Abt, F., Vogelsanger, E., Walter, A., Liebisch, F., 2021. Site-specific nitrogen management in winter wheat supported by low-altitude remote sensing and soil data. Precis. Agric. 22, 364–386.

Aversano, L., Bernardi, M.L., Cimitile, M., 2022. Water stress classification using Convolutional Deep Neural Networks. JUCS J. Univers. Comput. Sci. 28.

Baek, S.-H., Park, K.-H., Jeon, J.-S., Kwak, T.-Y., 2022. Using the CIELAB Color System for Soil Color Identification Based on Digital Image Processing. J. Korean Geotech. Soc. 38, 61–71.

Bagheri, N., Ahmadi, H., Alavipanah, S.K., Omid, M., 2013. Multispectral remote sensing for site-specific nitrogen fertilizer management. Pesqui. Agropecuária Bras. 48, 1394–1401.

Banerjee, B.P., Joshi, S., Thoday-Kennedy, E., Pasam, R.K., Tibbits, J., Hayden, M., Spangenberg, G., Kant, S., 2020. High-throughput phenotyping using digital and hyperspectral imaging-derived biomarkers for genotypic nitrogen response. J. Exp. Bot. 71, 4604–4615.

Bannerjee, G., Sarkar, U., Das, S., Ghosh, I., 2018. Artificial intelligence in agriculture: A literature survey. Int. J. Sci. Res. Comput. Sci. Appl. Manag. Stud. 7, 1–6.

Barnes, J., 2015. Azure machine learning. Microsoft Azur. Essentials. 1st ed, Microsoft.

Bartz, J.A., Brecht, J.K., 2002. Postharvest physiology and pathology of vegetables. Crc Press.

Basso, B., Fiorentino, C., Cammarano, D., Schulthess, U., 2016. Variable rate nitrogen fertilizer response in wheat using remote sensing. Precis. Agric. 17, 168–182.

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31.

Benton, T.G., Vickery, J.A., Wilson, J.D., 2003. Farmland biodiversity: is habitat heterogeneity the key? Trends Ecol. Evol. 18, 182–188.

Bhat, S.A., Huang, N.-F., 2021. Big data and ai revolution in precision agriculture: Survey and challenges. IEEE Access 9, 110209–110222.

Bisong, E., Bisong, E., 2019. Google automl: cloud vision. Build. Mach. Learn. Deep Learn. Model. Google Cloud Platf. A Compr. Guid. Beginners 581–598.

Blok, P.M., van Henten, E.J., van Evert, F.K., Kootstra, G., 2021. Image-based size estimation of broccoli heads under varying degrees of occlusion. Biosyst. Eng. 208, 213–233.

Boden, M., 1980. Artificial intelligence and natural man. Synthese 43.

Bodner, G., Nakhforoosh, A., Arnold, T., Leitner, D., 2018. Hyperspectral imaging: a novel approach for plant root phenotyping. Plant Methods 14, 1–17.

Bronzi, B., Brilli, C., Beone, G.M., Fontanella, M.C., Ballabio, D., Todeschini, R., Consonni, V., Grisoni, F., Parri, F., Buscema, M., 2020. Geographical identification of Chianti red wine based on ICP-MS element composition. Food Chem. 315, 126248.

Carranza, C., Lanchero, O., Miranda, D., Salazar, M.R., Chaves, B., 2008. Modelo simple de simulación de distribución de masa seca en brócoli (Brassica sp.) variedad Coronado y repollo (Brassica oleracea) híbrido Delus cultivados en la Sabana de Bogotá. Agron. Colomb. 26, 23–31.

Cecílio Filho, A.B., Carmona, V.M.V., Schiavon Junior, A.A., 2017. Broccoli growth and nutrient accumulation. Científica (Jaboticabal) 45, 95–104.

Chandel, N.S., Rajwade, Y.A., Dubey, K., Chandel, A.K., Subeesh, A., Tiwari, M.K., 2022. Water stress identification of winter wheat crop with state-of-the-art AI techniques and high-resolution thermal-RGB imagery. Plants 11, 3344.

Chengquan, Z., Hongbao, Y., Guohong, Y., Jun, H., Zhifu, X., 2020. A fast extraction method of broccoli phenotype based on machine vision and deep learning. Smart Agric. 2, 121.

Cheshkova, A.F., 2022. A review of hyperspectral image analysis techniques for plant disease detection and identif ication. Vavilov J. Genet. Breed. 26, 202.

Christ, A., Schmittmann, O., Lammers, P.S., 2021. Study on Deployment of a TrueColor Sensor Array for Dual Use-Weed Detection and N-Fertilizer Application. Agric. Eng. AgEng2021 484.

Cilia, C., Panigada, C., Rossini, M., Meroni, M., Busetto, L., Amaducci, S., Boschetti, M., Picchi, V., Colombo, R., 2014. Nitrogen status assessment for variable rate fertilization in maize through hyperspectral imagery. Remote Sens. 6, 6549–6565.

Commission, E., Environment, D.-G. for, 2017. Agri-environment schemes – Impacts on the agricultural environment. Publications Office. https://doi.org/doi/10.2779/633983

Concepcion II, R.S., Lauguico, S.C., Alejandrino, J.D., Bandala, A.A., Sybingco, E., Vicerra, R.R.P., Dadios, E.P., Cuello, J.L., 2021. Adaptive fertigation system using hybrid vision-based lettuce phenotyping and fuzzy logic valve controller towards sustainable aquaponics. J. Adv. Comput. Intell. Intell. Informatics 25, 610–617.

Conde, C., Delrot, S., Gerós, H., 2008. Physiological, biochemical and molecular changes occurring during olive development and ripening. J. Plant Physiol. 165, 1545–1562.

Conesa Martinez, A., Manera Bassa, F.J., Brotons Martinez, J.M., Fernández Zapata, J.C., Simón, I., Simon Grao, S., Alfonsea Simón, M., Martínez Nicolas, J.J., Valverde, J.M., García-Sánchez, F., 2019. Changes in the content of chlorophylls and carotenoids in the rind of Fino 49 lemons during maturation and their relationship with parameters from the CIELAB color space.

Council of Europe, 2024. History of Artificial Intelligence [WWW Document]. URL https://www.coe.int/en/web/artificial-intelligence/history-of-ai

Cozzolino, D., Restaino, E., Fassio, A., 2010. Discrimination of yerba mate (Ilex paraguayensis

St. Hil.) samples according to their geographical origin by means of near infrared spectroscopy and multivariate analysis. Sens. Instrum. Food Qual. Saf. 4, 67–72.

Cunniff, P., Washington, D., 1997. Official methods of analysis of AOAC International. J. AOAC Int 80, 127A.

Doerge, T.A., 1991. Nitrogen fertilizer management in Arizona.

Dou, X., Yang, Y., 2018. Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems. Comput. Electron. Agric. 148, 95–106.

Duan, S., Wu, S., Monier, E., Ullrich, P., 2022. AutoML-based Almond Yield Prediction and Projection in California. arXiv Prepr. arXiv2211.03925.

Dyrmann, M., Jørgensen, R.N., 2015. RoboWeedSupport: weed recognition for reduction of herbicide consumption, in: Precision Agriculture'15. Wageningen Academic Publishers, pp. 259–269.

El-Shikha, D.M., Waller, P., Hunsaker, D., Clarke, T., Barnes, E., 2007. Ground-based remote sensing for assessing water and nitrogen status of broccoli. Agric. water Manag. 92, 183–193.

El Sghair, M., Jovanovic, R., Tuba, M., 2017. An algorithm for plant diseases detection based on color features. Int J Agric Sci 2, 1–6.

Eli-Chukwu, N.C., 2019. Applications of artificial intelligence in agriculture: A review. Eng. Technol. Appl. Sci. Res. 9.

Erdem, T., Arın, L., Erdem, Y., Polat, S., Deveci, M., Okursoy, H., Gültaş, H.T., 2010. Yield and quality response of drip irrigated broccoli (Brassica oleracea L. var. italica) under different irrigation regimes, nitrogen applications and cultivation periods. Agric. Water Manag. 97, 681–688.

Espejo-Garcia, B., Malounas, I., Vali, E., Fountas, S., 2021. Testing the Suitability of Automated Machine Learning for Weeds Identification. Ai 2, 34–47.

European Environment Agency, 2020. Is Europe living within the limits of our planet? An assessment of Europe's environmental footprints in relation to planetary boundaries, EEA Report.

European Environment Agency, European Commission, D.C.A., 2022. Annual European Union greenhouse gas inventory 1990–2020 and inventory report 2022, National Inventory Reports.

Fayyaz, A.M., Al-Dhlan, K.A., Rehman, S.U., Raza, M., Mehmood, W., Shafiq, M., Choi, J.-G., 2022. Leaf Blights Detection and Classification in Large Scale Applications. Intell. Autom. Soft Comput. 31.

Ferdinand, Y., Al Maki, W.F., 2022. Broccoli leaf diseases classification using support vector machine with particle swarm optimization based on feature selection. Int. J. Adv. Intell. Informatics 8, 337–348.

Ferrer, A., Remón, S., Negueruela, A.I., Oria, R., 2005. Changes during the ripening of the very late season Spanish peach cultivar Calanda: Feasibility of using CIELAB coordinates as maturity indices. Sci. Hortic. (Amsterdam). 105, 435–446.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F., 2015. Efficient and robust automated machine learning. Adv. Neural Inf. Process. Syst. 28.

Folch-Fortuny, A., Prats-Montalbán, J.M., Cubero, S., Blasco, J., Ferrer, A., 2016. VIS/NIR

hyperspectral imaging and N-way PLS-DA models for detection of decay lesions in citrus fruits. Chemom. Intell. Lab. Syst. 156, 241–248.

Fraiwan, M., Faouri, E., Khasawneh, N., 2022. Multiclass classification of grape diseases using deep artificial intelligence. Agriculture 12, 1542.

Franzen, D., Mulla, D., 2015. A history of precision agriculture. Precis. Agric. Technol. Crop farming 1–20.

Gao, P., Zhang, Y., Zhang, L., Noguchi, R., Ahamed, T., 2019. Development of a recognition system for spraying areas from unmanned aerial vehicles using a machine learning approach. Sensors 19, 313.

Garcia-Manso, A., Gallardo-Caballero, R., Garcia-Orellana, C.J., Gonzalez-Velasco, H.M., Macias-Macias, M., 2021. Towards selective and automatic harvesting of broccoli for agri-food industry. Comput. Electron. Agric. 188, 106263.

Gebbers, R., Adamchuk, V.I., 2010. Precision agriculture and food security. Science (80-.). 327, 828–831.

Genc, L., Inalpulat, M., Kızıl, U., Mirik, M., Smith, S.E., Mendes, M., 2013. Determination of water stress with spectral reflectance on sweet corn (Zea mays L.) using classification tree (CT) analysis.

Glass, C.R., Gonzalez, F.J.E., 2022. Developing of New Technologies Driving Advances in Precision Agriculture to optimise inputs and reduce environmental footprint. C3-BIOECONOMY Circ. Sustain. Bioeconomy 69–75.

Goel, P.K., Prasher, S.O., Landry, J., Patel, R.M., Viau, A.A., 2003. Hyperspectral image classification to detect weed infestations and nitrogen status in corn. Trans. ASAE 46, 539.

Gómez-Casero, M.T., López-Granados, F., Pena-Barragán, J.M., Jurado-Expósito, M., García-Torres, L., Fernández-Escobar, R., 2007. Assessing nitrogen and potassium deficiencies in olive orchards through discriminant analysis of hyperspectral data. J. Am. Soc. Hortic. Sci. 132, 611–618.

Gomiero, T., Pimentel, D., Paoletti, M.G., 2011. Is there a need for a more sustainable agriculture? CRC. Crit. Rev. Plant Sci. 30, 6–23.

Gorretta, N., Nouri, M., Herrero, A., Gowen, A., Roger, J.-M., 2019. Early detection of the fungal disease" apple scab" using SWIR hyperspectral imaging, in: 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS). IEEE, pp. 1–4.

Gowen, A.A., O'Donnell, C.P., Cullen, P.J., Downey, G., Frias, J.M., 2007. Hyperspectral imaging–an emerging process analytical tool for food quality and safety control. Trends food Sci. Technol. 18, 590–598.

Goyal, N., Kumar, S., Saraswat, M., 2022. Detection of Unhealthy citrus leaves using Machine Learning Technique, in: 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, pp. 591–595.

Graeff, S., Pfenning, J., Claupein, W., Liebig, H.-P., 2008. Evaluation of Image Analysis to Determine the N-Fertilizer Demand of Broccoli Plants (Brassica oleracea convar. botrytis var. italica). Adv. Opt. Technol.

Gui, J., Wu, Z., Gu, M., Bao, X., 2019. Non-destructive detection of pesticide residues on broccoli based on NIR hyperspectral imaging. Int. Agric. Eng. J. 28, 289–294.

Guo, X., Ahlawat, Y.K., Liu, T., Zare, A., 2022. Evaluation of Postharvest Senescence of Broccoli via Hyperspectral Imaging. Plant Phenomics.

Gupta, S.B., Yadav, R.K., Hooda, R., Dhingra, S., Gupta, M., 2022. Analysis of Some Popular AI & ML Algorithms Used in Agriculture, in: 2022 International Conference on Computational Modelling, Simulation and Optimization (ICCMSO). IEEE, pp. 28–33.

Guru, D.S., Mallikarjuna, P.B., 2010. Spots and color based ripeness evaluation of tobacco leaves for automatic harvesting, in: Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia. pp. 198–202.

Hassan-Esfahani, L., Torres-Rua, A., McKee, M., 2015. Assessment of optimal irrigation water allocation for pressurized irrigation system using water balance approach, learning machines, and remotely sensed data. Agric. Water Manag. 153, 42–50.

Hawkins, D.M., 2004. The problem of overfitting. J. Chem. Inf. Comput. Sci. 44, 1–12.

Hayashi, M., Tamai, K., Owashi, Y., Miura, K., 2019. Automated machine learning for identification of pest aphid species (Hemiptera: Aphididae). Appl. Entomol. Zool. 54, 487–490.

Hernández-Hierro, J.M., Esquerre, C., Valverde, J., Villacreces, S., Reilly, K., Gaffney, M., González-Miret, M.L., Heredia, F.J., O'Donnell, C.P., Downey, G., 2014. Preliminary study on the use of near infrared hyperspectral imaging for quantitation and localisation of total glucosinolates in freeze-dried broccoli. J. Food Eng. 126, 107–112.

Hollberg, J.L., Schellberg, J., 2017. Distinguishing intensity levels of grassland fertilization using vegetation indices. Remote Sens. 9, 81.

Hosaka, A., Makino, Y., Kawagoe, Y., Oshita, S., 2012. Prediction of degradation rate of broccoli during storage by hyper spectral imaging., in: Post Harvest, Food and Process Engineering. International Conference of Agricultural Engineering-CIGR-AgEng 2012: Agriculture and Engineering for a Healthier Life, Valencia, Spain, 8-12 July 2012. CIGR-EurAgEng.

HU, L., ZHANG, X., ZHOU, Y., 2019. Farm size and fertilizer sustainable use: An empirical study in Jiangsu, China. J. Integr. Agric. 18, 2898–2909.

Huang, B., Sun, W., Zhao, Y., Zhu, J., Yang, R., Zou, Z., Ding, F., Su, J., 2007. Temporal and spatial variability of soil organic matter and total nitrogen in an agricultural ecosystem as affected by farming practices. Geoderma 139, 336–345.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., 2017. Speed/accuracy trade-offs for modern convolutional object detectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7310–7311.

Hughes, G., 1968. On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf. theory 14, 55–63.

Hussain, I., Hanjra, M.A., 2004. Irrigation and poverty alleviation: review of the empirical evidence. Irrig. Drain. 53, 1–15.

Ibáñez-Asensio, S., Marques-Mateu, A., Moreno-Ramón, H., Balasch, S., 2013. Statistical relationships between soil colour and soil attributes in semiarid areas. Biosyst. Eng. 116, 120–129.

IBM, 2024. What is AutoML [WWW Document]. URL https://www.ibm.com/topics/automl

Javaid, M., Haleem, A., Khan, I.H., Suman, R., 2023. Understanding the potential applications of Artificial Intelligence in Agriculture Sector. Adv. Agrochem 2, 15–30.

Jia, S., Li, H., Wang, Y., Tong, R., Li, Q., 2017. Hyperspectral imaging analysis for the classification of soil types and the determination of soil total nitrogen. Sensors 17, 2252.

Jiang, X., Yong, B., Garshasbi, S., Shen, J., Jiang, M., Zhou, Q., 2020. Crop and weed classification based on AutoML. arXiv Prepr. arXiv2010.14708.

Jin, H., Chollet, F., Song, Q., Hu, X., 2023. Autokeras: An automl library for deep learning. J. Mach. Learn. Res. 24, 1–6.

Kandpal, L.M., Lohumi, S., Kim, M.S., Kang, J.-S., Cho, B.-K., 2016. Near-infrared hyperspectral imaging system coupled with multivariate methods to predict viability and vigor in muskmelon seeds. Sensors Actuators B Chem. 229, 534–544.

Kanter, D.R., Zhang, X., Mauzerall, D.L., 2015. Reducing nitrogen pollution while decreasing farmers' costs and increasing fertilizer industry profits. J. Environ. Qual. 44, 325–335.

Karthickmanoj, R., Sasilatha, T., Padmapriya, J., 2021. Automated machine learning based plant stress detection system. Mater. Today Proc. 47, 1887–1891.

Kavuncuoğlu, E., Çetin, N., Yildirim, B., Nadimi, M., Paliwal, J., 2023. Exploration of Machine Learning Algorithms for pH and Moisture Estimation in Apples Using VIS-NIR Imaging. Appl. Sci. 13, 8391.

Khan, A., Vibhute, A.D., Mali, S., Patil, C.H., 2022. A systematic review on hyperspectral imaging technology with a machine and deep learning methodology for agricultural applications. Ecol. Inform. 69, 101678.

Khan, I.H., Liu, Haiyan, Li, W., Cao, A., Wang, X., Liu, Hongyan, Cheng, T., Tian, Y., Zhu, Y., Cao, W., 2021. Early detection of powdery mildew disease and accurate quantification of its severity using hyperspectral images in wheat. Remote Sens. 13, 3612.

Kim, D.-W., Jeong, S.J., Lee, W.S., Yun, H., Chung, Y.S., Kwon, Y.-S., Kim, H.-J., 2023. Growth monitoring of field-grown onion and garlic by CIE L* a* b* color space and region-based crop segmentation of UAV RGB images. Precis. Agric. 1–20.

Kim, Y., Glenn, D.M., Park, J., Ngugi, H.K., Lehman, B.L., 2010. Hyperspectral image analysis for plant stress detection, in: 2010 Pittsburgh, Pennsylvania, June 20-June 23, 2010. American Society of Agricultural and Biological Engineers, p. 1.

Koh, J.C.O., Spangenberg, G., Kant, S., 2021. Automated machine learning for high-throughput image-based plant phenotyping. Remote Sens. 13, 858.

Kong, W., Zhang, C., Liu, F., Nie, P., He, Y., 2013. Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis. sensors 13, 8916–8927.

Kotthoff, L., Thornton, C., Hoos, H.H., Hutter, F., Leyton-Brown, K., 2017. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. J. Mach. Learn. Res. 18, 1–5.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M. V, Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13, 8–17.

Kumar, J., Patel, N., Singh, R., Sahoo, P.K., Sudhishri, S., Sehgal, V.K., Marwaha, S., Singh, A.K., 2021. Development and evaluation of automation system for irrigation scheduling in broccoli (Brassica oleracea). Indian J. Agric. Sci. 91.

Kusumam, K., Krajník, T., Pearson, S., Duckett, T., Cielniak, G., 2017. 3D-vision based detection, localization, and sizing of broccoli heads in the field. J. F. Robot. 34, 1505–1518.

Lammel, J., Wollring, J., Reusch, S., 2001. Tractor based remote sensing for variable nitrogen fertilizer application. Plant Nutr. Food Secur. Sustain. agro-ecosystems through basic Appl. Res. 694–695.

Le, T.T., Fu, W., Moore, J.H., 2020. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics 36, 250–256.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

LeDell, E., Poirier, S., 2020. H2o automl: Scalable automatic machine learning, in: Proceedings of the AutoML Workshop at ICML. ICML San Diego, CA, USA.

Lee, C.-J., Yang, M.-D., Tseng, H.-H., Hsu, Y.-C., Sung, Y., Chen, W.-L., 2023. Single-plant broccoli growth monitoring using deep learning with UAV imagery. Comput. Electron. Agric. 207, 107739.

Lee, C.C., Koo, V.C., Lim, T.S., Lee, Y.P., Abidin, H., 2022. A multi-layer perceptron-based approach for early detection of BSR disease in oil palm trees using hyperspectral images. Heliyon 8.

Lee, L.C., Jemain, A.A., 2019. Predictive modelling of colossal ATR-FTIR spectral data using PLS-DA: empirical differences between PLS1-DA and PLS2-DA algorithms. Analyst 144, 2670–2678.

Lee, N.-Y., Na, I.-S., Lee, K.-W., Lee, D.-H., Kim, J.-W., Kook, M.-C., Hong, S.-J., Son, J.-Y., Lee, A.-Y., Om, A.-S., 2024. Detection of physical hazards from fruit processed products using hyperspectral imaging and prediction based on PLS-DA and logistic regression machine learning models. Appl. Food Res. 100506.

Liang, Z., Sang, M., Fan, P., Wu, B., Wang, L., Yang, S., Li, S., 2011. CIELAB coordinates in response to berry skin anthocyanins and their composition in Vitis. J. Food Sci. 76, C490–C497.

Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16, 321–332.

Lim, J., Fernández, C.A., Lee, S.W., Hatzell, M.C., 2021. Ammonia and Nitric Acid Demands for Fertilizer Use in 2050. ACS Energy Lett. 6, 3676–3685. https://doi.org/10.1021/acsenergylett.1c01614

Linaza, M.T., Posada, J., Bund, J., Eisert, P., Quartulli, M., Döllner, J., Pagani, A., G. Olaizola, I., Barriguinha, A., Moysiadis, T., 2021. Data-driven artificial intelligence applications for sustainable precision agriculture. Agronomy 11, 1227.

Link, A., Jasper, J., Olfs, H.W., 2004. Variable nitrogen fertilization by tractor-mounted remote sensing., in: Proceedings of the 7th International Conference on Precision Agriculture and Other Precision Resources Management, Hyatt Regency, Minneapolis, MN, USA, 25-28 July, 2004. Precision Agriculture Center, University of Minnesota, Department of Soil …, pp. 1413–1418.

Liu, D., Zeng, X.-A., Sun, D.-W., 2015. Recent developments and applications of hyperspectral imaging for quality evaluation of agricultural products: a review. Crit. Rev. Food Sci. Nutr. 55, 1744–1757.

Liu, F., He, Y., Wang, L., 2008. Determination of effective wavelengths for discrimination of fruit vinegars using near infrared spectroscopy and multivariate analysis. Anal. Chim. Acta 615, 10–17.

Liu, H., Li, M., Zhang, J., Gao, D., Sun, H., Zhang, M., Wu, J., n.d. Key Laboratory Of Modern

Precision Agriculture System Integration Research, MOEC, Key Laboratory Of Agricultural Information Acquisition Technology, MOAA, Beijing Key Laboratory Of Big Data Technology For Food Safety, BTAB, 2019. A novel wavelength select. Int. J. Agric. Biol. Eng 12, 149–155.

Loggenberg, K., Strever, A., Greyling, B., Poona, N., 2018. Modelling water stress in a Shiraz vineyard using hyperspectral imaging and machine learning. Remote Sens. 10, 202.

López-Granados, F., Jurado-Expósito, M., Atenciano, S., García-Ferrer, A., Sánchez de la Orden, M., García-Torres, L., 2002. Spatial variability of agricultural soil parameters in southern Spain. Plant Soil 246, 97–105.

Lu, R., Chen, Y.-R., 1999. Hyperspectral imaging for safety inspection of food and agricultural products, in: Pathogen Detection and Remediation for Safe Eating. SPIE, pp. 121–133.

Ma, Y., Liu, K., Guan, Z., Xu, X., Qian, X., Bao, H., 2018. Background augmentation generative adversarial networks (BAGANs): Effective data generation based on GAN-augmented 3D synthesizing. Symmetry (Basel). 10, 734.

Mahesh, S., Jayas, D.S., Paliwal, J., White, N.D.G., 2015. Hyperspectral imaging to classify and monitor quality of agricultural materials. J. Stored Prod. Res. 61, 17–26.

Makino, Y., Amino, G., 2020. Digitization of Broccoli freshness integrating external color and mass loss. Foods 9, 1305.

Makino, Y., Kousaka, Y., 2020. Prediction of degreening velocity of broccoli buds using hyperspectral camera combined with artificial neural networks. Foods 9, 558.

Malounas, I., Vierbergen, W., Kutluk, S., Zude-Sasse, M., Yang, K., Zhao, M., Argyropoulos, D., Van Beek, J., Ampe, E., Fountas, S., 2024. SpectroFood dataset: A comprehensive fruit and vegetable hyperspectral meta-dataset for dry matter estimation. Data Br. 110040.

Manheim, J., Doty, K.C., McLaughlin, G., Lednev, I.K., 2016. Forensic hair differentiation using attenuated total reflection Fourier transform infrared (ATR FT-IR) spectroscopy. Appl. Spectrosc. 70, 1109–1117.

Marquetti, I., Link, J.V., Lemes, A.L.G., dos Santos Scholz, M.B., Valderrama, P., Bona, E., 2016. Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee. Comput. Electron. Agric. 121, 313–319.

Marutani, M., Cruz, F., 1989. Influence of supplemental irrigation on development of potatoes in the tropics. HortScience 24, 920–923.

Mastrodimos, N., Lentzou, D., Templalexis, C., Tsitsigiannis, D.I., Xanthopoulos, G., 2022. Thermal and digital imaging information acquisition regarding the development of Aspergillus flavus in pistachios against Aspergillus carbonarius in table grapes. Comput. Electron. Agric. 192, 106628.

Mazzia, V., Khaliq, A., Salvetti, F., Chiaberge, M., 2020. Real-time apple detection system using embedded systems with hardware accelerators: An edge AI application. IEEE Access 8, 9102–9114.

McArthur, J.W., McCord, G.C., 2017. Fertilizing growth: Agricultural inputs and their effects in economic development. J. Dev. Econ. 127, 133–152.

Medic, T., 2023. Estimating Dry Matter and Total Soluble Content in Apples Using a Commercial Portable Hyperspectral Imaging System. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 48, 1757–1762.

Megeto, G.A.S., Silva, A.G. da, Bulgarelli, R.F., Bublitz, C.F., Valente, A.C., Costa, D.A.G. da, 2021. Artificial intelligence applications in the agriculture 4.0. Rev. Ciência Agronômica 51.

Mehta, S., Kukreja, V., Vats, S., 2023. Empowering Farmers with AI: Federated Learning of CNNs for Wheat Diseases Multi-Classification, in: 2023 4th International Conference for Emerging Technology (INCET). IEEE, pp. 1–6.

Michalski, R.S., Carbonell, J.G., Mitchell, T.M., 2013. Machine learning: An artificial intelligence approach. Springer Science & Business Media.

Mishra, P., Passos, D., 2022. Multi-output 1-dimensional convolutional neural networks for simultaneous prediction of different traits of fruit based on near-infrared spectroscopy. Postharvest Biol. Technol. 183, 111741.

Monteiro, S.T., Minekawa, Y., Kosugi, Y., Akazawa, T., Oda, K., 2007. Prediction of sweetness and amino acid content in soybean crops from hyperspectral imagery. ISPRS J. Photogramm. Remote Sens. 62, 2–12.

Montes, H.A., Cielniak, G., 2022. Multiple broccoli head detection and tracking in 3D point clouds for autonomous harvesting, in: AI for Agriculture and Food Systems.

Montes, H.A., Le Louedec, J., Cielniak, G., Duckett, T., 2020. Real-time detection of broccoli crops in 3D point clouds for autonomous robotic harvesting, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 10483–10488.

Mulla, D., Khosla, R., 2016. Historical evolution and recent advances in precision farming. Soil-specific farming Precis. Agric. 1–35.

Muruganantham, P., Samrat, N.H., Islam, N., Johnson, J., Wibowo, S., Grandhi, S., 2022. Rapid Estimation of Moisture Content in Unpeeled Potato Tubers Using Hyperspectral Imaging. Appl. Sci. 13, 53.

Nampally, T., Kumar, K., Chatterjee, S., Pachamuthu, R., Naik, B., Desai, U.B., 2023. StressNet: a spatial-spectral-temporal deformable attention-based framework for water stress classification in maize. Front. Plant Sci. 14, 1241921.

Nanni, M.R., Demattê, J.A.M., Rodrigues, M., Santos, G.L.A.A. dos, Reis, A.S., Oliveira, K.M. de, Cezar, E., Furlanetto, R.H., Crusiol, L.G.T., Sun, L., 2021. Mapping particle size and soil organic matter in tropical soil based on hyperspectral imaging and non-imaging sensors. Remote Sens. 13, 1782.

Nassau, K., 2024. Colour. Encycl. Br.

Natural Resources Conservation Service. U.S. Department of Agriculture, 1999. Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys.

Nelson, G.C., Rosegrant, M.W., Koo, J., Robertson, R., Sulser, T., Zhu, T., Ringler, C., Msangi, S., Palazzo, A., Batka, M., 2009. Climate change: Impact on agriculture and costs of adaptation. Intl Food Policy Res Inst.

Nguyen, H.D.D., Pan, V., Pham, C., Valdez, R., Doan, K., Nansen, C., 2020. Night-based hyperspectral imaging to study association of horticultural crop leaf reflectance and nutrient status. Comput. Electron. Agric. 173, 105458.

Nkaa, Fa., Nwokeocha, O.W., Ihuoma, O., 2014. Effect of phosphorus fertilizer on growth and yield of cowpea (Vigna unguiculata). IOSR J. Pharm. Biol. Sci. 9, 74–82.

Noé, S.R., Manuel, C.S.J., Jerzy, S.H.R., 2002. A yield sensing system for broccoli, in: 2002 ASAE Annual Meeting. American Society of Agricultural and Biological Engineers, p. 1.

Olympios, 2015. The technique of cultivation of outdoor vegetables. Stamouli, Athens.

Ondimu, S.N., Murase, H., 2008. Classification of water stress in Sunagoke moss using color texture and neural networks. Environ. Control Biol. 46, 21–29.

Pallottino, F., Menesatti, P., Figorilli, S., Antonucci, F., Tomasone, R., Colantoni, A., Costa, C., 2018. Machine vision retrofit system for mechanical weed control in precision agriculture applications. Sustainability 10, 2209.

Pan, W., Zhao, J., Chen, Q., 2015. Classification of foodborne pathogens using near infrared (NIR) laser scatter imaging system with multivariate calibration. Sci. Rep. 5, 9524.

Pandey, P., Ge, Y., Stoerger, V., Schnable, J.C., 2017. High throughput in vivo analysis of plant leaf chemical properties using hyperspectral imaging. Front. Plant Sci. 8, 1348.

Pandey, R., Gamit, N., Naik, S., 2014. Non-destructive quality grading of mango (Mangifera Indica L) based on CIELab colour model and size, in: 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies. IEEE, pp. 1246–1251.

Papadopoulos, G., Mavroeidis, A., Roussis, I., Kakabouki, I., Stavropoulos, P., Bilalis, D., 2023. Evaluation of tillage & fertilization in Carthamus tinctorius L. using remote sensing. Smart Agric. Technol. 4, 100158.

Pathak, H.S., Brown, P., Best, T., 2019. A systematic literature review of the factors affecting the precision agriculture adoption process. Precis. Agric. 20, 1292–1316.

Pathare, P.B., Opara, U.L., Al-Said, F.A.-J., 2013. Colour measurement and analysis in fresh and processed foods: a review. Food bioprocess Technol. 6, 36–60.

Pawlak, K., Kołodziejczak, M., 2020. The role of agriculture in ensuring food security in developing countries: Considerations in the context of the problem of sustainable food production. Sustainability 12, 5488.

Peng, Y., Nguy-Robertson, A., Arkebauer, T., Gitelson, A.A., 2017. Assessment of canopy chlorophyll content retrieval in maize and soybean: Implications of hysteresis on the development of generic algorithms. Remote Sens. 9, 226.

Pérez, S.M., 2021. Physicochemical Characterization of Pomegranate (Punica Granatum L.) Native to Jordan During Different Maturity Stages: Color Evaluation Using the Cielab and Cielch Systems. J. Ecol. Eng. 22, 214–221.

Pineda, M., Pérez-Bueno, M.L., Barón, M., 2022. Novel vegetation indices to identify broccoli plants infected with Xanthomonas campestris pv. campestris. Front. Plant Sci. 13, 790268.

Pisani, M., Zucco, M., 2009. Compact imaging spectrometer combining Fourier transform spectroscopy with a Fabry-Perot interferometer. Opt. Express 17, 8319–8331.

Playán, E., Mateos, L., 2006. Modernization and optimization of irrigation systems to increase water productivity. Agric. water Manag. 80, 100–116.

Poblete, T., Ortega-Farías, S., Moreno, M.A., Bardeen, M., 2017. Artificial neural network to predict vine water status spatial variability using multispectral information obtained from an unmanned aerial vehicle (UAV). Sensors 17, 2488.

Polder, G., Gowen, A., 2021. The hype in spectral imaging. Spectrosc. Eur.

Press, W.H., Teukolsky, S.A., 1990. Savitzky-Golay smoothing filters. Comput. Phys. 4, 669–672.

Psiroukis, V., Espejo-Garcia, B., Chitos, A., Dedousis, A., Karantzalos, K., Fountas, S., 2022. Assessment of different object detectors for the maturity level classification of broccoli crops

using uav imagery. Remote Sens. 14, 731.

Qin, J., Chao, K., Kim, M.S., Lu, R., Burks, T.F., 2013. Hyperspectral and multispectral imaging for evaluating food safety and quality. J. Food Eng. 118, 157–171.

Quemada, M., Gabriel, J.L., Zarco-Tejada, P., 2014. Airborne hyperspectral images and ground-level optical sensors as assessment tools for maize nitrogen fertilization. Remote Sens. 6, 2940–2962.

Rahimikhoob, H., Delshad, M., Habibi, R., 2023. Leaf area estimation in lettuce: Comparison of artificial intelligence-based methods with image analysis technique. Measurement 222, 113636.

Ramirez, R.A., 2006. Computer vision based analysis of broccoli for application in a selective autonomous harvester.

Rampáčková, E., Göttingerová, M., Kiss, T., Ondrášek, I., Venuta, R., Wolf, J., Nečas, T., Ercisli, S., 2021. CIELAB analysis and quantitative correlation of total anthocyanin content in European and Asian plums.

Redhu, N.S., Thakur, Z., Yashveer, S., Mor, P., 2022. Artificial intelligence: a way forward for agricultural sciences, in: Bioinformatics in Agriculture. Elsevier, pp. 641–668.

Rodríguez-Pulido, F.J., Gordillo, B., Heredia, F.J., González-Miret, M.L., 2021. CIELAB–Spectral image MATCHING: An app for merging colorimetric and spectral images for grapes and derivatives. Food Control 125, 108038.

Romero, M., Luo, Y., Su, B., Fuentes, S., 2018. Vineyard water status estimation using multispectral imagery from an UAV platform and machine learning algorithms for irrigation scheduling management. Comput. Electron. Agric. 147, 109–117.

Roy, S.K., Shibusawa, S., Okayama, T., 2006. Textural analysis of soil images to quantify and characterize the spatial variation of soil properties using a real-time soil sensor. Precis. Agric. 7, 419–436.

Ruett, M., Junker-Frohn, L.V., Siegmann, B., Ellenberger, J., Jaenicke, H., Whitney, C., Luedeling, E., Tiede-Arlt, P., Rascher, U., 2022. Hyperspectral imaging for high-throughput vitality monitoring in ornamental plant production. Sci. Hortic. (Amsterdam). 291, 110546.

Ruiz, G.R., Navas, L.M., Gil, J.G., 2009. Analysis of color variations on sunflower crop images, own-ing to changes in environmental illumination, in: 1st International Workshop on Computer Image Analysis in Agriculture.

Sabzi, S., Pourdarbani, R., Rohban, M.H., Fuentes-Penna, A., Hernández-Hernández, J.L., Hernández-Hernández, M., 2021. Classification of cucumber leaves based on nitrogen content using the hyperspectral imaging technique and majority voting. Plants 10, 898.

Sari, Y., Alkaff, M., 2020. Classification of rice leaf using fuzzy logic and hue saturation value (hsv) to determine fertilizer dosage, in: 2020 Fifth International Conference on Informatics and Computing (ICIC). IEEE, pp. 1–6.

Sarwar, N., Maqsood, M., Mubeen, K., Shehzad, M., Bhullar, M.S., Qamar, R., Akbar, N., 2010. Effect of different levels of irrigation on yield and yield components of wheat cultivars. Pak. J. Agri. Sci 47, 371–374.

Schanda, J., 2007. Colorimetry: understanding the CIE system. John Wiley & Sons.

Schelkanova, I., Pandya, A., Muhaseen, A., Saiko, G., Douplik, A., 2015. Early optical diagnosis of pressure ulcers, in: Biophotonics for Medical Applications. Elsevier, pp. 347–375.

Schellenberg, D.L., Bratsch, A.D., Shen, Z., 2009. Large single-head broccoli yield as affected by plant density, nitrogen, and cultivar in a plasticulture system. Horttechnology 19, 792–795.

Setyawan, T.A., Riwinanto, S.A., Nursyahid, A., Nugroho, A.S., 2018. Comparison of hsv and lab color spaces for hydroponic monitoring system, in: 2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE). IEEE, pp. 347–352.

Shahid, S., 2011. Impact of climate change on irrigation water demand of dry season Boro rice in northwest Bangladesh. Clim. Change 105, 433–453.

Sharma, A., Jain, A., Gupta, P., Chowdary, V., 2020. Machine learning applications for precision agriculture: A comprehensive review. IEEE Access 9, 4843–4873.

Shaw, D.J., 2007. World food summit, 1996, in: World Food Security: A History since 1945. Springer, pp. 347–360.

Sheahan, M., Barrett, C.B., 2017. Ten striking facts about agricultural input use in Sub-Saharan Africa. Food Policy 67, 12–25.

Shuai, L., Li, Z., Chen, Z., Luo, D., Mu, J., 2024. A research review on deep learning combined with hyperspectral Imaging in multiscale agricultural sensing. Comput. Electron. Agric. 217, 108577.

Siavoshi, M., Nasiri, A., Laware, S.L., 2011. Effect of organic fertilizer on growth and yield components in rice (Oryza sativa L.). J. Agric. Sci. 3, 217.

Siedliska, A., Baranowski, P., Pastuszka-Woźniak, J., Zubik, M., Krzyszczak, J., 2021. Identification of plant leaf phosphorus content at different growth stages based on hyperspectral reflectance. BMC Plant Biol. 21, 1–17.

Signoroni, A., Savardi, M., Baronio, A., Benini, S., 2019. Deep learning meets hyperspectral image analysis: A multidisciplinary review. J. Imaging 5, 52.

Silwal, A., Parhar, T., Yandun, F., Baweja, H., Kantor, G., 2021. A robust illumination-invariant camera system for agricultural applications, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 3292–3298.

Stamatas, G.N., Balas, C.J., Kollias, N., 2003. Hyperspectral image acquisition and analysis of skin, in: Spectral Imaging: Instrumentation, Applications, and Analysis II. SPIE, pp. 77–82.

Stivers, L.J., Jackson, L.E., Pettygrove, G.S., 1993. Use of nitrogen by lettuce, celery, broccoli, and cauliflower: a literature review.

Sun, G., Xie, H., Sinnott, R.O., 2017. A crop water stress monitoring system utilising a hybrid e-infrastructure, in: Proceedings of The10th International Conference on Utility and Cloud Computing. pp. 161–170.

Sun, T., Zhang, W., Miao, Z., Zhang, Z., Li, N., 2023. Object localization methodology in occluded agricultural environments through deep learning and active sensing. Comput. Electron. Agric. 212, 108141.

Sun, X., Subedi, P., Walker, R., Walsh, K.B., 2020. NIRS prediction of dry matter content of single olive fruit with consideration of variable sorting for normalisation pre-treatment. Postharvest Biol. Technol. 163, 111140.

Taghizadeh, M., Gowen, A., O'Donnell, C.P., 2009. Prediction of white button mushroom (Agaricus bisporus) moisture content using hyperspectral imaging. Sens. Instrum. Food Qual. Saf. 3, 219–226.

Taylor, J., 2023. Precision agriculture.

Templalexis, C., Giorni, P., Lentzou, D., Mozzoni, F., Battilani, P., Tsitsigiannis, D.I., Xanthopoulos, G., 2023. IoT for Monitoring Fungal Growth and Ochratoxin A Development in Grapes Solar Drying in Tunnel and in Open Air. Toxins (Basel). 15, 613.

Thomas, J., Coors, S., Bischl, B., 2018. Automatic gradient boosting. arXiv Prepr. arXiv1807.03873.

Thompson, T.L., Doerge, T.A., Godin, R.E., 2002. Subsurface drip irrigation and fertigation of broccoli: I. Yield, quality, and nitrogen uptake. Soil Sci. Soc. Am. J. 66, 186–192.

Tremblay, N., Bélec, C., Jenni, S., Fortier, E., Mellgren, R., 2008. The Dualex–a new tool to determine nitrogen sufficiency in broccoli, in: International Symposium on Application of Precision Agriculture for Fruits and Vegetables 824. pp. 121–132.

Tsenkova, R., 2010. Aquaphotomics: Water in the biological and aqueous world scrutinised with invisible light. Spectrosc. Eur. 22, 6.

Tu, Q., Yang, T., Qu, Y., Gao, S., Zhang, Z., Zhang, Q., Wang, Y., Wang, J., He, L., 2019. In situ colorimetric detection of glyphosate on plant tissues using cysteamine-modified gold nanoparticles. Analyst 144, 2017–2025.

Tunny, S.S., Kurniawan, H., Amanah, H.Z., Baek, I., Kim, M.S., Chan, D., Faqeerzada, M.A., Wakholi, C., Cho, B.-K., 2023. Hyperspectral imaging techniques for detection of foreign materials from fresh-cut vegetables. Postharvest Biol. Technol. 201, 112373.

Ugarte Fajardo, J., Bayona Andrade, O., Criollo Bonilla, R., Cevallos-Cevallos, J., Mariduena-Zavala, M., Ochoa Donoso, D., Vicente Villardon, J.L., 2020. Early detection of black Sigatoka in banana leaves using hyperspectral images. Appl. Plant Sci. 8, e11383.

USDA, 2016. The commercial storage of fruits, vegetables and florist and nursery stocks. [WWW Document]. URL www.ars.usda.gov/arsuserfiles/oc/np/commercialstorage/commercialstorage.pdf

Vågen, I.M., Skjelvåg, A.O., Bonesmo, H., 2004. Growth analysis of broccoli in relation to fertilizer nitrogen application. J. Hortic. Sci. Biotechnol. 79, 484–492.

Vasilakakis, M., 2006. Μετασυλλεκτική φυσιολογία, μεταχείριση οπωροκηπευτικών και τεχνολογία. Διαιτητική αξία οπωροκηπευτικών. Εκδόσεις Γαρταγάνη, Θεσσαλονίκη, Ελλάς. ΕΕ.

Vermeulen, S.J., Challinor, A.J., Thornton, P.K., Campbell, B.M., Eriyagama, N., Vervoort, J.M., Kinyangi, J., Jarvis, A., Läderach, P., Ramirez-Villegas, J., 2013. Addressing uncertainty in adaptation planning for agriculture. Proc. Natl. Acad. Sci. 110, 8357–8362.

Vieira, T.F., Makimori, G.Y.F., dos Santos Scholz, M.B., Zielinski, A.A.F., Bona, E., 2020. Chemometric approach using ComDim and PLS-DA for discrimination and classification of commercial yerba mate (Ilex paraguariensis St. Hil.). Food Anal. Methods 13, 97–107.

Vignesh, T., Thyagharajan, K.K., Murugan, D., n.d. Land Use and Land Cover Classification Using CIELAB Color Space, PCNN and SOM.

Virnodkar, S.S., Pachghare, V.K., Patil, V.C., Jha, S.K., 2020. Remote sensing and machine learning for crop water stress determination in various crops: a critical review. Precis. Agric. 21, 1121–1155.

Vivó-Truyols, G., Schoenmakers, P.J., 2006. Automatic selection of optimal Savitzky− Golay smoothing. Anal. Chem. 78, 4598–4608.

Wakchaure, M., Patle, B.K., Mahindrakar, A.K., 2023. Application of AI techniques and robotics in agriculture: A review. Artif. Intell. Life Sci. 100057.

Wang, H., Peng, J., Xie, C., Bao, Y., He, Y., 2015. Fruit quality evaluation using spectroscopy technology: a review. Sensors 15, 11889–11927.

Wang, Y., Hu, X., Hou, Z., Ning, J., Zhang, Z., 2018. Discrimination of nitrogen fertilizer levels of tea plant (Camellia sinensis) based on hyperspectral imaging. J. Sci. Food Agric. 98, 4659–4664.

Wieme, J., Mollazade, K., Malounas, I., Zude-Sasse, M., Zhao, M., Gowen, A., Argyropoulos, D., Fountas, S., Van Beek, J., 2022. Application of hyperspectral imaging systems and artificial intelligence for quality assessment of fruit, vegetables and mushrooms: A review. Biosyst. Eng. 222, 156–176.

Wien, H.C., Wurr, D.C.E., 1997. Cauliflower, broccoli, cabbage and Brussels sprouts. Physiol. Veg. Crop.

Williams, D., Karley, A., Britten, A., McCallum, S., Graham, J., 2023. Raspberry plant stress detection using hyperspectral imaging. Plant Direct 7, e490.

Wipf, D., Nagarajan, S., 2007. A new view of automatic relevance determination. Adv. Neural Inf. Process. Syst. 20.

Wold, S., Sjostrom, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratoary Systems 58: 109–130.

Xia, J., Zhang, J., Zhao, Y., Huang, Y., Xiong, Y., Min, S., 2019. Fourier transform infrared spectroscopy and chemometrics for the discrimination of paper relic types. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 219, 8–14.

Xie, C., Yang, C., He, Y., 2017. Hyperspectral imaging for classification of healthy and gray mold diseased tomato leaves with different infection severities. Comput. Electron. Agric. 135, 154–162.

Xu, Y., Knudby, A., Shen, Y., Liu, Y., 2018. Mapping monthly air temperature in the Tibetan Plateau from MODIS data based on machine learning methods. IEEE J. Sel. Top. Appl. earth Obs. Remote Sens. 11, 345–354.

Yakushev, V., Kanash, E., 2016. Evaluation of wheat nitrogen status by colorimetric characteristics of crop canopy presented in digital images. J. Agric. Informatics 7.

Yakushev, V.P., Kanash, E. V, 2011. Evaluation of wheat nitrogen status by colorimetric characteristics of crop canopy from digital images, in: Precision Agriculture 2011. pp. 341–351.

Yang, C., Akimoto, Y., Kim, D.W., Udell, M., 2019. OBOE: Collaborative filtering for AutoML model selection, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1173–1183.

Yi, Q.-X., Huang, J.-F., Wang, F.-M., Wang, X.-Z., Liu, Z.-Y., 2007. Monitoring rice nitrogen status using hyperspectral reflectance and artificial neural network. Environ. Sci. Technol. 41, 6770–6775.

Yu, H., Kong, B., Hou, Y., Xu, X., Chen, T., Liu, X., 2022. A critical review on applications of hyperspectral remote sensing in crop monitoring. Exp. Agric. 58.

Yuan, B.-Z., Nishiyama, S., Kang, Y., 2003. Effects of different irrigation regimes on the growth and yield of drip-irrigated potato. Agric. water Manag. 63, 153–167.

Yuan, L., Yan, P., Han, W., Huang, Y., Wang, B., Zhang, J., Zhang, H., Bao, Z., 2019. Detection of anthracnose in tea plants based on hyperspectral imaging. Comput. Electron. Agric. 167, 105039.

Zhang, B., Ou, Y., Yu, S., Liu, Yuchen, Liu, Ying, Qiu, W., 2023. Gray mold and anthracnose disease detection on strawberry leaves using hyperspectral imaging. Plant Methods 19, 1–13.

Zhang, J., Tian, Y., Yan, L., Wang, B., Wang, L., Xu, J., Wu, K., 2021. Diagnosing the symptoms of sheath blight disease on rice stalk with an in-situ hyperspectral imaging technique. Biosyst. Eng. 209, 94–105.

Zhang, W., Zhang, WeiXin, Yang, Y., Hu, G., Ge, D., Liu, H., Cao, H., 2021. A cloud computing-based approach using the visible near-infrared spectrum to classify greenhouse tomato plants under water stress. Comput. Electron. Agric. 181, 105966.

Zhang, Y., Nock, J.F., Al Shoffe, Y., Watkins, C.B., 2019. Non-destructive prediction of soluble solids and dry matter contents in eight apple cultivars using near-infrared spectroscopy. Postharvest Biol. Technol. 151, 111–118.

Zheng, H., Lu, H., 2012. A least-squares support vector machine (LS-SVM) based on fractal analysis and CIELab parameters for the detection of browning degree on mango (Mangifera indica L.). Comput. Electron. Agric. 83, 47–51.

Zhou, C., Hu, J., Xu, Z., Yue, J., Ye, H., Yang, G., 2020. A monitoring system for the segmentation and grading of broccoli head based on deep learning and neural networks. Front. Plant Sci. 11, 402.

Zhou, C., Ye, H., Sun, D., Yue, J., Yang, G., Hu, J., 2022. An automated, high-performance approach for detecting and characterizing broccoli based on UAV remote-sensing and transformers: A case study from Haining, China. Int. J. Appl. Earth Obs. Geoinf. 114, 103055.

Zimmermann, B., Kohler, A., 2013. Optimizing Savitzky–Golay parameters for improving spectral resolution and quantification in infrared spectroscopy. Appl. Spectrosc. 67, 892–902.

Zou, K., Ge, L., Zhang, C., Yuan, T., Li, W., 2019. Broccoli seedling segmentation based on support vector machine combined with color texture features. IEEE Access 7, 168565–168574.

Zou, K., Ge, L., Zhou, H., Zhang, C., Li, W., 2021. Broccoli seedling pest damage degree evaluation based on machine learning combined with color and shape features. Inf. Process. Agric. 8, 505–514.