Poznań University of Technology

Politechnika Poznańska

Faculty of Control, Robotics and Electrical Engineering
Institute of Robotics and Machine Intelligence

Wydział Automatyki, Robotyki i Elektrotechniki
Instytut Robotyki i Inteligencji Maszynowej

Doctoral dissertation

Multispectral imaging utilising machine learning techniques for agricultural and environmental protection applications

Obrazowanie wielospektralne z wykorzystaniem technik uczenia maszynowego w rolnictwie i ochronie środowiska

mgr inż. Przemysław Aszkowski

Supervisor: dr hab. inż. Paweł Drapikowski
Auxiliary supervisor: dr inż. Marek Kraft

Poznań, 2026

mgr inż. Przemysław Aszkowski
Multispectral imaging utilising machine learning techniques for agricultural and environmental protection applications
Doctoral dissertation, Poznań, 2026
Supervisor: dr hab. inż. Paweł Drapikowski
Auxiliary supervisor: dr inż. Marek Kraft

Poznań University of Technology
Institute of Robotics and Machine Intelligence
Faculty of Control, Robotics and Electrical Engineering

Abstract

Computer vision, encompassing both image acquisition and image analysis, is often the first stage of perception and control pipelines in robotics and decision-support systems. While RGB imagery in the visible spectrum remains the most commonly used data source and benefits from the most mature processing methods, multispectral imaging can provide additional information beyond human vision. In many applications, including agriculture and environmental protection, these additional spectral bands contain valuable information that cannot be captured using RGB imagery alone, particularly for plant stress monitoring. Multispectral imaging can support a wide range of agricultural and environmental applications, such as disease detection, plant segmentation, or estimation of soil parameters. Such capabilities are particularly important in precision agriculture, where they can contribute to improved yields, reduced fertiliser usage, and more informed decision-making.

Relevant vision systems operate across a wide range of spatial scales and platforms, from Earth-observation satellites monitoring large regions of the planet, through unmanned aerial vehicles surveying individual fields, to agricultural robots observing individual plants. This thesis discusses the challenges associated with processing multispectral data acquired from such diverse systems. These challenges include environmental variability resulting from geographical location, climate, weather conditions, or crop type, as well as the spectral and spatial characteristics of the acquired data. Furthermore, each imaging platform introduces its own constraints and trade-offs, requiring dedicated approaches to data acquisition and analysis. The thesis further analyses machine learning techniques applicable to multispectral data and investigates how they can address the aforementioned challenges. Particular attention is given to the progress achieved in multispectral image analysis in comparison with traditional RGB-based approaches.

The research includes the development and evaluation of machine learning methods for agricultural and environmental applications. The conducted experiments, using both multispectral and RGB imagery, demonstrate the impact of

multispectral information on model performance, considering not only predictive performance but also deployment-related constraints, particularly in resource-limited robotic systems. The research further investigates how multispectral data can be utilised to reduce the effort required for manual annotation. In addition, experiments on model explainability and spectral band selection are presented, aiming to simplify deployment and improve the practical applicability of multispectral systems.

As publicly available datasets and open-source software are essential for reproducibility and scientific progress, this thesis contributes not only through the use of existing datasets but also through the collection and curation of new datasets. Furthermore, several open-source tools were developed and released as part of the research. One of these tools has attracted interest from the research community and helped to popularise machine learning techniques among specialists from related disciplines.

Overall, this thesis demonstrates the value of combining multispectral imaging with modern machine learning methods, particularly deep learning, for agricultural and environmental applications. Beyond improving predictive performance, the presented research highlights approaches that facilitate practical deployment under real-world hardware constraints.

Streszczenie

Wizja komputerowa, obejmująca zarówno pozyskiwanie, jak i analizę obrazów, stanowi często pierwszy etap procesów percepcji i sterowania w robotyce oraz systemach wspomagania decyzji. Chociaż obrazy RGB w paśmie widzialnym pozostają najczęściej wykorzystywanym źródłem danych i korzystają z najbardziej rozwiniętych metod przetwarzania, obrazowanie wielospektralne może dostarczyć dodatkowych informacji wykraczających poza możliwości ludzkiego wzroku. W wielu zastosowaniach, w tym w rolnictwie i ochronie środowiska, te dodatkowe pasma spektralne zawierają cenne informacje, których nie da się uchwycić przy użyciu wyłącznie obrazów RGB, zwłaszcza w przypadku monitorowania stresu roślin.

Obrazowanie wielospektralne może wspierać szeroki zakres zastosowań rolniczych i środowiskowych, takich jak wykrywanie chorób, segmentacja roślin lub szacowanie parametrów gleby. Takie możliwości są szczególnie ważne w rolnictwie precyzyjnym, gdzie mogą przyczynić się do poprawy plonów, zmniejszenia zużycia nawozów i podejmowania bardziej świadomych decyzji.

Systemy wizyjne wykorzystywane w tych zastosowaniach działają w szerokim zakresie skal przestrzennych i platform, od satelitów obserwujących znaczne obszary Ziemi, poprzez bezzałogowe statki powietrzne obrazujące pojedyncze pola uprawne, aż po roboty rolnicze obserwujące pojedyncze rośliny. Niniejsza rozprawa omawia wyzwania związane z przetwarzaniem danych wielospektralnych pozyskanych z tak różnorodnych systemów. Wyzwania te obejmują zmienność środowiskową wynikającą z położenia geograficznego, klimatu, warunków pogodowych czy rodzaju upraw, a także charakterystykę spektralną i przestrzenną pozyskanych danych. Ponadto każda platforma obrazowania wiąże się z własnymi ograniczeniami i kompromisami, wymagając dedykowanych metod pozyskiwania oraz analizy danych. W rozprawie przeanalizowano również techniki uczenia maszynowego mające zastosowanie do danych wielospektralnych oraz zbadano, w jaki sposób mogą one pomagać w rozwiązywaniu wspomnianych problemów. Szczególną uwagę poświęcono postępom osiągniętym w analizie obrazów wielospektralnych w porównaniu z!tradycyjnymi podejściami opartymi na RGB.

Badania obejmują opracowanie i ocenę metod uczenia maszynowego do zastosowań rolniczych i środowiskowych. Przeprowadzone eksperymenty z wykorzystaniem zarówno obrazów wielospektralnych, jak i RGB pokazują wpływ informacji wielospektralnej na skuteczność modeli, biorąc pod uwagę nie tylko skuteczność predykcji, ale także ograniczenia związane z wdrożeniem, szczególnie w systemach robotycznych o ograniczonych zasobach. Zbadano również, w jaki sposób dane wielospektralne mogą być wykorzystane do zmniejszenia nakładu pracy wymaganego do ręcznego adnotowania danych. Ponadto przedstawiono eksperymenty dotyczące wyjaśnialności modeli i selekcji pasm spektralnych, których celem było uproszczenie wdrażania i zwiększenie praktycznej użyteczności systemów wielospektralnych.

Ponieważ publicznie dostępne zbiory danych i oprogramowanie wolnoźródłowe mają kluczowe znaczenie dla odtwarzalności wyników i rozwoju nauki, rozprawa wnosi wkład nie tylko poprzez wykorzystanie istniejących zbiorów danych, ale także poprzez pozyskanie i opracowanie nowych zbiorów danych. Ponadto w ramach badań opracowano i udostępniono kilka narzędzi open source. Jedno z tych narzędzi wzbudziło znaczne zainteresowanie społeczności badawczej i pomogło spopularyzować techniki uczenia maszynowego wśród specjalistów z pokrewnych dyscyplin.

Podsumowując, rozprawa potwierdza zasadność połączenia obrazowania wielospektralnego z nowoczesnymi metodami uczenia maszynowego, w szczególności głębokiego uczenia, w zastosowaniach rolniczych i środowiskowych. Oprócz poprawy skuteczności predykcji, przedstawione badania podkreślają podejścia, które ułatwiają praktyczne wdrożenie w warunkach rzeczywistych ograniczeń sprzętowych.

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisor, Paweł Drapikowski, for his scientific guidance, continuous support, and meticulous attention to detail throughout the course of this work. I am also deeply grateful to my co-supervisor, Marek Kraft, who guided me through every stage of the doctoral journey – from the initial idea, through the research process, to the completion of this dissertation. His support extended far beyond academic advice – he consistently offered encouragement, inspiration, and kindness.

I would also like to thank my family, especially my wife, for her unwavering support during the years of my doctoral studies, and my parents, whose upbringing and encouragement made it possible for me to pursue and achieve my life goals.

Last but not least, I would like to thank my colleagues, especially Bartosz and Dominik. Without you, this journey would not have been possible. Thank you for your collaboration, inspiration, support, and motivation along the way.

Contents

List of acronyms

BOA Bottom-of-Atmosphere.

CLIP Contrastive Language-Image Pretraining.

CNN Convolutional Neural Network.

CRP Calibrated Reflectance Panel.

FLOPs Floating-Point Operations.

FPS Frames Per Second.

GIS Geographic Information System.

GPU Graphics Processing Unit.

Grad-CAM Gradient-weighted Class Activation Mapping.

GSD Ground Sampling Distance.

IoU Intersection over Union.

kNN k-Nearest Neighbours.

LiDAR Light Detection and Ranging.

LIME Local Interpretable Model-agnostic Explanations.

LSTM Long Short-Term Memory.

MSE Mean Squared Error.

NDVI Normalized Difference Vegetation Index.

NIR Near-Infrared.

ONNX Open Neural Network Exchange.

PCA Principal Component Analysis.

R-CNN Region-based Convolutional Neural Network.

RAM Random Access Memory.

RNN Recurrent Neural Network.

SAM Segment Anything Model.

SAR Synthetic Aperture Radar.

SHAP SHapley Additive exPlanations.

SVM Support Vector Machine.

SVR Support Vector Regression.

SWIR Short-Wave Infrared.

TOA Top-of-Atmosphere.

TPU Tensor Processing Unit.

UAV Unmanned Aerial Vehicle.

ViT Vision Transformer.

VRAM Video Random Access Memory.

XAI Explainable Artificial Intelligence.

XGBoost Extreme Gradient Boosting.

1 Introduction

The field of computer vision has evolved rapidly in recent decades, largely driven by advances in machine learning techniques, especially deep learning models such as CNN (Convolutional Neural Network), which have outperformed traditional methods in many applications [1]. Alongside the development of data processing techniques, sensor technology has also steadily advanced, leading to smaller, cheaper, and more accessible hardware. This applies not only to commonly used RGB imaging systems, but even more so to multispectral and hyperspectral technologies, which are now employed across a wide range of platforms, from satellites to ground-level survey robots [2]. Although the cost of multispectral equipment remains significantly higher than that of RGB cameras, it provides information beyond the range of human vision and has been shown to be superior in certain applications, especially when individual wavelength bands have a known physical interpretation.

These developments create new opportunities but also introduce challenges, both from the perspective of scientific methodology and practical engineering. To build applications with a meaningful impact on society, many factors must be considered throughout the entire process, including both model development and later deployment. These factors include, among others, the cost of hardware, the amount of data and expert knowledge required to build reliable models, as well as the reliability and explainability of the final results. The latter is especially important for building trust among end users. In this context, multispectral imaging has attracted increasing attention, both in research activity and in the growing availability of public datasets. For example, the Sentinel-2 satellite missions have already collected tens of petabytes of multispectral imagery, a large proportion of which is publicly accessible.

One of the fields that can particularly benefit from recent advances in remote sensing, and multispectral imaging in particular, is agriculture and environmental protection [3, 4, 5]. Within this domain, an important topic is precision agriculture, which can be understood as a computer-assisted management approach aimed at improving yields, reducing financial costs, and increasing sustainability. By enabling both local and global improvements to agricultural processes, such as optimised use of fertilisers and water, precision agriculture can reduce environmental pressure while maintaining or improving productivity. At the same time, systematic monitoring of environmental conditions can further support agricultural decision-making and resilience. These technological developments are closely aligned with sustainability objectives that have gained increasing importance in recent decades, particularly within the European Union, as reflected in policy frameworks such as the Farm to Fork Strategy and the European Green Deal. In this sense, precision agriculture can be viewed not only as a technological advancement, but also as a key policy instrument for achieving broader sustainable development goals.

Due to the wide range of tasks, spanning from global-scale monitoring of the Earth to analysis of individual plants, such as monitoring photosynthetic activity [6], both remote sensing and ground-level multispectral surveys can be applied. While these approaches often rely on similar machine learning techniques, they also present unique challenges related to data characteristics and scale. Despite the significant progress made in recent years and the numerous opportunities for practical applications, the use of multispectral imaging to support reliable and scalable decision-making systems remains a substantial challenge [7, 8, 9, 10].

1.1 Problem description and motivation

As remote imaging is an important topic in agriculture and environmental protection, it has a direct impact on society. Machine learning systems can help in monitoring, early response, support decision-making, and management processes. Although precision agriculture encompasses both crop farming and animal husbandry, this research focuses on the former. The topics of interest range from crop growth and health monitoring, through the assessment of soil parameters and fertiliser usage, to environmental monitoring, such as evaluating fire risk. In a broader context, these applications may contribute to food security and crisis management.

Within this context, monitoring and decision-making remain major bottlenecks. Traditional agricultural monitoring methods, such as the inspection of parts of a field by a farmer or other domain expert, have already been enhanced by new technologies, including field observation using UAV (Unmanned Aerial Vehicle). However, these approaches often still rely on humans to analyse the collected imagery. The general challenges include monitoring scale, reliability and bias, cost, and response delay. Therefore, to tackle these challenges, automated data processing is necessary, at least as a means of supporting human experts. In certain tasks where the consequences of errors are limited, such as weed removal by agricultural robots, decision-making can even be fully automated.

Nevertheless, machine learning is not a silver bullet capable of solving the above problems. On the contrary, it introduces a broad range of new difficulties, often affecting algorithmic systems more than human operators. Multispectral images often provide limited intuitive value when observed directly by humans, as human perception is adapted to the RGB spectrum. However, because multispectral imaging is often superior to RGB imagery, it is particularly well suited for use in conjunction with machine learning techniques. Additionally, since multispectral data sources range from satellite-based Earth observation to autonomous agricultural robots, different types of data introduce distinct challenges. Common challenges, both scientific and technical, include data variability and heterogeneity, differences in image parameters, the complexity of preprocessing and analysis pipelines, and the cost of practical deployment. While some of these challenges are well known and remain active topics of research, others are often addressed only marginally or even omitted entirely. In particular, practical aspects related to deployment and cost often receive limited attention. A review of the literature allowed the identification of the following research gaps:

- limited generalisation of existing models, which often overfit to specific geographic regions or crop types and perform significantly worse in new environments, especially for multispectral data without uniform radiometric protocols [11, 12, 13];
- high cost and complexity of multispectral equipment, which are frequently ignored and limits practical adoption in real-world tasks [14, 7];
- lower maturity of multispectral data processing techniques compared with methods based on RGB imagery, resulting in the underutilisation of multispectral equipment and data [14, 15];
- insufficient consideration of data annotation cost and effort, requiring excessive manual labelling [12, 16];
- neglect of hardware-related constraints, with embedded processing and storage capabilities frequently ignored [12, 13];
- lack of model explainability and uncertainty estimation, reducing trust in automated systems [11, 17];
- limited availability of integrated, openly accessible tools and datasets, especially open-source solutions [11, 13].

Although it is not possible to address all of these gaps, this research focuses on those most closely related to practical applications, especially with regard to cost and hardware constraints. These considerations are especially important for autonomous robotic systems, which operate under limited computational resources and require near real-time responses. As computer vision typically represents the first step in perception and control pipelines, its constraints must be considered from the earliest stages of system design. Similarly, in the context of space technologies, particularly

Earth observation satellites, which constitute around 20% of all satellites [18], embedded hardware capabilities play a critical role. In such cases, preprocessing raw imagery data prior to transmission to Earth may be advantageous, as it can save bandwidth and reduce storage requirements [19].

1.2 Goals of the thesis

Based on the above analysis, the following twofold research hypothesis is proposed:

- ***The use of multispectral imaging combined with modern machine learning techniques will improve predictive performance and model explainability in selected agricultural and environmental protection applications, compared with analogous methods based solely on visible-spectrum imagery or simple vegetation indices.***

- ***Additionally, separation and restriction of the processing to the key spectral bands will allow the algorithms to work at a higher speed without significant deterioration of the quality of the obtained results for selected applications.***

The above hypothesis allows the following principal goals of this thesis to be defined:

- To analyse and improve existing applications of multispectral imaging in agriculture through the application of new machine learning methods.

- To develop new machine learning tools and methods (with a focus on deep learning) for multispectral imaging applications in agriculture using publicly available and custom datasets.

- To improve the performance and reduce the cost of multispectral imaging applications by reducing the amount of data required for acquisition (both spatial and spectral).

1.3 Contribution of the thesis

The main contributions of the conducted research are:

- **Comparison, analysis, and advancement of machine learning methods.** The conducted studies focus particularly on reducing the cost of practical applications and addressing challenges related to efficient data curation for training using multispectral images.
- **Creation and curation of new datasets.** The availability of new datasets enables other researchers to advance their work and facilitates the comparative evaluation of multispectral imaging methods.
- **Experimental validation of the proposed research hypotheses.** Empirical evidence was collected to support the stated hypotheses, and a critical discussion of the limitations of multispectral imaging was also provided.
- **Development of software tools and code repositories.** Open-source tools are essential for the advancement of science. They enable other researchers to reproduce the results and attract scientists from other fields to apply modern machine learning techniques. This has a real societal impact, creating more opportunities for practical applications in agriculture and environmental protection.

1.4 Author activities

1.4.1 List of grants

The following grants were obtained and carried out by the author during the course of the doctoral studies:

- Research project PRELUDIUM 22, funded by the National Science Centre, titled "Development of weak supervision methods for multispectral image processing in the field of precision agriculture". The research focused on utilising weak supervision to reduce manual data-labelling requirements, as well as on the creation of a custom dataset.

- Mobility scholarship within the TERRINet Call 10 project, conducted at the Institut de Robòtica i Informàtica Industrial at the Universitat Politècnica de Catalunya in Spain. The research focused on creating a new method for discovering the spatial topology of camera networks, managing their activities, and tracking people within them using Graph Neural Networks.

1.4.2 List of projects and workshops

The most important projects in which the author participated, and workshops co-organised by the author, are listed below.

- Co-organisation of the workshop titled *Explainable AI in Space* (EASi) at the ECAI 2025 conference in Bologna. As Technical and Scientific Co-Chair, the author participated in the preparation of a dataset for the public challenge, as well as in the assessment of participants' submissions.

- Participation as a researcher in the PUT-ISS project, sent to the International Space Station under the LeopardISS project, one of the experiments selected for Poland's inaugural space mission, the IGNIS mission. The author developed a robotic vision system deployed on a space-qualified embedded computer.

1.4.3 List of publications

Publications authored during the course of doctoral studies relevant to this dissertation:

- Deepness: Deep neural remote sensing plugin for QGIS / **Przemysław Aszkowski**, Bartosz Ptak, Marek Kraft, Dominik Pieczyński, Paweł Drapikowski // SoftwareX - 2023, vol. 23, s. 101495-1 - 101495-6 **[200 pts]**
- Estimation of corn crop damage caused by wildlife in UAV images / **Przemysław Aszkowski**, Marek Kraft, Paweł Drapikowski, Dominik Pieczyński // Precision Agriculture - 2024, vol. 25, s. 2505–1 - 2505–26 **[140 pts]**
- Streamlining Crop Segmentation with Multispectral Imaging and Foundation Models: Minimizing Manual Annotation / **Przemysław Aszkowski**, Marek Kraft // W: IEEE 20th International Conference on Intelligent Computer Communication and Processing (ICCP 2024): IEEE, 2024 - s. 1-8 **[20 pts]**
- Challenges of crop classification from satellite imagery with Eurocrops dataset / **Przemysław Aszkowski**, Marek Kraft // W: Progress in Polish Artificial Intelligence Research 4 / red. Adam Wojciechowski, Piotr Lipiński - Łódź, Polska: Wydawnictwo Politechniki Łódzkiej, 2023 - s. 25-30 **[20 pts]**

Additional publications authored during doctoral studies that are not directly related to the scope of this thesis:

- ISO-compatible personal temperature measurement using visual and thermal images with facial region of interest detection / Bartosz Ptak, **Przemysław**

Aszkowski, Joanna Weissenberg, Marek Kraft, Michał Weissenberg // IEEE Access - 2024, vol. 12, s. 44262-44277 **[100 pts]**

- A fast, lightweight deep learning vision pipeline for autonomous UAV landing support with added robustness / Dominik Pieczyński, Bartosz Ptak, Marek Kraft, Mateusz Piechocki, **Przemysław Aszkowski** // Engineering Applications of Artificial Intelligence - 2024, vol. 131, s. 107864-1-107864-13 **[140 pts]**
- Efficient People Counting in Thermal Images: The Benchmark of Resource-Constrained Hardware / Mateusz Piechocki, Marek Kraft, Tomasz Pajchrowski, **Przemysław Aszkowski**, Dominik Pieczyński // IEEE Access - 2022, vol. 10, s. 124835 - 124847 **[100 pts]**
- Thermo Presence: The Low-resolution Thermal Image Dataset and Occupancy Detection UsingEdge Devices / **Przemysław Aszkowski**, Mateusz Piechocki // W: Proceedings of the 3rd Polish Conference on Artificial Intelligence PP-RAI'2022, April 25-27, 2022, Gdynia, Poland - Gdynia, Polska: Uniwersytet Morski w Gdyni, 2022 - s. 49-52 **[20 pts]**

1.5 Thesis structure

Chapter 2 describes different types of multispectral vision systems and their applications in agriculture and environmental protection, followed by an overview of tools helpful in this process. The chapter continues with a detailed discussion of the challenges associated with multispectral imaging in practical scenarios.

Chapter 3 focuses on machine learning methods for processing multispectral data, from historical approaches to modern deep learning techniques, together with a description of the evaluation metrics used throughout the thesis. Different strategies for model development are presented, with a focus on the amount of manual annotation required. Next, the topics of band explainability and spectral band reduction are discussed, highlighting their role in enabling practical applications. Finally, hardware constraints are examined, and their relationship to the previously discussed topics is outlined.

Chapter 4 presents the quantitative results and improvements achieved by the proposed approaches, as well as a description of the implemented applications and open-source tools. It begins with an approach showing how multispectral data can be used in the training process of a model for an agricultural robot. Next, the results of research on satellite data are presented, with a focus on explainability. The chapter continues with example application of multispectral data and a description of the open-source plugin for QGIS created as part of the research. The chapter concludes with an analysis of the spectral band reduction results.

Chapter 5 provides a summary of the research and implementation results presented in the thesis. It also discusses the limitations of multispectral imaging, acknowledging its constraints and weaknesses. Finally, a list of suggestions for further work is presented.

2 Computer vision systems for agriculture and environmental protection applications

Determining an appropriate multispectral system for a given problem is crucial but requires an understanding of its possibilities and limitations. This includes selecting suitable cameras and platforms while considering their technical parameters, as well as identifying the type of task to be addressed through automated processing. This chapter addresses these aspects, with particular attention given to the practical application of multispectral systems, especially in relation to hardware costs.

2.1 Spectral sensing principles

2.1.1 Imaging modalities

Spectral imaging refers to the acquisition of image data in multiple bands across the electromagnetic spectrum. The most widespread conventional camera captures light in three wavelength ranges within the spectrum visible to the human eye, and is commonly referred to as an RGB camera, as the captured bands correspond to the red, green, and blue regions of the visible spectrum. However, spectral imaging may also involve non-visible wavelengths, such as ultraviolet, infrared, or X-ray radiation. In agricultural and environmental contexts, the utilised spectrum usually remains within the 300-2500 nm range, as presented in fig. 2.1.

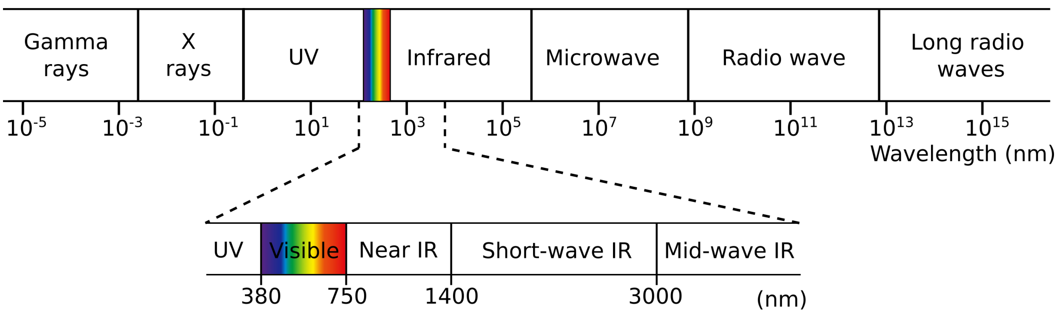

Figure 2.1: Spectral ranges of imaging modalities within the electromagnetic spectrum, highlighting the regions typically used in agricultural and environmental applications (approximately 300-2500 nm).

Multispectral data refers to imagery acquired in multiple discrete spectral bands, typically ranging from 4 to 15 bands. Among these bands, those referred to as red, green, and blue are often included. However, they are not equivalent to the bands of a conventional RGB camera. Instead of spanning hundreds of nanometres of the spectrum and overlapping each other, bands in multispectral sensors are narrower and usually well separated.

Hyperspectral imaging acquires data in a much larger number of bands, up to hundreds. These bands are typically narrow and contiguous across the spectrum. This high spectral resolution enables detailed characterisation of material spectral signatures.

Figure 2.2 presents the conceptual differences between these types of imaging. Both multispectral and hyperspectral imaging provide richer spectral information than RGB imagery, but this comes at the cost of higher data volumes and greater equipment complexity, resulting in the need for an informed trade-off between information and cost. Moreover, the higher capabilities of such cameras bring more complex calibration procedures and increased sensitivity to noise and illumination variability. Consequently, despite the conceptual distinction between multispectral and hyperspectral technologies, their applications often overlap or even complement each other. In research, especially within review articles, the term multispectral imaging sometimes includes hyperspectral imaging, and vice versa. Nevertheless, for many real-world applications in small-scale agriculture, hyperspectral imaging systems remain impractical because of their high cost.

A single-band camera, commonly referred to as a monochrome camera, captures data in only one spectral band, which is typically achieved using additional optical filters fitted to measure radiation within a specific wavelength range. Monochrome cameras are typically less expensive than multispectral cameras, offering greater flexibility in sensor configuration, as the wavelength bandwidth may be targeted at specific physical measurements. By combining multiple monochrome cameras into a single hardware system, it is possible to achieve results similar to those of a single multispectral camera. As equipment manufacturers rarely differentiate between these approaches or provide clear information about them, this thesis treats both solutions equivalently. Nevertheless, there may be practical differences with respect to spectral band alignment, geometric calibration, and time synchronisation.

2.1.2 Camera system resolution

Typical parameters characterising a camera system differ depending on the camera type and the broader system in which it operates. A hyperspectral pushbroom

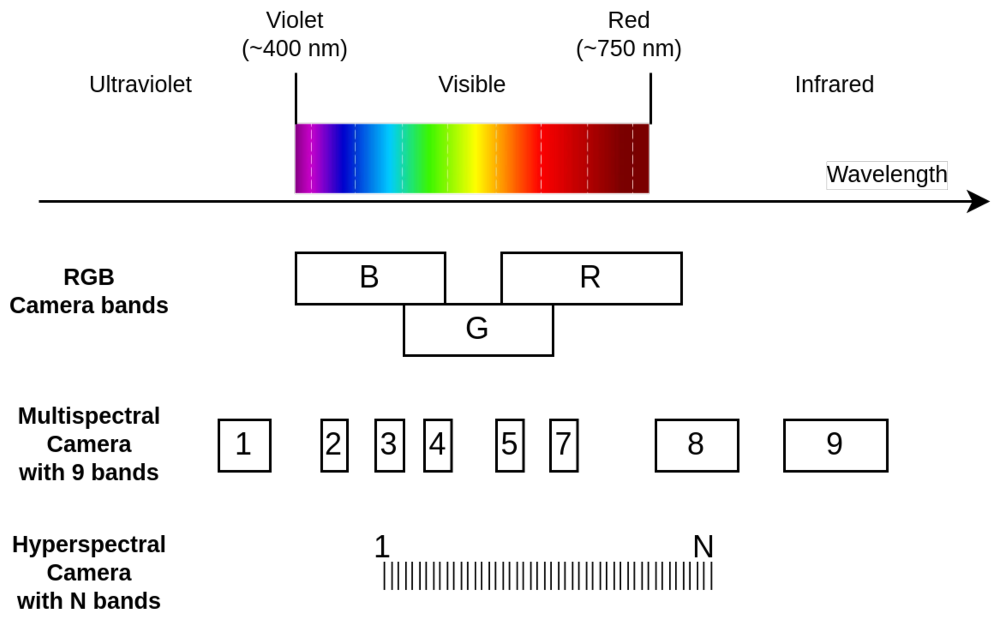

Figure 2.2: Comparison of spectral sampling strategies in RGB, multispectral, and hyperspectral imaging, illustrating differences in the number, width, and continuity of spectral bands across the electromagnetic spectrum.

camera mounted on a satellite cannot be described using the same parameters as a multispectral camera mounted on an agricultural robot. Nevertheless, several parameters can be generalised across different imaging systems.

Spectral resolution describes the ability to distinguish between different wavelengths of electromagnetic radiation. It is determined by the spectral bandwidth of each band and the spectral separation between bands. For hyperspectral images, it may be represented by a simple mathematical measure, whereas for multispectral images, it may require each band to be defined individually.

Spatial resolution, also referred to as camera resolution, describes the system's ability to distinguish fine spatial details. In remote sensing, spatial resolution is commonly expressed as the GSD (Ground Sampling Distance), which is determined by the imaging distance, sensor resolution, and optical parameters such as focal length and sensor size. For camera systems without a fixed distance to the observed object, spatial resolution may instead be described simply by the pixel resolution of the acquired images.

Temporal resolution defines how frequently the same area is observed. It is usually used in the context of satellite systems, where temporal resolution is expressed as revisit time, typically measured in days. For a UAV-based system, temporal resolution may correspond to the number of days between consecutive flights. In the case of a stationary camera observing a single object, it may instead be understood as the acquisition frequency, expressed in FPS (Frames Per Second).

Table 2.1 presents a comparison of different camera types using exemplary models.

Table 2.1: Comparison of representative imaging systems used in UAV remote sensing. Specifications are approximate and may vary with configuration.

	RunCam Thumb Pro	**MicaSense RedEdge-P Dual**	**HAIP BlackBird V2**	**Cubert ULTRIS 5**
Modality	RGB	Multispectral	Hyperspectral (pushbroom)	Hyperspectral (snapshot)
Spatial resolution	3840×2160	1456×1088	540×540	290×275
Acquisition rate	30–120 fps	up to 3 fps	0.33 cube/s; 180 line/s	15 fps
Number of spectral bands	3	10	100	51
Field of view	150°	38–50°	33–37°	15°
Spectral coverage	Visible (RGB)	444–842 nm (10 discrete bands)	500–1000 nm (5 nm step)	450–850 nm
Dimensions	54×26×21 mm	132×88×97 mm	80x60x90 mm	29x29x56 mm
Weight	16 g	745 g	790 g	126 g
Approx. price	∼100 USD	∼14,000 USD	∼33,000 USD	∼15,000 USD

2.1.3 Thermal and radar sensing

Multispectral imaging is traditionally associated with optical measurements of reflected solar radiation, whereas the operating principles of thermal and radar sensors differ fundamentally. Thermal imaging measures radiation emitted by the observed object itself, typically in the 8-14 μm wavelength range, which allows surface temperature to be estimated. The resulting data can still be interpreted as a single-band image. Due to this distinction, some authors explicitly state that they use multispectral imagery comprising visible and thermal infrared data, or even list them separately, for example "[...] thermal-and multispectral-imaging" [20]. Nevertheless, in a broader, application-oriented sense, thermal data are often included under the multispectral umbrella [21].

On the other hand, radar sensing differs more substantially from optical imaging and is generally not considered multispectral data. Radar systems, such as those used in the Sentinel-1 satellite, do not rely on solar illumination, but acquire data using C-band SAR (Synthetic Aperture Radar), which provides multi-polarisation backscatter measurements. These capture different scattering mechanisms related to surface roughness, structure, and moisture content at a wavelength of approximately 5.6 cm [22].

2.1.4 Physical meaning of spectral bands

A key advantage of multispectral imaging compared with RGB imaging lies in the physical interpretability of individual spectral bands. As different wavelengths interact in distinct ways with specific elements of the environment, they enable the extraction of meaningful and physically grounded information.

For example, SWIR (Short-Wave Infrared) bands are useful for estimating the moisture content of vegetation and soil [23], while bands near 443 nm are useful for aerosol optical thickness estimation [24]. Similarly, plant health is strongly related to chlorophyll content, which strongly absorbs red and blue light, while reflectance in the green band remains higher [25]. Healthy vegetation also shows higher reflectance in the NIR (Near-Infrared) band due to internal leaf structure. Figure 2.3 shows simplified illustrative reflectance curves of healthy and stressed plants, and how the differences can be observed within bands at specific wavelengths. This behaviour enables multispectral sensors to capture physiological changes that are invisible to the human eye and cannot be reliably detected using RGB imagery alone.

A direct comparison of pixel values in individual spectral bands may be misleading, as measured reflectance varies over time and depends strongly on illumination conditions, making repeatable measurements in open environments challenging. This limitation can be addressed by combining information from multiple bands, for example through the use of vegetation indices [26].

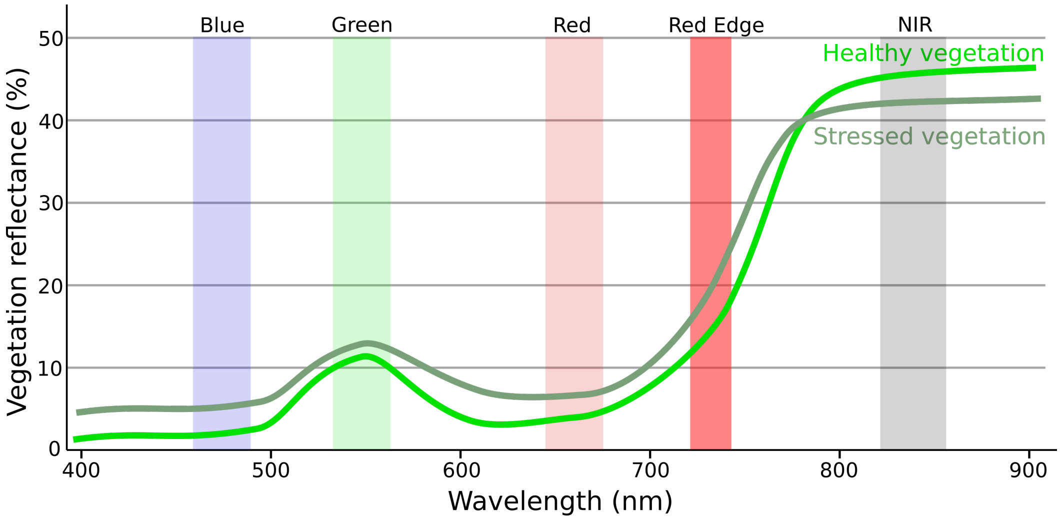

Figure 2.3: Illustrative spectral reflectance curves of healthy and stressed vegetation, showing characteristic differences across wavelengths, including increased reflectance in the red region and decreased reflectance in the NIR region for stressed plants.

2.2 Multispectral camera architectures

Literature [27, 28, 29] commonly classifies multispectral and hyperspectral cameras based on the method used to acquire the spectral data cube. The most widely used architectures include whiskbroom, pushbroom, wavelength-scanning, and snapshot systems.

Point-scan (whiskbroom) cameras acquire spectral information sequentially for individual pixels and require precise mechanical scanning, involving controlled movement of either the sample or the detector. This method enables high spectral resolution and signal quality. However, the acquisition process is relatively slow and sensitive to movement, making it difficult to apply outside controlled environments.

Line-scan (pushbroom) cameras capture a single image line across multiple spectral bands simultaneously, typically projected onto a 2D sensor array, offering a good compromise between acquisition speed and spectral and spatial resolution. Movement of either the sample or the detector is still required, but only in one direction, with accurate knowledge of the motion necessary for reconstruction of the data cube.

Plane-scan (wavelength-scanning) cameras capture a full spatial image at once while sequentially iterating across wavelengths, usually employing tunable optical filters. Such a design allows accurate spatial registration within individual bands. However, it is sensitive to motion and illumination changes occurring during the capture of consecutive wavelengths, which may introduce spectral misalignment artefacts.

Single-shot (snapshot) cameras capture all spectral bands simultaneously over the full spatial field of view. Their primary advantage is the ability to capture dynamic scenes without distortions caused by motion, making them attractive for robotic applications in less controlled environments. However, this often comes at the expense of spatial or spectral resolution, as well as a lower signal-to-noise ratio due to optical constraints. Systems composed of multiple independent cameras, each dedicated to a particular spectral band, may also be classified as snapshot systems.

In addition to these standard designs, the literature [30] describes alternative hybrid architectures, which offer new advantages and disadvantages compared with the typical methods listed above. Nevertheless, the consequences of the capture method are often similar, such as band misalignment in moving systems, challenges in maintaining proper image geometry, or noise in measured reflectance.

As a result, different camera types are inherently better suited to specific tasks. For example, a line-scan system is a natural choice for a platform moving at a constant velocity over the target, such as an Earth observation satellite. On the other hand, a snapshot camera would be preferable for an agricultural robot with an irregular movement pattern. Another important factor in camera selection is its weight, physical dimensions, and compatibility with other equipment, particularly the connectors used to trigger data capture from an onboard UAV computer.

2.3 Data acquisition platforms and sources

Multispectral data used in agricultural and environmental protection applications can be acquired using distinct types of platform, as reported in both the literature and industrial applications [31, 32, 33, 7]. Each platform type is best suited to specific applications, while simultaneously introducing unique challenges related to both data acquisition and subsequent processing. This section categorises the main acquisition platforms into satellite-based, airborne, UAV-based, and ground-based robotic systems. Figure 2.4 presents exemplary data acquired from these platforms.

2.3.1 Satellite-based platforms

Satellite platforms provide the most widely used source of large-scale multispectral data for agricultural and environmental monitoring, operating at regional

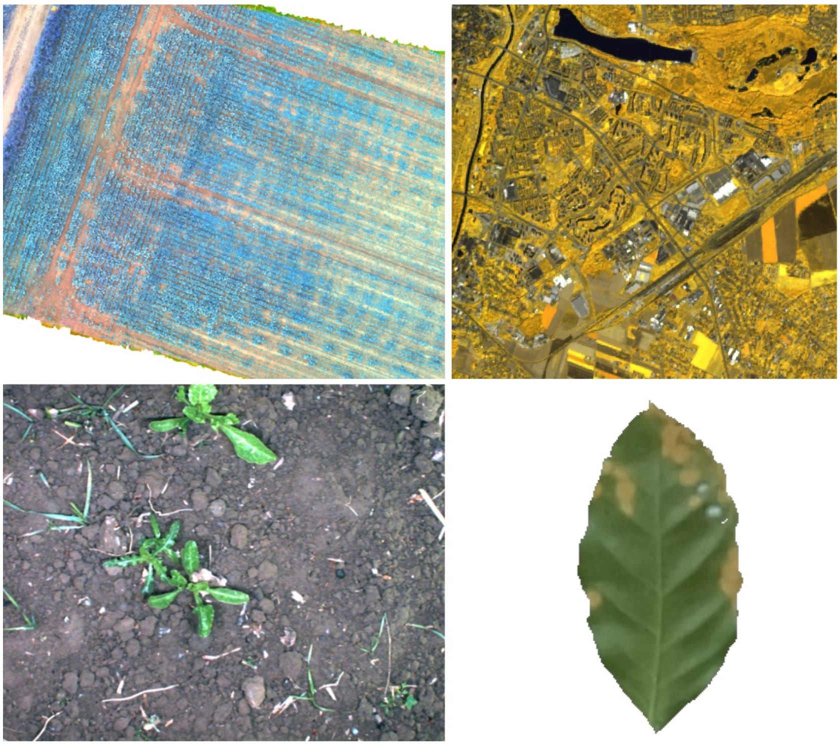

Figure 2.4: Examples of multispectral data sources: from a UAV, Sentinel-2 satellite, a ground-based robotic system, and individual objects, illustrating differences in spatial resolution, coverage, and perspective. Visualisation using false colours.

or even global scales. The captured images are typically characterised by relatively low spatial resolution, commonly in the range of 10-50 metres per pixel for global Earth observation satellites. As a consequence of this resolution, a single pixel may contain a mixture of reflectance from heterogeneous sources, such as roads, soil, and vegetation within the same area. Despite the coarse spatial resolution, the volume of generated data is on the order of terabytes per day. The temporal resolution of satellite imagery ranges from several days to weeks, depending on the number of satellites in a mission and their configuration. One of the major limitations of multispectral satellite platforms is their dependence on atmospheric conditions, as cloud cover often prevents the acquisition of reliable data. Table 2.2 provides a comparison of popular multispectral and hyperspectral satellite missions with public data access for research purposes.

2.3.2 Airborne platforms

Airborne platforms primarily consist of manned aerial vehicles, typically small fixed-wing aircraft. Typical flight altitudes range from 1 to 10 km, enabling monitoring on a regional scale and covering areas of up to thousands of square kilometres per mission, with a GSD typically in the range of 0.5-5 m per pixel. Airborne platforms

Table 2.2: Comparison of selected satellite missions providing publicly available multispectral and hyperspectral data.

Name	Spatial resolution	Temporal resolution	Number of bands	Wavelength range	Country
Sentinel-2	10–60 m	5–10 days	13	∼442–2202 nm	European Union
Landsat 8–9	30 m	8–16 days	11	∼430–2290 nm	USA
MODIS	250–1000 m	1–2 days	36	∼405–14385 nm	USA
PRISMA	30 m	7–29 days	∼250	∼400–2500 nm	Italy
EnMAP	30 m	4–27 days	∼230	∼420–2450 nm	Germany

allow for flexible sensor payload configurations, but the operational and logistical costs are relatively high.

2.3.3 UAV platforms

Although UAV platforms are sometimes classified as a subset of airborne systems, in this thesis UAVs are understood as relatively small flying vehicles, typically multirotors, with a maximum mass of up to 25 kg and an operational altitude of 120 m, operated within visual line of sight according to regulations issued by the European Aviation Safety Agency. UAV-based multispectral imaging enables the acquisition of very high spatial resolution data, typically ranging from 0.2 to 10 cm per pixel, making these platforms particularly suitable for precision agriculture and plot-level analysis. The minimum safe flight altitude is often constrained by rotor downwash, especially when monitoring tall or delicate crops, as the induced airflow causes plant movement. For example, during rapeseed crop monitoring conducted as part of this research, the minimum practical altitude was approximately 13 m for a small UAV (DJI Mavic 3M) and exceeded 20 m for a larger platform (DJI Matrice 600 Pro). Figure 2.5 shows photographs of both UAVs equipped with multispectral cameras that were used to create new datasets during this research.

The survey is usually performed using mission planning software, which generates a trajectory enabling the capture of individual images on a regular grid, allowing the mapping of a surface area with a desired GSD and enabling subsequent orthophoto reconstruction. Therefore, cameras are most often oriented in a nadir-

Figure 2.5: Examples of UAV platforms used for multispectral data acquisition: a DJI Matrice 600 Pro equipped with an external MicaSense multispectral camera (left), and a DJI Mavic 3 Multispectral with an integrated multispectral sensor (right).

looking configuration. Mission parameters must be adapted to the characteristics of the surveyed area, especially the availability of distinct image features. These include flight speed, front and side overlap, and image acquisition frequency. Figure 2.6 presents an example of a mission planned and carried out with DJI equipment.

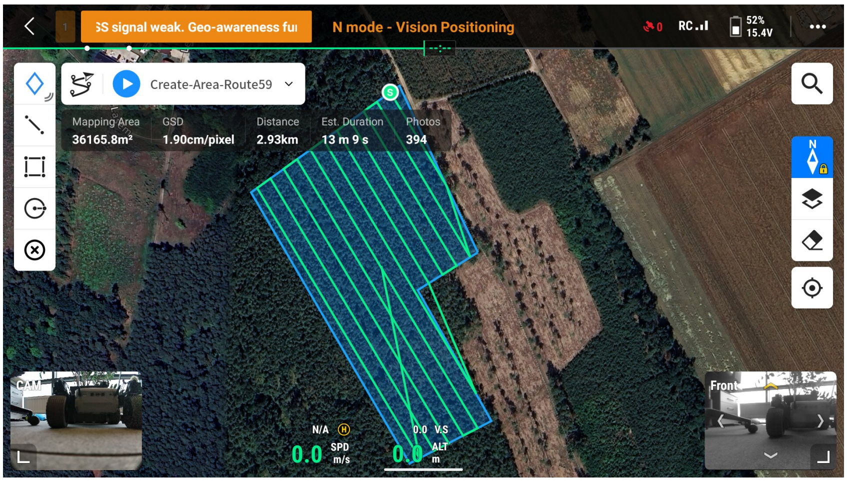

Figure 2.6: Example of a UAV mapping mission planned in DJI software, illustrating flight path planning and its parameters.

As flights are performed at low altitude, positioning accuracy becomes important, necessitating the use of onboard RTK GNSS systems, which allow centimetre-level georeferencing accuracy to be achieved. Key limitations of UAV platforms include high susceptibility to wind and precipitation, as well as short flight durations, typically limited to around 30 minutes by battery capacity. The adoption rate of UAV platforms has been further constrained by new regulations enforced by regional governments, particularly in recent years, partly due to the dual-use nature of UAV technology. While UAV operation remains relatively accessible for agricultural applications, it is often significantly more restricted in urban environments and in the vicinity of airports or military areas.

2.3.4 Ground-based robotic platforms

Ground-based robotic systems, also referred to as field robots, are capable not only of monitoring fields and capturing image data, but also of processing these data onboard in real time in order to perform targeted actions, such as selective weed removal [34]. This requires a near real-time processing pipeline, as perception and decision-making need to occur immediately before actuation. Since data are typically captured and processed as individual images, the camera may be oriented in any direction, depending on the task. In addition to spectral imagery, depth or other three-dimensional perception sensors are frequently integrated, as they provide critical distance information required for precise actuation.

Field robots usually operate at the scale of individual fields and, as they operate near ground level, they enable the acquisition of imagery with high spatial resolution, often below 1 mm per pixel. This level of detail allows the observation of individual plant structures, which is particularly beneficial in applications requiring high spatial precision, such as the identification of small insects or the monitoring of rice leaf area, as demonstrated in [35].

Despite their technical potential, agricultural robots are still predominantly confined to research and pilot deployments, largely due to high system and operational costs. Nevertheless, they remain an active area of research, with a broad spectrum of robotic platforms, as presented in [36]. Some robots, such as BoniRob, have been used across multiple applications, including plant classification [37]. Figure 2.7 shows examples of field robots used in agricultural research.

Figure 2.7: Examples of ground-based robotic platforms for agricultural applications. Left: BoniRob used for weed control (image from [38]); right: Harvest CROO robot with manipulators to pick strwaberries (image from harvestcroorobotics.com website).

2.4 Preprocessing and calibration

Calibration and preprocessing are crucial steps in ensuring the reliability of acquired multispectral imagery, even in carefully controlled environments [28].

Preprocessing strongly influences all downstream processing stages, including the performance of machine learning models, as errors introduced at early stages tend to propagate throughout the entire pipeline. This section focuses on post-acquisition preprocessing, that is, the operations applied to data already output by the camera system. While some preprocessing steps may be performed internally by modern cameras, the boundaries of responsibility are not fixed. The most important preprocessing steps are briefly described below.

Radiometric preprocessing corrects sensor-related artefacts that affect pixel intensities. It typically includes dark current correction, sensor non-uniformity correction, and vignetting correction.

Reflectance calibration converts raw sensor measurements into physically meaningful reflectance values that are comparable across space and time. For satellite imagery, this typically involves atmospheric correction, whereas for UAV-based systems, CRP (Calibrated Reflectance Panel) with known spectral properties is often used.

Geometric preprocessing ensures spatial consistency across images and spectral bands. It typically starts with lens distortion correction using camera intrinsic parameters. Band co-registration aligns individual spectral bands, which are often spatially misaligned due to camera optical design and the lack of time synchronisation between sensors.

Mosaicking and orthophoto reconstruction combine individual georeferenced images into a single spatially consistent product. For UAV data, this usually involves feature detection and matching, internal 3D point cloud representation, and texture mapping.

Importantly, preprocessing steps differ across platforms. Selected steps and products in processing pipelines for Sentinel-2 and UAV DJI Mavic 3M data are described below.

2.4.1 Sentinel-2 products

Sentinel-2 data are distributed in several standardised products at different processing levels [39, 40].

Level-1B products provide radiometrically corrected measurements, but still in sensor geometry, tied to the acquisition geometry.

Level-1C products provide TOA (Top-of-Atmosphere) reflectance, already after radiometric and geometric correction.

Level-2A products provide BOA (Bottom-of-Atmosphere) surface reflectance, further processed with atmospheric correction and scene classification, including the detection of clouds and cloud shadows. This product level is generally preferred for most agricultural and environmental protection tasks. The Sen2Cor processor is the official atmospheric correction tool used to convert Level-1C products into Level-2A products, while also generating auxiliary scene classification layers.

2.4.2 UAV multispectral pipeline (DJI Mavic 3M)

The DJI Mavic 3M is equipped with an RGB camera, a dedicated sensor for four multispectral bands, and an additional upward-facing sunlight sensor on the top of the aircraft. This sensor measures incident illumination during flight, providing metadata that can later be used to compensate for short-term changes in lighting conditions during the flight. Despite the availability of this sensor, the use of a CRP is still recommended. In a typical workflow, an image of the panel is taken before or after the flight, enabling the conversion of pixel values to absolute reflectance. Geometric calibration is also required, as the individual spectral bands are significantly misaligned. Figure 2.8 shows an example false-colour image acquired by the Mavic 3M after reflectance calibration but without spectral band co-registration. Band co-registration is therefore necessary to enable accurate pixel-wise analysis. Even with a proper algorithm, artefacts may still persist. Moreover, additional, although similar, processing is required to jointly use multispectral and RGB imagery, as RGB images are captured using a different optical geometry and sensor configuration.

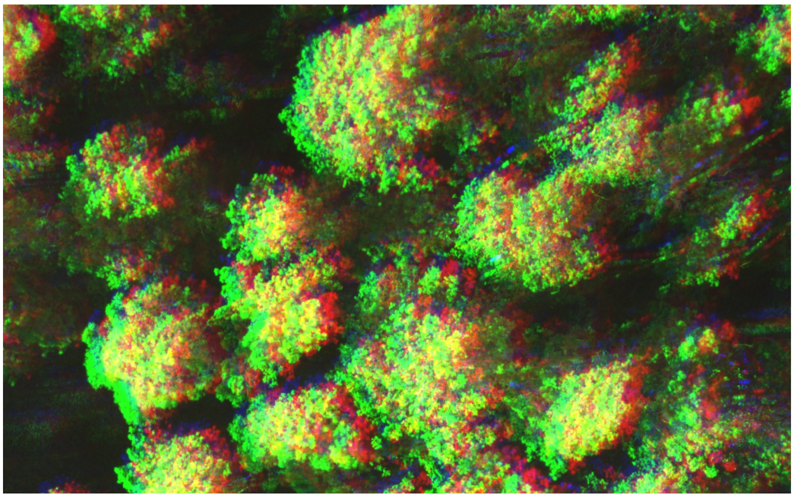

Figure 2.8: False-colour composite illustrating spatial misalignment among multispectral bands (NIR, Red Edge, and Red) acquired by the DJI Mavic 3 Multispectral prior to band co-registration, captured over pine trees infested with mistletoe.

Figure 2.9 presents an overview of the preprocessing pipeline from raw UAV data to processing-ready individual images, without orthophoto mosaicking. When building orthophotos, some tools integrate parts of this processing internally. For individual image preprocessing, open-source scripts have been developed, as described in the availability section.

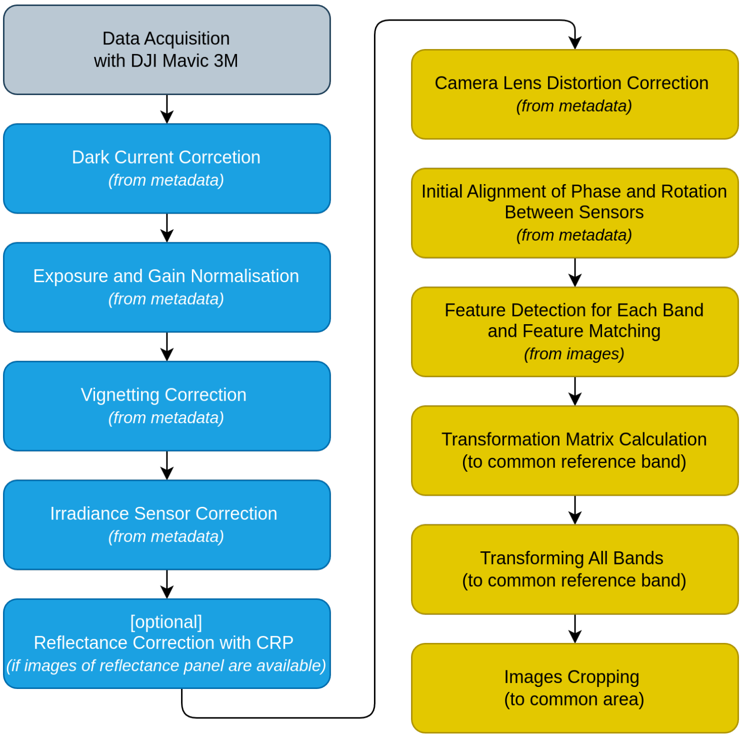

Figure 2.9: Overview of the preprocessing workflow for multispectral imagery acquired with the DJI Mavic 3M.

2.5 Data visualisation and analysis tools

This section describes approaches and tools used to visualise and analyse multispectral and remote sensing data by humans, without employing dedicated machine learning pipelines. Such tools play an important role in exploratory and initial analysis, annotation, and visualisation prior to decision-making.

Image-level and orthophoto-based visualisation. Individual multispectral images can be visualised in a similar manner to RGB images using false-colour representations, where multispectral bands are mapped to RGB channels. Consequently, the number of bands visualised at once is limited to three. In Earth observation tasks, the most commonly used representation is a georeferenced orthophoto [41], stitched from multiple overlapping images. However, this representation is inherently lossy, as individual images from multiple angles are reduced to a single orthogonal view. This limitation was also observed during this research, where individual mistletoe plants growing on pine trees were visible only in oblique views in individual images, but not in the orthophoto representation.

GIS (Geographic Information System). GIS software is particularly well suited to handling georeferenced raster data, as it manages underlying spatial reference systems and coordinate transformations while allowing the overlay of multiple data sources, including vector data. Widely used tools, including the open-source QGIS and the commercially available ArcGIS, provide a rich set of generic analytical operations, such as raster algebra, band composition, or index calculation. For integration with machine learning methods, GIS software usually needs to be extended through programmatic access, frequently available in Python, or through custom processing pipelines using libraries such as *GeoPandas*, *GDAL* or *Rasterio*, facilitating access to geospatial data.

Photogrammetry software. The generation of orthophotos from aerial and UAV imagery typically relies on dedicated photogrammetry software. Well-established software packages include the open-source WebODM as well as the commercially available Pix4D and DJI Terra. These tools often implement a complete processing pipeline, including reflectance calibration, band alignment, and point cloud generation [42], although support for RGB imagery is usually superior to multispectral capabilities.

Alternative data representations. Beyond 2D raster imagery, alternative representation methods include 3D point clouds or textured models. While these representations preserve more information than orthophotos, they are less widespread in Earth observation due to their greater complexity. Moreover, 3D representations tend to be more useful for large, rigid objects such as buildings or terrain features, whereas their advantages are limited for small, deformable objects such as plants. Another popular composite processing format is an elevation or height map, which can be visualised as a greyscale raster. In agricultural contexts, it may be useful for particular tasks, such as estimating crop height.

2.6 Machine learning tasks

From an agricultural perspective, farming activities can be broadly grouped into pre-harvesting, harvesting, and post-harvesting stages [43]. Each stage involves a wide range of tasks related to crops, soil, and the surrounding environment. From a machine learning perspective, tasks can be grouped according to the relationship between the input data and the expected output format. In the context of agricultural and environmental applications, the most frequently reported tasks in the literature [44, 43, 45] include detection, segmentation, classification, and regression. These tasks differ not only in their output representations, but also in computational complexity, annotation requirements, and suitability for different data acquisition platforms. Figure 2.10 presents exemplary results for these tasks.

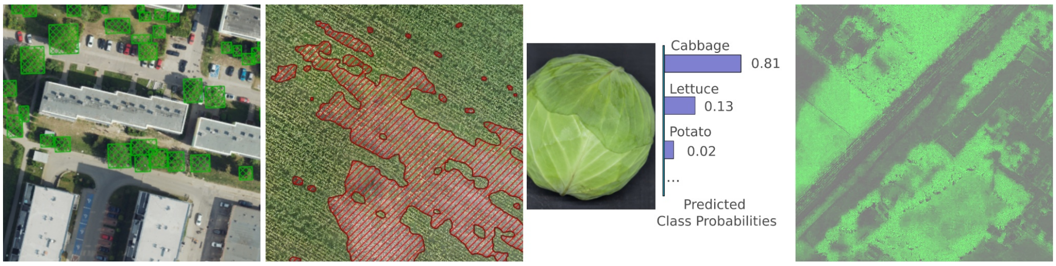

Figure 2.10: Illustrative examples of common machine learning tasks in agricultural and environmental applications: detection (of trees), segmentation (of field crop damage), classification (of plants), and regression (biomass estimation).

Detection. Object detection aims to localise individual objects by predicting bounding boxes and assigning them to predefined categories. This task is widely used when individual entities are spatially separated from one another and approximate localisation is sufficient. Typical applications include the detection of individual plants or fruits, plant diseases, or pests. In a broader sense, detection can also refer to change or anomaly detection, where the objective is not to identify a discrete physical object, but rather an event or effect, such as flooding or sudden vegetation loss in a given area.

Segmentation. Semantic segmentation assigns to each pixel in the image a probability of belonging to a predefined class. Instance segmentation further extends this concept by grouping pixels belonging to the same object, producing a separate mask for each individual instance identified in the image. Instance segmentation therefore provides richer information than object detection alone. Typical applications include identifying the boundaries of individual plants and land cover mapping into classes such as field, water, and road.

Classification. Classification aims to assign a single label to an image. Multi-class classification sorts data into three or more classes, whereas binary classification distinguishes between only two mutually exclusive categories, usually the presence or absence of a given condition. Typical applications include sorting individual crops based on their quality or detecting environmental hazards, such as river contamination or methane leaks.

Regression. Regression predicts a single continuous value of a target variable, either for an entire image or for each pixel, allowing quantitative estimation of physical variables. Typical applications include estimation of crop yield across a field, soil moisture and nutrient content, or the assessment of fire risk.

Some of the above tasks may be used interchangeably depending on the application. For example, plant disease monitoring can be achieved either through segmentation of infected regions or detection of individual diseased plants. Similarly, regression problems can be reformulated as classification tasks by discretising continuous variables into a limited number of classes.

In addition to the primary tasks described above, supportive tasks are frequently employed to enhance data quality, either to improve visual interpretability or to serve as inputs for further processing. Notable examples include super-resolution and pansharpening. Super-resolution aims to increase spatial resolution without acquiring additional data. For instance, the authors in [46] performed super-resolution of Sentinel-2 imagery using a transformer-based deep learning model, achieving a resolution of 5 m for the upsampled images. Pansharpening combines high-spatial-resolution panchromatic or RGB images with simultaneously acquired lower-resolution multispectral data. The authors in [47] proposed a Laplacian pyramid pansharpening network with a resolution ratio of 4.

2.7 Practical challenges and limitations

The use of multispectral imaging in agricultural and environmental protection contexts involves a significant number of challenges and limitations, as reported by the authors in [11, 48, 49]. These challenges affect both the practical feasibility of monitoring systems and downstream data processing techniques, including machine learning methods. The most important limitations are discussed in the subsections below.

2.7.1 Real-world variability and data distribution

Agricultural and environmental monitoring systems operate in highly variable and largely uncontrolled conditions, which strongly affect the acquired multispectral data. This variability arises from both environmental factors and biological processes influencing plant development.

Environmental conditions can change substantially within hours, sometimes even during a single mission, such as mapping an individual field using a UAV. First, variations in solar irradiance caused by changes in sun angle, cloud cover, and shading lead to non-uniform illumination conditions across acquired images. This is especially important for multispectral data, as the measured pixel values must be converted into actual surface reflectance, as described in section 2.4. Furthermore, weather conditions such as haze, clouds, and aerosol concentration may vary significantly between consecutive acquisitions, especially in the case of satellite imagery.

Crop, soil, and landscape variability represents an inherent source of complexity. Agricultural crops exhibit high variability even within a single species due to differences in cultivation conditions, growth stage, fertilisation, and treatment. For example, a field infested with weeds looks different from a field treated with herbicides to eradicate everything except the crops. Such high intra-class variability may result in different spectral responses for visually similar plants. Soil variability also plays an important role, not only by affecting plant growth but also by altering the background reflectance, especially during early growth stages or in areas with sparse vegetation. In addition, geographical variability constitutes an additional source of complexity, as agricultural fields from different countries may differ substantially in climate, terrain, and farming practices. Even within relatively small geographic regions, local variations in temperature and weather patterns can significantly influence crop appearance and spectral characteristics.

As a consequence, all of the above factors contribute to the emergence of a distribution shift between training and deployment data. A model trained on data from one field may perform poorly when applied to another, which was particularly observed in this research for crop classification using Sentinel-2 imagery. This highlights the need for model robustness, adaptability to varying environments, and good generalisation. It is worth noting that some of these challenges may be mitigated in ground-based agricultural robots, which can operate under more controlled conditions, for example through the use of artificial illumination.

2.7.2 Camera and monitoring system trade-offs

The selection of camera systems and monitoring platforms involves multiple trade-offs that directly influence the characteristics of the acquired data, the complexity of the processing pipeline, and the overall system cost.

First, spatial resolution, expressed by GSD, must be selected appropriately for a given task. High spatial resolution enables detailed monitoring of individual plants, but reduces coverage area and increases data volume. In UAV-based mapping, higher resolution also results in longer mapping times. Therefore, for some applications, it may be impractical to monitor the entire area, making a sampling strategy with selected regions of interest a more viable solution. Next, it may not always be possible to provide the temporal resolution required to capture crop growth dynamics or rapid stress responses. In some cases, temporal resolution is limited by the fixed revisit time of existing satellites, while in others, such as UAV-based monitoring, it can be increased at the cost of additional flights.

Spectral resolution also introduces important trade-offs. Cameras with a higher number of bands provide richer spectral information, but this does not always translate into improved performance for a given task. This is particularly evident in hyperspectral data, where adjacent bands often exhibit high redundancy. Similarly, in multispectral imaging, not all bands are equally informative, and their relevance depends on the underlying physical properties of the monitored objects. This highlights the need to determine the number of bands required for a specific application. Finally, increased spatial and spectral resolution directly affects the size and cost of a given platform, imposing economic and operational constraints on practical deployments.

2.7.3 Data processing and analysis

The processing of multispectral data typically involves multi-stage pipelines, which introduce several methodological and practical challenges. Even prior to the actual analysis, camera calibration procedures are required, the complexity of which depends on the specific sensor. Any errors introduced during preprocessing may propagate through subsequent processing stages and negatively affect the final results.

Acquired data are often heterogeneous. For example, in Sentinel-2 imagery, different bands are provided at different spatial resolutions. In addition, the presence of clouds can obscure the monitored area in satellite data, effectively reducing temporal resolution and requiring algorithms capable of handling missing observations. Moreover, the integration of multiple data sources, including not only multispectral imagery but also auxiliary modalities such as weather data or soil moisture measurements, may further improve analysis results, while requiring more complex processing algorithms.

Raw multispectral data also pose significant challenges in terms of storage, transmission, and computational requirements. These challenges can be alleviated through the use of cloud-based processing platforms and data hubs, such as Google Earth Engine or NASA Earth Exchange.

From a machine learning perspective, many applications require vast amounts of high-quality labelled data to train models, which in turn requires annotations from domain experts. This issue is frequently compounded by class imbalance [50]. For instance, datasets may contain significantly fewer samples of diseased plants than healthy ones, with disease occurrences limited to specific locations and environmental conditions.

Finally, high predictive accuracy alone is insufficient for practical deployment. Every application requires end-user trust, creating the need for model explainability and the assessment of prediction uncertainty.

2.8 Summary and outlook

This chapter examined the role of computer vision in agricultural and environmental protection applications, with emphasis on sensing modalities, data characteristics, and the range of tasks considered in this work, including classification, segmentation, and regression. The diversity of data sources, ranging from satellite platforms to agricultural robots, directly determines the types of real-world problems that can be effectively addressed, as each platform imposes specific spatial and temporal constraints. Particular attention was given to environmental variability and the resulting requirements for model generalisation, highlighting data availability and quality as important factors limiting the development of machine learning models. Additionally, practical deployment considerations, such as hardware performance and system cost, must be taken into account.

In the following chapters, these aspects will guide the selection of appropriate machine learning techniques for specific tasks, with particular emphasis on validation strategies in the context of geospatial data.

3 Machine learning techniques for processing multispectral data

Multispectral imaging constitutes an important modality for agricultural and environmental protection applications, with inherent advantages over standard RGB imagery, as it provides access to the physical and biochemical properties of plants, soil, and land cover. While this additional spectral information enables more accurate monitoring and analysis, it also introduces significant challenges related to data heterogeneity, high dimensionality, and the limited availability of high-quality annotated ground truth data. As a result, machine learning techniques are essential for extracting meaningful patterns across spectral, spatial, and temporal dimensions at scale.

This chapter presents an overview of machine learning methods used for processing multispectral imagery in agricultural and environmental contexts. It begins with a review of the historical evolution from pixel-wise and feature-engineering approaches, through classical machine learning methods, to modern deep learning and recently emerged foundation models. Differences between processing RGB and multispectral images are highlighted, along with the architectural adaptations required to address multispectral characteristics. Subsequently, learning paradigms and approaches beyond fully supervised learning are discussed, which can help manage the vast amount of unlabelled and noisy data. Finally, practical considerations critical for real-world applications are examined, including limited and imbalanced datasets, computational constraints, and the need for uncertainty awareness and explainability, leading to the concept of band reduction.

3.1 Historical evolution of multispectral image analysis

The analysis of multispectral imagery, similarly to the broader field of computer vision, has undergone a gradual evolution driven by advances in sensing technology, computational resources, and image processing methods. Early approaches were significantly limited by computational power, resulting in methods operating at the level of individual pixels or small neighbourhoods. Over time, the focus shifted towards more context-aware techniques that utilise the spatial and spectral structure present in multispectral data.

3.1.1 Data representations for multispectral learning

The choice of data representation strongly affects the types of algorithms that can be applied. Multispectral data can be represented either as images with multiple spectral channels or as tabular data, depending on whether spatial context can be meaningfully exploited for a given task. When both spatial and spectral dimensions are considered, the data can be interpreted as a two-dimensional image with bands treated as channels, resulting in a three-dimensional data cube. With the inclusion of a temporal dimension, such as repeated observations over time, the representation extends to a four-dimensional spatio-spectral-temporal data cube.

In a pixel-based representation, each pixel is treated as an independent sample and described by a one-dimensional vector, where each element corresponds to reflectance at a specific wavelength. When temporal information is included, this representation may be extended to a two-dimensional matrix. This approach is especially common for imagery with low GSD, particularly from satellites, where the relevance of spatial features may be reduced for specific tasks. For example, in crop type classification task, the spatial characteristics of individual plants are usually not discernible at metre-level resolution, making a pixel-wise approach a sufficient and practical representation.

3.1.2 Pixelwise approaches

Early multispectral image analysis was predominantly pixel-wise, relying on simple rules and criteria applied to individual pixels or their neighbourhoods. Threshold-based methods were commonly used to separate vegetation from soil by determining optimal threshold values at specific wavelengths. Similarly, watershed algorithms [51] and contour-based processing techniques were used to aggregate pixels into larger segmented regions. Review articles from the mid-2000s and early 2010s [52, 53] discuss approaches such as clustering, per-pixel classifiers, averaging of pixel values over individual fields, and the selection of optimal wavelength bands for vegetation identification and mapping. During this period, simple convolutional operations for edge detection and filtering were also applied [54].

While these methods are computationally efficient and easily interpretable, they rely on handcrafted rules and are unable to capture broader spatial context. Furthermore, they are inherently sensitive to environmental variations, such as changes in illumination, and often fail to generalise. Nevertheless, pixel-wise methods can still be effectively combined with machine learning techniques, which help mitigate some of these limitations by automatically learning decision boundaries from data.

3.1.3 Classical feature extraction

To overcome some of the limitations of pixel-wise processing, researchers gradually incorporated spatial features originally developed for classical RGB image processing. These classical features generally refer to generic image and signal processing techniques, without a clear relation to the properties of the multispectral domain.

Early efforts employed geometric image moments to describe both spectral and spatial patterns, for example in illumination-invariant texture classification [55]. Subsequently, transform-based features such as derivatives, Fourier transforms, and wavelet-based decompositions were utilised for spectral analysis [56]. These methods are especially useful for hyperspectral imagery, where the larger number of spectral bands allows richer spectral signal characterisation.

In parallel, feature detectors such as SURF and FAST were applied to multispectral images to identify salient local structures. Together with corresponding feature descriptors and matching techniques, these methods were shown to be applicable across different bands [57], although challenges arising from nonlinear relationships between pixel intensities across bands were noted.

3.1.4 Domain-specific feature engineering

An especially influential class of features in multispectral image processing consists of domain-specific features that exploit prior scientific and physical knowledge about the observed phenomena. Among these, vegetation indices are the most widely used, as they are derived from relationships between the biophysical properties of vegetation and reflectance at specific wavelengths. One of the most prominent examples is NDVI (Normalized Difference Vegetation Index), computed using reflectance values from the red (600-700 nm) and NIR (700-1300 nm) bands, which is widely used for quantifying the health and growth of vegetation [58]. Its physical interpretation is rooted in plant physiology: chlorophyll pigments in healthy leaves absorb radiation in the red region, while reflectance in the NIR region is high due to light scattering within the spongy mesophyll structure of the leaf [59, 60].

Due to their normalised formulation, most vegetation indices, including NDVI, are dimensionless and inherently more robust to noise such as variations in solar irradiance and certain atmospheric conditions. However, NDVI is known to suffer from saturation in areas of dense and very green vegetation and to be sensitive to background soil colour. These limitations have motivated the development of alternative, specialised indices, such as SAVI, designed to reduce the impact of bare soil during early plant growth. Table 3.1 provides a summary of popular vegetation indices and their applications.

Table 3.1: Comparison of commonly used vegetation indices and their applications.

Name	Formula with names and Sentinel-2 wavelengths [nm]	Example applications	Reference
Normalised Difference Vegetation Index (NDVI)	(NIR − Red) / (NIR + Red) = (842 − 665) / (842 + 665)	General vegetation health and mapping, biomass estimation	[61, 62]
Green Normalised Difference Vegetation Index (GNDVI)	(NIR − Green) / (NIR + Green) = (842 − 560) / (842 + 560)	Early crop stress detection, nitrogen estimation, precision fertilisation	[63, 61]
Normalised Difference Red Edge (NDRE)	(NIR − Red Edge) / (NIR + Red Edge) = (842 − 705) / (842 + 705)	Dense vegetation and forest mapping, yield prediction	[64, 65]
Leaf Chlorophyll Index (LCI)	(NIR − Red Edge) / (NIR + Red) = (842 − 705) / (842 + 665)	Chlorophyll estimation, dense crop monitoring	[66, 67]
Optimised Soil Adjusted Vegetation Index (OSAVI)	(NIR − Red) / (NIR + Red + 0.16) × 1.16 = (842 − 665) / (842 + 665 + 0.16) × 1.16	Sparse vegetation and early growth stages	[68, 61]
Normalised Difference Water Index (NDWI)	(NIR − SWIR) / (NIR + SWIR) = (842 − 1610) / (842 + 1610)	Plant water stress, drought assessment, fire risk analysis	[69, 61]

Beyond vegetation-focused indices, Earth observation employs various other multispectral indices. Examples include MNDWI, which emphasises surface water and moisture content, and NDBI, which is useful in urban and land-use mapping.

3.1.5 Importance of feature selection

As the spectral and spatial resolution of multispectral systems increases, the selection of informative features remains essential, even in the presence of modern machine learning techniques. Using a subset of features, or artificially constructed derived features, can simplify model architecture, reduce training time, and improve prediction accuracy, generalisation, and model stability [70]. Although methods for automatic feature selection have advanced significantly, especially with deep learning, domain-specific features continue to play an important role. In agricultural applications, physics-informed features such as vegetation indices, especially NDVI, remain important [71].

3.2 Classical machine learning

Machine learning encompasses a broad family of algorithms that learn patterns from data and generalise to unseen samples. In the context of agricultural and environmental monitoring, these methods have historically played a central role in the processing of multispectral imagery, supporting a wide range of tasks, from feature extraction and clustering to regression and classification [72, 73]. This section focuses on classical machine learning approaches developed prior to the emergence of deep learning, providing a non-exhaustive overview of the most commonly used methods [74], with emphasis on their application to multispectral data in agriculture and environmental protection.

Feature preprocessing and dimensionality reduction. Unlike deep learning approaches, classical machine learning methods do not rely on learning hierarchical representations directly from raw inputs. Instead, the statistical properties and quality of the input feature space play a crucial role. Consequently, preprocessing of raw data, such as normalisation of numerical ranges, as well as feature extraction, are often essential initial steps in processing pipelines. PCA (Principal Component Analysis) is a widely used dimensionality reduction technique that projects the original feature space onto a set of orthogonal components while preserving the maximum variance present in the data. In computer vision, PCA is generally applied to feature vectors or flattened image representations. However, in the context of multispectral imagery, PCA is particularly effective when applied along the spectral dimension, where it reduces redundancy between highly correlated bands with minimal loss of information [75]. Beyond improving computational efficiency, dimensionality reduction can simplify model complexity, enhance interpretability, and help with visualisation of high-dimensional multispectral data.

Distance-based classifiers assign labels to new samples by measuring their similarity to labelled training data in the feature space, without explicitly learning complex decision boundaries. Among these methods, the kNN (k-Nearest Neighbours) algorithm is one of the most commonly used. It assigns a new sample to the class most frequently represented among its k nearest labelled neighbours. Authors in [76] employed kNN for pixel-wise classification of multispectral imagery to identify leaf diseases. Although distance-based methods are conceptually simple and computationally efficient, they rely on assumptions of local similarity in the feature space and are therefore sensitive to noise and shifts in data distribution.

Margin-based methods, most notably SVM (Support Vector Machine), have been widely applied across computer vision tasks due to their strong generalisation capabilities, particularly in scenarios characterised by a limited number of training samples and high-dimensional feature spaces. The SVM algorithm seeks to identify an optimal hyperplane that maximises the margin between classes in the feature space. For data that is not linearly separable, SVM employs the so-called *kernel trick*, to implicitly map input features into a higher-dimensional space, where a separating hyperplane can be found. These properties enable robust performance in data-scarce settings, but with limited scalability when applied to high-resolution imagery or large-scale datasets. As an example, the authors in [77] employed SVM to identify different varieties of rice seeds using a combination of spectral information and spatial features. SVR (Support Vector Regression) operates on principles similar to those of SVM but replaces discrete class separation with margin-based regression, enabling the prediction of continuous variables instead of categorical labels.

Tree-based and ensemble methods constitute an especially popular class of algorithms in Earth observation due to their simplicity and interpretability, particularly for classification tasks. The decision tree represents the most fundamental method in this family. It operates by recursively partitioning the feature space into smaller regions by selecting optimal features and thresholds at each step to maximise a purity criterion. This mechanism is especially well suited to multispectral data, where reflectance values across spectral bands often exhibit nonlinear interactions. However, individual decision trees are prone to overfitting, particularly when trained on limited or noisy data. To mitigate this limitation, ensemble methods such as Random Forests aggregate predictions from a large number of individual decision trees, resulting in improved generalisation performance and robustness. Gradient Boosted Decision Tree methods further extend the ensemble paradigm by sequentially training trees to correct the errors of previous models, with notable implementations such as XGBoost (Extreme Gradient Boosting) and AdaBoost. While boosting often improves model accuracy, it also introduces additional hyperparameters to tune. As an example, the authors in [78] benchmarked decision tree and random forest classifiers for wetland mapping using different combinations of spectral bands and identified Random Forests as the model of choice, with additional performance gains achieved through boosting. Tree-based methods can also be applied to regression tasks by changing the split criteria and output aggregation mechanisms to regression-oriented ones, such as minimising the MSE (Mean Squared Error). In this setting, the resulting models are commonly referred to as regression trees and random forest regression, and predict continuous variables instead of discrete class labels.

Shallow neural networks, typically comprising one or two hidden layers, represent an intermediate class of models between traditional statistical learning methods and deep neural networks. In the context of multispectral imagery, they can be applied to classification, regression, and low-dimensional feature extraction tasks. Their limited depth can be advantageous when training data are scarce, as it reduces the risk of overfitting, but these models generally lack the capacity to capture complex patterns.

Many classical machine learning methods operate on individual pixels or on tabular representations, and therefore ignore spatial structure by design, relying primarily on spectral information and manually engineered features. Moreover, these methods often struggle to scale to very large training datasets. Due to their relative simplicity, classical methods are typically less computationally demanding, which makes them well suited to resource-constrained systems, with the additional advantage of model interpretability. Furthermore, the widespread availability of mature and easy-to-use software libraries, such as the `scikit-learn` package in Python, has further reinforced their popularity and applicability. In many studies, multiple classical machine learning methods are applied to the same data and their performance is compared. For example, [79] evaluated several algorithms for detecting Citrus Huanglongbing disease using multispectral imagery, while [80] compared different methods for estimating soil salinity.

A systematic literature review by [74] presented the distribution of machine learning methods used in agricultural research publications. Artificial neural networks were employed in only 25% of the reviewed studies, while random forests and SVM accounted for 19% and 16%, respectively. It should be noted that this review covered publications from 2019 to 2023 and may therefore include works published prior to the wider adoption of deep learning, potentially introducing bias. In summary, although classical machine learning methods are generally inferior to deep learning approaches for high-resolution imagery, given enough training data, where spatial context is critical, they continue to play an important role in the analysis of multispectral data, especially in Earth observation applications involving lower-resolution imagery.

3.3 Deep learning in computer vision

Deep learning refers to a class of machine learning methods that utilise neural networks consisting of a large number of layers. Since the introduction of AlexNet [81] in 2012, which represented a major breakthrough, deep learning has undergone rapid development and has been applied to a wide range of computer vision tasks across many domains [82].

The success of deep learning-based methods stems from their ability to automatically learn hierarchical feature representations directly from raw data, often reducing or eliminating the need for manual, domain-specific feature engineering. As a result, deep learning approaches have demonstrated superior results in many tasks compared with classical machine learning techniques. However, they also introduce new challenges, particularly in the context of multispectral and hyperspectral data.

This section reviews deep learning architectures most frequently used in agricultural and environmental protection tasks [13, 83, 7], with a focus on methods developed for, or adapted to, multispectral and hyperspectral data processing [84]. In addition, common challenges associated with applying deep learning to multispectral imagery are briefly discussed. The use of deep learning for auxiliary tasks, such as image generation, is addressed in subsequent sections.

3.3.1 CNN architectures

CNNs are the most widely used deep learning models, utilised in the majority of computer vision tasks in agriculture, including RGB, multispectral, and hyperspectral imaging, as reported in recent review studies [85, 86]. At their core, CNNs exploit local spatial correlations in the data through the use of convolutional filters. The most commonly used CNN architectures focus on two-dimensional (2D) convolutions, which are effective at revealing spatial patterns and textures in images. However, one-dimensional (1D) CNNs can also be utilised, especially in the context of multispectral and hyperspectral data, to analyse purely spectral features independently of spatial context. In addition, three-dimensional (3D) convolutions enable the joint learning of spectral-spatial features by simultaneously operating across both spectral and spatial dimensions.

A typical CNN architecture consists of convolutional layers, pooling layers, non-linear activation functions, normalisation layers, and fully connected layers, which are usually placed at the end of the network. Modern architectures incorporate additional mechanisms such as residual connections [87], attention modules [88], and multi-scale feature fusion [89]. Through different combinations of these components, a wide range of popular CNN architectures have been developed and are commonly used as backbones for custom neural networks. A backbone refers to the part of the network responsible for feature extraction and can often be adapted by modifying the first input layer to accommodate a different number of bands.

Popular backbone architectures for spatial feature extraction include ResNet [87], which introduced residual (skip) connections, EfficientNet [90], which proposed compound scaling of network depth, width, and resolution and ConvNeXt [91], which incorporated design principles inspired by ViT (Vision Transformer). It is possible to achieve classification and regression tasks by attaching a task-specific network head to the backbone, typically implemented as a small number of fully connected layers. An example CNN architecture for classification is illustrated in fig. 3.1.

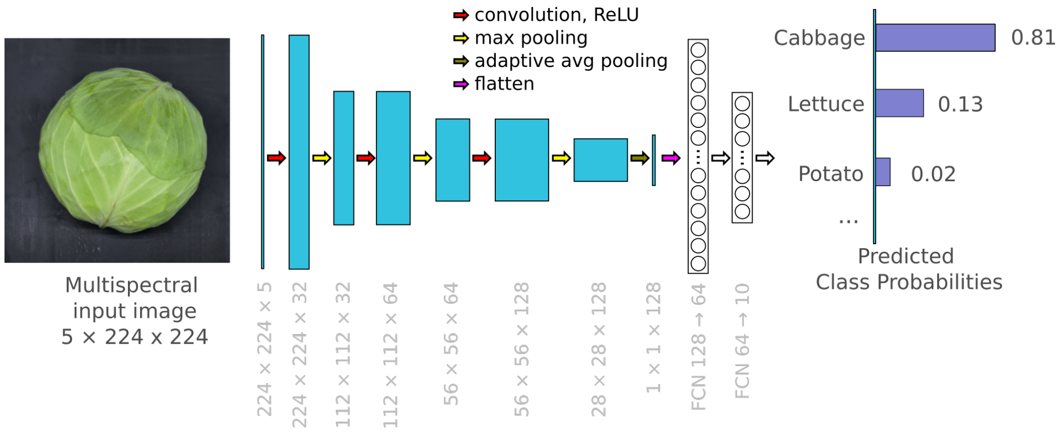

Figure 3.1: Example of a simple CNN architecture for image classification, illustrating a feature extraction backbone followed by a task-specific classification head with fully connected layers.

Encoder-decoder architectures for segmentation For semantic segmentation tasks, where the model output typically has the same spatial resolution as the input image, encoder-decoder architectures based on fully convolutional networks are particularly effective. The encoder component is responsible for feature extraction, while the decoder generates dense prediction maps by projecting high-dimensional feature representations back to the original size [92]. The U-Net architecture [93] introduced long skip connections that link encoder and decoder layers at corresponding spatial resolutions, enabling the preservation of spatial information. The success of U-Net inspired a large family of related models, including UNet++ [94], which introduced dense skip connections, and U-Net 3+ [95], which uses full-scale skip connections. nnU-Net [96] extends this concept further by providing a self-contained framework that automatically optimises the U-Net architecture and training pipeline for a given dataset rather than relying on a single fixed network design.

Other widely used encoder-decoder models include DeepLab [97], DeepLabV3+ [98], and LinkNet [99]. An example of a typical encoder-decoder architecture for image segmentation is illustrated in fig. 3.2. Although many segmentation models were originally developed for medical imaging tasks, they have been successfully adapted to multispectral remote sensing data. For example, the authors in [100] evaluated the performance of different deep learning models for tree segmentation in airborne multispectral imagery.

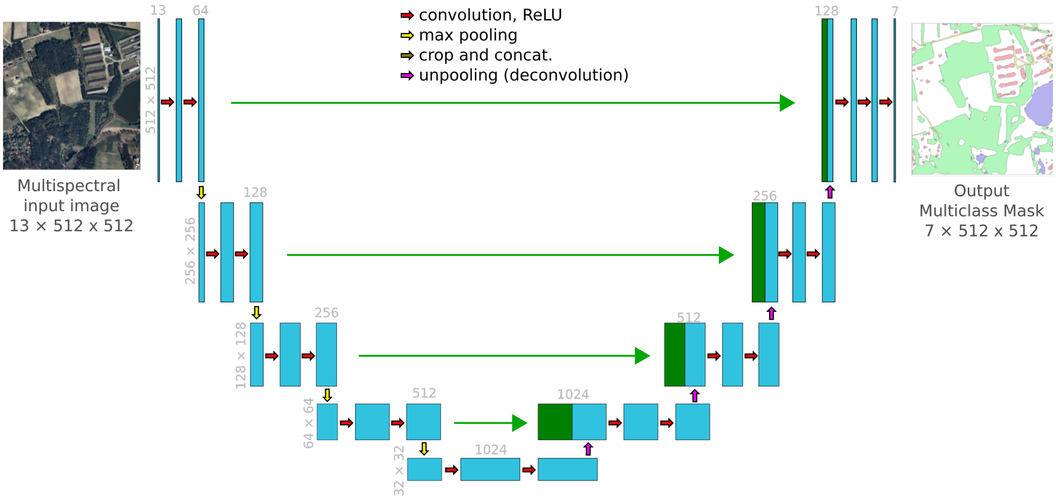

Figure 3.2: Example of an encoder-decoder CNN architecture for semantic segmentation, illustrating the downsampling encoder, upsampling decoder, and skip connections.

Object detection and instance segmentation Object detection aims to localise and classify individual objects within an image, typically by providing bounding boxes together with associated class labels. In more advanced variants, detection models may additionally produce a segmentation mask for each object instance, a task commonly referred to as instance segmentation. Early methods predominantly relied on two-stage architectures. In these approaches, a first stage generates candidate object regions, while a second stage refines them and performs classification. Although this design often achieves high detection accuracy, it is computationally expensive. Representative examples include R-CNN (Region-based Convolutional Neural Network) [101] and its subsequent extensions, such as Faster R-CNN [102] and Mask R-CNN [103].

To address the efficiency limitations of two-stage detectors, one-stage detectors were introduced, integrating object localisation and classification into a single neural network. Notable examples include the Single Shot Detector (SSD) [104] and the YOLO family of models [105], which have been widely adopted in practical applications. YOLO-based models directly regress bounding box coordinates and class probabilities in a single forward pass, enabling high inference speed. More recent YOLO variants have incorporated architectural and training improvements, as well as support for instance segmentation. As of early 2026, some of the most recent YOLO variants and derivatives include YOLOv11 and YOLO26. Due to licensing constraints for commercial use associated with certain YOLO releases, alternative models such as YOLOX [106] have also gained popularity. While YOLO-based detectors are highly computationally efficient, they may exhibit reduced performance when detecting small objects, which is a common task in Earth observation and aerial imagery [107]. In many applications, object detection also serves as a first step in subsequent tasks such as object tracking.

3.3.2 Transformer-based architectures

Transformer-based models represent a major paradigm shift in computer vision compared with CNNs. Originally developed for natural language processing, transformers also proved useful for vision tasks. The introduction of the ViT [108] in 2020 marked the first successful application of transformer architectures to image data. Unlike CNNs, which rely on local receptive fields and hierarchical spatial aggregation, transformers capture global contextual relationships through a self-attention mechanism that allows each element of the input to attend to all other elements. In ViT architectures, patch-based tokenisation is typically used to process images, dividing the input image into many fixed-size spatial patches. Each patch is projected into a feature embedding, which serves as input to the transformer encoder. To preserve spatial information, positional embeddings must also be added.

SegFormer [109] introduced a significant evolution of the traditional encoder-decoder approach for image segmentation by employing a transformer as the encoder and a lightweight multilayer perceptron as the decoder. More recent architectures such as Mask2Former [110] and OneFormer [111] further extend transformer-based segmentation by supporting multiple tasks, including semantic, instance, and panoptic segmentation, and have become widely adopted in computer vision research [112].

The Swin Transformer [113] introduced a hierarchical architecture operating across multiple scales to capture both local and global context efficiently [114].

The Detection Transformer (DETR) [115] was specifically designed for end-to-end object detection, predicting all bounding boxes and class labels in a single image pass. This family of architectures continues to evolve, and as of early 2026, the RF-DETR-XXL model has been reported as state of the art, surpassing YOLO26x on the COCO object detection benchmark.

Another important family of transformer-based models is DINO, with DINOv3 [116] being the most recent version. It relies on self-supervised training with a large-scale unlabelled dataset. DINO models demonstrate strong performance across a wide range of tasks, from segmentation, through image classification to depth estimation. Extensions such as MaskDINO [117] adapt the DINO architecture specifically for semantic and instance segmentation.

Although transformer-based models often surpass CNN architectures, especially in their ability to capture global contextual information, they remain computationally expensive and typically rely on large-scale training datasets. As a result, their adoption in Earth observation and agricultural applications has so far been limited, although they are becoming increasingly popular in these domains [118].

3.3.3 Temporal architectures

Temporal information is sometimes essential for crop monitoring and environmental change detection. RNN (Recurrent Neural Network), particularly LSTM (Long Short-Term Memory) networks, can be used for sequence modelling by design, thanks to the gating mechanisms in their architecture, which control the flow of information by adding or removing content from the cell state. As LSTM networks operate on one-dimensional input vectors, they are often utilised in conjunction with other feature extractors. For example, the authors in [119] combined CNNs with LSTM units to extract spatiotemporal features for crop yield estimation, while the authors in [120] proposed a shallow LSTM network integrated with a Random Forest model to improve wheat yield prediction using UAV-based multispectral imagery. Although transformer-based models also enable temporal information extraction [121], LSTM networks still remain used and, in some cases, achieve comparable or even superior performance. For instance, the authors in [122] reported no significant advantage of transformer-based approaches over RNNs for land-cover classification using Sentinel satellite time series.

3.3.4 General challenges

While deep learning methods often provide state-of-the-art performance, they also introduce new practical challenges. First, unlike classical methods, where feature engineering is used to explicitly encode domain knowledge, deep learning places greater emphasis on network architecture design, the training process, validation strategies, and data quality. Although many high-quality frameworks and libraries that abstract much of the engineering complexity are readily available, most notably PyTorch and TensorFlow, substantial expertise and effort are still required from the user. An important aspect of training custom deep learning models is hyperparameter tuning, including the selection of learning rates and their adaptation during training, the number of training epochs and early stopping criteria, the choice of loss functions and optimisation algorithms, and batch size configuration. Moreover, significant computational resources are often required for both training and inference, particularly for more advanced models. Deep learning methods still focus primarily on two-dimensional (2D) spatial image representations. Three-dimensional (3D) representations are also utilised in some applications [123]. For example, a neural radiance field-based approach has been used for fruit counting, producing point cloud representations from a set of posed images acquired from different viewpoints, as presented in [124]. However, 3D representations remain more complex, computationally expensive, and difficult to scale for agricultural and environmental monitoring tasks.

3.3.5 Architectural challenges specific to multispectral data

Techniques developed for RGB image processing often require adaptation to handle multispectral data, frequently with some loss of performance. One of the primary challenges is channel mismatch. For some architectures, extending the number of input channels from three to a higher number of spectral bands is sufficient, but this modification often comes at a cost. Models pretrained on large-scale RGB datasets cannot be fully utilised, as parts of the pretrained weights, especially in the first convolutional layers, are discarded. Furthermore, when processing multispectral data, models should not rely solely on spatial feature extraction using two-dimensional convolutions, but should also account for spectral features or joint spectral-spatial features. A similar challenge applies to transformer-based models, where the attention mechanism, originally designed to capture spatial relationships, can be extended to capture spectral features.

Moreover, multispectral data are often heterogeneous, with different spatial resolutions across bands, which is typically addressed by upsampling or downsampling selected bands. The presence of a higher number of highly correlated bands can additionally lead to overfitting and computational inefficiencies. When temporal information is required, data may not be available at regular intervals, for example due to cloud cover in satellite imagery. This irregularity further complicates model design or necessitates preprocessing workarounds, such as interpolating missing observations using temporally adjacent data. To address this issue, the authors in [125] proposed a dedicated approach for reconstructing missing satellite data using a unified spatial-temporal-spectral CNN.

Some popular model architectures include limited support for multispectral data, but often yield inferior results. For example, object detection models such as YOLO, when applied to multispectral imagery, provide inferior data augmentation strategies and cannot fully utilise pretrained weights. Consequently, domain-specific model adaptation, through targeted architectural modifications, often achieves better results than a naive extension of the input channels [126]. For instance, the authors in [127] proposed YOLO-MS, introducing substantial modifications to the YOLOv5 backbone, while the authors in [128] presented MOD-YOLO, reporting improved detection accuracy over baseline YOLO models.

Although deep neural networks are effective at learning feature representations, the incorporation of domain-specific features, especially vegetation indices, often leads to improved performance compared with directly utilising all available bands. For example, the authors in [129] achieved the highest accuracy in monitoring grain crop nitrogen status by using RGB, NIR, and NDVI together as input channels.

3.4 Evaluation metrics and validation strategies

Accurate and fair evaluation of models is essential for assessing their performance and enabling reliable comparison between different approaches. In the context of multispectral data analysis, particularly for Earth observation tasks, evaluation is influenced not only by the choice of performance metrics but, more importantly, by the adopted validation strategy, as inappropriate experimental design may lead to misleading conclusions [130]. Furthermore, the training process of modern machine learning models is characterised by a high degree of randomness, arising primarily from random parameter initialisation and data shuffling during training. Therefore, to ensure that observed performance differences are statistically significant, models should be evaluated using multiple runs with different random seeds and supplemented with appropriate statistical analysis.

3.4.1 Universal metrics

TP (True Positives) denotes the number of correctly identified instances. In the context of classification, it corresponds to the number of objects correctly assigned to a given class, whereas in pixel-wise segmentation it represents the number of pixels correctly identified as belonging to that class. Analogously, *FP* (False Positives) denotes the number of instances incorrectly assigned to a given class, *TN* (True Negatives) represents the number of instances correctly identified as not belonging to that class, and *FN* (False Negatives) corresponds to the number of instances that belong to the class but are incorrectly classified otherwise.

Based on the above, the following metrics can be defined:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3.1}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3.2}$$

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3.3}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3.4}$$

$$\text{Dice} = \frac{2TP}{2TP + FP + FN}. \tag{3.5}$$

The F1-score, defined as the harmonic mean of Precision and Recall, is widely used in classification tasks. In the context of binary segmentation, it is mathematically equivalent to the Dice coefficient, therefore, these terms are often used interchangeably.

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \tag{3.6}$$

The IoU (Intersection over Union) metric is particularly useful in segmentation tasks, as well as in object detection, where it is used to evaluate the overlap between predicted and ground truth bounding boxes.

In the presence of class imbalance, evaluation metrics must be interpreted with caution. In such cases, it is often beneficial to compute metrics on a per-class basis and report their macro-averaged or weighted average values, where the weights correspond to the class frequencies in the dataset. This approach ensures that minority classes are adequately represented and that dominant classes do not bias the overall performance assessment.

3.4.2 Efficiency and deployment-related metrics

From a deployment perspective, training time and the computational resources required during model optimisation are of secondary importance, as training is typically performed offline. Consequently, the key evaluation metrics focus on inference-time efficiency and resource requirements, and include:

Computational complexity and resource consumption, describing the cost of running the model during inference on the target hardware. This is often expressed in terms of FLOPs (Floating-Point Operations) per inference cycle. However, with modern quantised models relying heavily on integer arithmetic, throughput is often expressed in terms of operations per second instead. Furthermore, due to modern hardware acceleration architectures such as GPU (Graphics Processing Unit) and TPU (Tensor Processing Unit), evaluation increasingly emphasises RAM (Random Access Memory) or VRAM (Video Random Access Memory) consumption, as well as power usage, typically measured through power draw in watts or energy consumption in joules per inference.

Model size, understood as the storage footprint required to store the trained model in non-volatile memory. It is typically expressed as the size of the model weights in megabytes (MB) or gigabytes (GB), or as the total number of learnable parameters.

Inference time, defined as the time required for the model to process a single input sample. It is typically measured in milliseconds (ms) or reported as FPS. In real-time systems, the primary focus is often on latency, defined as the time elapsed from image acquisition to the generation of a control signal, rather than on throughput, which measures how many samples can be processed per unit time when parallel execution is possible.

3.4.3 Dataset splitting and validation strategies

Earth observation data are frequently spatially correlated, such that images acquired from geographically proximate areas are more similar to each other than observations from distant locations [131]. This effect is often even more pronounced for multispectral data than for RGB imagery due to the richer spectral information. Both spatial and temporal autocorrelation violate the standard machine learning assumption that data samples are independent and identically distributed. This highlights the need for appropriate validation strategies. Dataset splitting into training, validation, and test sets must be performed with particular care. Importantly, validation is not only used to estimate model performance, but also defines what type of generalisation a model is expected to achieve.

Improper dataset splitting can lead to over-optimistic performance estimates [132] or result in models that fail to generalise to new regions or time periods. In particular, random splitting is usually methodologically incorrect for Earth observation data due to geospatial proximity, as nearby pixels or image patches are highly correlated and similar samples may appear in both training and test sets. To mitigate spatial autocorrelation and prevent leakage of training data into validation or test sets, several simple strategies can be employed:

- using spatially isolated regions for training and testing, such as individual fields or entire geographic regions;
- creating a buffer zone between training and test areas, where samples are excluded from all sets;
- applying distance-based exclusion rules, for example by excluding from training all samples within a predefined radius of test samples.

Similarly, to address temporal autocorrelation, models trained on multi-year datasets may be evaluated using leave-one-year-out validation, ensuring that the test data represent unseen temporal conditions.

Standard k-fold cross-validation with randomly shuffled data is therefore of limited use in this context, as it does not account for spatial or temporal dependencies. To address this limitation, several studies have proposed spatial and spatiotemporal resampling strategies. For example, the authors in [133] and [134] presented algorithms not only for performing spatiotemporal resampling, but also for gaining insight into underlying data correlations. The authors in [135] demonstrated that standard cross-validation can lead to performance estimates that are up to 40% more optimistic compared with spatial cross-validation in real-world scenarios. The authors in [132] proposed a spatially fair train-test split that accounts for spatial autocorrelation and aligns test-set difficulty using kriging variance as a proxy for prediction difficulty. The authors in [136] introduced a spatial+ cross-validation strategy that jointly accounts for spatial autocorrelation and differences in feature space, resulting in more reliable performance estimates than random or purely spatial cross-validation. Additional care is required when the spatial distribution of target variables is highly imbalanced. In such cases, stratified splitting strategies may be necessary to balance the distribution of predicted values while preserving spatial separation between training and test sets.

Interestingly, these precautions do not always apply. For example, the authors in [137] argue that spatial cross-validation does not provide accuracy benefits over standard cross-validation for biomass mapping across the entire study area, that is, when calculating an aggregate result over the full map.

In summary, there is no single universally “correct” validation strategy. Validation procedures must be aligned with the task, deployment scenario, and expected generalisation behaviour, with explicit consideration of spatial and temporal autocorrelation. Statistical analysis is recommended to justify dataset splitting choices, although such analyses are still frequently omitted in the literature.

3.5 Learning paradigms and techniques

Earth observation data in agricultural and environmental protection applications are characterised by extensive spatial and temporal coverage and high intrinsic variability, while reliable ground-truth annotations remain scarce and costly. Similar challenges arise at smaller spatial scales, such as field-level monitoring, where environmental heterogeneity is high and manual labelling remains a significant burden. These characteristics limit the applicability of purely supervised learning approaches and motivate the adoption of more advanced learning paradigms that reduce dependence on extensive human annotation.

This section reviews the principal learning paradigms and associated techniques applicable to multispectral data, highlighting their underlying assumptions, advantages, limitations, and representative applications in agricultural and environmental monitoring.

Fully supervised learning. In fully supervised learning, models are trained using pairs of multispectral inputs and corresponding ground-truth annotations, such as class labels or pixel-level segmentation masks. Ground-truth labels are required for all training samples and for the evaluation of model performance, which often makes annotation a critical bottleneck in many applications. The creation of high-quality labels is typically expensive and time-consuming, as it requires domain expertise and careful manual effort. To mitigate these constraints, pretrained models are frequently used for initialisation, enabling faster convergence. In this setting, fully supervised learning is often applied as a fine-tuning stage following other training paradigms, such as self-supervised or semi-supervised pretraining.

As an example, the authors in [138] used 5000 manually annotated UAV images to detect and classify 13 pest species in soybean fields, achieving an overall accuracy of 94%.

Despite its reliance on extensive labelled data, fully supervised learning remains the dominant paradigm in agricultural and environmental protection applications, particularly in scenarios where high-quality annotations are available and high predictive accuracy is required.

Weakly supervised learning. Weakly supervised learning relaxes the requirement for precise annotations by allowing the use of coarse, incomplete, or noisy labels. In this setting, supervisory information is often imprecise and may not fully correspond to the spatial extent of the target objects.

In agricultural and environmental applications, weak supervision commonly arises when only image-level labels are available, when target objects occupy only a portion of the image, or when annotations are derived from indirect sources such as regional statistics or administrative records. For example, in a plant segmentation task, an entire image may be labelled as containing a specific crop, even though the crop occupies only part of the scene and is surrounded by weeds. In contrast, a fully supervised approach would require a pixel-level segmentation mask of the plant.

By reducing annotation effort, weakly supervised learning offers a practical compromise between labelling cost, data availability, and model performance. While such methods typically achieve lower accuracy than fully supervised learning given the same amount of data, they enable the effective use of larger and more weakly annotated datasets.

Semi-supervised learning. Semi-supervised learning is a hybrid approach that combines a small set of labelled samples with a substantially larger collection of unlabelled data. While the human-annotated samples provide reliable supervisory information, the unlabelled data contribute to shaping decision boundaries.

The authors in [139] introduced a large-scale agricultural image dataset for object detection consisting of 100032 images, of which only 960 were manually labelled. They demonstrated that semi-supervised training improved detection performance, increasing the mAP of a YOLO-based model from 0.322 to 0.347. Similarly, the authors in [140] proposed MSMatch, a semi-supervised scene classification method for multispectral satellite imagery, achieving performance comparable to fully supervised approaches while using only five labelled examples per class, compared with hundreds of labels required under full supervision.

Semi-supervised methods are frequently combined with knowledge distillation strategies, in which a smaller student model is trained to mimic the outputs of a larger teacher model. Due to its higher capacity, the teacher model can learn features more effectively from a limited set of labelled data, while the student model is further trained using pseudo-labels generated by the teacher on unlabelled samples [141].

Unsupervised learning. Unsupervised learning operates without ground-truth annotations, with patterns learnt solely from unlabelled data. In this setting, the model learns internal representations based on the statistical structure of the data, identifying similarities and differences between samples without access to semantic class information. Unsupervised methods are primarily applied to structure discovery tasks such as clustering, dimensionality reduction, change detection, and representation learning for subsequent model training.

Models trained in this manner can be used, for example, in land-cover analysis by clustering pixels from satellite imagery into groups with similar spectral characteristics, without prior knowledge of class definitions. The authors in [142] utilised this approach for detecting land-cover changes from satellite data. As no manual annotation is required, dataset preparation can be less costly. However, unsupervised methods are often more challenging to design and implement, and typically require large amounts of unlabelled data.

Self-supervised learning. Self-supervised learning is formally a subset of unsupervised learning, but its objectives and methodology differ sufficiently to warrant separate consideration. The central idea is to generate supervisory signals directly from the data itself. It typically involves pretext tasks that encourage the model to learn domain-relevant feature representations. For segmentation or general representation learning, these may include predictive tasks (e.g., jigsaw puzzles, slice-order prediction, or rotation recognition), generative tasks (e.g., image colourisation, denoising, or inpainting), or contrastive learning, in which the model is trained to minimise the distance between positive sample pairs while maximising the distance between negative pairs [143]. In the context of multispectral imagery, these approaches can be extended with spectral-specific tasks, such as the reconstruction of individual spectral bands. The purpose of pretext tasks is not the task itself, but the learning of transferable feature representations.

Self-supervised learning has become a core part of the training of large-scale foundation models. In Earth observation applications, self-supervision enables the generation of feature embeddings for satellite imagery, supporting geo-semantic search and retrieval. Recent efforts aim to apply this approach at a global scale. A prominent example is the Google Satellite Embedding dataset, a precomputed output generated by the AlphaEarth Foundations model [144], in which each 10-metre pixel is represented by a 64-dimensional embedding vector encoding spatial and temporal information.

In remote sensing, particularly for multispectral imagery, autoencoder-based approaches have been widely used [145]. More recently, transformer-based architectures have been increasingly adopted in state-of-the-art self-supervised models. Although self-supervised learning does not require manual annotations, it typically relies on very large volumes of unlabelled data and substantial computational resources for model training.

Active learning. Active learning is a human-in-the-loop approach aimed at reducing the cost of manual annotation by iteratively requesting the most informative samples for human labelling. Instead of annotating the entire dataset, the model actively queries human experts for labels on data points that are expected to provide the greatest performance improvement.

An important aspect of this approach is the sample selection strategy, which aims to identify instances that are more valuable than those obtained through random sampling. The most commonly used approach is uncertainty-based sampling [146], in which the model selects samples for which its predictions are least confident. Another widely used strategy is committee-based sampling, where multiple models are employed and samples are selected based on the rate of disagreement among the ensemble models.

The authors in [147] applied active learning to vegetation type classification using UAV multispectral imagery, using spectral-spatial features to discriminate between similar classes. The proposed method achieved 93% accuracy while using only 20% of the labelled data, compared with 95% accuracy obtained with full annotation.

Few-shot and zero-shot learning. Few-shot learning aims to address the challenge of model generalisation when only a limited number of training examples are available for new classes. This approach is particularly valuable in scenarios where data are scarce or expensive to obtain [148], such as in the case of rare environmental events. For classification tasks, few-shot learning is commonly implemented using metric learning, where samples are mapped into an embedding space in which instances belonging to the same class are located closer together according to a distance metric. The authors in [149] applied few-shot learning to hyperspectral imagery for the classification of frost damage severity in tomato plants. By using drought-stressed plants as a source domain, the method achieved 70% accuracy with only eight target samples.

Zero-shot learning extends this concept further by enabling the prediction of previously unseen classes without any task-specific labelled examples. This approach typically relies on auxiliary information, such as class attributes or textual descriptions, and operates within a shared embedding space that links visual and semantic representations. Such mechanisms are widely used in vision-language models. The authors in [150] introduced the Llama3-MS-CLIP vision-language model, pretrained on large-scale multispectral Sentinel-2 imagery with automatically generated captions, demonstrating improved performance over RGB-based vision-language models in zero-shot classification and retrieval tasks.

Transfer learning. Transfer learning is a paradigm in which knowledge gained during training on a source domain is reused to improve performance on a related task or target domain, enabling the reuse of previously learnt feature representations. In practice, transfer learning is most commonly implemented through model fine-tuning, where pretrained network weights are further optimised using data from the target dataset [151]. Fine-tuning can be performed using relatively small amounts of labelled data and may involve updating all network parameters or only a subset of layers while keeping the remaining weights frozen. From a domain perspective, transfer learning allows for domain adaptation, which aims to mitigate distribution differences between the source and target domains.

In the context of multispectral learning, models pretrained on RGB imagery can provide useful spatial feature representations, but additional training is needed for spectral information extraction and for adapting the model to the specific task. The authors in [152] highlight transfer learning as a key enabler for agricultural machine learning, where models must generalise across changing domains under limited labelled data. For example, a model trained to detect crop types in one country may require adaptation before being applied to another region with different climatic and agricultural characteristics.

Continual (lifelong) learning. Continual learning aims to mimic the gradual nature of human learning by enabling models to absorb new knowledge while remaining operational and continuing to perform their primary tasks. Newly acquired information is not simply added to the existing model but may also revise previous knowledge. This paradigm is particularly promising for long-term Earth observation applications, such as environmental monitoring and change detection, where gradual yet continuous changes in ecosystem conditions must be taken into account.

In recent decade, continual learning has become an active research area, with a growing number of proposed methods [153]. However, its practical adoption in agricultural applications remains limited due to complexity and operational constraints. One of the main challenges is catastrophic forgetting, where model performance on previously learned categories deteriorates as new data are incorporated. This issue is especially pronounced in deep neural networks, whose parameters are highly interdependent. In addition, achieving an appropriate balance between stability and plasticity remains difficult, particularly when it is unclear whether observed changes reflect domain shift or the emergence of new knowledge. For example, plant stress caused by drought may be interpreted either as a change in environmental conditions or as a new class.

3.6 Training strategies under real-world constraints

Different learning paradigms described in section 3.5 focus on specific data limitations, particularly when only limited ground-truth annotations are available. However, in agricultural and environmental applications, additional challenges arise from the high variability of natural conditions, as discussed in section 2.7. As a consequence, models often struggle to generalise to unseen acquisition conditions, geographic regions, and local environmental variability. To further address these limitations, a range of complementary training strategies can be employed.

One of the issues requiring special attention is class imbalance, which is primarily a training-time problem affecting the design of the model training pipeline. Rare classes, such as disease outbreaks or endangered plant species, may be significantly underrepresented in the dataset but are still required to be detected reliably, whereas standard training procedures tend to bias predictions towards dominant classes. The most widely used techniques for addressing class imbalance include [50]:

- **Algorithm-level strategies**, such as loss reweighting, where higher weights are assigned to underrepresented classes to increase their influence during model optimisation.
- **Data-level strategies**, such as undersampling of majority classes and oversampling of minority classes. A wide range of methods exists that allow more informed sample selection than random sampling, for example based on distance in feature space.
- **Threshold moving**, where class-specific decision thresholds are fine-tuned, for example by decreasing the classification threshold for rare classes.

These approaches can be used individually or in combination to balance model sensitivity and reduce bias towards frequent classes. It should be noted, however, that model bias may sometimes be intentional and desirable, depending on the operational objective. For example, enforcing class balance may lead to overprediction of rare classes or an increased rate of false-positive detections.

To make more efficient use of limited data and reduce annotation requirements, data augmentation techniques are indispensable. Classical augmentation methods artificially expand the training set through random image transformations such as rotations, contrast adjustments, or illumination effects. A wide range of implementations exists for RGB imagery, primarily focusing on preserving spatial features. For multispectral data, improved performance may be achieved by additionally considering radiometric and spectral features. An example is the MixChannel method proposed by the authors in [154], which mixes images of the same location acquired at different times to reduce spatiotemporal overfitting, significantly improving generalisation for forest type mapping.

Another approach involves physically grounded simulation, which allows the generation of entirely new data using digital models. While graphical rendering engines are widely available for RGB imagery, simulation for multispectral data is more challenging because reflectance values in individual spectral bands must be derived from physical properties, such as chlorophyll content. Furthermore, simplified physical assumptions may fail to capture the full complexity and variability of real-world environments.

Recent work also explores generative and domain adaptation techniques, including diffusion models and generative adversarial networks (GANs), allowing the synthesis of training samples for new domains [155]. These methods enable scene or patch synthesis, modification of background conditions (e.g., soil type), and targeted generation of rare-class examples. The authors in [156] introduced a diffusion-based framework, DODA, allowing the generation of realistic labelled data from a few reference images without retraining the generative model.

In many practical scenarios, camera cost or deployment constraints limit the availability of certain spectral modalities at inference time. For example, multispectral cameras may be easily available during data collection for training, whereas only RGB sensors are feasible for operational deployment. This constraint can be addressed with multimodal training strategies, in which richer spectral information is used during training, requiring less annotations, while the deployed model operates using a reduced set of modalities.

To summarise, in multispectral agricultural and environmental applications, training strategies and data augmentation techniques serve not only to expand the training set but also to improve model robustness to the environmental variability encountered in real-world scenarios. Consequently, physically informed and domain-aware augmentation are important tools for achieving reliable generalisation.

3.7 Foundation models

In the computer vision field, foundation models are large pretrained deep learning models capable of capturing rich semantic and contextual representations, including relationships between objects, their spatial configuration, and environmental variability [157]. These models are characterised by broad transferability across a wide range of downstream tasks. They often integrate multiple modalities, most commonly vision and text, and support human prompting mechanisms. Foundation models have attracted increasing interest in recent years, enabled by the availability of vast amounts of data and large-scale computational resources, often requiring tens of thousands to millions of GPU hours for training on datasets of terabyte scale, using either annotated or unlabelled data.

One of the most prominent examples is SAM (Segment Anything Model) [158] and its newer version, SAM2, which enable zero-shot segmentation. The model was trained on over one billion masks from 11 million images and uses a ViT as the image encoder. The model is promptable, primarily through point or bounding-box inputs, and its zero-shot performance on specific tasks can be competitive with fully supervised models trained for those tasks. Although the model is designed for RGB images, the input channels can be replaced by a single-channel vegetation index such as NDVI to facilitate tasks such as biomass segmentation. However, performance degradation may occur when transferring the model to such non-RGB modalities.

Among the most widely used vision-language foundation models is the CLIP (Contrastive Language-Image Pretraining) family, commonly applied to downstream tasks such as image classification, image-text retrieval, or as guidance for diffusion models. Although originally developed for RGB imagery, extensions such as the Llama3-MS-CLIP model [150] have been proposed to support multispectral data. This model was trained on approximately one million multispectral image-caption pairs. Experimental results demonstrate significant improvements in zero-shot classification and retrieval compared with RGB-based vision-language models.

Foundation models have also emerged for agricultural and environmental protection applications, particularly in Earth observation, where large volumes of unlabelled data are available [159]. These models support a wide range of tasks, from parcel boundary extraction to flood detection in multispectral imagery [160]. Remote sensing foundation models often integrate multiple sensing modalities, including optical data, SAR, LiDAR (Light Detection and Ranging), and thermal imagery, to address the heterogeneous nature of geospatial data [161]. Such multimodal integration has been shown to improve performance in tasks such as land-cover classification, object detection, and change detection.

SpectralGPT [162] is a foundation model specifically designed for spectral imagery, using a 3D generative transformer to jointly model spatial and spectral information. It was trained on more than one million multispectral images and supports diverse Earth observation tasks, including scene classification, semantic segmentation, and change detection. In contrast, the SkySense foundation model [163] focuses primarily on multimodal spatiotemporal modelling rather than explicit spectral sequence modelling and was trained on 21.5 million temporal sequences. The Prithvi-EO-2.0 model [164] jointly models spectral and spatiotemporal information and was pretrained on 4.2 million global multispectral time-series samples from NASA's Harmonized Landsat and Sentinel-2 archive, incorporating spatial, temporal, and location embeddings. As of 2025, according to the authors, it achieved state-of-the-art performance in disaster response, land-cover classification, and crop mapping tasks within the Geo-Bench benchmark [165].

TerraMind [166] is a generative multimodal Earth observation foundation model pretrained on aligned multispectral Sentinel-2 data together with SAR, geospatial metadata, and textual information. It models cross-modal relationships and supports any-to-any translation across data modalities. A key architectural feature is its dual-scale early-fusion strategy, which jointly learns token-level representations.

Despite their advantages, multispectral foundation models face several challenges. Models trained on data from one sensor often do not generalise well to others due to differences in spectral band configurations. Large-scale labelled multispectral datasets remain limited, particularly for vision-language training. Furthermore, multispectral Earth observation data are inherently complex and require joint modelling of spectral, spatial, and temporal dimensions, while many existing models address only one or two of these aspects. Finally, standardised benchmarks and leaderboards need to be further extended and widely adopted to enable reliable quantitative comparison between models. Another limitation is the high computational cost of foundation models, which may restrict their use in operational systems and edge deployments.

3.8 Practical constraints - efficiency and deployment

In many applications, multispectral data can be stored after acquisition and processed offline, for example using cloud-based infrastructure, where computational resources and processing time are less constrained. This approach is suitable for many large-scale analysis tasks, such as seasonal crop assessment or estimating crop nutrient requirements across a field.

However, in numerous agricultural and environmental protection scenarios, machine learning models must operate in near-real-time or even real time and therefore are deployed on embedded or edge devices. For example, immediate decisions are necessary for onboard processing on UAVs, e.g. to adjust the flight trajectory while performing targeted spraying only over infested plants, as demonstrated in [167]. Other examples include continuous wildlife monitoring and detection using autonomous field sensors, as well as onboard satellite processing, where preliminary filtering (e.g., cloud or change detection) may be necessary due to limited transmission bandwidth.

Deployment in such environments introduces system-level constraints, including latency, throughput, memory footprint, and energy consumption, all of which directly affect operational cost and feasibility, particularly in remote sensing applications [168]. Therefore, model evaluation should also consider efficiency-related metrics, discussed in section 3.4.2, alongside predictive performance metrics such as accuracy or F1-score. In practical systems, a slight reduction in predictive performance is often acceptable if it is accompanied by substantial improvements in inference speed or energy efficiency.

Achieving such trade-offs requires careful selection of the model architecture. Lightweight CNNs, or in some cases classical machine learning methods, may provide sufficient performance while requiring significantly fewer resources than computationally intensive transformer-based models. Additional efficiency gains can be obtained through model compression techniques such as knowledge distillation, pruning, and weight quantisation [169].

Finally, deployment performance is strongly influenced by the underlying hardware platform. General-purpose CPUs are increasingly complemented or replaced by hardware optimised for machine learning, such as GPUs, TPUs, NPUs, or dedicated edge accelerators. The joint optimisation of model architecture and hardware configuration is therefore essential for practical multispectral applications operating under real-world constraints.

3.9 Explainability, uncertainty and band reduction

Explainability has become an important aspect of machine learning, including agricultural and environmental applications where model outputs are used to support operational decisions. In this context, users may need to understand the factors influencing model predictions in order to build trust. This requirement is especially important in high-stakes applications. Despite its importance, the adoption of explainability remains relatively low in agricultural applications. According to the review article [170], only 5.2% of studies utilising machine learning in agriculture published between 2018 and 2025 include explainability in their results, with attention-based visualisation being the most commonly used.

The need for explainability is driven mainly by the increasing complexity of machine learning models. While simple models such as decision trees offer a degree of intrinsic interpretability, deep learning networks are typically treated as black boxes. This lack of transparency becomes even more pronounced for multimodal models [171]. XAI (Explainable Artificial Intelligence) methods aim to provide insight into the knowledge acquired by the model and its reasoning process, and can help expose hidden model biases or spurious correlations. This is especially relevant in multispectral remote sensing, where models may rely on subtle spectral or contextual cues that are not physically meaningful.

XAI tools can be broadly categorised according to the stage at which they are applied to the model [172]. Ante-hoc methods incorporate interpretability directly into the model architecture, providing explanations after training. In contrast, post-hoc methods are applied once the model is trained and treat it as a black box, often in a model-agnostic manner. Depending on the model input, these methods may analyse the importance of specific image regions, spectral bands, engineered features, or even the internal state of the network, such as learnt weights or attention patterns.

The literature lacks a clear and coherent grouping of XAI methods. Nevertheless, the following categories are often distinguished [173, 174, 175, 176]:

- **Perturbation-based** methods, which evaluate the effect of systematically modifying or occluding parts of the input, for example using a sliding window to create occlusion maps.

- **Gradient-based** approaches, such as saliency maps and Grad-CAM (Gradient-weighted Class Activation Mapping). These methods are specific to neural networks and estimate the importance of input features by backpropagating gradient signals.

- **Model-agnostic feature attribution** methods, which quantify the contribution of individual input variables, such as spectral bands or vegetation indices, to a given prediction. The popular SHAP (SHapley Additive exPlanations) method assigns each feature a contribution value based on Shapley theory, quantifying its impact on the model prediction. The LIME (Local Interpretable

Model-agnostic Explanations) method, on the other hand, explains a prediction by approximating the model locally with a simple interpretable model.

- **Model-intrinsic visualisation** techniques applicable to deep networks. In CNNs, this may include visualisation of learnt filters at specific layers, whereas in transformer architectures it often involves attention-based visualisation methods. Beyond raw attention visualisation, techniques such as attention rollout aggregate attention weights across layers to estimate the contribution of image patches, which are input tokens, to the final prediction. In self-supervised transformer models, such as DINO, attention patterns often align with semantically meaningful image structures, allowing attention weights to serve as an implicit explanation of model behaviour, although their faithfulness remains debated [177].

Several recent studies demonstrate the practical use of explainability in multispectral agricultural applications. The authors in [178] estimated temporal crop stress using vegetation indices derived from Sentinel-2 imagery and employed SHAP to identify the most important features for each month's prediction, together with Grad-CAM heatmaps to highlight regions that most influenced the learnt representations. Similarly, the authors in [179] applied SHAP to identify the most important channels among spectral bands and vegetation indices for change detection using a U-Net-based model, concluding that the NDVI index was the most informative source of information.

Further advancements in LLMs and their multimodal variants may enable new approaches to computer vision explainability, for example through reasoning-style explanations analogous to textual reasoning chains. The authors in [180] proposed unified architectures that tightly integrate visual perception with language-based reasoning, demonstrating improved semantic grounding. Additionally, [171] distinguish graph-based XAI approaches for structured vision tasks, where features and modalities are represented as nodes and their relationships as weighted edges, allowing intuitive visualisation and analysis of their contribution to model predictions. Nevertheless, the internal reasoning mechanisms of such models remain largely opaque and require further research.

When interpreting XAI results, it is important to recognise that most methods reflect statistical importance rather than physical causality. Features identified as important may correspond to correlations present in the training set rather than physically meaningful causality, especially when the model generalises poorly to unseen data. This issue is particularly relevant in multispectral remote sensing, where spatial context and environmental factors may be strongly correlated but physically irrelevant in the real world. For example, a vegetation classification model may rely on soil background characteristics instead of plant spectral characteristic. Therefore, explainability results should be interpreted with caution and in conjunction with domain knowledge.

Explainability is closely related to uncertainty estimation, as both provide complementary information. While XAI methods aim to explain why a prediction was made, uncertainty estimation quantifies the model's confidence in that prediction. This is especially relevant for deep learning models, which often tend to be overconfident. Reliable uncertainty estimation is essential in decision-support systems, as it helps identify situations in which the model lacks confidence and human intervention may be required. Two main types of uncertainty are commonly distinguished [181]. Aleatoric uncertainty represents irreducible noise inherent in the data, such as sensor noise, whereas epistemic uncertainty reflects model uncertainty due to insufficient model capacity or training data, suggesting potentially poor generalisation to unseen data distributions. Several approaches for estimating uncertainty in deep learning models have been proposed, including Bayesian neural networks, Monte Carlo dropout, and deep ensembles. These techniques approximate the distribution of model predictions rather than relying on a single deterministic output.

Explainability techniques are also closely related to spectral band importance analysis and band reduction. Multispectral and hyperspectral sensors often provide many highly correlated bands, some of which may be redundant for a given task. Identifying the most informative bands enables an informed trade-off between system complexity and predictive performance. Reducing the number of bands may enable cheaper sensor design, simplify the model, and reduce computational requirements. This issue is particularly pronounced in hyperspectral imaging, where the large number of bands can introduce significant redundancy. As highlighted by [182], increasing the number of spectral bands does not necessarily provide additional information but instead increases the complexity of the classification task. Consequently, many hyperspectral band selection methods rely on statistical analysis to identify an optimal subspace of informative bands [183].

Several strategies exist for band selection. Physics-based approaches rely on domain knowledge to select wavelength ranges relevant to a specific task, such as short-wave infrared bands for methane detection. Wrapper methods evaluate model performance using different subsets of bands for training. XAI-based approaches use feature attribution results to rank bands according to their contribution to model predictions. For example, authors in [184] report that NIR bands are often the most informative for general vegetation-related tasks. In contrast, [185] showed that soybean yield prediction depends on multiple spectral bands, explaining why models based solely on vegetation indices, which use only a limited subset of bands, may result in reduced performance. Beyond classical band selection, feature extraction techniques may further reduce data dimensionality, sometimes more effectively than direct band selection. However, this comes at the cost of reduced physical interpretability, as extracted features no longer correspond to specific wavelengths that can be directly measured by optical filters or sensors. Alternative perspectives on band importance have also been proposed. For instance, [186] introduced a visual and statistical framework that shifts the focus from identifying a single optimal subset to understanding the relative importance of individual bands. By repeating band selection experiments and tracking the frequency with which each band is selected, they constructed band frequency histograms that reveal consistently important spectral regions. The authors in [187] proposed an evolutionary approach, employing a genetic algorithm with an embedded 3D convolutional neural network as a fitness function to guide band selection.

Although band selection is especially important for hyperspectral imagery due to its high dimensionality, it remains relevant for multispectral data as well. Importantly, hyperspectral band reduction can serve as a bridge towards practical multispectral system design. Once the most informative bands are identified, they can inform the design of lower-cost multispectral sensors that capture only the most relevant wavelengths, facilitating real-world deployment in agricultural and environmental monitoring applications.

3.10 Summary and outlook

This chapter has reviewed the current state-of-the-art machine learning techniques relevant to multispectral image processing. In the next chapter, deep learning models will be employed predominantly for data characterised by high spatial resolution, while simpler approaches will be considered for satellite-based observations. The study will focus primarily on supervised learning paradigms, complemented by the use of foundation models and transfer learning strategies. As highlighted, dataset variability will be considered in both model design and evaluation, with careful dataset partitioning and validation to avoid overoptimistic performance estimates. Explainability and spectral band selection will be investigated as key components for interpreting model behaviour, building trust, and reducing system complexity, thereby supporting practical deployment in environmental applications.

4 Implementation and results

4.1 Minimising manual annotations

Accurate segmentation of crops and weeds is an important task in precision agriculture. As deep learning models typically require large amounts of pixel-level annotations, the preparation of which is costly, reducing the demand for manual labelling while maintaining high segmentation accuracy is an important challenge.

The approach presented in this section addresses this problem by utilising multispectral imagery available only during the training phase, while restricting the final model to standard RGB images. This design allows to benefit from richer spectral information without increasing the operational cost and complexity of the deployed hardware. To achieve this, two complementary learning strategies are applied. First, a model is trained in a fully supervised manner directly on multispectral data. Afterwards, a semi-supervised teacher-student approach is used. A second model, operating only on RGB images, is trained using segmentation masks synthetically generated by the first multispectral model. Although an RGB-only model could also be trained in a fully supervised manner, achieving comparable performance would require substantially more manually labelled data. This behaviour is expected due to the richer feature representation provided by the additional spectral bands. The overall concept of this dual approach is illustrated in fig. 4.1.

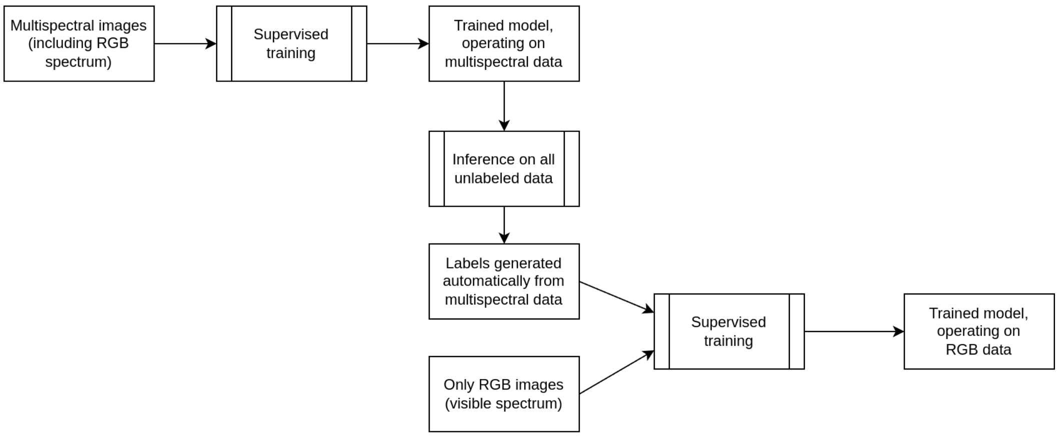

Figure 4.1: Diagram illustrating the utilisation of multispectral data for automatic label generation used to train the model operating exclusively on RGB data.

In addition to reducing the number of labelled images, the potential of foundation models to further simplify the labelling process is also investigated. Instead of manually creating full pixel-wise segmentation masks, an annotator may provide only a bounding box, while the segmentation mask is automatically generated by a foundation model and, if necessary, manually refined.

The following subsections describe the dataset, model architectures, and experiments used to evaluate the proposed methodology, along with the obtained results.

4.1.1 Dataset

The experiments were conducted using the Sugar Beet Field Dataset, which was introduced in [38] and used for plant classification and localisation tasks. The data were collected over a three-month period during the spring growing season using the field robot BoniRob on a sugar beet farm near Bonn, Germany. The robotic platform was equipped with multiple sensors, including a four-channel multispectral camera and an RGB-D sensor. In this research, a subset of the dataset corresponding to a single acquisition day was used. From this subset, only RGB and NIR imagery, together with the associated manually annotated segmentation masks, were considered. The NIR values captured by the JAI AD-130GE camera are sensor intensities. Although the dataset does not describe any radiometric calibration procedure, these intensities may be assumed to correspond to reflectance values to some extent, as the lighting conditions were consistent across images. Nevertheless, this relationship is not critical for the purposes of this research. To simplify the task, segmentation masks corresponding to weeds were discarded and treated as background. Example RGB images, corresponding NIR data, and ground-truth masks are presented in fig. 4.2.

The data were partitioned into four subsets:

- **labelled training set**, containing 169 images with segmentation masks. This set was further subdivided into variants containing only 50%, 10%, and 5% of the data, yielding 84, 16, and 8 images, respectively;
- **unlabelled training set**, containing 1865 images without segmentation masks, used for semi-supervised training;
- **validation set**, containing 57 images with segmentation masks, with a subset of 8 samples used in selected experiments;
- **test set**, containing 57 images with segmentation masks, used exclusively for final performance evaluation.

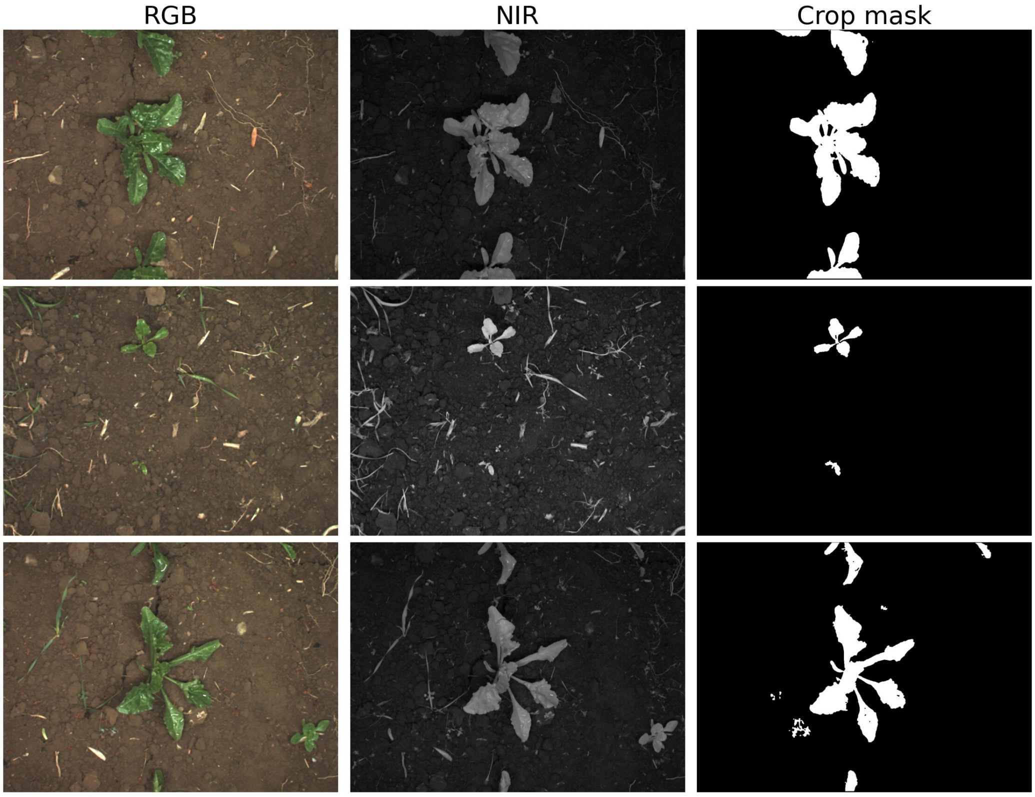

Figure 4.2: Three examples of labelled data, showing an RGB image, the corresponding NIR image, and the manually labelled mask in the last column.

The original dataset was divided based on spatial position within the field, ensuring that partially overlapping images did not appear in different subsets and that spatial correlation between subsets was minimised. Although NIR imagery can be used directly as an additional spectral channel, domain-specific vegetation indices are expected to provide more explicit information for plant discrimination. Therefore, NDVI was additionally calculated. Figure 4.3 presents an example comparison of RGB, NIR, and the corresponding NDVI images.

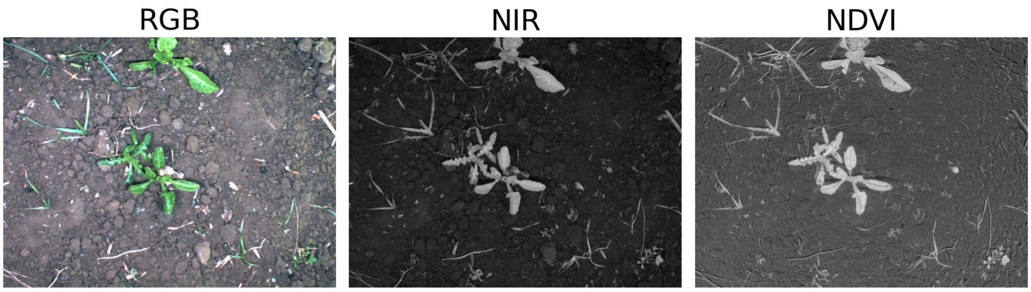

Figure 4.3: Comparison of input modalities, including RGB imagery, the corresponding NIR channel, and the derived NDVI representation.

4.1.2 Methods

To determine an appropriate segmentation model for the subsequent experiments, several deep learning architectures were evaluated. The compared models included U-Net, U-Net++, DeepLabv3, PAN, and a Mix Vision Transformer (MiT) encoder combined with a U-Net-style decoder. All convolutional models employed ImageNet-pretrained encoders (ResNet18 or ResNet50) to improve convergence and performance. Training was performed on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU (24 GB memory). The batch size was set between 2 and 8, depending on the memory requirements of the specific architecture. The models were trained using the cross-entropy loss function and the AdamW optimiser. The learning rate was controlled using a reduce-on-plateau strategy, with the initial value manually adjusted according to the size of the training dataset, typically starting from 10^{-4}, with early stopping applied after 30 epochs without improvement in the validation metric. Prior to training, all images were normalised using the mean and standard deviation calculated over the entire training set. Because some experimental configurations involved small labelled subsets, with as few as eight images, extensive data augmentation was applied to improve model generalisation. The augmentation pipeline included random scaling (±20%), random cropping (to 1280×960 pixels), horizontal and vertical flips ($p = 0.5$), random rotations (±15°), brightness, contrast, and saturation adjustments (±0.05), Gaussian noise addition (variance 0.02), and affine transformations involving rotations of up to ±5° and translations of up to ±10 pixels. Model performance was assessed using standard pixel-wise segmentation metrics, including IoU, accuracy, precision, and recall.

4.1.3 Experiments and results

A series of experiments was conducted to evaluate the proposed approach. In the first stage, several model architectures were trained using the full labelled training set (100%) in order to establish a performance baseline and select a network for further analysis. The results of this comparison are presented in table 4.1.

The evaluated architectures achieved comparable performance levels. The highest accuracy was obtained with the MiT-B3 model. However, this architecture is significantly larger, with a weight size of 543 MB compared with 165 MB for U-Net. Considering the intended application in resource-constrained environments, such as real-time operation on agricultural robots, as well as the computational limitations of this study, a more lightweight solution was preferred. The U-Net architecture was therefore selected for subsequent experiments, as it offered competitive segmentation accuracy while maintaining substantially lower computational requirements. The IoU metric values for the crop class are comparable to, or higher than, those reported

Table 4.1: Comparison of different model architectures trained on 100% of labelled RGB data.

Model	IoU crop	Precision crop	Accuracy crop
U-Net ResNet18	0.8569	0.9157	0.9934
U-Net ResNet50	0.8669	0.9216	0.9939
U-Net++ ResNet18	0.8661	0.9294	0.9939
MiT-B1 (with U-Net decoder)	0.8734	0.9336	0.9943
MiT-B3 (with U-Net decoder)	0.8955	0.9431	0.9953
FPN ResNet18	0.8660	0.9294	0.9939
PAN ResNet18	0.8363	0.8880	0.9922
DeepLabv3	0.8656	0.9373	0.9939

in related studies, such as 0.80 in [188] and 0.837 in [189]. It should be noted, however, that these works used larger versions of the dataset.

In the next stage, the selected architecture was trained using reduced subsets of the labelled training data containing 5%, 10%, 50%, and 100% of the available samples. For each subset, training was repeated using different combinations of input channels, including RGB, NIR, and NDVI. These experiments were designed to analyse the influence of both the amount of labelled data and the type of spectral information on the final performance. When very small training sets were used (5% and 10%), the impact of stochastic factors such as weight initialisation and data shuffling became more significant. To reduce this effect, each configuration was trained five times with different random seeds, and the median IoU value was calculated. For the smallest subset containing 16 images (10%), the observed IoU standard deviation across runs was approximately 0.01. The results are summarised in table 4.2.

An important observation is that a model trained on the full dataset using RGB inputs alone achieves performance comparable to that of a model trained on only 5% of the data when additional spectral information in the form of RGB+NDVI is provided. Furthermore, models trained with RGB+NDVI consistently outperform those using RGB+NIR. Although NDVI is computed from the NIR and red channels, it represents a domain-specific feature that directly highlights vegetation characteristics. Based on these results, the RGB+NDVI configuration was selected for the subsequent experiments.

Table 4.2: IoU scores achieved on the test set, evaluated for different training set sizes and input modalities.

	IoU for crop (on test set)			
Training set percentage	**RGB**	**NIR**	**RGB + NIR**	**RGB + NDVI**
5%	0.7926	0.8038	0.8432	0.8558
10%	0.8137	0.8436	0.8623	0.8738
50%	0.8301	0.8448	0.8720	0.8820
100%	0.8569	0.8791	0.9046	<u>0.9098</u>

It should also be noted that the validation set contained 57 images and remained unchanged even for experiments using very small training subsets (as few as eight images). In practical scenarios, a reduction in training data would typically be accompanied by a smaller validation set. While the validation size does not affect the training procedure itself, it influences model selection through early stopping and performance monitoring. To investigate this effect, additional experiments were performed using reduced validation sets. The IoU metric was calculated at each training epoch for both large and small validation subsets. As expected, the overall learning trends were similar. However, smaller validation sets resulted in noticeably higher variability of the metric across epochs, which may lead to the selection of slightly overperforming or underperforming models and, on average, tends to result in marginally inferior model selection [190, 191]. To ensure a fair comparison in the main experiments and to isolate the effect of training set size, the validation set size was kept constant across all configurations.

Models trained on a reduced portion of the labelled dataset, operating on both RGB and NDVI input data, were next used to generate synthetic segmentation masks for the entire training set, including previously unlabelled images. Although such automatically generated annotations are expected to be less precise than manually prepared ground truth, they can still support the training of a model operating exclusively on RGB imagery. This strategy resembles the principle of knowledge distillation. The model responsible for producing a large number of artificial labels is referred to as the teacher, while the RGB-only model trained using these labels is called the student. The student model was trained using both the synthetic masks generated for the full unlabelled set (1865 images) and the original 10% subset of manually annotated data, which had already been used to train the teacher. While the use of synthetic data may introduce irreversible defects in the trained model [192], it is expected that the additional training data will improve the performance of the RGB-only model compared with training using the limited labelled subset alone.

Table 4.3 presents the performance of RGB student models trained with artificially generated masks, together with results obtained by the teacher model and by RGB models trained exclusively on manually annotated data. In addition to standard binary metrics, the multiclass IoU is also reported, where the background is treated as a separate class.

Table 4.3: Comparison of selected crop segmentation models, trained with different supervision strategies.

Model	**IoU multi-class**	**IoU crop**	**Precision crop**	**Accuracy crop**	**Recall crop**
RGB 10%	0.9023	0.8137	0.8907	0.9912	0.9040
RGB 100%	0.9250	0.8569	0.9157	0.9934	0.9303
Teacher (RGB and NDVI) 10%	0.9338	0.8737	0.9223	0.9942	0.9431
Student RGB (on unlabelled data) and 10% labelled	0.9214	0.8501	0.9085	0.9930	0.9297

The best student model achieved an IoU of 0.8501, which is close to the performance of the RGB model trained on the full labelled dataset (IoU = 0.8569) and substantially higher than the result obtained using only 10% of the labelled data (IoU = 0.8137). This demonstrates that the proposed approach enables an almost tenfold reduction in manual annotations while maintaining comparable segmentation accuracy. Example predictions of the best-performing student model, alongside the teacher model outputs and the ground-truth masks, are shown in fig. 4.4.

As a further step towards reducing annotation effort, the use of the foundation model SAM (*sam_vit_h*) was investigated. Instead of manually delineating full pixel-level masks, the annotator provides only bounding boxes enclosing individual crops. These bounding boxes, combined with RGB, NIR, or NDVI imagery, were used as prompts to automatically generate segmentation masks using SAM. The generated masks were subsequently treated as ground truth for training crop segmentation models. Since SAM expects three-channel RGB input, single-channel NIR and NDVI images were converted to pseudo-RGB format by duplicating the channel three times. Bounding box annotations were prepared for 10% of the training set, resulting in a total of 35 annotated boxes. Segmentation masks were generated separately using RGB, NIR, and NDVI inputs, and each variant was used to train a U-Net model. The results, presented in table 4.4, show performance obtained using only the 10% subset and without any original, manually created segmentation masks.

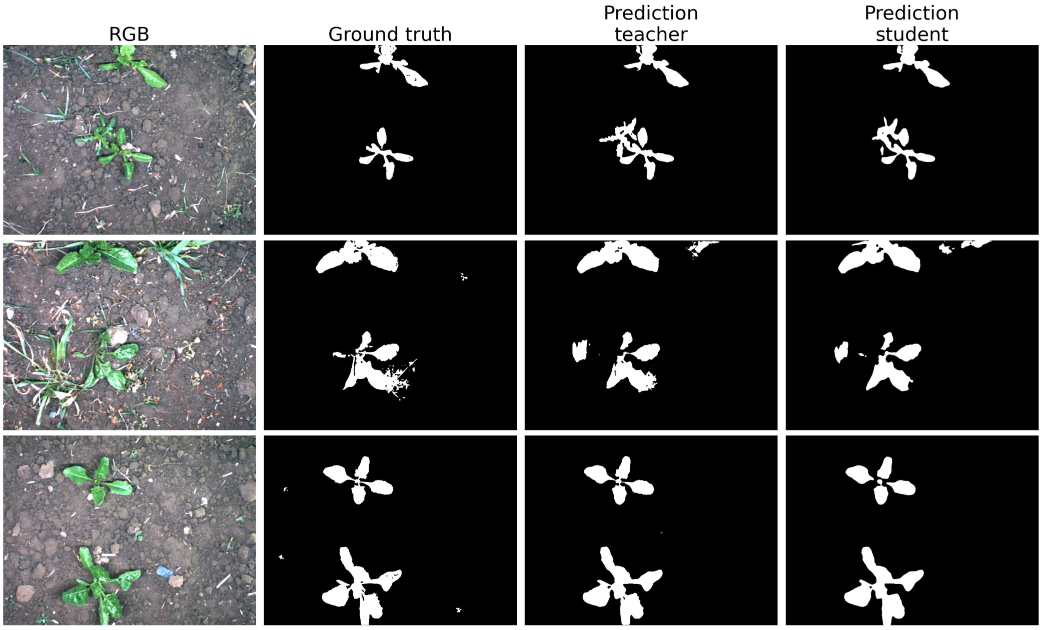

Figure 4.4: RGB image, ground truth segmentation mask, and model predictions on three example test images. The student model utilises only RGB images and was trained on 10% of the labelled training set (16 images) alongside 1865 images annotated by the teacher model. The teacher model, trained on the same 10% of labelled training data, incorporated NDVI data.

Models trained using SAM outputs derived from NIR or NDVI achieved better results than those based on RGB-derived masks. This suggests that, despite being trained primarily on RGB imagery, SAM can benefit from pseudo-RGB inputs originating from the NIR band, which contain more discriminative information for vegetation detection than RGB data alone. The obtained values can be directly compared with earlier experiments using original ground-truth masks, as presented in table 4.2. Notably, the performance gap between SAM-generated and manually annotated masks is relatively small. For example, for RGB+NDVI input, the model trained on SAM-generated masks achieved IoU = 0.8642, compared with 0.8737 when trained on original annotations.

Table 4.4: IoU for the crop class achieved by U-Net models trained using SAM outputs as ground truth.

	SAM inference on:		
Model	**RGB image**	**NIR image**	**NDVI image**
U-Net trained on RGB (10%)	0.8065	0.8113	0.8141
U-Net trained on RGB+NDVI (10%)	0.8361	0.8642	0.8662

Overall, the results of the conducted experiments demonstrate that multispectral imaging can be used not only to improve predictive performance directly, but also to substantially reduce the amount of manual annotation required to achieve a given level of segmentation accuracy compared with approaches based solely on RGB imagery. Furthermore, the proposed teacher-student framework showed that a multispectral model trained using a limited set of manually annotated data can leverage a larger set of unlabelled images to generate synthetic labels for training an RGB-only model. This approach enabled an almost tenfold reduction in the amount of required manual annotations while maintaining comparable performance. These findings show that multispectral data may provide significant benefits even when the final deployment system is restricted to RGB sensors due to practical constraints such as hardware cost. In such scenario, multispectral imagery can be used solely during the model development phase, while the deployed model operates exclusively on RGB data. Additionally, the experiments involving the SAM foundation model demonstrated its potential to streamline the manual annotation workflow. Overall, these results support the broader objectives of this thesis by improving the practical applicability of machine learning methods in agriculture through the reduction of annotation costs while maintaining high predictive performance.

4.2 Soil monitoring from space

Monitoring of agricultural crops and soil parameters remains an active research area, particularly with the increasing availability of multispectral and hyperspectral data. The high dimensionality and volume of such data introduce both opportunities and challenges. Beyond developing predictive models, it is essential to understand the underlying data structure and ensure model interpretability.

Soil characteristics encompass both physical and chemical properties and play a crucial role in agricultural decision-making, including fertilisation strategies and early contamination detection. Traditional methods for monitoring soil properties rely on in-field sampling followed by laboratory analysis. Although these methods provide high accuracy, they are inherently difficult to scale, resulting in limited spatial coverage. Remote sensing approaches that utilise spectral information offer an alternative for estimating soil parameters [193]. However, their accuracy remains limited, and existing research is relatively scarce, with particular emphasis on challenges related to generalisation and robustness. For instance, the authors in [194], hyperspectral imagery with 115 bands was used to estimate soil nitrogen content across 1297 samples, achieving a relative root mean square error of approximately 24%. In contrast, the authors in [195] employed only five multispectral bands to estimate soil organic carbon at the field level, achieving an error of around 10%. The authors in [196] investigated the behaviour of regression models for soil parameter estimation using hyperspectral satellite data, demonstrating that these models often rely on a small subset of spectral features, and that the number of utilised features can be significantly reduced without substantial loss in predictive performance. There is also ongoing research on related approaches that infer environmental contamination indirectly through plant monitoring rather than direct soil analysis. For example, the authors in [197] highlighted that combining hyperspectral imaging with machine learning enables non-destructive estimation and classification of heavy metal stress in plants.

To advance research in this area, a new dataset was acquired and curated, consisting of laboratory measurements of topsoil micronutrient content paired with multispectral and hyperspectral imagery. In total, the dataset includes 4098 ground-truth samples from agricultural fields located in the Greater Poland and Opole Voivodeships.

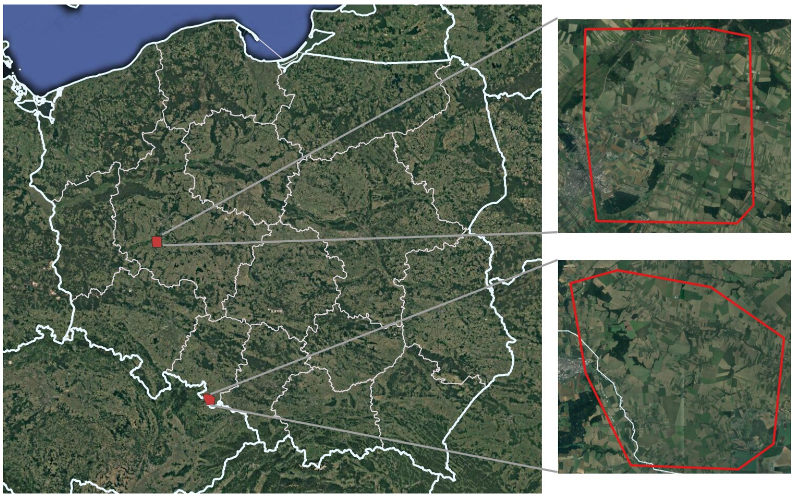

Figure 4.5: Location of measurement areas (red polygons) on the satellite map of Poland, with zoom-in on the area.

Ground-truth laboratory measurements cover concentrations of iron (Fe), zinc (Zn), boron (B), copper (Cu), sulphur (S), and manganese (Mn). For each field, 12 soil samples were collected across multiple sampling areas at a depth of less than 15 cm. These were subsequently mixed and analysed in the laboratory using the Mehlich 3 methodology. Corresponding imagery was acquired from multispectral (MSI) and hyperspectral (HSI) satellite platforms, complemented by aerial hyperspectral data collected from an aircraft cruising at a speed of 62 m/s, at an altitude of 2760 m above sea level. Detailed specifications of these sensing modalities are provided in table 4.5. The majority of the dataset was released as part of the Hyperview2 challenge [1], organised as part of the EASi workshop at the ECAI 2025 conference.

[1] https://challenges.philab.esa.int/portfolio/easi-workshop-hyperview2/

Both the imagery and the laboratory measurements were obtained in November, in order to minimise the impact of biomass.

Table 4.5: Comparison of sensing modalities used in the dataset and their specifications.

Property	HSI Satellite	MSI Satellite	HSI Aerial
Average patch size (pixels)	2.12	5.02	16.76
Number of bands	230	12	430
Spectral range [nm]	402–2490	443–2190	414–2357
Spatial resolution (GSD) [m]	30	10	2
Sensor name(s)	PRISMA payload	Sentinel-2 MSI	HySpex VNIR-1800 / SWIR-384

Example imagery of a selected field is presented in fig. 4.6.

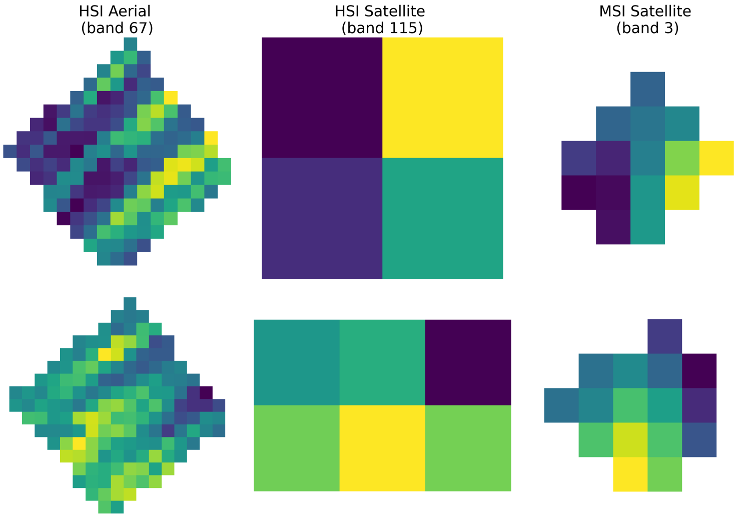

Figure 4.6: Example spectral imagery for selected fields, illustrating spatial variability using a single spectral band.

As an initial step, statistical analysis was performed to clean and prepare the dataset. Since the primary objective is regression-based estimation of element concentrations, which is sensitive to extreme values, outliers were removed. A sample was considered an outlier if it exceeded four standard deviations from the mean. Although the distributions of ground-truth measurements are heavy-tailed and deviate from normality, this threshold was found empirically to be effective. For example, sulphur concentration has a mean of 26.21 and a standard deviation of 14.19, yet values as high as 797.27 were observed. Such extreme values are likely attributable to laboratory errors rather than true anomalies. Even if valid, their rarity makes them unsuitable for effective model training. Additionally, duplicate samples from the same fields were removed. In total, 87 samples were excluded, resulting in a final dataset of 4014 samples. Figure 4.7 illustrates the distributions of iron and sulphur concentrations after outlier removal.

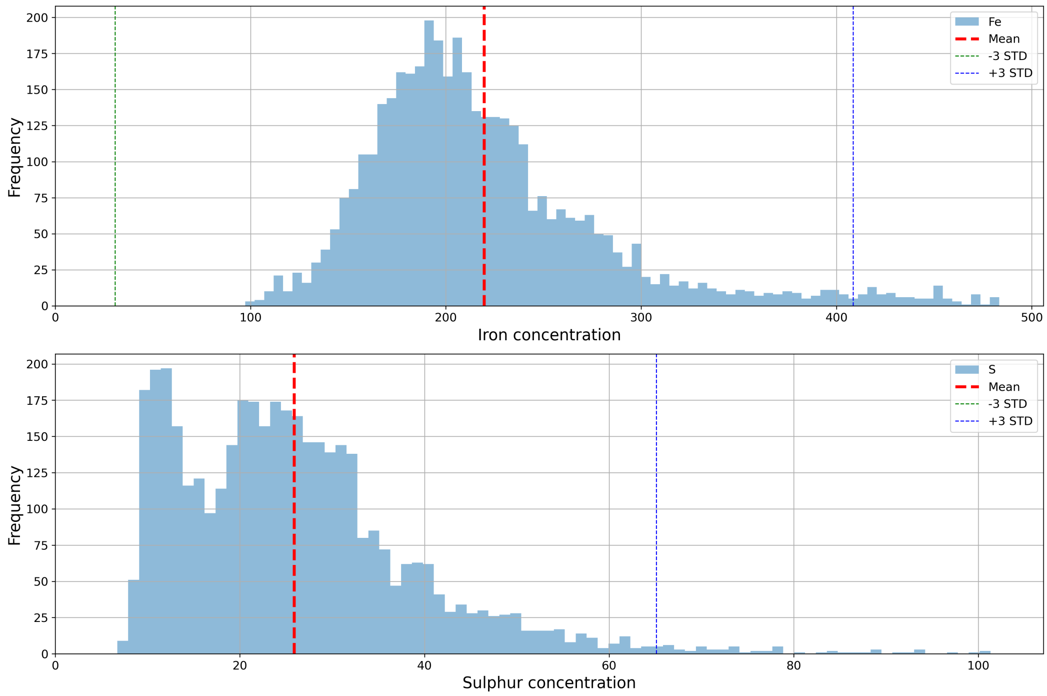

Figure 4.7: Distribution of ground-truth values for selected soil properties after outlier removal, shown for iron and sulphur concentrations.

As discussed in section 3.4.3, an essential aspect of dataset usability is the division into training, validation, and test sets. Several splitting strategies were therefore evaluated. A simple random split was deemed unsuitable, as it allows models to overfit by implicitly learning spatial correlations between samples and their associated target values. Although the monitored fields exist in a two-dimensional spatial domain, the dataset provides only a one-dimensional representation of spatial ordering, reflecting the sequence of sample acquisition. This representation limits the ability to fully eliminate spatial dependencies between the training and test sets. Nevertheless, it still captures meaningful locality information, allowing partial mitigation of spatial leakage when designing dataset splits.

A straightforward approach involves partitioning the dataset based on distinct spatial clusters while attempting to preserve the distribution of target values. This method ensures strong spatial separation between the training and test sets but introduces distribution shifts. Figure 4.8 illustrates this effect for iron concentration. In this setup, the validation and test sets share spatial clusters and are randomly split between them. While complete independence between validation and test sets is desirable, it is less critical than ensuring independence between the training and test sets, especially given the limited hyperparameter tuning performed.

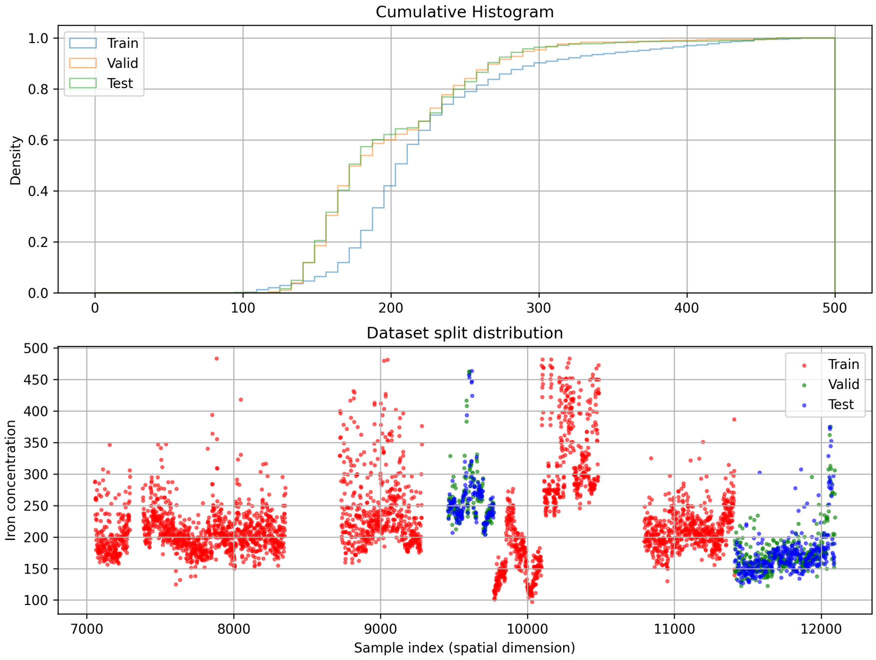

Figure 4.8: Dataset split based on spatial clusters. Top: cumulative distributions of iron concentration across training, validation, and test sets. Bottom: sequential ordering of samples with corresponding iron concentrations and assigned splits.

To mitigate the distribution mismatch, a stratification approach based on target values was introduced. This method aims to minimise differences in value distributions across splits while at the same time maximising spatial separation in the one-dimensional domain. The results of this strategy are shown in fig. 4.9. Although this approach yields well-balanced distributions, it introduces partial spatial leakage, as nearby samples may be assigned to different splits. This reflects a trade-off between statistical consistency and spatial independence.

For model training, both reflectance values and target concentrations were normalised to the range [0, 1] using statistics computed from the training set. Previous work from the first Hyperview challenge [198, 199] demonstrated that feature engineering can significantly improve performance, particularly for simpler models. Inspired by these findings, additional features on top of raw spectral reflectance values were constructed, including spatial averages of reflectance, spectral gradients, and representations derived from wavelet and Fourier transforms.

The experimental design aimed to evaluate the impact of input modalities, model architectures, dataset splitting strategies, and feature engineering. A range of

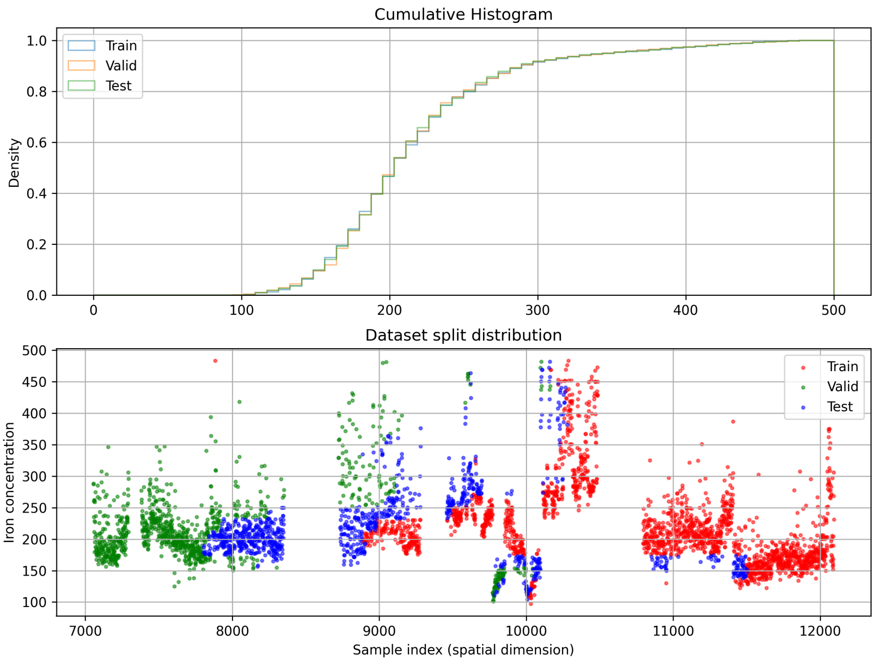

Figure 4.9: Dataset split using value-based stratification. Top: cumulative distributions of iron concentration across sets. Bottom: sequential ordering of samples with corresponding iron concentrations and assigned splits.

models was considered, including Random Forest, SVM, XGBoost, and transformer-based architectures. To assess predictive performance across all target elements, a composite score based on MSE was calculated according to the formula:

$$\text{score} = \frac{\sum_{i=1}^{6} \left(\frac{\text{MSE}_i}{\text{MSE}_{i-\text{base}}} \right)}{6}$$

where $\text{MSE}_{i-\text{base}}$ denotes the mean squared error of a baseline model that predicts, for each element independently, the mean value computed from the training data. Results for models using extended feature sets are presented in table 4.6 and 4.7.

Table 4.6: Model performance across data modalities using value-stratified dataset splitting.

Model	MSI Satellite	HSI satellite	HSI aerial
SVM (SVR with RBF)	0.621	0.515	0.515
Random Forest	0.457	**0.436**	0.409
XGBoost	0.516	0.475	0.475
ViT-L-14	**0.448**	0.462	**0.401**

Table 4.7: Model performance across data modalities using cluster-based dataset splitting.

Model	MSI Satellite	HSI satellite	HSI aerial
SVM (SVR with RBF)	0.863	0.924	0.815
Random Forest	0.816	**0.776**	**0.762**
XGBoost	0.802	0.798	0.814
ViT-L-14	**0.794**	0.811	0.767

Hyperparameters for each model class were selected using a coarse grid search, based on validation performance. Across models, copper (Cu) was consistently the most challenging element to predict, while boron (B) yielded the lowest errors. For example, a Random Forest model trained on MSI satellite data achieved an overall score of 0.457, with element-wise scores of 0.423 (Fe), 0.452 (Zn), 0.379 (B), 0.594 (Cu), 0.538 (S), and 0.439 (Mn). Figure 4.10 presents a comparison between predicted and ground-truth values for all elements obtained using a Random Forest model trained on MSI satellite data.

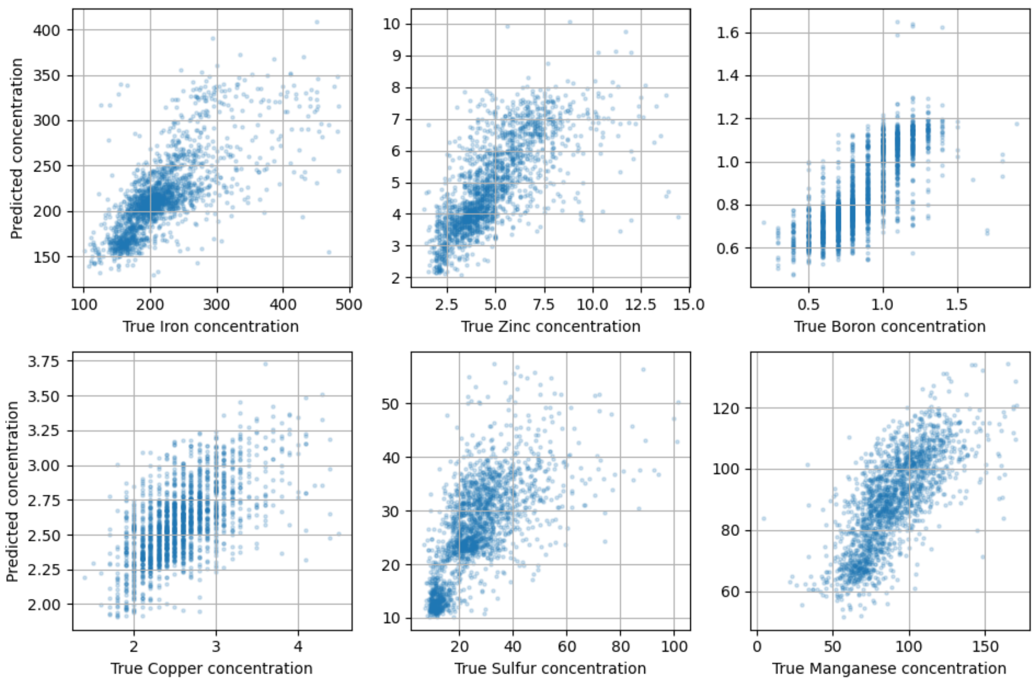

Figure 4.10: Predicted versus ground-truth concentrations for all elements using a Random Forest model trained on MSI satellite data with value-stratified splitting.

Combining MSI with HSI satellite and aerial data improved performance, reducing the Random Forest score from 0.409 to 0.387. Although transformer-based models achieved slightly better results, their computational cost was significantly higher. Given the limited spatial information available in the dataset, simpler models such as Random Forest remained competitive and were therefore selected for further analysis.

Feature engineering also proved to be beneficial. For a Random Forest model trained on HSI satellite data, incorporating engineered features improved the score from 0.489 to 0.436. This suggests that further feature design could yield additional gains. Performance on value-stratified splits was comparable to results achieved by participants in the Hyperview2 challenge, where top-performing teams reported scores between 0.364 and 0.494. These teams often employed additional optimisation techniques, such as training separate models for different spatial resolutions.

Finally, better performance was observed for value-stratified splits compared with cluster-based splits. This is likely due to a combination of partial data leakage and more similar distributions between training and test sets. Determining the relative contribution of these factors would require a larger dataset with more diverse spatial coverage and explicit two-dimensional spatial information, rather than relying on a single spatial index. This observation aligns with findings from [200], which highlight the limited availability and small size of publicly available soil datasets as a key challenge in improving predictive accuracy.

For the explainability analysis, the trained Random Forest model was used, with evaluation performed on the test set. Figure 4.11 presents the contribution of individual multispectral bands to model predictions, aggregated across all predicted elements. The analysis was conducted for both dataset splitting strategies (cluster-based and value-stratified) and without the use of extended features, relying solely on raw reflectance values. For both splitting approaches, bands 1 and 11 exhibited the highest importance, corresponding to wavelengths of 490 nm and 1610 nm, respectively. While the importance of bands varies across different target elements, a consistent subset of bands appears to be most informative. To further validate this observation, models were trained using subsets of bands ranked by their SHAP importance. A model trained on the most important bands (1, 11, 12, and 2) achieved a score of 0.54, whereas a model trained on the least important bands (9, 7, 8, and 5) achieved a substantially worse score of 0.76.

Figure 4.12 presents a similar analysis for hyperspectral satellite data, focusing on a single target variable (iron concentration). In this case, the feature set included both reflectance values and spectral gradients, with band importance aggregated across all derived features. The most influential bands were identified as bands 3, 99, 4, and 222, corresponding to wavelengths of 419 nm, 1339 nm, 427 nm, and 2435 nm. A model trained using only this subset of bands achieved a score of 0.40 for iron prediction, compared with 0.38 obtained when using all bands without extended features. The average reflectance curve reveals a noticeable anomaly around 1380

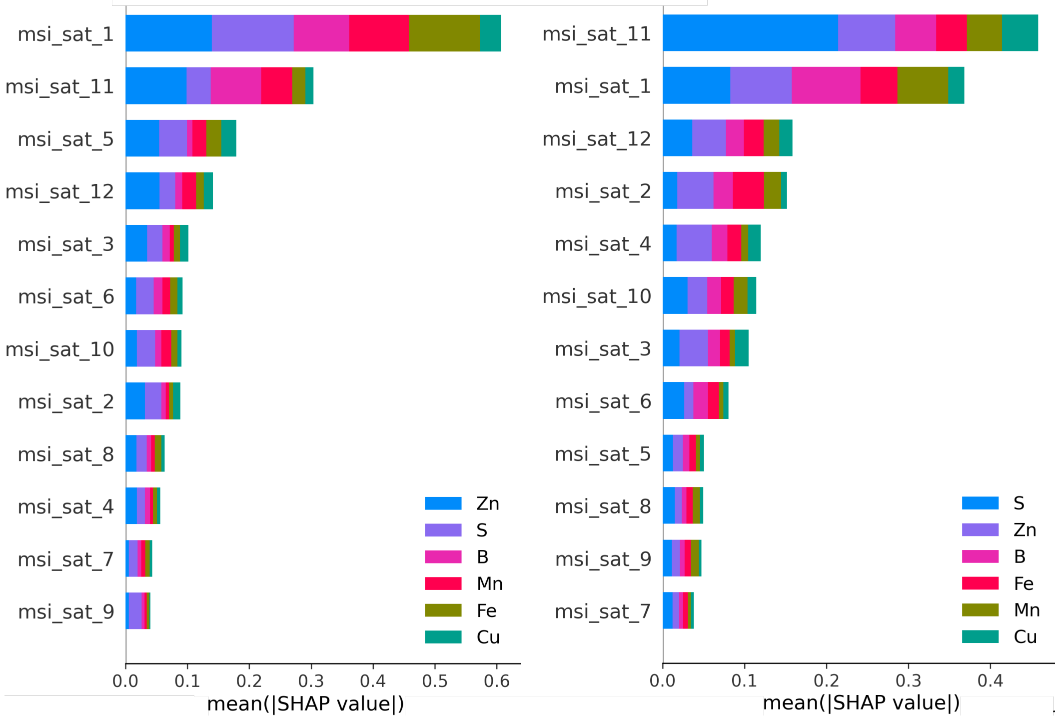

Figure 4.11: Mean absolute SHAP values of multispectral bands for the Random Forest model, aggregated across all predicted elements. Results are shown for cluster-based (left) and value-stratified (right) dataset splits, using raw reflectance features only.

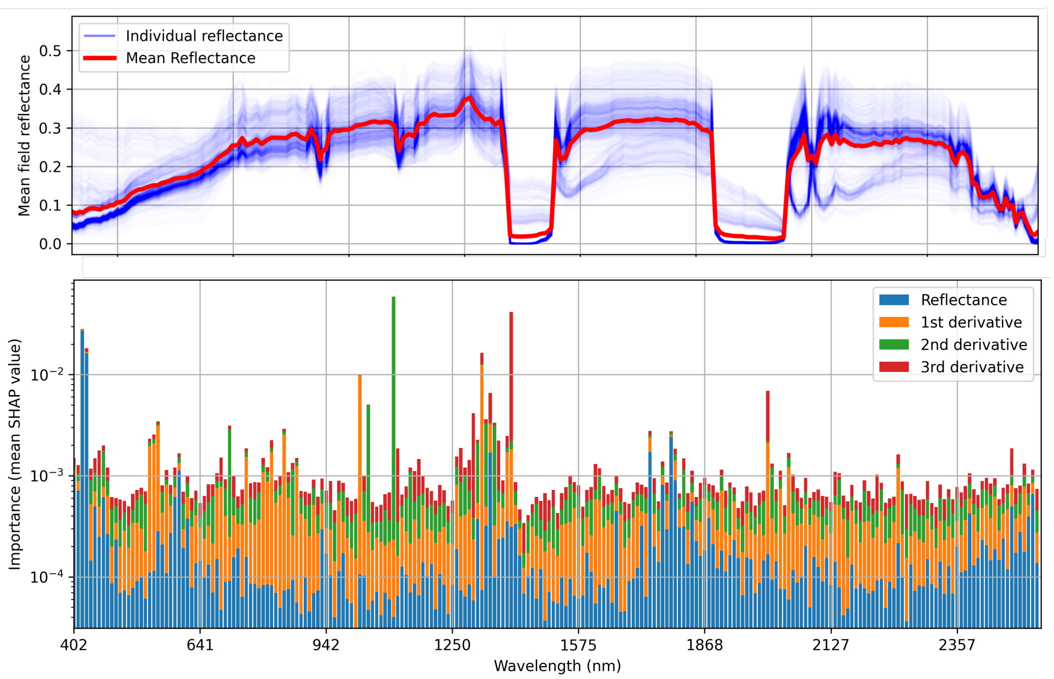

Figure 4.12: Average reflectance spectra (top) and corresponding mean absolute SHAP values (bottom) for hyperspectral satellite data, aggregated across reflectance and gradient features for iron concentration prediction.

nm and 1890 nm, which is most likely caused by water absorption effects. In these wavelength ranges, reflectance values for most fields are close to zero, suggesting that these bands carry limited useful information, which is consistent with their relatively low importance in the model. However, bands located at the boundaries of these regions appear to have higher importance, although this may be due to leakage between the training and test sets.

The above experiments demonstrate that both multispectral and hyperspectral satellite imagery can be utilised for soil parameter estimation. The obtained results indicate that spectral data contain information relevant to the prediction of soil micronutrient concentrations, thereby supporting the broader hypothesis that spectral imaging combined with machine learning constitutes a valuable tool for agricultural monitoring. Nevertheless, the observed sensitivity of model performance to the dataset splitting strategy highlights the need for larger and more geographically diverse datasets that better capture spatial dependencies and mitigate distribution shifts. Consequently, model generalisation remains one of the principal challenges in the development of soil monitoring systems.

Furthermore, due to the relatively low spatial resolution of the satellite imagery, simpler machine learning models were found to achieve predictive performance comparable to that of more complex deep neural network architectures, while simultaneously offering partial interpretability, albeit at the cost of additional feature engineering. Model explainability was investigated using SHAP values, which revealed that only a limited subset of spectral bands contributes substantially to predictive performance. The identification of the most informative spectral bands made it possible to train a model using only a reduced subset of bands while maintaining performance close to that achieved when using the full spectral data. This finding supports the second part of the research hypothesis. It should, however, be noted that the interpretation of feature importance in spectral data is not always unambiguous, as neighbouring spectral bands are often highly correlated.

4.3 Data processing tool - Deepness

As the creation and use of neural network models with geospatial data is not a trivial task for scientists without programming expertise, the capabilities of modern machine learning are often underutilised by GIS specialists. Although GIS tools offer extensive visualisation and spatial analysis capabilities, the integration of machine learning and deep learning methods for remote sensing into typical GIS workflows remains limited, particularly outside expensive proprietary software.

To address this gap, an open-source plugin for QGIS called Deepness was developed [201]. The main idea behind this tool is to democratise access to the potential created by deep learning by facilitating its application to remote sensing data without requiring expert knowledge of machine learning or programming. The plugin enables the application of deep neural network models directly within the GIS environment, allowing users to perform model inference on any georeferenced multispectral raster data without leaving the familiar QGIS workflow. The operating principle of the plugin is presented in fig. 4.13.

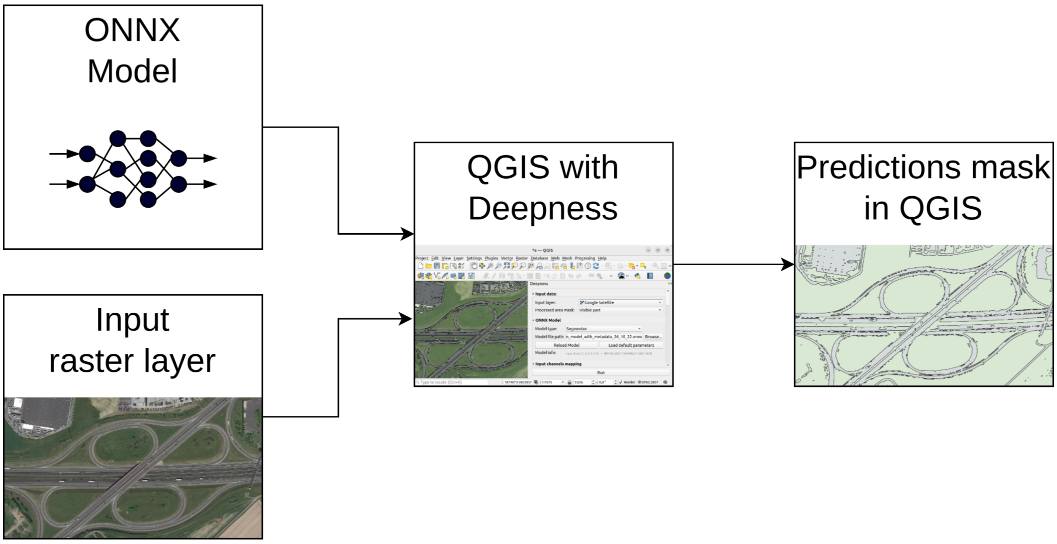

Figure 4.13: Deepness plugin operation principle diagram. A georeferenced raster and a trained deep learning model (ONNX format) are used as inputs to generate a prediction layer.

The processing pipeline can be summarised as follows:

- The user loads a georeferenced raster layer in QGIS (e.g., a custom orthophoto).
- The user selects a trained neural network model and specifies the processing parameters.
- The plugin divides the raster data into smaller tiles compatible with the model input size.
- The model performs inference on each tile.
- The predictions are merged and saved as a new QGIS layer while preserving geospatial information.
- The user can further interact with and analyse the results within QGIS.

A wide range of model types is supported, enabling the execution of the tasks described in section 2.6, most notably object detection, segmentation, regression as well as superresolution. Object detection presents a particular challenge due to the large number of incompatible models from the YOLO family that are commonly used. These models may produce different types of outputs, including rectangular bounding boxes, oriented bounding boxes, or instance segmentation masks. Models are required to be provided in the ONNX (Open Neural Network Exchange) format, which also allows embedding metadata used during processing, such as the expected GSD expressed in centimetres per pixel, tile overlap ratio or confidence thresholds. To improve computational performance, the plugin also supports GPU-accelerated processing. Figure 4.14 presents example results obtained using several trained models.

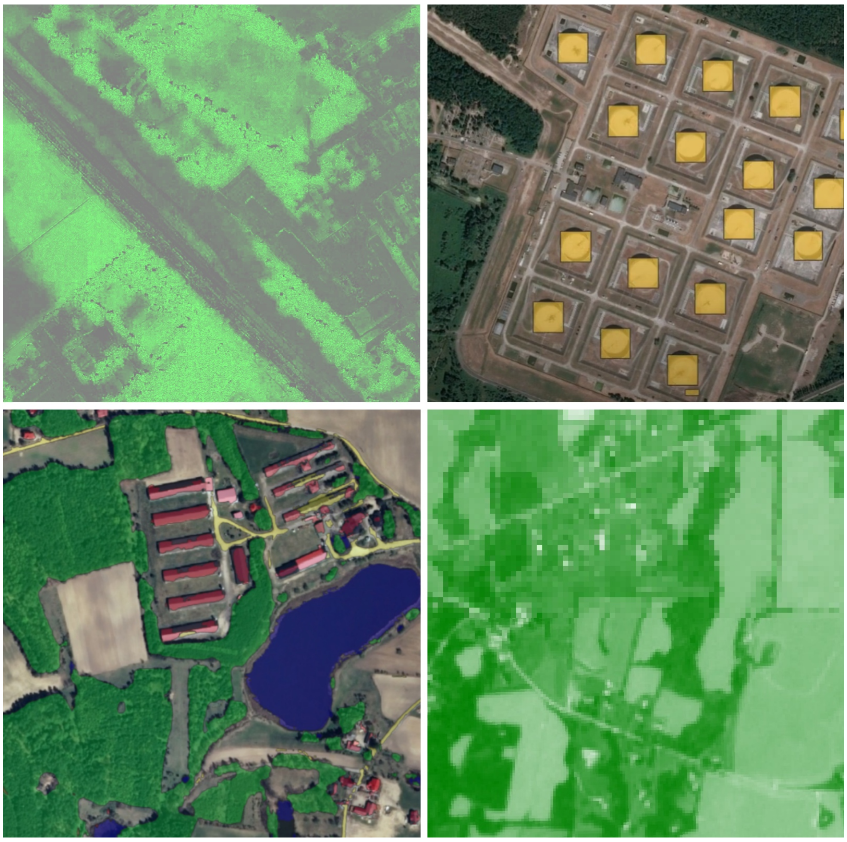

Figure 4.14: Example outputs generated with the Deepness plugin: biomass regression, oil tank detection, satellite land-cover segmentation and Sentinel-2 superresolution.

The plugin automatically handles the complexity associated with processing georeferenced data. This includes coordinate system transformations, spatial scaling, tile generation, result stitching, tile overlap handling, prediction thresholding, merging of results, and conversion to QGIS-compatible layers. Any combination of input channels is supported, with configurable mapping between multispectral image channels and model inputs, allowing for flexibility in model design and data sources, which often provide spectral channels in different orders and formats. The architecture of the plugin allows processing of large datasets, potentially reaching terabyte scale, by dividing the input data into tiles that are processed sequentially. The input data source may be either a local file or data streamed from a remote server. From the perspective of practical deployment, the plugin helps minimise the operational costs associated with machine learning models by reducing the effort required for users with large collections of geospatial data to integrate such models into their existing workflows.

The Deepness plugin is easy to install within the QGIS ecosystem, free to use (under the Apache 2.0 licence), and accompanied by extensive documentation for both end users and developers creating custom models. Researchers and developers are also encouraged to extend the plugin with new functionality, and many have already contributed improvements to the project.

The plugin has been enthusiastically received by the GIS community worldwide, reaching over 75000 downloads from the QGIS plugin repository and nearly 200 stars on GitHub. Alongside the plugin, a registry of publicly available trained models is maintained, including models for applications such as land-cover segmentation, tree detection, ship detection, road segmentation, and solar panel detection. Some of the models work with multispectral data, primarily Sentinel-2 imagery, although RGB-based models constitute the majority.

Apart from enabling model inference, the plugin documentation also provides detailed instructions on how to prepare custom models. These guidelines include examples implemented in Python, as well as workflows that allow model preparation without programming knowledge using web-based tools, based on instructions contributed by one of the users. The plugin has also encouraged researchers from a wide range of disciplines to adopt machine learning techniques in their work. Reported applications include counting penguin nests, monitoring tree health, and detecting archaeological structures. Some users have also contributed their own trained models, extending the available collection within the model zoo.

Beyond its technical capabilities, the Deepness plugin helps bridge the gap between advances in machine learning research and their practical application in agriculture and environmental protection, where the analysed data originate from satellite or UAV platforms, supporting both multispectral and RGB-based models. By reducing the level of programming expertise required to utilise machine learning models, it makes state-of-the-art methods more accessible to GIS specialists and other domain experts. The plugin supports efficient inference through GPU acceleration while accounting for the hardware constraints of desktop computers, enabling the processing of large geospatial datasets on typical personal computers. Furthermore, the accompanying model registry promotes research reproducibility and comparative evaluation by allowing users to rapidly verify and benchmark models developed by other researchers on new datasets and application scenarios.

4.4 Crop damage estimation

Corn crop damage caused by wildlife constitutes a significant problem in agriculture, as it causes crop and yield loss. Accurate assessment of such damage is important not only for estimating economic losses but also for monitoring wildlife activity and supporting compensation procedures. Traditional approaches rely primarily on manual ground-level inspection of fields. However, these methods are time-consuming and often inaccurate, particularly when crops are fully developed and the damage cannot be easily assessed from the ground. Imagery collected using UAVs provides a promising alternative for large-scale crop monitoring. UAV-based imaging enables rapid data acquisition over large areas while maintaining high spatial resolution. Nevertheless, when the analysis relies on manual annotation of the damaged areas in aerial imagery, the process remains labour-intensive. The use of deep learning image segmentation offers an opportunity to automate this process. This subsection describes an automated approach for detecting wildlife-induced crop damage using deep learning-based segmentation of aerial imagery. The objective of the method is to identify damaged areas within corn fields and estimate the total percentage of affected crops. The research was conducted using a corn crop damage dataset. The dataset consists of data from 14 individual maize fields in western Poland that were partially damaged by wildlife. The dataset includes both RGB and NDVI orthophotos for selected fields, captured using a Sentera multispectral camera mounted on a UAV. The flights were conducted at an altitude of 100 m, resulting in a GSD of approximately 3 cm per pixel. The dataset was used to assess the extent of crop damage for insurance and compensation purposes. Ground-truth data were provided as manual annotations of the damaged areas. Two types of annotations were used: polygonal regions representing larger damaged areas and point annotations representing small isolated damage spots. For the purpose of training segmentation models, these annotations were converted into a common binary mask, representing the presence of crop damage. Several deep learning architectures were evaluated for the segmentation task. Among the tested models, the U-Net++ architecture and the transformer-based SegFormer model achieved the best predictive performance. Example predictions of the trained models compared with manual annotations are presented in fig. 4.15.

The influence of tile size used during training was also investigated. The experiments showed that the best performance was obtained for tiles of size 512 pixels, while smaller (256 px) and larger (768 px) tiles resulted in inferior segmentation metrics. This result suggests that the selected tile size provides a suitable balance between capturing sufficient spatial context and limiting model complexity, while still providing an adequate number of training samples.

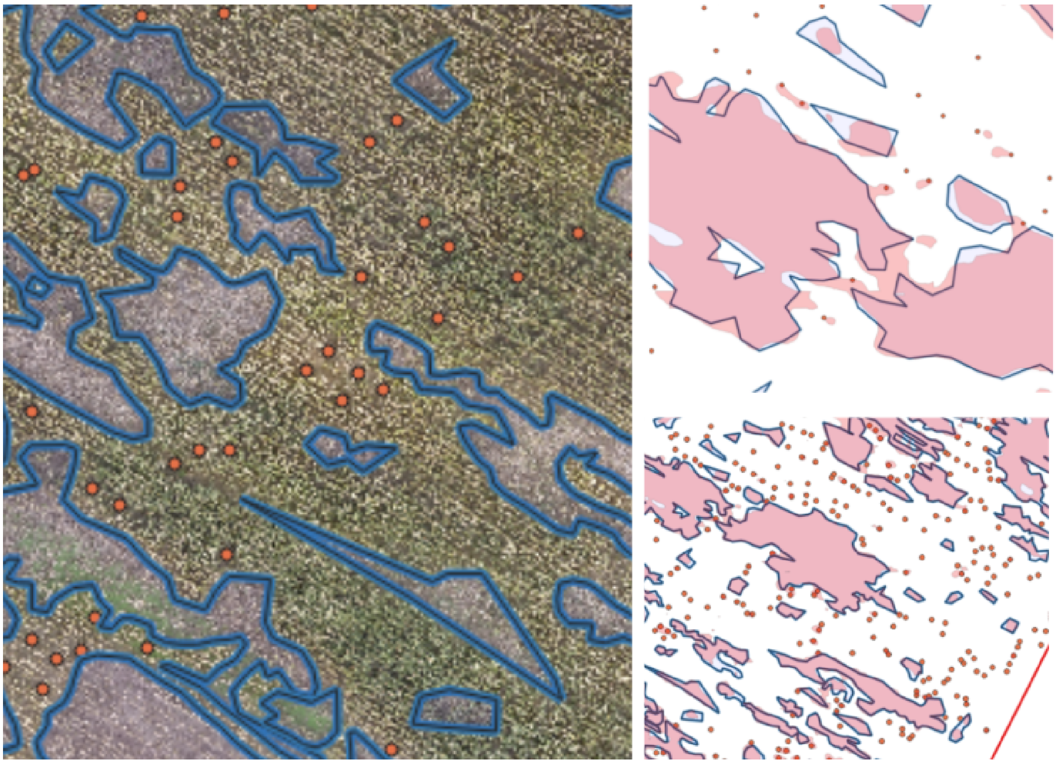

Figure 4.15: Example predictions of crop damage segmentation (pink polygons) compared with ground-truth annotations (azure polygons and red dots).

Another aspect analysed in this study was the impact of spectral information on segmentation performance. The models were trained using three different input configurations: RGB imagery only, NDVI imagery only, and a combination of RGB and NDVI channels. The results of these experiments are presented in table 4.8.

Table 4.8: Comparison of segmentation performance for different input channel configurations.

Input	IoU (total area)	F1-score (damaged area)	Precision (damaged area)	Recall (damaged area)
Only RGB (3 channels)	0.835	0.335	0.423	0.372
Only NDVI (1 channel)	0.797	0.278	0.311	0.288
RGB and NDVI (4 channels)	0.868	0.349	0.680	0.311

The combination of RGB and NDVI data achieved the best IoU metric, although the improvement over RGB-only input was relatively modest. It should be noted that the NDVI data available in the dataset had a significantly lower spatial resolution, with a GSD of approximately 50 cm per pixel, compared with 3 cm per pixel for the RGB imagery. As a result, the NDVI layer required interpolation to match the resolution of the RGB data before it could be used as an additional input channel. The limited usefulness of NDVI-only input can be explained by the underlying interpretation of this vegetation index, as it primarily indicates the presence of live green vegetation, and therefore it may fail to distinguish recently damaged crops from healthy plants. Consequently, relying solely on this index may be effective mainly for identifying areas where vegetation damage occurred earlier and the plants have already deteriorated.

During the experiments, several challenges related to real-world agricultural conditions were encountered. These issues, discussed in more detail in section 2.7.1, include not only differences between fields but also variability within a single field, such as different stages of crop growth or varying time since the crop damage occurred. Moreover, the manual annotations were not always fully consistent. In some areas, it was difficult to distinguish actual crop damage from locations where the plants had either not grown or had developed poorly. In addition, the clearances between crop rows were only occasionally annotated as damaged areas. In other cases, particularly for recent and relatively minor damage, it was difficult to determine an unambiguous boundary between damaged and healthy crops, as the transition between them was gradual. Examples of such challenging cases are presented in fig. 4.16.

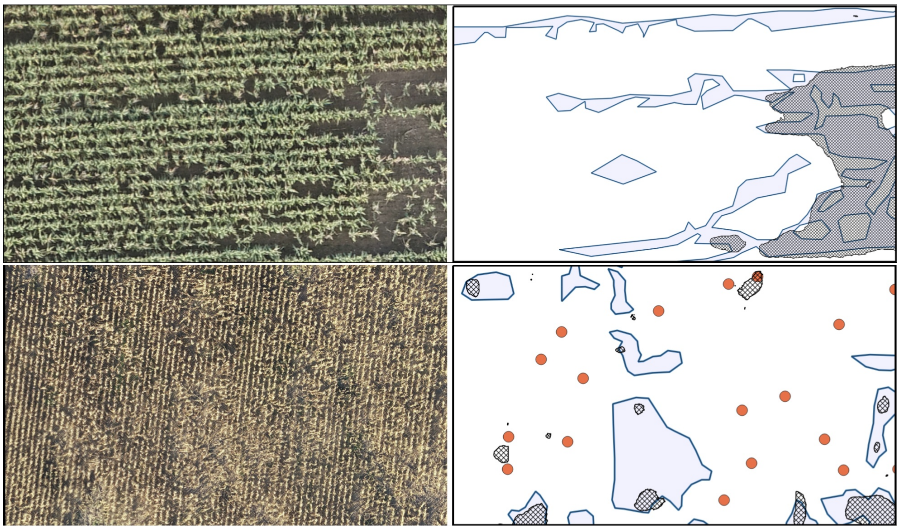

Figure 4.16: Examples of fields with annotation inconsistencies and areas where crop damage is difficult to distinguish. Left: fragment of a problematic field; right: manual area annotations (azure polygons), smaller damage regions marked as red points, and model predictions of damaged crops represented with black hatching.

From a practical perspective, the most important metric is the accuracy of estimating the total percentage of damaged crops. In an evaluation scenario where the model was trained on all fields except one and then tested on the unseen field, the proposed method achieved an average relative error of approximately 18% in estimating the damaged area.

These findings suggest that, for this particular application, high-resolution RGB imagery combined with deep learning-based segmentation can already provide strong performance. Multispectral information, represented here by the NDVI channel, can further improve the results, even when the multispectral data are available at a lower spatial resolution. At the same time, the results highlight the importance of evaluating the trade-off between potential performance improvements and the increased cost and complexity associated with multispectral sensing systems.

4.5 Band reduction

This section focuses on the practical aspects of multispectral band selection and evaluates its impact on a binary image classification task. In particular, the study compares the performance of models trained on full multispectral data against those using reduced band combinations, including RGB-only inputs.

The dataset used in this research was introduced by [202]. It consists of 6726 multispectral images of coffee leaves, annotated for the presence or absence of lesions caused by the fungus *Hemileia vastatrix* (coffee rust). The data were collected from Colombian farms with the aim of supporting the development of predictive models for early disease detection, potentially reducing economic losses for farmers. Example RGB images are shown in fig. 4.17. While some infected leaves exhibit clear visual symptoms, others remain difficult to distinguish.

Figure 4.17: Example RGB images from the coffee leaf dataset, showing healthy leaves (top) and leaves infected with coffee rust (bottom).

All images were captured in a controlled environment with consistent lighting conditions and precise alignment between RGB and multispectral modalities, eliminating the need for co-registration. The dataset includes five multispectral bands:

blue (450-500 nm), green (500-620 nm), red (620-750 nm), red-edge (~840 nm), and NIR (750-900 nm). RGB images are treated as a three-channel input, resulting in a total of eight input channels. The spatial resolution is approximately 171 × 294 pixels, with varying leaf coverage across images. A comparison of different modalities is presented in fig. 4.18.

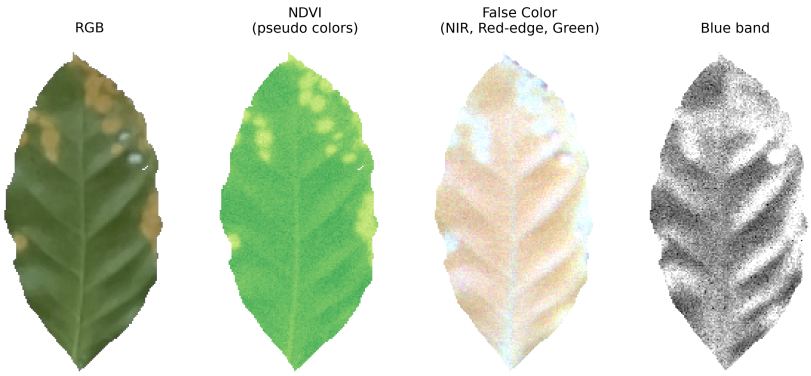

Figure 4.18: Comparison of image modalities for a single coffee leaf affected by rust.

The experiments aim to assess both the importance of multispectral information and the effect of band reduction on model performance and training stability. Due to the relatively small dataset size and the binary nature of the labels, repeated training was necessary. As the dataset is imbalanced, both loss functions and evaluation metrics were weighted according to class prevalence. Data augmentation was applied consistently across all bands. Geometric transformations were jointly applied to all channels to preserve spatial alignment, while intensity jitter and noise were applied independently to each band, followed by normalisation. Training employed early stopping and cosine annealing of the learning rate, with initial learning rates tuned per architecture. For models with more than three input channels, pretrained weights in the first convolutional layer were adapted by averaging across channels. Although suboptimal, this approach provided better initialisation than random weights.

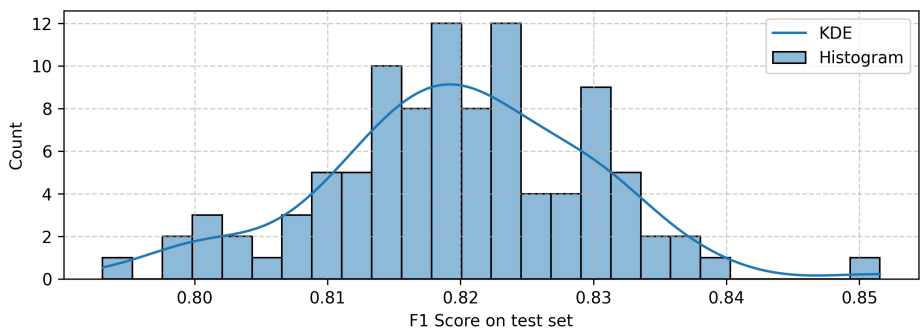

Figure 4.19: Distribution of test F1-scores over 100 training runs with different random seeds for the Swin-Tiny model using all input bands.

Preliminary experiments revealed that training randomness (e.g., initialisation, data order, augmentation) significantly affects model performance. To quantify this effect, 100 independent training runs were performed for a fixed dataset split using different random seeds for training. As shown in fig. 4.19, the Swin-Tiny model trained on all bands achieved a mean F1-score of 0.8198 with a standard deviation of 0.0099. This variability confirms that single-run evaluations are unreliable for this dataset. Therefore, all subsequent experiments employed 20-fold cross-validation. Additionally, for each fold, training was repeated with two different random seeds, resulting in 40 independent runs per configuration. Due to the lack of previously published benchmarks for this dataset, direct comparison with other works is not possible.

As an initial step, several modern architectures with comparable parameter counts were evaluated. Both the mean F1-score and its variance across runs were considered. Figure 4.20 presents the F1-scores achieved using all eight bands. The Swin-Tiny model achieved the best overall performance, with both the highest mean score and the lowest variance. Similar trends were observed for RGB-only inputs. However, ConvNeXT-Tiny exhibited notably lower variance in this setting, indicating more stable training with fewer input channels. Model sizes and inference times are summarised in table 4.9. Based on these results, the Swin-Tiny architecture was selected for further experiments.

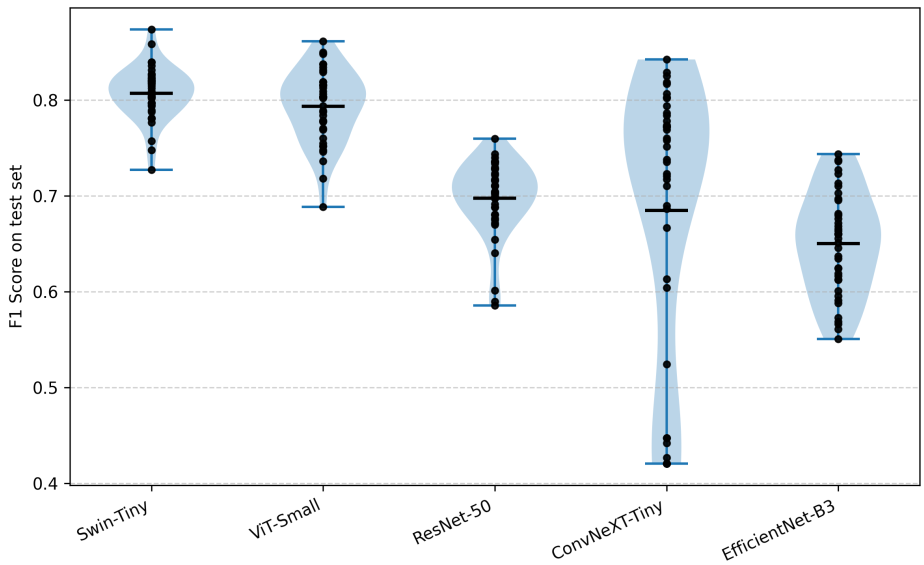

Figure 4.20: Comparison of test F1-scores for different model architectures trained on all input bands, evaluated using 20-fold cross-validation with two seeds per fold.

To analyse spectral band importance using theoretical methods, an average-performing Swin-Tiny model trained on all bands (F1-score of 0.821) was selected. The corresponding confusion matrix is shown in table 4.10. Figure 4.21 presents occlusion maps and attention rollout visualisations, illustrating the regions of the input image that contribute most to the model's predictions.

Table 4.9: Comparison of model architectures in terms of parameter count, inference throughput, and single-image latency.

Model	Params (M)	Throughput (img/s)	Single-image (ms)
ConvNeXT-Tiny	27.83	2170.7 ± 15.6	5.20 ± 0.149
EfficientNet-B3	10.70	2732.7 ± 93.7	12.24 ± 0.45
ResNet-50	23.53	2648.5 ± 16.0	5.72 ± 0.32
Swin-Tiny	27.53	1646.5 ± 73.9	9.48 ± 0.27
ViT-Small	22.16	2614.0 ± 46.5	4.55 ± 0.15

Table 4.10: Confusion matrix for the average-performing Swin-Tiny model trained on all spectral bands. Rows: true labels.

True / Predicted	No Rust	Rust
No Rust	37	9
Rust	16	106

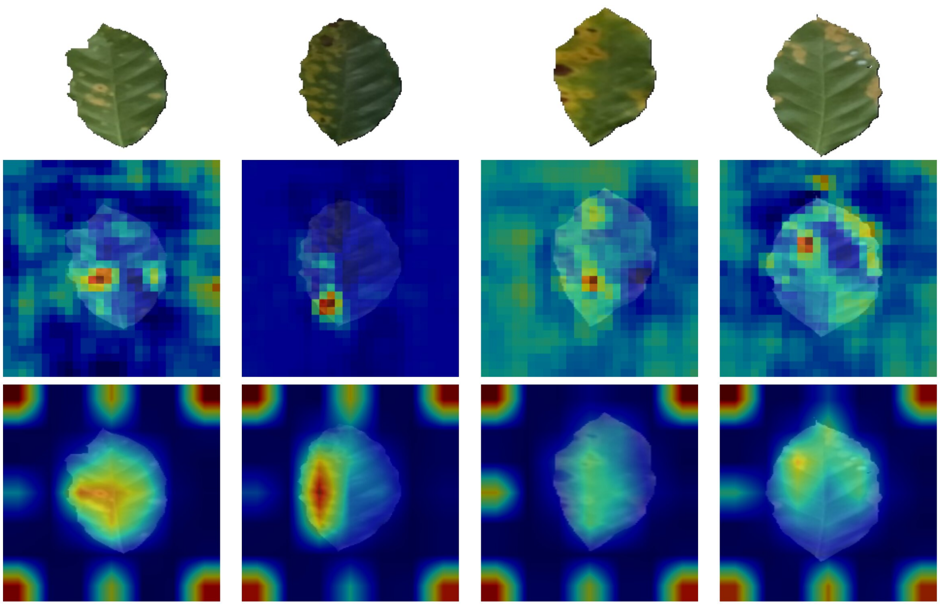

Figure 4.21: XAI visualisations for rust classification using the Swin-Tiny model trained on all spectral bands. The top row shows the RGB input images, the middle row presents occlusion maps, and the bottom row displays attention rollout highlighting regions contributing to model predictions.

Band importance was evaluated using four methods:

- band ablation, measuring performance degradation when a band is removed;

- gradient-based saliency, indicating local sensitivity of the output to band variation;
- integrated gradients, estimating contribution to predictions;
- SHAP values, treating each channel as a feature group to measure its importance.

The results are presented in fig. 4.22. RGB channels were treated as a single feature group (despite consisting of three channels), which explains their dominant importance. Among the multispectral bands, NIR and green showed the highest relevance. However, these methods assume a degree of feature independence, which is violated due to strong correlations between bands. Consequently, they do not directly provide an optimal band selection strategy. Additionally, importance rankings varied slightly across different training runs, further limiting their reliability.

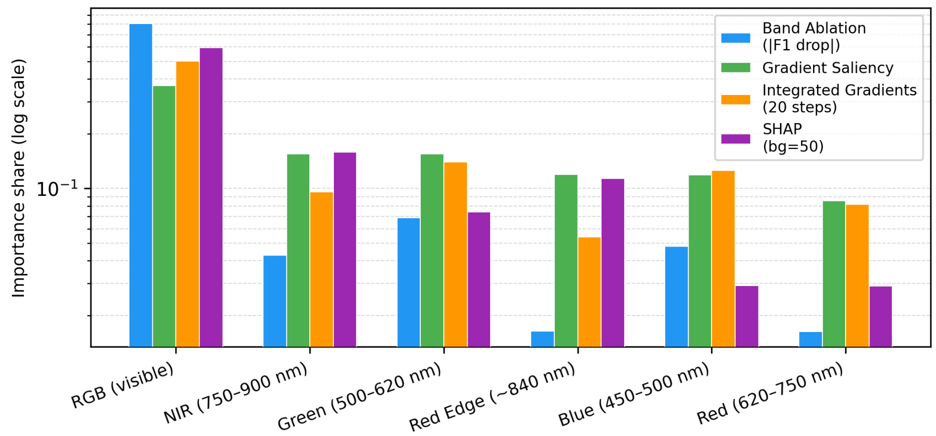

Figure 4.22: Comparison of spectral band importance estimated using four methods for the Swin-Tiny model trained on all bands. The RGB channels are treated as a single feature group.

To obtain more tangible evidence of band importance, models were trained on different combinations of bands. Similarly to the earlier experiments, each configuration was evaluated using 20-fold cross-validation with two seeds per fold. The tested configurations included: full multispectral input, RGB combined with one or two additional bands, multispectral-only inputs, and NDVI derived from other bands. Quantitative results are presented in table 4.11, with a visual comparison in fig. 4.23.

The best performance was achieved using all bands augmented with NDVI. Notably, models trained on only four channels (RGB + one multispectral band) achieved performance comparable to that of the model trained on all eight bands. All combinations that included both RGB and at least one additional band outperformed RGB-only models. In contrast, models trained solely on multispectral bands performed worse, with results comparable to those obtained using NDVI alone. These

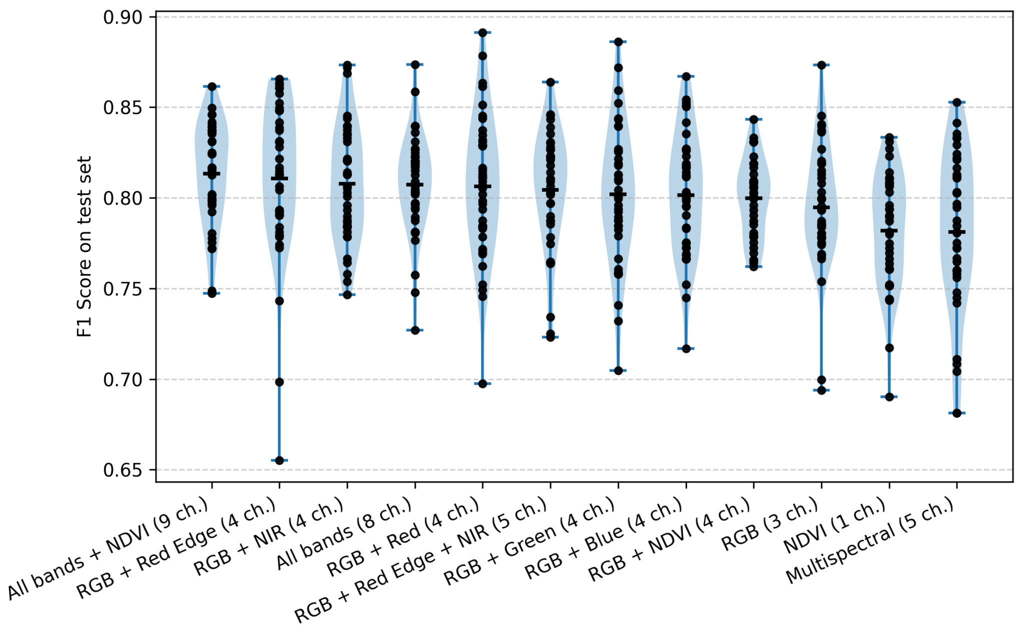

Figure 4.23: Comparison of test F1-scores for the Swin-Tiny model across different input band combinations, evaluated using 20-fold cross-validation with two random seeds per fold.

Table 4.11: Performance comparison of selected input band configurations for the Swin-Tiny model. Reported metrics include the mean and standard deviation over 20-fold cross-validation with two seeds per fold.

Input Bands	F1-Score	Accuracy	Precision	Recall
All bands + NDVI (9 ch.)	<u>0.813</u> ± 0.028	<u>0.853</u> ± 0.022	<u>0.801</u> ± 0.030	<u>0.843</u> ± 0.032
RGB + Red Edge (4 ch.)	0.811 ± 0.044	0.851 ± 0.030	0.798 ± 0.046	0.839 ± 0.043
RGB + NIR (4 ch.)	0.808 ± 0.032	0.847 ± 0.024	0.794 ± 0.037	0.840 ± 0.031
All bands (8 ch.)	0.807 ± 0.028	0.852 ± 0.019	0.798 ± 0.026	0.828 ± 0.039
RGB + Red (4 ch.)	0.806 ± 0.039	0.846 ± 0.029	0.794 ± 0.040	0.840 ± 0.042
RGB + Green (4 ch.)	0.802 ± 0.038	0.843 ± 0.032	0.791 ± 0.042	0.833 ± 0.037
RGB + Blue (4 ch.)	0.802 ± 0.033	0.845 ± 0.027	0.793 ± 0.038	0.827 ± 0.033
RGB (3 ch.)	0.795 ± 0.034	0.836 ± 0.022	0.782 ± 0.035	0.831 ± 0.041
Multispectral (5 ch.)	0.781 ± 0.043	0.840 ± 0.026	0.791 ± 0.043	0.791 ± 0.057

findings show that reducing the number of multispectral bands is possible without significantly degrading the results. In some cases, reduced input dimensionality even improved performance, suggesting that the model can learn more effectively from the remaining bands.

The reduction in the number of input channels also decreases inference time, both in batch-processing and single-image latency scenarios. The exact speedup depends on the model architecture, implementation details, and hardware configuration, particularly the dominant system bottlenecks. Nevertheless, all measurements were performed on the same hardware setup that was used during training. For CNN-based architectures, where the reduction in feature channels primarily affects the initial convolutional layers, decreasing the number of input channels from eight (all channels) to four (RGB + Red Edge) resulted in an approximately 11% reduction in inference time. However, for the Swin-Tiny architecture, the impact of reducing the number of input channels was smaller, reaching approximately 3%. This is because, in transformer-based models, the input projection stage constitutes only a small fraction of the overall computational cost compared to the transformer blocks, particularly the self-attention and feed-forward multi-layer perceptron operations. Moreover, reducing the number of input channels decreases not only the inference time itself, but also the data loading time from storage or camera equipment and the memory transfer time to the GPU. In these cases, the achieved speedup is nearly proportional to the reduction in the number of channels. This aspect is particularly relevant for embedded platforms, where data transmission and memory bandwidth limitations may contribute more to the total processing time than the model inference itself. Furthermore, reducing the number of input channels creates additional opportunities for model simplification. A lower-dimensional input space may enable the extraction of meaningful features using smaller and less computationally demanding architectures, potentially leading to further reductions in inference time.

The observed behaviour is consistent with the curse of dimensionality, also known as the Hughes phenomenon, which states that for a fixed number of training samples, the predictive performance of a classifier initially increases with the number of input dimensions but, beyond a certain point, begins to deteriorate. This effect is commonly observed in hyperspectral data [203, 204]. Moreover, the spectral bands are highly correlated and therefore partially redundant. This property has also been exploited in prior work [204], where inter-band correlation was used to guide the selection of the most informative bands. There is also a noticeable correspondence between band importance estimated using gradient-based saliency methods and the results obtained from training models on different band combinations. However, as discussed earlier, these theoretical importance measures do not directly translate into final model performance. The accuracy of classifying a single leaf alone is likely insufficient for practical applications, and would likely be even lower in real-world conditions than in the controlled environment used in this dataset. However, practical systems could leverage information from multiple nearby leaves to improve overall prediction reliability.

The conducted experiments demonstrated that combining multispectral bands with RGB imagery leads to superior predictive performance compared with models relying solely on RGB data for the classification of individual coffee leaves. Spectral band importance was analysed using both explainability methods and empirical evaluation through repeated model training. This analysis made it possible to identify an effective four-band configuration that retained most of the relevant information and even achieved slightly higher predictive performance than model trained using all available bands. Consequently, the proposed band reduction strategy decreases sensing, storage, and computational requirements, which directly supports the second part of the research hypothesis. Deep learning models are well suited for this type of task, as they can effectively exploit the rich spatial information present in leaf imagery. Furthermore, explainability techniques can be employed to identify the image regions that contribute most strongly to model predictions, thereby increasing the transparency and building trust in such systems.

5 Discussion

The findings presented in this dissertation support the main research hypothesis, demonstrating that multispectral imaging can provide measurable improvements, as well as new opportunities, compared with conventional RGB imaging, particularly when combined with machine learning methods. At the same time, the results indicate that these benefits are strongly task-dependent. In some applications, the performance gains obtained from multispectral information may not justify the additional complexity associated with sensor hardware, calibration requirements, data storage, and processing pipelines.

Although this dissertation does not directly target specific agricultural and environmental protection applications, it develops the technological foundations necessary to support them. The emphasis on multispectral imaging and machine learning provides tools that can be adapted to a wide range of real-world scenarios, including those briefly outlined in this work. In this context, the research contributes to broader efforts in sustainable agriculture and environmental monitoring.

Within agricultural and environmental protection applications, spectral bands such as NIR and red-edge are especially valuable because they are physically meaningful and contain information unavailable in standard RGB imagery. These wavelengths can increase model sensitivity to subtle phenomena, such as vegetation stress or moisture conditions. In addition, vegetation indices derived from such bands remain useful even in the era of deep learning, as they represent domain-informed feature engineering. Although models can theoretically learn equivalent transformations directly from raw inputs, explicitly providing composite channels such as NDVI may simplify the learning process and, in some cases, improve predictive performance.

The dissertation also highlights the importance of selecting machine learning methods according to the nature of the data. Deep learning techniques are particularly effective at extracting complex relationships and rich spatial patterns that would be difficult to model manually. This is especially relevant in high-resolution imagery acquired from UAVs or ground-based robotic platforms, where texture and local spatial context are highly informative. The majority of publications employ deep learning methods for imagery with a spatial resolution finer than 10 m per pixel [44]. On the other hand, for data with low spatial resolution, such as many satellite products with a GSD of approximately 10-20 m, spectral and temporal features may be more informative than fine spatial structure. Under such conditions, simpler machine learning models may achieve performance comparable to more complex deep learning approaches while offering lower computational cost and greater interpretability. Explainability is particularly important for building user trust, which remains a prerequisite for adoption in agricultural and environmental decision-making. Moreover, predictive accuracy alone should not be the sole criterion in system design. Practical considerations such as hardware cost, scalability, inference speed, and explainability are important. Consequently, the choice of imaging modality and modelling approach inevitably involves trade-offs.

The literature review and experimental studies further emphasised the importance of scientific rigour in dataset preparation and train-test splitting strategies. In remote sensing applications, improper partitioning can easily lead to spatial or temporal leakage, resulting in overoptimistic estimates of generalisation performance. Reliable benchmarking therefore requires carefully designed evaluation procedures. This is particularly important for small datasets, where the training process often needs to be repeated with different random seeds in order to obtain statistically reliable performance metrics.

Several case studies illustrated the practical value of multispectral imaging across scales. The beetroot crop segmentation study demonstrated that multispectral data can be particularly valuable during the training stage, reducing the amount of manual annotation required to train a model that ultimately uses RGB imagery only. The soil monitoring study showed the potential of satellite imaging for large-scale environmental assessment, although substantial challenges remain before such systems can routinely support farmers or policy-makers. While large volumes of satellite imagery are already available, the scarcity of reliable ground-truth measurements remains a major challenge, as these still largely rely on manual annotation or on-site measurements. At the local scale, the crop-damage segmentation case study showed that UAV imagery can directly support farm-management workflows by reducing manual inspection effort. Although monitoring data obtained from UAVs, satellites, or ground-level robots may seem to represent competing sources of information, they should instead be viewed as complementary. They operate at different spatial, temporal, and economic scales, and therefore address different classes of problems while also presenting distinct technical challenges. Finally, the band-reduction experiments demonstrated that careful spectral band selection can enable smaller models and simpler hardware configurations without substantial performance loss. This can reduce both system cost and inference latency, not only by decreasing the computational cost of model inference, but also by reducing the time required to load or transmit data from the camera to the processing unit. In addition to band reduction, feature engineering, both manual and automatic, remains important for lower-cost deployments. Such optimisation is especially relevant for hyperspectral imaging, where neighbouring bands are often highly correlated and even a small subset may contain sufficient task-relevant information.

The thesis also highlights the importance of open science. Publicly available datasets and open-source implementations are essential for reproducibility and for accelerating progress in the field. During the course of this research, two additional multispectral datasets not described in the thesis were collected and publicly released, captured using multispectral UAV imagery for mistletoe detection and rapeseed field damage caused by European bison (links in section 5.3). The development of the Deepness software further showed that practical adoption of machine learning depends not only on algorithmic advances, but also on open-source, easily accessible tools that integrate with established expert workflows, in this case within the QGIS ecosystem. Ongoing development of the software and collaboration with domain specialists also showed the importance of interdisciplinary cooperation in creating solutions with real societal impact.

Overall, this thesis contributes both methodological findings and open-source implementations, helping to bridge the gap between research and practice. The most important results of this work can be summarised as follows:

- Demonstration that machine learning methods applied to multispectral imagery outperform conventional approaches in agricultural and environmental monitoring tasks.

- Identification and validation of optimal spectral modalities, features, and explainability strategies, supported by real-world case studies.

- Development and release of open-source tools and datasets to facilitate adoption by both research and practitioner communities.

5.1 Limitations

Several limitations should be discussed. Firstly, the datasets used throughout the thesis are geographically constrained. However, as discussed in the dissertation, generalisation across environments remains one of the central challenges for practical deployment. Agricultural and environmental systems vary substantially with season, climate, soil type, and crop species. Much of the current research literature remains focused on local-scale studies that may report overly optimistic results. Similarly, experiments conducted in controlled environments can overestimate performance relative to real field conditions. High-quality ground-truth data remain a major bottleneck for both methodological progress and operational deployment. In agriculture especially, annotations are often uncertain, subjective, or noisy, while direct measurements may be expensive and time-consuming to obtain.

A further limitation concerns the relative maturity of methods based on conventional RGB images. RGB imaging has received substantially greater research attention than multispectral imaging, and many state-of-the-art architectures are optimised or pre-trained for three-channel inputs. As a result, multispectral workflows often require additional adaptation and engineering effort. Multispectral data processing is also inherently more demanding than RGB imaging. RGB imagery is standardised for human visual perception and often exhibits greater tolerance to illumination variability. In contrast, multispectral imaging requires stronger radiometric consistency so that pixel values correspond to meaningful physical reflectance. Improper normalisation may remove subtle but important spectral features. This challenge is compounded by differences in sensor response, calibration procedures, and optical characteristics, which can reduce transferability across platforms. Moreover, spectral datasets may be particularly vulnerable to unintended information leakage between training and testing subsets. If not properly controlled, models may exploit spurious correlations related to acquisition conditions rather than physically meaningful patterns.

It should be noted that multispectral and hyperspectral data are not universally beneficial, not only in terms of increased system complexity but also in terms of predictive performance. Several studies have demonstrated that the use of additional spectral information does not necessarily lead to improved results when compared to standard RGB imagery. For example, the authors in [205] reported comparable performance between RGB-based and multispectral vegetation indices for monitoring vegetation health in palm tree cultivation. Similarly, [206] obtained only slightly more stable results using multispectral imagery than RGB imagery for vineyard water status prediction. At the same time, the authors noted the higher spatial resolution of RGB images, which may provide advantages for precise plant segmentation tasks. The authors in [207] performed tree species classification and demonstrated that incorporating multi-temporal information significantly improved classification accuracy for RGB imagery, which was not the case for multispectral images, ultimately yielding better results for RGB imagery than for multispectral data. During the preparation of this dissertation, cases were also encountered in which multispectral imaging did not improve performance. In one UAV-based study on mistletoe detection in pine trees, RGB-based approaches outperformed those based on multispectral data. A likely explanation is imperfect co-registration between spectral bands caused by the sequential acquisition of multispectral images during

UAV motion, resulting in spatial misalignment, whereas this issue does not arise in standard RGB imagery. Another possibility is that the spectral responses of pine trees and mistletoe are too similar within the limited set of available bands. A study using hyperspectral equipment could help clarify this issue.

Finally, advanced machine learning models, especially deep neural networks, may behave as black boxes, offering limited interpretability of their predictions. They may also fail to meet the computational constraints of edge devices or real-time systems requiring low-latency inference. In some scenarios, other sensing modalities such as LiDAR or radar imagery may be more appropriate than optical multispectral imaging, for example, where cloud-penetrating observations are required.

5.2 Future work

The continued development of precision agriculture is expected to accelerate, driven both by the increasing availability of sensors, platforms, and data, and by global environmental and food-security pressures. Consequently, wider deployment of operational machine learning systems can be expected, increasing the importance of robustness and model generalisation. To support this transition, further research should focus on larger multi-year and multi-region datasets, together with standardised public benchmarks that allow fair comparison between methods. Closely related priorities include domain adaptation and transfer learning techniques that improve model portability across sensors, locations, and growing seasons.

Continuous learning approaches may also become increasingly important, allowing models to adapt to evolving environmental conditions and management practices over time. These developments are likely to coincide with continued advances in Earth observation systems, including more frequent satellite revisit times, improved radiometric quality, and denser global monitoring coverage.

Future work should also place stronger emphasis on explainability and uncertainty quantification. In particular, models should aim to distinguish between aleatoric and epistemic uncertainty. Such capabilities are valuable not only for interpretation but also for identifying situations in which the model should defer to human expertise, thereby building trust in the technology.

Another promising direction is the integration of large-scale foundation models, especially multimodal systems capable of combining imagery, text, and tabular data. Although Earth-observation foundation models have already emerged, their practical use in agriculture and environmental protection remains limited. Nevertheless, if current progress continues at the pace observed in recent years, such models may become transformative tools for monitoring and decision support. While full-scale foundation models are currently impractical for real-time edge deployment, they may still be valuable for offline analysis, few-shot learning, or knowledge distillation into smaller deployable models. For example, a large multimodal model may identify a rare plant disease from a short textual description or a few visual examples, after which its knowledge could be transferred to a lightweight, field-deployable model.

Finally, regardless of progress in modelling techniques, the continued development of open-source software and openly available benchmark datasets will remain essential to ensure that future advances are reproducible, accessible, and practically useful.

5.3 Data availability statement

The collected and curated datasets referenced in this thesis are publicly available at:

- Hyperview2: `https://platform.ai4eo.eu/hyperview2`,
- Mistletoe on pines: `https://doi.org/10.5281/zenodo.20572271`,
- Rapeseed field damage: `https://doi.org/10.5281/zenodo.20560943`,

5.4 Declaration of code availability

To support the reproducibility of the findings presented in this thesis, the associated code repositories are provided:

- `https://github.com/PUTvision/qgis-plugin-deepness`,
- `https://github.com/PUTvision/coffee-rust`,
- `https://github.com/PUTvision/mavic-3m-calibration`
- `https://github.com/PUTvision/multispectral-beetroots`

List of Figures

List of Tables

Bibliography

[1] J. Chai, H. Zeng, A. Li, and E. W. Ngai. „Deep learning in computer vision: A critical review of emerging techniques and application scenarios". In: *Machine Learning with Applications* 6 (2021), p. 100134 (cit. on p. 5).

[2] S. Mukhtar, A. Arbabi, and J. Viegas. „Compact Spectral Imaging: A review of miniaturized and integrated systems". In: *Laser & Photonics Reviews* 19.21 (2025), e01042 (cit. on p. 5).

[3] S. Ghazal, A. Munir, and W. S. Qureshi. „Computer vision in smart agriculture and precision farming: Techniques and applications". In: *Artificial Intelligence in Agriculture* 13 (2024), pp. 64–83 (cit. on p. 5).

[4] B. G. Ram, P. Oduor, C. Igathinathane, K. Howatt, and X. Sun. „A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects". In: *Computers and Electronics in Agriculture* 222 (2024), p. 109037 (cit. on p. 5).

[5] S. Khanal, K. Kc, J. P. Fulton, S. Shearer, and E. Ozkan. „Remote sensing in agriculture—accomplishments, limitations, and opportunities". In: *Remote sensing* 12.22 (2020), p. 3783 (cit. on p. 5).

[6] J. Wiland-Szymańska, E. Kazimierczak-Grygiel, P. Drapikowski, K. Borowiak, and M. Drapikowska. „Does a gender of Welwitschia mirabilis plants influence their photosynthetic activity?" In: *Plos one* 18.9 (2023), e0291122 (cit. on p. 6).

[7] S. Zhang, X. Wang, H. Lin, and Z. Qiang. „A review of the application of UAV multispectral remote sensing technology in precision agriculture". In: *Smart Agricultural Technology* (2025), p. 101406 (cit. on pp. 6, 7, 19, 40).

[8] M. Padhiary, D. Saha, R. Kumar, L. N. Sethi, and A. Kumar. „Enhancing precision agriculture: A comprehensive review of machine learning and AI vision applications in all-terrain vehicle for farm automation". In: *Smart Agricultural Technology* 8 (2024), p. 100483 (cit. on p. 6).

[9] U. Khan, M. Khan, M. Latif, et al. „A systematic literature review of machine learning and deep learning approaches for spectral image classification in agricultural applications using aerial photography". In: *Computers, Materials, & Continua* 78.3 (2024), p. 2967 (cit. on p. 6).

[10] H. Yu, B. Kong, Y. Hou, et al. „A critical review on applications of hyperspectral remote sensing in crop monitoring". In: *Experimental Agriculture* 58 (2022), e26 (cit. on p. 6).

[11] M. Albahar. „A survey on deep learning and its impact on agriculture: challenges and opportunities“. In: *Agriculture* 13.3 (2023), p. 540 (cit. on pp. 7, 29).

[12] A. Kamilaris and F. X. Prenafeta-Boldú. „Deep learning in agriculture: A survey“. In: *Computers and electronics in agriculture* 147 (2018), pp. 70–90 (cit. on p. 7).

[13] D. Wang, W. Cao, F. Zhang, et al. „A review of deep learning in multiscale agricultural sensing“. In: *Remote Sensing* 14.3 (2022), p. 559 (cit. on pp. 7, 40).

[14] A. Caroppo, D. Giovanni, and A. Leone. „A systematic review of available multispectral UAV image datasets for precision agriculture applications“. In: *Remote Sensing* 18.4 (2026), p. 659 (cit. on p. 7).

[15] J. U. M. Akbar, S. F. Kamarulzaman, A. J. M. Muzahid, M. A. Rahman, and M. Uddin. „A comprehensive review on deep learning assisted computer vision techniques for smart greenhouse agriculture“. In: *IEEE Access* 12 (2024), pp. 4485–4522 (cit. on p. 7).

[16] J. Li, D. Chen, X. Qi, et al. „Label-efficient learning in agriculture: A comprehensive review“. In: *Computers and Electronics in Agriculture* 215 (2023), p. 108412 (cit. on p. 7).

[17] A. Cartolano, A. Cuzzocrea, and G. Pilato. „Analyzing and assessing explainable AI models for smart agriculture environments“. In: *Multimedia Tools and Applications* 83.12 (2024), pp. 37225–37246 (cit. on p. 7).

[18] R. Wilkinson, M. Mleczko, R. Brewin, et al. „Environmental impacts of earth observation data in the constellation and cloud computing era“. In: *Science of The Total Environment* 909 (2024), p. 168584 (cit. on p. 8).

[19] G. Guerrisi, F. Del Frate, and G. Schiavon. „Artificial intelligence based on-board image compression for the Φ-Sat-2 mission“. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), pp. 8063–8075 (cit. on p. 8).

[20] J. Bendig, A. Bolten, and G. Bareth. „Introducing a low-cost mini-UAV for thermal-and multispectral-imaging“. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 39 (2012), pp. 345–349 (cit. on p. 17).

[21] F. Qingyun and W. Zhaokui. „Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery“. In: *Pattern Recognition* 130 (2022), p. 108786 (cit. on p. 17).

[22] R. Torres, P. Snoeij, D. Geudtner, et al. „GMES Sentinel-1 mission“. In: *Remote sensing of environment* 120 (2012), pp. 9–24 (cit. on p. 17).

[23] Q. Gao, M. Zribi, M. J. Escorihuela, and N. Baghdadi. „Synergetic use of Sentinel-1 and Sentinel-2 data for soil moisture mapping at 100 m resolution“. In: *Sensors* 17.9 (2017), p. 1966 (cit. on p. 17).

[24] M. Á. Obregón, G. Rodrigues, M. J. Costa, M. Potes, and A. M. Silva. „Validation of ESA Sentinel-2 L2A aerosol optical thickness and columnar water vapour during 2017–2018“. In: *Remote Sensing* 11.14 (2019), p. 1649 (cit. on p. 17).

[25] B. Boiarskii and H. Hasegawa. „Comparison of NDVI and NDRE indices to detect differences in vegetation and chlorophyll content“. In: *J. Mech. Contin. Math. Sci* 4 (2019), pp. 20–29 (cit. on p. 17).

[26] R. D. Jackson and A. R. Huete. „Interpreting vegetation indices“. In: *Preventive veterinary medicine* 11.3-4 (1991), pp. 185–200 (cit. on p. 17).

[27] G. ElMasry, N. Mandour, S. Al-Rejaie, E. Belin, and D. Rousseau. „Recent applications of multispectral imaging in seed phenotyping and quality monitoring—An overview“. In: *Sensors* 19.5 (2019), p. 1090 (cit. on p. 18).

[28] D. Wu and D.-W. Sun. „Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review—Part I: Fundamentals“. In: *Innovative Food Science & Emerging Technologies* 19 (2013), pp. 1–14 (cit. on pp. 18, 23).

[29] T. Adão, J. Hruška, L. Pádua, et al. „Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry“. In: *Remote sensing* 9.11 (2017), p. 1110 (cit. on p. 18).

[30] A. Morales, R. Guerra, P. Horstrand, et al. „A multispectral camera development: From the prototype assembly until its use in a UAV system“. In: *Sensors* 20.21 (2020), p. 6129 (cit. on p. 19).

[31] S. K. Phang, T. H. A. Chiang, A. Happonen, and M. M. L. Chang. „From satellite to UAV-based remote sensing: A review on precision agriculture“. In: *Ieee Access* 11 (2023), pp. 127057–127076 (cit. on p. 19).

[32] F. Ma, M. Yuan, and I. Kozak. „Multispectral imaging: Review of current applications“. In: *Survey of Ophthalmology* 68.5 (2023), pp. 889–904 (cit. on p. 19).

[33] F. Samadzadegan, A. Toosi, and F. Dadrass Javan. „A critical review on multi-sensor and multi-platform remote sensing data fusion approaches: current status and prospects“. In: *International journal of remote sensing* 46.3 (2025), pp. 1327–1402 (cit. on p. 19).

[34] F. Visentin, S. Cremasco, M. Sozzi, et al. „A mixed-autonomous robotic platform for intra-row and inter-row weed removal for precision agriculture“. In: *Computers and Electronics in Agriculture* 214 (2023), p. 108270 (cit. on p. 23).

[35] Y. Li, X. Qi, Y. Cai, et al. „A Rice Leaf Area Index Monitoring Method Based on the Fusion of Data from RGB Camera and Multi-Spectral Camera on an Inspection Robot“. In: *Remote Sensing* 16.24 (2024), p. 4725 (cit. on p. 23).

[36] A. Comparetti, A. Fagiolini, S. Fountas, and V. Cascio. „Field robots for precision agriculture“. In: *AIMS Agriculture and Food* 10.4 (2025), pp. 885–916 (cit. on p. 23).

[37] I. Sa, M. Popović, R. Khanna, et al. „WeedMap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming“. In: *Remote Sensing* 10.9 (2018), p. 1423 (cit. on p. 23).

[38] N. Chebrolu, P. Lottes, A. Schaefer, et al. „Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields“. In: *The International Journal of Robotics Research* 36.10 (2017), pp. 1045–1052 (cit. on pp. 23, 66).

[39] A. Gatti and A. Bertolini. „Sentinel-2 products specification document“. In: *Rapport technique* (2015), pp. 4–7 (cit. on p. 24).

[40] M. Main-Knorn, B. Pflug, J. Louis, et al. „Sen2Cor for sentinel-2“. In: *Image and signal processing for remote sensing XXIII*. Vol. 10427. SPIE. 2017, pp. 37–48 (cit. on p. 24).

[41] C. Delgado-Vera, M. Aguirre-Munizaga, M. Jiménez-Icaza, N. Manobanda-Herrera, and A. Rodríguez-Méndez. „A photogrammetry software as a tool for precision agriculture: A case study“. In: *International Conference on Technologies and Innovation*. Springer. 2017, pp. 282–295 (cit. on p. 26).

[42] I. Ioja, V. Nedeff, M. Agop, F. M. Nedeff, and C. Tomozei. „Software uses in precision agriculture based on drone image processing–A review“. In: *2024 9th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE)*. IEEE. 2024, pp. 1–6 (cit. on p. 27).

[43] V. Meshram, K. Patil, V. Meshram, D. Hanchate, and S. Ramkteke. „Machine learning in agriculture domain: A state-of-art survey“. In: *Artificial Intelligence in the Life Sciences* 1 (2021), p. 100010 (cit. on pp. 27, 28).

[44] L. Ma, Y. Liu, X. Zhang, et al. „Deep learning in remote sensing applications: A meta-analysis and review“. In: *ISPRS journal of photogrammetry and remote sensing* 152 (2019), pp. 166–177 (cit. on pp. 28, 98).

[45] I. Attri, L. K. Awasthi, and T. P. Sharma. „Machine learning in agriculture: a review of crop management applications“. In: *Multimedia Tools and Applications* 83.5 (2024), pp. 12875–12915 (cit. on p. 28).

[46] A. Sharifi and M. M. Safari. „Enhancing the spatial resolution of sentinel-2 images through super-resolution using transformer-based deep learning models“. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2025) (cit. on p. 29).

[47] C. Jin, L.-J. Deng, T.-Z. Huang, and G. Vivone. „Laplacian pyramid networks: A new approach for multispectral pansharpening“. In: *Information Fusion* 78 (2022), pp. 158–170 (cit. on p. 29).

[48] V. Sineglazov, R. Koniushenko, and S. Dolgorukov. „Problems of Multispectral Image Processing in Agriculture“. In: *Electronics and Control Systems* 1.83 (), pp. 18–27 (cit. on p. 29).

[49] R. C. d. Oliveira and R. D. d. S. e. Silva. „Artificial intelligence in agriculture: benefits, challenges, and trends“. In: *Applied Sciences* 13.13 (2023), p. 7405 (cit. on p. 29).

[50] T. Miftahushudur, H. M. Sahin, B. Grieve, and H. Yin. „A survey of methods for addressing imbalance data problems in agriculture applications“. In: *Remote Sensing* 17.3 (2025), p. 454 (cit. on pp. 31, 55).

[51] P. Li and X. Xiao. „Multispectral image segmentation by a multichannel watershed-based approach“. In: *International Journal of Remote Sensing* 28.19 (2007), pp. 4429–4452 (cit. on p. 34).

[52] E. Adam, O. Mutanga, and D. Rugege. „Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review“. In: *Wetlands ecology and management* 18.3 (2010), pp. 281–296 (cit. on p. 34).

[53] T. N. Tran, R. Wehrens, and L. M. Buydens. „Clustering multispectral images: a tutorial“. In: *Chemometrics and Intelligent Laboratory Systems* 77.1-2 (2005), pp. 3–17 (cit. on p. 34).

[54] A. Rydberg and G. Borgefors. „Integrated method for boundary delineation of agricultural fields in multispectral satellite images“. In: *IEEE Transactions on Geoscience and Remote Sensing* 39.11 (2002), pp. 2514–2520 (cit. on p. 34).

[55] L. Wang and G. Healey. „Using Zernike moments for the illumination and geometry invariant classification of multispectral texture“. In: *IEEE Transactions on Image Processing* 7.2 (1998), pp. 196–203 (cit. on p. 35).

[56] P.-H. Hsu. „Feature extraction of hyperspectral images using wavelet and matching pursuit“. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 62.2 (2007), pp. 78–92 (cit. on p. 35).

[57] C. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo. „Multispectral image feature points“. In: *Sensors* 12.9 (2012), pp. 12661–12672 (cit. on p. 35).

[58] S. Huang, L. Tang, J. P. Hupy, Y. Wang, and G. Shao. „A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing“. In: *Journal of forestry research* 32.1 (2021), pp. 1–6 (cit. on p. 35).

[59] R. B. Myneni, F. G. Hall, P. J. Sellers, and A. L. Marshak. „The interpretation of spectral vegetation indexes“. In: *IEEE Transactions on Geoscience and remote Sensing* 33.2 (1995), pp. 481–486 (cit. on p. 35).

[60] R. Myneni, C. Tucker, G. Asrar, and C. Keeling. „Interannual variations in satellite-sensed vegetation index data from 1981 to 1991“. In: *Journal of Geophysical Research: Atmospheres* 103.D6 (1998), pp. 6145–6160 (cit. on p. 35).

[61] J. Xue and B. Su. „Significant remote sensing vegetation indices: A review of developments and applications“. In: *Journal of sensors* 2017.1 (2017), p. 1353691 (cit. on p. 36).

[62] W. Rouse. „Monitoring vegetation system in the great plain with ERTS“. In: *3rd ERTS symposium, NASA Washington DC, 1973*. Vol. 1. 1973, pp. 309–317 (cit. on p. 36).

[63] A. A. Gitelson, Y. J. Kaufman, and M. N. Merzlyak. „Use of a green channel in remote sensing of global vegetation from EOS-MODIS“. In: *Remote sensing of Environment* 58.3 (1996), pp. 289–298 (cit. on p. 36).

[64] A. Gitelson and M. N. Merzlyak. „Spectral reflectance changes associated with autumn senescence of Aesculus hippocastanum L. and Acer platanoides L. leaves. Spectral features and relation to chlorophyll estimation“. In: *Journal of plant physiology* 143.3 (1994), pp. 286–292 (cit. on p. 36).

[65] T. V. Tran, R. Reef, and X. Zhu. „A review of spectral indices for mangrove remote sensing“. In: *Remote Sensing* 14.19 (2022), p. 4868 (cit. on p. 36).

[66] B. Datt. „Visible/near infrared reflectance and chlorophyll content in Eucalyptus leaves“. In: *International Journal of Remote Sensing* 20.14 (1999), pp. 2741–2759 (cit. on p. 36).

[67] B. Zebarth, M. Younie, J. Paul, and S. Bittman. „Evaluation of leaf chlorophyll index for making fertilizer nitrogen recommendations for silage corn in a high fertility environment“. In: *Communications in soil science and plant analysis* 33.5-6 (2002), pp. 665–684 (cit. on p. 36).

[68] G. Rondeaux, M. Steven, and F. Baret. „Optimization of soil-adjusted vegetation indices“. In: *Remote sensing of environment* 55.2 (1996), pp. 95–107 (cit. on p. 36).

[69] S. K. McFeeters. „The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features“. In: *International journal of remote sensing* 17.7 (1996), pp. 1425–1432 (cit. on p. 36).

[70]D. Theng and K. K. Bhoyar. „Feature selection techniques for machine learning: a survey of more than two decades of research". In: *Knowledge and Information Systems* 66.3 (2024), pp. 1575–1637 (cit. on p. 36).

[71]A. Nițu, C. Florea, M. Ivanovici, and A. Racoviteanu. „NDVI and Beyond: Vegetation Indices as Features for Crop Recognition and Segmentation in Hyperspectral Data". In: *Sensors (Basel, Switzerland)* 25.12 (2025), p. 3817 (cit. on p. 37).

[72]A. A. Khan, A. A. Laghari, and S. A. Awan. „Machine learning in computer vision: A review." In: *EAI Endorsed Transactions on Scalable Information Systems* 8.32 (2021) (cit. on p. 37).

[73]S. V. Mahadevkar, B. Khemani, S. Patil, et al. „A review on machine learning styles in computer vision—techniques and future directions". In: *Ieee Access* 10 (2022), pp. 107293–107329 (cit. on p. 37).

[74]S. O. Araujo, R. S. Peres, J. C. Ramalho, F. Lidon, and J. Barata. „Machine learning applications in agriculture: current trends, challenges, and future perspectives". In: *Agronomy* 13.12 (2023), p. 2976 (cit. on pp. 37, 39).

[75]H. R. Shahdoosti and H. Ghassemian. „Combining the spectral PCA and spatial PCA fusion methods by an optimal filter". In: *Information Fusion* 27 (2016), pp. 150–160 (cit. on p. 37).

[76]S. D. Bauer, F. Korč, and W. Förstner. „The potential of automatic methods of classification to identify leaf diseases from multispectral images". In: *Precision Agriculture* 12.3 (2011), pp. 361–377 (cit. on p. 37).

[77]W. Liu, C. Liu, and L. Zheng. „Discrimination in varieties of rice seeds with multispectral imaging using support vector machine". In: *Transactions of the Chinese Society of Agricultural Engineering* 30.10 (2014), pp. 145–151 (cit. on p. 38).

[78]T. M. Berhane, C. R. Lane, Q. Wu, et al. „Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory". In: *Remote sensing* 10.4 (2018), p. 580 (cit. on p. 38).

[79]Y. Lan, Z. Huang, X. Deng, et al. „Comparison of machine learning methods for citrus greening detection on UAV multispectral images". In: *Computers and electronics in agriculture* 171 (2020), p. 105234 (cit. on p. 39).

[80]A. Zarei, M. Hasanlou, and M. Mahdianpari. „A comparison of machine learning models for soil salinity estimation using multi-spectral earth observation data". In: *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences* 3 (2021), pp. 257–263 (cit. on p. 39).

[81]A. Krizhevsky, I. Sutskever, and G. E. Hinton. „Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012) (cit. on p. 39).

[82]M. Z. Alom, T. M. Taha, C. Yakopcic, et al. „The history began from alexnet: A comprehensive survey on deep learning approaches". In: *arXiv preprint arXiv:1803.01164* (2018) (cit. on p. 40).

[83]A. Sharma, A. Jain, P. Gupta, and V. Chowdary. „Machine learning applications for precision agriculture: A comprehensive review". In: *IEEe Access* 9 (2020), pp. 4843–4873 (cit. on p. 40).

[84] L. Shuai, Z. Li, Z. Chen, D. Luo, and J. Mu. „A research review on deep learning combined with hyperspectral Imaging in multiscale agricultural sensing“. In: *Computers and Electronics in Agriculture* 217 (2024), p. 108577 (cit. on p. 40).

[85] M. F. Guerri, C. Distante, P. Spagnolo, F. Bougourzi, and A. Taleb-Ahmed. „Deep learning techniques for hyperspectral image analysis in agriculture: A review“. In: *ISPRS Open Journal of Photogrammetry and Remote Sensing* 12 (2024), p. 100062 (cit. on p. 40).

[86] M. A. Jabed and M. A. A. Murad. „Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability“. In: *Heliyon* 10.24 (2024) (cit. on p. 40).

[87] K. He, X. Zhang, S. Ren, and J. Sun. „Deep residual learning for image recognition“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 40, 41).

[88] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. „Cbam: Convolutional block attention module“. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19 (cit. on p. 40).

[89] T.-Y. Lin, P. Dollár, R. Girshick, et al. „Feature pyramid networks for object detection“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125 (cit. on p. 40).

[90] M. Tan and Q. Le. „Efficientnet: Rethinking model scaling for convolutional neural networks“. In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114 (cit. on p. 41).

[91] Z. Liu, H. Mao, C.-Y. Wu, et al. „A convnet for the 2020s“. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986 (cit. on p. 41).

[92] Y. Yu, C. Wang, Q. Fu, et al. „Techniques and challenges of image segmentation: A review“. In: *Electronics* 12.5 (2023), p. 1199 (cit. on p. 41).

[93] O. Ronneberger, P. Fischer, and T. Brox. „U-net: Convolutional networks for biomedical image segmentation“. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241 (cit. on p. 41).

[94] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. „Unet++: A nested u-net architecture for medical image segmentation“. In: *International workshop on deep learning in medical image analysis*. Springer. 2018, pp. 3–11 (cit. on p. 41).

[95] H. Huang, L. Lin, R. Tong, et al. „Unet 3+: A full-scale connected unet for medical image segmentation“. In: *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Ieee. 2020, pp. 1055–1059 (cit. on p. 41).

[96] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. „nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation“. In: *Nature methods* 18.2 (2021), pp. 203–211 (cit. on p. 41).

[97] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. „Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs“. In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848 (cit. on p. 41).

[98] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. „Encoder-decoder with atrous separable convolution for semantic image segmentation“. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818 (cit. on p. 41).

[99] A. Chaurasia and E. Culurciello. „Linknet: Exploiting encoder representations for efficient semantic segmentation“. In: *2017 IEEE visual communications and image processing (VCIP)*. IEEE. 2017, pp. 1–4 (cit. on p. 41).

[100] I. Ulku, E. Akagündüz, and P. Ghamisi. „Deep semantic segmentation of trees using multispectral images“. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), pp. 7589–7604 (cit. on p. 41).

[101] R. Girshick, J. Donahue, T. Darrell, and J. Malik. „Rich feature hierarchies for accurate object detection and semantic segmentation“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587 (cit. on p. 42).

[102] S. Ren, K. He, R. Girshick, and J. Sun. „Faster r-cnn: Towards real-time object detection with region proposal networks“. In: *Advances in neural information processing systems* 28 (2015) (cit. on p. 42).

[103] K. He, G. Gkioxari, P. Dollár, and R. Girshick. „Mask r-cnn“. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969 (cit. on p. 42).

[104] W. Liu, D. Anguelov, D. Erhan, et al. „Ssd: Single shot multibox detector“. In: Springer. 2016, pp. 21–37 (cit. on p. 42).

[105] T. Diwan, G. Anirudh, and J. V. Tembhurne. „Object detection using YOLO: challenges, architectural successors, datasets and applications“. In: *multimedia Tools and Applications* 82.6 (2023), pp. 9243–9275 (cit. on p. 42).

[106] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. „Yolox: Exceeding yolo series in 2021“. In: *arXiv preprint arXiv:2107.08430* (2021) (cit. on p. 43).

[107] S. Gui, S. Song, R. Qin, and Y. Tang. „Remote sensing object detection in the deep learning era—a review“. In: *Remote Sensing* 16.2 (2024), p. 327 (cit. on p. 43).

[108] A. Dosovitskiy. „An image is worth 16x16 words: Transformers for image recognition at scale“. In: *arXiv preprint arXiv:2010.11929* (2020) (cit. on p. 43).

[109] E. Xie, W. Wang, Z. Yu, et al. „SegFormer: Simple and efficient design for semantic segmentation with transformers“. In: *Advances in neural information processing systems* 34 (2021), pp. 12077–12090 (cit. on p. 43).

[110] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. „Masked-attention mask transformer for universal image segmentation“. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 1290–1299 (cit. on p. 43).

[111] J. Jain, J. Li, M. T. Chiu, et al. „Oneformer: One transformer to rule universal image segmentation“. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 2989–2998 (cit. on p. 43).

[112] X. Li, H. Ding, H. Yuan, et al. „Transformer-based visual segmentation: A survey“. In: *IEEE transactions on pattern analysis and machine intelligence* (2024) (cit. on p. 43).

[113] Z. Liu, Y. Lin, Y. Cao, et al. „Swin transformer: Hierarchical vision transformer using shifted windows“. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022 (cit. on p. 43).

[114] G. A. Pereira and M. Hussain. „A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships“. In: *arXiv preprint arXiv:2408.15178* (2024) (cit. on p. 43).

[115] N. Carion, F. Massa, G. Synnaeve, et al. „End-to-end object detection with transformers“. In: *European conference on computer vision*. Springer. 2020, pp. 213–229 (cit. on p. 43).

[116] O. Siméoni, H. V. Vo, M. Seitzer, et al. „Dinov3“. In: *arXiv preprint arXiv:2508.10104* (2025) (cit. on p. 43).

[117] F. Li, H. Zhang, H. Xu, et al. „Mask dino: Towards a unified transformer-based framework for object detection and segmentation“. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 3041–3050 (cit. on p. 44).

[118] Y.-M. Qin, Y.-H. Tu, T. Li, et al. „Deep Learning for sustainable agriculture: A systematic review on applications in lettuce cultivation“. In: *Sustainability* 17.7 (2025), p. 3190 (cit. on p. 44).

[119] S. M. M. Nejad, D. Abbasi-Moghadam, A. Sharifi, and A. Tariq. „Capsular attention Conv-LSTM network (CACN): A deep learning structure for crop yield estimation based on multispectral imagery“. In: *European Journal of Agronomy* 161 (2024), p. 127369 (cit. on p. 44).

[120] Y. Shen, B. Mercatoris, Z. Cao, et al. „Improving wheat yield prediction accuracy using LSTM-RF framework based on UAV thermal infrared and multispectral imagery“. In: *Agriculture* 12.6 (2022), p. 892 (cit. on p. 44).

[121] X. Su, P. Howard, and S. Bethard. „Transformer-based temporal information extraction and application: A review“. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025, pp. 28810–28829 (cit. on p. 44).

[122] L. Zhao and S. Ji. „CNN, RNN, or ViT? An evaluation of different deep learning architectures for spatio-temporal representation of sentinel time series“. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2022), pp. 44–56 (cit. on p. 44).

[123] J. Li, X. Qi, S. H. Nabaei, et al. „A survey on 3D reconstruction techniques in plant phenotyping: from classical methods to neural radiance fields (NeRF), 3D gaussian splatting (3DGS), and beyond“. In: *Plant Phenomics* (2025), p. 100137 (cit. on p. 45).

[124] L. Meyer, A. Gilson, U. Schmid, and M. Stamminger. „Fruitnerf: A unified neural radiance field based fruit counting framework“. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 1–8 (cit. on p. 45).

[125] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei. „Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network“. In: *IEEE Transactions on Geoscience and Remote Sensing* 56.8 (2018), pp. 4274–4288 (cit. on p. 45).

[126] J. E. Gallagher and E. J. Oughton. „Surveying You Only Look Once (YOLO) multispectral object detection advancements, applications and challenges“. In: *IEEE Access* (2025) (cit. on p. 46).

[127] Y. Xie, L. Zhang, X. Yu, and W. Xie. „YOLO-MS: Multispectral object detection via feature interaction and self-attention guided fusion“. In: *IEEE Transactions on Cognitive and Developmental Systems* 15.4 (2023), pp. 2132–2143 (cit. on p. 46).

[128] Y. Shao, Q. Huang, et al. „MOD-YOLO: Multispectral object detection based on transformer dual-stream YOLO“. In: *Pattern Recognition Letters* 183 (2024), pp. 26–34 (cit. on p. 46).

[129] I. Blekanov, A. Molin, D. Zhang, et al. „Monitoring of grain crops nitrogen status from uav multispectral images coupled with deep learning approaches“. In: *Computers and Electronics in Agriculture* 212 (2023), p. 108047 (cit. on p. 46).

[130] W. Cardoso, T. Barros, G. Gonçalves, C. Premebida, and U. J. Nunes. „Multispectral Image Segmentation in Agriculture: Evaluating Deep Learning Models with Train-Test Split and Cross-Validation Strategies“. In: *2024 7th Iberian Robotics Conference (ROBOT)*. IEEE. 2024, pp. 1–6 (cit. on p. 46).

[131] R. Beigaitė, M. Mechenich, and I. Žliobaitė. „Spatial cross-validation for globally distributed data“. In: *International Conference on Discovery Science*. Springer. 2022, pp. 127–140 (cit. on p. 48).

[132] J. J. Salazar, L. Garland, J. Ochoa, and M. J. Pyrcz. „Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy“. In: *Journal of Petroleum Science and Engineering* 209 (2022), p. 109885 (cit. on pp. 48, 49).

[133] P. Schratz, M. Becker, M. Lang, and A. Brenning. „Mlr3spatiotempcv: Spatiotemporal resampling methods for machine learning in R“. In: *Journal of Statistical Software* 111 (2024), pp. 1–36 (cit. on p. 49).

[134] R. Valavi, J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. „blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models“. In: *Biorxiv* (2018), p. 357798 (cit. on p. 49).

[135] J. Pohjankukka, T. Pahikkala, P. Nevalainen, and J. Heikkonen. „Estimating the prediction performance of spatial models via spatial k-fold cross validation“. In: *International Journal of Geographical Information Science* 31.10 (2017), pp. 2001–2019 (cit. on p. 49).

[136] Y. Wang, M. Khodadadzadeh, and R. Zurita-Milla. „Spatial+: A new cross-validation method to evaluate geospatial machine learning models“. In: *International Journal of Applied Earth Observation and Geoinformation* 121 (2023), p. 103364 (cit. on p. 49).

[137] A. M.-C. Wadoux, G. B. Heuvelink, S. De Bruin, and D. J. Brus. „Spatial cross-validation is not the right way to evaluate map accuracy“. In: *Ecological Modelling* 457 (2021), p. 109692 (cit. on p. 49).

[138] E. C. Tetila, B. B. Machado, G. Astolfi, et al. „Detection and classification of soybean pests using deep learning with UAV images“. In: *Computers and Electronics in Agriculture* 179 (2020), p. 105836 (cit. on p. 50).

[139] G. Tseng, K. Sinkovics, T. Watsham, D. Rolnick, and T. C. Walters. „Semi-supervised object detection for agriculture“. In: *2nd AAAI Workshop on AI for Agriculture and Food Systems*. 2023 (cit. on p. 51).

[140]P. Gómez and G. Meoni. „MSMatch: Semisupervised multispectral scene classification with few labels“. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), pp. 11643–11654 (cit. on p. 51).

[141]J. Gou, B. Yu, S. J. Maybank, and D. Tao. „Knowledge distillation: A survey“. In: *International journal of computer vision* 129.6 (2021), pp. 1789–1819 (cit. on p. 51).

[142]B. P. Salmon, J. C. Olivier, K. J. Wessels, et al. „Unsupervised land cover change detection: Meaningful sequential time series analysis“. In: *IEEE Journal of selected topics in applied earth observations and remote sensing* 4.2 (2010), pp. 327–335 (cit. on p. 52).

[143]T. Akilan, N. Jahan, and W. Zhang. „Self-Supervised Learning for Image Segmentation: A Comprehensive Survey“. In: *arXiv preprint arXiv:2505.13584* (2025) (cit. on p. 52).

[144]C. F. Brown, M. R. Kazmierski, V. J. Pasquarella, et al. „Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data“. In: *arXiv preprint arXiv:2507.22291* (2025) (cit. on p. 52).

[145]Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu. „Self-supervised learning in remote sensing: A review“. In: *IEEE Geoscience and Remote Sensing Magazine* 10.4 (2022), pp. 213–247 (cit. on p. 52).

[146]M. Wu, C. Li, and Z. Yao. „Deep active learning for computer vision tasks: Methodologies, applications, and challenges“. In: *Applied Sciences* 12.16 (2022), p. 8103 (cit. on p. 53).

[147]S. Miao, C. Wang, G. Kong, et al. „Utilizing active learning and attention-CNN to classify vegetation based on UAV multispectral data“. In: *Scientific Reports* 14.1 (2024), p. 31061 (cit. on p. 53).

[148]S. Thota, S. Chitta, V. K. R. Vangoor, C. S. Ravi, and V. S. M. Bonam. „Few-ShotLearning in Computer Vision: Practical Applications and Techniques“. In: *Human-Computer Interaction* 3.1 (2023) (cit. on p. 53).

[149]S. Ruan, H. Cang, H. Chen, et al. „Hyperspectral classification of frost damage stress in tomato plants based on few-shot learning“. In: *Agronomy* 13.9 (2023), p. 2348 (cit. on p. 53).

[150]C. T. Marimo, B. Blumenstiel, M. Nitsche, J. Jakubik, and T. Brunschwiler. „Beyond the visible: Multispectral vision-language learning for earth observation“. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2025, pp. 359–375 (cit. on pp. 53, 57).

[151]X. Li, Y. Grandvalet, F. Davoine, et al. „Transfer learning in computer vision tasks: Remember where you come from“. In: *Image and Vision Computing* 93 (2020), p. 103853 (cit. on p. 54).

[152]M. I. Hossen, M. Awrangjeb, S. Pan, and A. A. Mamun. „Transfer learning in agriculture: a review“. In: *Artificial Intelligence Review* 58.4 (2025), p. 97 (cit. on p. 54).

[153]H. Qu, H. Rahmani, L. Xu, B. Williams, and J. Liu. „Recent advances of continual learning in computer vision: An overview“. In: *IET Computer Vision* 19.1 (2025), e70013 (cit. on p. 54).

[154] S. Illarionova, S. Nesteruk, D. Shadrin, et al. „MixChannel: Advanced augmentation for multispectral satellite images“. In: *Remote Sensing* 13.11 (2021), p. 2181 (cit. on p. 56).

[155] A. Heschl, M. Murillo, K. Najafian, and F. Maleki. „SynthSet: generative diffusion model for semantic segmentation in precision agriculture“. In: *European Conference on Computer Vision*. Springer. 2024, pp. 267–283 (cit. on p. 56).

[156] S. Xiang, P. M. Blok, J. Burridge, H. Wang, and W. Guo. „DODA: Adapting Object Detectors to Dynamic Agricultural Environments in Real-Time with Diffusion“. In: *arXiv preprint arXiv:2403.18334* (2024) (cit. on p. 56).

[157] M. Awais, M. Naseer, S. Khan, et al. „Foundation models defining a new era in vision: a survey and outlook“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025) (cit. on p. 57).

[158] A. Kirillov, E. Mintun, N. Ravi, et al. „Segment anything“. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 4015–4026 (cit. on p. 57).

[159] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, et al. „Vision foundation models in remote sensing: A survey“. In: *IEEE Geoscience and Remote Sensing Magazine* (2025) (cit. on p. 57).

[160] Y. Li, B. Dang, F. Wei, J. Tan, and Y. Lin. „Domain Knowledge-Aware Remote Sensing Foundation Model for Flood Detection in Multi-Spectral Imagery“. In: *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2024, pp. 785–789 (cit. on p. 57).

[161] Z. Huang, H. Yan, Q. Zhan, et al. „A survey on remote sensing foundation models: From vision to multimodality“. In: *arXiv preprint arXiv:2503.22081* (2025) (cit. on p. 57).

[162] D. Hong, B. Zhang, X. Li, et al. „SpectralGPT: Spectral remote sensing foundation model“. In: *arXiv preprint arXiv:2311.07113* (2023) (cit. on p. 58).

[163] X. Guo, J. Lao, B. Dang, et al. „Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 27672–27683 (cit. on p. 58).

[164] D. Szwarcman, S. Roy, P. Fraccaro, et al. „Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications“. In: *IEEE Transactions on Geoscience and Remote Sensing* (2025) (cit. on p. 58).

[165] A. Lacoste, N. Lehmann, P. Rodriguez, et al. „Geo-bench: Toward foundation models for earth monitoring“. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 51080–51093 (cit. on p. 58).

[166] J. Jakubik, F. Yang, B. Blumenstiel, et al. „Terramind: Large-scale generative multimodality for earth observation“. In: *arXiv preprint arXiv:2504.11171* (2025) (cit. on p. 58).

[167] S. Khan, M. Tufail, M. T. Khan, et al. „Real-time recognition of spraying area for UAV sprayers using a deep learning approach“. In: *Plos one* 16.4 (2021), e0249436 (cit. on p. 59).

[168]P. Miralles, K. Thangavel, A. F. Scannapieco, et al. „A critical review on the state-of-the-art and future prospects of Machine Learning for Earth Observation Operations“. In: *Advances in Space Research* 71.12 (2023), pp. 4959–4986 (cit. on p. 59).

[169]J. Kim, S. Chang, and N. Kwak. „PQK: model compression via pruning, quantization, and knowledge distillation“. In: *arXiv preprint arXiv:2106.14681* (2021) (cit. on p. 59).

[170]N. Ahmed and W. Farooqi. „Deep Learning in Remote Sensing–Driven Precision Agriculture: A Comparative Review of Models, Data Modalities, and Environmental Applications“. In: *Data Modalities, and Environmental Applications* () (cit. on p. 60).

[171]A. Günther, H. Najjar, and A. Dengel. „Explainable multimodal learning in remote sensing: Challenges and future directions“. In: *IEEE Geoscience and Remote Sensing Letters* 21 (2024), pp. 1–5 (cit. on pp. 60, 61).

[172]C. O. Retzlaff, A. Angerschmid, A. Saranti, et al. „Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists“. In: *Cognitive Systems Research* 86 (2024), p. 101243 (cit. on p. 60).

[173]S. Stassin, A. Englebert, G. Nanfack, et al. „An experimental investigation into the evaluation of explainability methods for computer vision“. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2023, pp. 91–106 (cit. on p. 60).

[174]U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim. „Towards a rigorous evaluation of XAI methods on time series“. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE. 2019, pp. 4197–4201 (cit. on p. 60).

[175]Z. Cheng, Y. Wu, Y. Li, L. Cai, and B. Ihnaini. „A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision“. In: *Sensors* 25.13 (2025), p. 4166 (cit. on p. 60).

[176]A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek. „Explainable AI methods-a brief overview“. In: *International workshop on extending explainable AI beyond deep models and classifiers*. Springer. 2020, pp. 13–38 (cit. on p. 60).

[177]R. Kashefi, L. Barekatain, M. Sabokrou, and F. Aghaeipoor. „Explainability of vision transformers: A comprehensive review and new perspectives“. In: *Multimedia Tools and Applications* 85.2 (2026), p. 115 (cit. on p. 61).

[178]N. M. James, H. Kansara, and R. Agrawal. „Temporal Crop Stress Analysis on Multi-Spectral Vegetation Indices using Contrastive Learning and Explainable Ensemble Models“. In: *2025 5th Asian Conference on Innovation in Technology (ASIANCON)*. IEEE. 2025, pp. 1–8 (cit. on p. 61).

[179]Y. Naceur, S. Bouzidi, and M. Zaouali. „U-Net for Remote Sensing: A Spectral Index-Based Approach With Explainable AI for Robust Change Detection“. In: *IEEE Transactions on Geoscience and Remote Sensing* (2025) (cit. on p. 61).

[180]L. Forsythe. „From Vision to Reasoning: Leveraging Deep Learning for Enhancing Large Language Models in Multimodal Understanding“. In: *Transactions on Computational and Scientific Methods* 5.10 (2025) (cit. on p. 61).

[181]E. Hüllermeier and W. Waegeman. „Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods“. In: *Machine learning* 110.3 (2021), pp. 457–506 (cit. on p. 62).

[182] S. De Backer, P. Kempeneers, W. Debruyn, and P. Scheunders. „A band selection technique for spectral classification". In: *IEEE geoscience and remote sensing letters* 2.3 (2005), pp. 319–323 (cit. on p. 62).

[183] W. Sun and Q. Du. „Hyperspectral band selection: A review". In: *IEEE Geoscience and Remote Sensing Magazine* 7.2 (2019), pp. 118–139 (cit. on p. 62).

[184] T. Barros, P. Conde, G. Gonçalves, et al. „Multispectral vineyard segmentation: A deep learning comparison study". In: *Computers and electronics in agriculture* 195 (2022), p. 106782 (cit. on p. 62).

[185] H. Najjar, M. Miranda, M. Nuske, R. Roscher, and A. Dengel. „Explainability of Sub-Field Level Crop Yield Prediction using Remote Sensing". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2025) (cit. on p. 62).

[186] A. Ifarraguerri and M. W. Prairie. „Visual method for spectral band selection". In: *IEEE Geoscience and Remote Sensing Letters* 1.2 (2004), pp. 101–106 (cit. on p. 63).

[187] M. Esmaeili, D. Abbasi-Moghadam, A. Sharifi, A. Tariq, and Q. Li. „Hyperspectral image band selection based on CNN embedded GA (CNNeGA)". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), pp. 1927–1950 (cit. on p. 63).

[188] M. Fawakherji, A. Youssef, D. Bloisi, A. Pretto, and D. Nardi. „Crop and weeds classification for precision agriculture using context-independent pixel-wise segmentation". In: *2019 third IEEE international conference on robotic computing (IRC)*. IEEE. 2019, pp. 146–152 (cit. on p. 69).

[189] A. Milioto, P. Lottes, and C. Stachniss. „Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs". In: *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2018, pp. 2229–2235 (cit. on p. 69).

[190] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson. „Machine learning algorithm validation with a limited sample size". In: *PloS one* 14.11 (2019), e0224365 (cit. on p. 70).

[191] J. M. Durden, B. Hosking, B. J. Bett, D. Cline, and H. A. Ruhl. „Automated classification of fauna in seabed photographs: The impact of training and validation dataset size, with considerations for the class imbalance". In: *Progress in Oceanography* 196 (2021), p. 102612 (cit. on p. 70).

[192] I. Shumailov, Z. Shumaylov, Y. Zhao, et al. „The curse of recursion: Training on generated data makes models forget". In: *arXiv preprint arXiv:2305.17493* (2023) (cit. on p. 70).

[193] H. Yu, B. Kong, Q. Wang, X. Liu, and X. Liu. „Hyperspectral remote sensing applications in soil: a review". In: *Hyperspectral remote sensing* (2020), pp. 269–291 (cit. on p. 73).

[194] Y.-Q. Song, X. Zhao, H.-Y. Su, et al. „Predicting spatial variations in soil nutrients with hyperspectral remote sensing at regional scale". In: *Sensors* 18.9 (2018), p. 3086 (cit. on p. 73).

[195] J. Reyes, W. Wiedemann, A. Brand, J. Franke, and M. Ließ. „Predictive monitoring of soil organic carbon using multispectral UAV imagery: a case study on a long-term experimental field". In: *Spatial Information Research* 32.6 (2024), pp. 683–696 (cit. on p. 73).

[196] V. Zaigrajew, H. Baniecki, L. Tulczyjew, et al. „Red teaming models for hyperspectral image analysis using explainable AI“. In: *arXiv preprint arXiv:2403.08017* (2024) (cit. on p. 74).

[197] Y. Zhai, L. Zhou, H. Qi, P. Gao, and C. Zhang. „Application of visible/near-infrared spectroscopy and hyperspectral imaging with machine learning for high-throughput plant heavy metal stress phenotyping: A review“. In: *Plant Phenomics* 5 (2023), p. 0124 (cit. on p. 74).

[198] J. Nalepa, B. Le Saux, N. Longépé, et al. „The hyperview challenge: Estimating soil parameters from hyperspectral images“. In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022, pp. 4268–4272 (cit. on p. 77).

[199] R. S. Kuzu, F. Albrecht, C. Arnold, R. Kamath, and K. Konen. „Predicting soil properties from hyperspectral satellite images“. In: *2022 ieee international conference on image processing (icip)*. IEEE. 2022, pp. 4296–4300 (cit. on p. 77).

[200] S. Jain, D. Sethia, and K. C. Tiwari. „A critical systematic review on spectral-based soil nutrient prediction using machine learning“. In: *Environmental Monitoring and Assessment* 196.8 (2024), p. 699 (cit. on p. 80).

[201] P. Aszkowski, B. Ptak, M. Kraft, D. Pieczyński, and P. Drapikowski. „Deepness: Deep neural remote sensing plugin for QGIS“. In: *SoftwareX* 23 (2023), p. 101495 (cit. on p. 83).

[202] J. L. Aroca-Trujillo and A. Perez-Ruiz. „Colombian coffee tree leaves multispectral images dataset“. In: *Data in Brief* 59 (2025), p. 111421 (cit. on p. 89).

[203] A. Salimi, M. Ziaii, A. Amiri, et al. „Using a Feature Subset Selection method and Support Vector Machine to address curse of dimensionality and redundancy in Hyperion hyperspectral data classification“. In: *The Egyptian Journal of Remote Sensing and Space Science* 21.1 (2018), pp. 27–36 (cit. on p. 95).

[204] D. Deb and U. Verma. „Correlation-Based Band Selection for Hyperspectral Image Classification“. In: *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2025, pp. 8865–8869 (cit. on p. 95).

[205] A. Panthakkan, S. Anzar, K. Sherin, S. A. Mansoori, and H. Al-Ahmad. „Evaluation of UAV-Based RGB and Multispectral Vegetation Indices for Precision Agriculture in Palm Tree Cultivation“. In: *arXiv preprint arXiv:2505.07840* (2025) (cit. on p. 100).

[206] P. López-García, D. Intrigliolo, M. A. Moreno, et al. „Machine learning-based processing of multispectral and RGB UAV imagery for the multitemporal monitoring of vineyard water status“. In: *Agronomy* 12.9 (2022), p. 2122 (cit. on p. 100).

[207] H. Grybas and R. G. Congalton. „A comparison of multi-temporal RGB and multispectral UAS imagery for tree species classification in heterogeneous New Hampshire Forests“. In: *Remote Sensing* 13.13 (2021), p. 2631 (cit. on p. 100).