National Technical University of Athens

School of Rural, Surveying and Geoinformatics Engineering

MSc Geoinformatics

Thesis Name:

Maturity Estimation of Open-field Brassicaceae crops using Remote Sensing and Deep Convolutional Networks

Vasilis Psiroukis

13 December, 2021

Supervisor: Karatzalos Konstantinos

Thesis Evaluation Committee

Karantzalos Konstantinos

Associate Professor, National Technical University of Athens
Laboratory of Remote Sensing
Department of Topography
School of Rural, Surveying and Geoinformatics Engineering

Karathanasi Vasilia

Professor, National Technical University of Athens
Laboratory of Remote Sensing
Department of Topography
School of Rural, Surveying and Geoinformatics Engineering

Fountas Spyros

Associate Professor, Agricultural University of Athens
Laboratory of Agricultural Engineering
Department of Natural Resources Management and Agricultural Engineering

Abstract

Broccoli is an example of a high-value crop that requires delicate handling throughout the growing season and during its post-harvesting treatment. As the broccoli heads can be easily damaged, resulting in visible stains, it is thus still harvested by hand using handheld knives. On top of that, it allows for a very strict time window of "optimal maturity" when the high-end quality broccoli heads should be harvested, before they remain exposed for too long in high humidity conditions and become susceptible to fungal infections and quality degradation. Even slight delays from this time window can result in major losses in final production, while manual harvesting is a very laborious task, not only for the process of harvesting itself, but for the scouting required to initially identify the field segments where several broccoli plants have reached this maturity level. The aim of this study is to automate this process, by using a state-of-the-art Object Detection deep convolutional neural network model, YOLOv5, trained on multispectral UAV images collected from low altitude flights, and assess its capacity to effectively detect and classify broccoli heads based on their maturity level. The experiment took place in Marathon region, Greece, in a commercial organic vegetable production unit. This region is specifically known for its horticultural production, being the main vegetable provider for Athens, the capital of Greece, while the timing of the data acquisition flights was specifically designed to be performed a few hours prior to the first wave of selective harvesting. The training of the Object Detection model was conducted using various training hyperparameters and dataset (multispectral layers) configurations. The results of the training and validation experiments indicated that the model was able to perform very well for the task of automated maturity detection. All experimental iterations maintained an F-1 score higher than 0.81 and a MAP@0.5 higher than 0.865, which are solid performances considering the open-field nature of the datasets.

Περίληψη

Το οργανικό μπρόκολο είναι ένα παράδειγμα καλλιέργειας υψηλής αξίας που απαιτεί λεπτούς χειρισμούς καθ' όλη τη διάρκεια της καλλιεργητικής περιόδου και κατά τη μετασυλλεκτική του επεξεργασία. Καθώς τα κεφάλια του μπρόκολου, τα οποία αποτελούν το εμπορεύσιμο μέρος του φυτού, μπορούν εύκολα να υποστούν ζημιά, με αποτέλεσμα να δημιουργούνται ορατοί τραυματισμοί και χρωματισμοί στην επιφάνειά τους, η συγκομιδή γίνεται ακόμη με το χέρι χρησιμοποιώντας μαχαίρια. Συν τοις άλλοις, επιτρέπει ένα πολύ αυστηρό χρονικό παράθυρο "βέλτιστης ωριμότητας" κατά το οποίο πρέπει να συγκομιστούν τα κεφάλια μπρόκολου υψηλής ποιότητας, πριν παραμείνουν εκτεθειμένα για πολύ καιρό σε συνθήκες υψηλής υγρασίας και γίνουν ευάλωτα σε μυκητολογικές μολύνσεις και υποβάθμιση της ποιότητας. Ακόμη και μικρές καθυστερήσεις από αυτό το χρονικό παράθυρο μπορούν να οδηγήσουν σε μεγάλες απώλειες στην τελική παραγωγή, ενώ η χειρωνακτική συγκομιδή είναι μια πολύ επίπονη εργασία, όχι μόνο για την ίδια τη διαδικασία συγκομιδής, αλλά και για την ανίχνευση που απαιτείται για τον αρχικό εντοπισμό των τμημάτων του χωραφιού όπου πολλά φυτά μπρόκολου έχουν φτάσει σε αυτό το επίπεδο ωριμότητας. Στόχος της παρούσας μελέτης είναι η αυτοματοποίηση αυτής της διαδικασίας, με τη χρήση ενός μοντέλου βαθιού συνελικτικού νευρωνικού δικτύου ανίχνευσης αντικειμένων, του YOLOv5, το οποίο εκπαιδεύτηκε σε πολυφασματικές εικόνες που συλλέχθηκαν από πτήσεις χαμηλού υψομέτρου με Συστήματα Μη-Επανδρωμένων Αεροσκαφών (ΣΜηΕΑ), και η αξιολόγηση της ικανότητάς του να ανιχνεύει και να ταξινομεί αποτελεσματικά τα κεφάλια μπρόκολου με βάση το επίπεδο ωριμότητάς τους. Το πείραμα πραγματοποιήθηκε στην περιοχή του Μαραθώνα, σε μια εμπορική μονάδα παραγωγής βιολογικών λαχανικών. Η περιοχή αυτή είναι ιδιαίτερα γνωστή για την κηπευτική της παραγωγή, καθώς αποτελεί τον κύριο προμηθευτή λαχανικών για την Αθήνα, την πρωτεύουσα της Ελλάδας, ενώ ο χρόνος των πτήσεων συλλογής δεδομένων σχεδιάστηκε ειδικά ώστε να πραγματοποιείται λίγες ώρες πριν από το πρώτο κύμα επιλεκτικής συγκομιδής, ώστε να επιτύχει τα εξής: 1) να διασφαλιστεί ότι ολόκληρος ο αγρός ήταν άθικτος, μεγιστοποιώντας τον αριθμό της πυκνότητας των δειγμάτων σε κάθε εικόνα και το παραγόμενο ορθομωσαϊκό του αγρού και 2) για να διασφαλιστεί ότι μεμονωμένα φυτά διαφορετικών επιπέδων ωριμότητας υπήρχαν σε ολόκληρο τον αγρό, καθώς τότε ξεκινούσε η περίοδος συγκομιδής. Η εκπαίδευση του μοντέλου ανίχνευσης αντικειμένων διεξήχθη χρησιμοποιώντας διάφορους συνδυασμούς υπερπαραμέτρων εκπαίδευσης και συνόλου δεδομένων (φασματικά επίπεδα). Τα αποτελέσματα των πειραμάτων εκπαίδευσης και επαλήθευσης έδειξαν ότι το μοντέλο ήταν σε θέση να εκτελέσει πολύ καλά το έργο της αυτοματοποιημένης ανίχνευσης ωριμότητας. Όλες οι πειραματικές επαναλήψεις διατήρησαν F-1 score μεγαλύτερο από 0,81 και MAP@0.5 μεγαλύτερη από 0,865.

Table of Contents

1. Introduction

1.1 Precision Agriculture and the Food Safety Problem

The most important challenge of modern agriculture is to achieve food security for the constantly increasing global population, estimated to reach 10 billion people by 2050 (FAO, 2017). To assist in this endeavor, during the last five years alone, the total volume of investments in the agricultural sector has increased by 80%, aiming to achieve a productivity growth of 70% by 2050 under the scenario of climate change and an expected reduction of cultivated agricultural lands (Tsouros et al., 2019).

In traditional industrial agriculture, growers manage their fields uniformly, by performing similar operations and applying inputs at the same rate and frequency across the entire cultivated area. However, they are aware of underlying differences within their fields, as specific regions produce higher and better yield than others. Micro-differences in topological, chemical, and biological parameters of the crops' growing environment directly affect nutrient availability and overall plant health, thus creating this variability. For instance, field slope can affect soil water and nutrient retention, while fungal infections tend to manifest in colder and more humid segments of the field. The same pattern applies in crop diseases and weed infestations. Hence, growing conditions and stress factors will not be consistent across a growing area, and crop productivity will naturally also vary.

When these inherent differences are not considered appropriately, agricultural inputs and operations are often mismanaged, by committing more resources to the "wrong" zones, unnecessarily adding to the cost of food production and increasing the environmental impact of the sector. To address these issues, a new field management system emerged during the 1990s', when tools such as positioning systems (i.e. the Global Positioning System/GPS that became fully operational in 1993) that enabled the monitoring of the aforementioned variability, became widely accessible. The fundamental challenge was to optimally manage all zones within a field according to their different production capacities. This way, a new farming management concept, that was based upon observing, measuring and responding to inter and intra-field variability in crops emerged, and became widely known as Precision Agriculture (PA). Many definitions of PA exist and many people have different ideas of what PA should encompass. One of the most widely used definitions comes from the US House of Representatives (1997), referring to PA as "an integrated information and production-based farming system that is designed to increase long term, site-specific and whole farm production efficiency, productivity and profitability while minimizing unintended impacts on wildlife and the environment".

In general, PA is a crop management system that, instead of managing whole fields as a single unit, attempts to divide them into smaller segments, and then distribute inputs and resources according to their actual needs. The ultimate objective of PA is to optimise production efficiency and yield quality while minimizing environmental impact and food safety risk. This is achieved through the deployment of novel technological components that generate data streams of multiple environmental, soil and crop parameters, thus enabling a data-driven decision making production approach. These technologies vary, from field-level stationary or machinery-mounted sensors, to aerial systems and satellite platforms.

1.2 Yield and Maturity Estimation

Crop yield is the most important piece of information for crop management in precision agriculture. Early crop yield estimations are valuable insights that allow farmers to optimize farm-level decisions, like the upcoming farming operations scheduling or farm management decisions such as whether to sell or store the final product. At the same time, yield forecasts can act as an assessment tool for the expected income of each production unit, while also enabling an efficient transfer of the production from the farm to the food supply chain.

Traditionally, crop yield estimations were performed by in-field sampling, such as destructive methods that include biomass weighting and grain size measurements during the latter growth stages of the crops. Naturally, limitations such as the difficulty in applying manual sampling in large fields and the high level of uncertainty in the estimations due to the low volume of data provided set such methods as ineffective and unreliable. The development of crop-specific models provides an accessible solution that can drastically decrease uncertainty while minimizing the efforts required for their applications. Yield and maturity forecasts using agrometeorological data as inputs for a statistical regression is a common technique that has been used in many experiments (i.e. Lobell et al., 2009) and research programs (NASS, 2006). After data from multiple years has been collected and used to create a matrix of historic yield data and recordings of several agrometeorological parameters, a regression equation is derived that describes yield or maturity as a function of multiple agronometeorological parameters. Additionally, during the development phase, potential stress instances are identified and registered as risk factors which are also integrated into the models. This way, the models become robust and capable of providing accurate forecasts even in unusual situations, such as the occurrence of a heat stress or frost during certain growth stages that ultimately affect yield. Such models naturally become more accurate and robust as the volume of data provided as inputs is increased. The recent boom in agricultural data caused by the widespread accessibility of sensors in the agricultural sector can allow for more complex models to be adopted, by providing constant datastreams generated from sensing networks in each field separately throughout the cultivation periods.

Crop-specific yield and maturity prediction models are developed by identifying which parameters influence crop growth and development the most. These are the factors that cause large spatial yield variability and crop development within a single growing season, and include weather conditions, soil properties and farm management procedures such as tillage and irrigation. For maturity detection approaches using imagery data, the critical parameters are often distinguishable through visual characteristics of the crops. These factors include but are not limited to changes in spectral reflectance for cases were low resolution imagery data are used for field-level estimations (i.e. satellite data), or crop-level development stage identification, such as blooming, in cases were higher resolution data are used (i.e. low altitude aerial or proximal imagery). Once these factors and their level of influence on crop growth have been determined through experimental trials, the weight of each factor is assigned to the respective parameter in the model's mathematical equations. As a result, in the following growing seasons these factors are then used as inputs in the model, generating estimations for crop growth, development and ultimately yield based on data from the current season (Hogenboom et al., 2004).

As yield is highly correlated with biomass, it is used to estimate yield to be harvested on the fields several weeks in advance, providing a valuable tool for harvest planning and farm economics to the farmers. In the case of remote sensing based models, field-level measurements collected in the experimental sites are used as reference data for validation of the remote imagery, as well as a calibration factor of the models. In case a crop-type demonstrates rapid change in its characteristics, the correlation of a certain parameter with final yield is considered significant only after a certain growth stage (i.e. the flowering of Brassicaceae plants), with this period being identified as a critical stage, and data during this period are assigned a higher weight due to their direct correlation with maturity level and final yield.

1.3 Remote Sensing

1.3.1 Remote Sensing in Agriculture

Remote sensing is the ability to collect information of an object without any physical contact with the object itself (Lillesand and Keifer, 1994). More specifically, it is the process of detecting and monitoring the physical characteristics of an area or target by measuring its reflected and/or emitted radiation from a distance. Therefore, remote sensing sources in general context include any form of data, collected from several meters (Unmanned Aerial Vehicles), up to several hundred kilometers (satellite platforms) above the target. Remote sensors can be either passive or active. Passive sensors respond to external stimuli and record the energy that is reflected or emitted from the observed target. The most common source of radiation detected by passive sensors is reflected sunlight. On the other hand, active sensors use internal stimuli to collect data, by projecting energy to the target and measuring a specific property of the process. For instance, in the case of a laser sensor, the measured attribute is the time that it takes for the energy to reflect back to its sensor and be recorded. Human vision is a form of remote sensing, and more specifically, a passive one. Through our eyes, the target's reflection is converted into a set of information such as its color and shape. Visible light in the form we perceive it, also known as the visible spectrum, is described by a wave function, the frequency of which determines what we perceive as "color" for everything we see. However, this is but a very small part of the electromagnetic spectrum.

Depending upon the wavelength of the energy and characteristics of an object, energy gets reflected, absorbed, or transmitted. Generally, all objects have a high reflectance in the color spectrum we see, as they absorb energy in all spectrums of the visible light, except the one that reaches our eyes. If an object changes color, it becomes distinguishable because essentially the change in reflection took place within the visible spectrum. For example, an unripe green-yellow tomato has ripened and it now appears red, because in its current state it reflects more light in the range of 650-730 nm, which corresponds to the red color. Such observations have led to the following questions: since changes in the visible spectrum are immediately perceptible to us, and have an effect on the properties of the object, what happens to changes in spectra that we cannot see with our eyes, and of course, what kind of information can we to take if we can measure them?

Plants have a special behavior in terms of light reflectance, as photosynthetically-active tissues demonstrate high absorption of Red light, while reflecting the largest portion in the Green spectrum -

and therefore appear green to us (Tucker, 1979). Photosynthetically active radiation (PAR) is the spectral range of solar radiation from 400 to 700 nm that photosynthetic organisms are able to use in the process of photosynthesis. However, there is another unique ability of vegetation: it also reflects most of the Near-Infrared (NIR) light that hits it, due to the spongy mesophyll (Tucker, 1979). In this area, sturdy plants have very high rates of reflection, and under stress plants under reduction show a decrease in its reflection. Therefore, healthy plants with high photosynthetic activity normally reflect most of the NIR light that hits their leaves (700-800 nm), while they absorb most of the Red one (630-670 nm). If this is not the case, it is an indicator that something prevents the plants from performing their biological functions properly. Such factors can vary, from a source of stress, to a disease infestation (Tucker et al., 1980). These deviations can be detected quicker by identifying variations in reflectance, before they become visible to the human eye, via specific symptoms such as chlorosis, when the leaves no longer appear green to us. This reflectance behaviour or "profile" is called spectral signature, and is the basis for remote sensing applications in agriculture as it can provide direct information about crop health and overall condition. Remote sensing leverages on this spectral signature to estimate the properties of plants through non-destructive processes in a fast and accurate way (Moran et al., 1997), and has a wide range of applications in agriculture, including but not being limited to crop growth monitoring, crop yield and quality estimation, identification of irrigation needs, as well as biotic and abiotic damage such as pests infestations, disease infections, hail damage, flood and drought damage (Mondal and Basu, 2009).

The characteristic property of the spectral signature and the potential it held in all these applications in agriculture sparked the rapid development of numerical indices resulting from simple formulas using the percentage reflections in selected "critical" spectra. These indices are called vegetation indices (VIs), and are mathematical quantitative combinations of the absorption and scattering rates of plants in different bands of the electromagnetic spectrum used to detect, quantify and monitor specific crop-related parameters. VIs provide a simple yet elegant method for measuring plant responses throughout the season, exploiting the basic differences between soil and plant spectra, and are often calculated as a type of relationship between reflected light in the visible and NIR bands. Many such indices were rapidly created to cover a variety of different applications since satellite imagery became widely accessible, starting in 1972 (Cihlar et al., 1991), with the launch of Landsat 1. In 1973, however, Rousse introduced the Normalized Difference Vegetation Index (NDVI), calculated as the ratio of the difference and the sum of the reflectance in the NIR and red bands, which remains the most widely used index to this day. As of today, several hundreds of VIs have been developed, to address specific problems in the agricultural sector alone, such as crop monitoring, disease detection, irrigation and input applications optimisation, crop identification, as well as yield estimation and maturity identification (Anastasiou et al., 2018). Naturally, the first remote sensing datasources integrated to yield forecasting models where satellite imagery derived VIs (Tucker et al., 1980), due to the ability of spectral data (and especially in the NIR-based VIs) to calculate biomass and crop vigour, parameters directly related to final yield. High image acquisition costs and low spatial resolution were always an obstacle in many applications and adoption of satellite imagery, however, recent open-source satellite platforms such as ESA’s Sentinel-2 can provide imagery of very high resolution without any cost to the user. Nevertheless, inherent limitations of satellite data still exist to this day.

1.3.2 Levels of Remote Sensing

The collection of remote sensing imagery data is achieved with a common principle, "isolating" specific spectral bands using sensors that allow the collection of these data, and is performed in three (3) levels, based on the proximity of the platform that carries the sensors (and therefore sensing distance) to the target or observed area. The first level consists of satellite platforms, orbiting at 700-800 km above the earth's surface. Although modern satellite instruments can offer very high resolution, large-scale multispectral imagery even in the form of open data (Sentinel-2 offers 10x10 m2 resolution for the most commonly used bands), problems such as cloud coverage and the revisiting frequency (the interval between each time a satellite platform of the constellation captures data of an area during its orbit) are inherent to this sensing form and will most likely continue to remain for as long this remote sensing method is used (Zheng et al., 2017). The major advantage of satellite remote sensing is that they require no physical presence on the observed area whatsoever, and in the case of open data, there are practically zero costs associated with their acquisition, as no sensing equipment is required (Segarra et al., 2020). Furthermore, they often require little expertise and technical knowledge to perform simple tasks, such as the generation of a vegetation index map and a simple statistical analysis of an agricultural field. Finally, satellites provide long-lasting archived data time-series, e.g. in the case of Landsat, reaching up to 50 years in the past when the first Landsat satellite was launched to orbit.

The second level is the oldest remote sensing application, and it involves the collection of data from the ground at a proximal level. This method naturally offers the highest possible accuracy, reaching sub-meter sensing distance from the target, but with data collection being more time consuming, laborious and challenging. This form of data collection not only requires physical presence at the observed area, but also at each sampling location itself. There is some confusion on whether this method of data collection can truly be considered "remote", but if we follow the fundamental term of remote sensing, then it naturally falls under this category as no direct contact with observed targets occurs in any form during data acquisition. In the case of agricultural proximal remote sensing, the process is usually performed with the sensing system mounted on a suitable terrestrial vehicle that cruises across the field, but can often be performed on foot, mostly in cases of horticultural crops where the risk of soil compaction and high vegetation coverage in late crop growth stages do not allow vehicles to cross the field. Furthermore, for mountainous terrains such as orchards, vineyards or grasslands with steep slopes, this form of sensing is not only very difficult to perform, but also dangerous. Generally, collection of reflectance data on a proximal level is performed as ground truth data, to validate the data of other, more distant remote sensing sources (Zhou et al., 2018).

At the final level we encounter aerial remote sensing. This category can be further divided into two (2) sub-categories, based on the platforms used, and once again, the proximity - in this case, the altitude of the platform. The first one is aerial photography performed by manned aircrafts, which was the first form of aerial surveying, and usually takes place at altitudes of around 500 m above ground. The first instance of this remote sensing form was performed by Gaspard-Félix Tournachon, a French photographer better known with his artistic name Nadar, who captured photographs of Paris, France, from a hot air balloon in 1858 (Holmes, 2013). In modern manned-aircraft aerial imagery, the data collection process involves light aircrafts carrying the sensing systems, which consist of the sensing

devices (cameras) accompanied by stabilisation components as well as positioning and orientation instruments. This method naturally offers significantly higher resolution and overall data quality compared to satellite imagery, although the high cost acquisitions and the sheer difficulty of both data acquisition and processing have resulted in this method becoming less and less popular in agriculture, as other, easier to deploy options such as UAVs become available and more accessible. However, as manned aircrafts are more weather tolerant and wind resilient, they still often remain the only viable option when imagery of areas with sustained high wind speeds and strong gusts should be collected (Liknes et al., 2010). In the second sub-category of aerial imagery we encounter UAVs, the most recent form of remote sensing.

The second sub-category of aerial imagery consists of the most recent type of remote sensing platforms, UAVs, which in recent years have become a hot spot for practical research and applications across the agricultural sector. UAVs have a unique ability to collect high quality imagery data in a very efficient way, enabling the conduction of a broad range of surveying operations and numerous applications in cultivated fields. UAV remote sensing provides a nondestructive and cost-effective way for rapid monitoring of agricultural fields using accessible and cost-effective platforms, capable of flying at low altitudes and capturing images of unparalleled spatial resolution (Matese et al., 2015). Moreover, as the flights are performed by human operators on demand and are much easier to be deployed than manned aircrafts, UAVs offer a potentially very high temporal resolution. The only barriers are the environmental conditions that may not allow for the aircrafts safe operation (Burkart et al., 2015). The choice of acquisition timing and frequency is an essential characteristic for multiple agricultural applications with strict time windows. Moreover, the ability of UAVs to acquire data close to the surface and the integration of correction systems (i.e. sunlight correction sensors) makes UAV imagery collection unaffected by cloud coverage (Berni et al., 2009), similarly to manned aerial flights. Another advantage of UAVs is their capacity to carry multiple sensors, as different sensing systems can easily be mounted and unmounted through aircraft modifications and gimbals. Naturally, different sensors cannot perform optimal data collection during a single mission flight, as their required flight parameters may vary, thus individual flights for each sensor may be demanded. Nevertheless, the possibility of a single aircraft being capable of generating multiple datasets is still a major advantage, especially when taking into consideration the time efficiency of UAV flights and the relatively low acquisition costs of both the aircrafts and their sensing components.

From what has been described so far, it would appear that UAVs maintain all the advantages of other remote sensing platforms, without any of their limitations. This is obviously not the case, as there are several drawbacks associated with both their operation and data handling. Several studies have investigated the UAVs' advantages and disadvantages over other platforms (Laliberte et al., 2007; Zhang and Kovacs, 2012; Lucieer et al., 2014; Colomina and Molina, 2014; Pajares, 2015; Mogili and Deepak, 2018) as well as their inherent limitations (Jones et al., 2006; Hardin and Hardin, 2010; Hunt et al., 2010). Initially, although UAVS are highly flexible in their deployment, a ground operator is still required on the observed area. On top of that, the ground operator should be a highly trained, and often licensed, pilot. Despite UAV imagery data being considered of higher-quality compared to satellite, they demand management of large volumes of data and pre-processing, and the generated datasets are

limited to the data collected by the user himself. On the other hand, satellite data time-series are easily available on web-based platforms and are generally ready to analyze. Furthermore, UAV provides data at a resolution unreachable by satellite but cannot rival the latter's observed extent and remains constrained in particular territories by national and/or international legislation. As UAVs are remotely guided, they require a constant link with a ground control system, and this direct connection is vulnerable to interferences that can easily disrupt connection, leading to a loss of control over the aircraft. Finally, challenges regarding autonomy and flight duration, aircraft stability, susceptibility to weather conditions, regulations and restrictions regarding their operation, platform and sensing components' failures and payload capacity are all active problems that are currently optimised by the UAV manufacturers. An analysis of the advantages and limitations of each remote sensing method is presented in the following table, based on the analysis of Delavarpour et al. (2021) (Table 1).

Table 1. Characteristics of different remote sensing platforms (Delavarpour et al., 2021).

Specification	Ground-Based	Satellite	Manned Aircraft	UAVs
Cost	Low	Highest	High	Lowest
Operating environment	Indoor/outdoor	Outdoor	Outdoors	Indoor/outdoor
Time-consuming	Long	Shortest	Short	Short
Labor-intensity	Highest	Low	High	Medium
Operational risk	Low	Moderate	High	Low
Trained pilot requirement	No	No	Yes	No
Automatic crop spraying	No	No	No	Yes
Spatial resolution	Highest	Low	Moderate	Highest
Spatial accuracy	Moderate	Low	High	High
Temporal advantage	No	No	No	Yes
Adaptability	Low	Low	Low	High
Maneuverability	Limited	Limited	Moderate	High
Deployability	Moderate	Difficult	Complex	Easy
Susceptibility to weather	Yes	Yes	Yes	No
Repeatability rate	Minutes	Day	Hours	Minutes
Feasibility for small areas	Yes	No	No	Yes
Autonomy and sociability	Low	Low	Low	High
Real-time data availability	No	No	Yes	Yes
Limited to specific hours	No	Yes	No	No
Running at low altitude	No	No	No	Yes
Ground coverage	Smallest	Large	Medium	Small
Observation range	Local	Worldwide	Regional	Local
Operational complexity	Simple	Complex	Complex	Simplest

1.4 Unmanned Aerial Vehicles

1.4.1 UAVs in Agriculture

UAVs, or simply drones, are typically light weight, low airspeed remotely operated aircrafts, established as a highly efficient platform for remotely sensing data gathering missions across multiple domains and scientific disciplines. UAVs can carry a varied array of sensors and electronics, ranging from image acquisition sensors (visible, multispectral and hyperspectral cameras) to active sensing systems (e.g. LiDAR) and integrated automation mechanisms (e.g. real-time actuators or spraying systems). UAVs are the up-and-coming tools that have developed the most in recent years for use in the agrifood sector. By 2030, the Agriculture UAV Market is projected to reach over $32 Billion, with a CAGR of 7.1% during the following years, making it the second biggest addressable market for drone solutions, after construction (FAO, 2018; PWC, 2017). The main reason behind the widespread use and interest for UAVs is their wide

range of applications along with the unparalleled efficiency and flexibility they offer, for each and every specific situation (Tsouros et al., 2019; Yang et al., 2017). They have a unique capacity to cover large areas, provide data of very high spatial resolution at a high temporal frequency, without damaging the growing environment or disrupting the observed ecosystems. For these reasons, UAVs have also become a heavily utilised tool for many emerging applications in the area of ecological monitoring and biodiversity conservation, domains which are closely related to agriculture. UAV derived aerial imagery data can pick out anomalies that are difficult to be spotted while on the ground, providing a complete image of cultivated fields at a bird's eye view. The capabilities of UAVs in agriculture, however, are not limited to just locating problem areas. They are useful in monitoring several crop growth parameters throughout the season, while the emergence of data science and computer vision applications in agriculture has enabled very sophisticated applications, such as crop stress and disease identification, weed infestation mapping and plant-level yield and maturity identification.

Finally, concerns about operation safety and flight regulations emerged as the number of UAVs operating across the globe, for both recreational and professional purposes, increased rapidly. Drones by themselves can cause injuries to the operators and nearby individuals, because of their quick movements and their rotating parts, sharp blades and edges. Naturally, to address these issues, regulatory frameworks that prevent any potential risk associated to the use of UAVs such as forbidden flight areas and/or requirement for UAV operator certifications have been created across the world, however, they greatly vary among countries, while a number of UAV operators are completely unaware of the regulations in their counties (Stocker et al., 2017). One of the most common regulatory strategies across Europe is the requirement of different certifications from UAV pilots, based on the type of flights they execute, but most importantly, based on the characteristics of the aircraft they operate.

1.4.2 UAV Types

Considering the sizes, configurations, and characteristics, of UAVs, several classifications have been suggested in different studies (Shakhatreh et al., 2019, Tsouros et al., 2019) and different regulatory frameworks exist in different countries or regions, with the take-off weight of the aircraft being often what determines what set of regulations it falls within. Modifications and customizations of the structure of both fixed-wings and rotary-wings are common practices in agriculture to adapt UAVs in unstructured environments of field conditions (Gatti and Giulietti, 2013). A common modification in agricultural UAVs involves the removal of some redundant structural material and their replacement with carbon-fiber-composite parts, to reduce the mass and increase strength (Hunt et al., 2015). This allows superior performance and enhanced agility in take-off and landings operations, while also enabling longer flight autonomy and area coverage, but often at the cost of aircraft stability (D'sa et al., 2016). Regardless of modifications, however, UAVs should maintain limited width and high agility to respond quickly to sudden external disturbances such as gusts of wind to minimise the drift, allowing them to hover in place until the risk is minimised. These adjustments should also be considered during mission planning and flight designing, as they affect the no-payload and maximum take-off weight of the aircraft. There are three (3) major types of UAVs frequently used in PA applications: 1) fixed-wing, 2) rotary-winged and 3) hybrid Vertical Take-Off and Landing (VTOL) UAVs.

Fixed-wing UAVs are typically launched from a launcher catapult-like ramp as they require building momentum in order to take-off. Some very light aircrafts, such as the SenseFly eBee can be launched manually, without the need of a ramp or additional mechanism (Figure 1). Regardless of takeoff method, they require runways for landing, as there is very little control during this phase of the mission. The navigation control is succeeded through actuation of certain surfaces in the wings, known as aileron (controls pitch), elevator (control roll) and rudder (controls yaw) (Figure 2). Fixed-wing UAVs offer the greatest autonomy out of the aforementioned aircraft types, being capable of traveling several kilometers from the launch point and covering larger areas much more efficiently. These characteristics make them excellent platforms when obtaining fast and detailed information of large agricultural fields is the objective of the mission. However, as these aircrafts are not propelled vertically by any mechanism that allows them to hover or levitate, they should maintain a minimum air-speed throughout the mission flight to ensure that the generated lift cancels out the weight and the aircraft does not lose altitude. Therefore, fixed-wing UAVs are not capable of performing slow air-speed missions, and as a result, they are required to fly at higher cruise altitudes, not only for safety reasons, but also to achieve the desired overlaps on the collected image dataset (described in detail in the following chapter). Naturally, this type of high altitude flights decreases the maximum data resolution that can be obtained as well. Overall, fixed-wing UAVs are preferred for heavier/denser payloads and missions that require longer flight endurance, due to their higher autonomy and payload capacity [48].

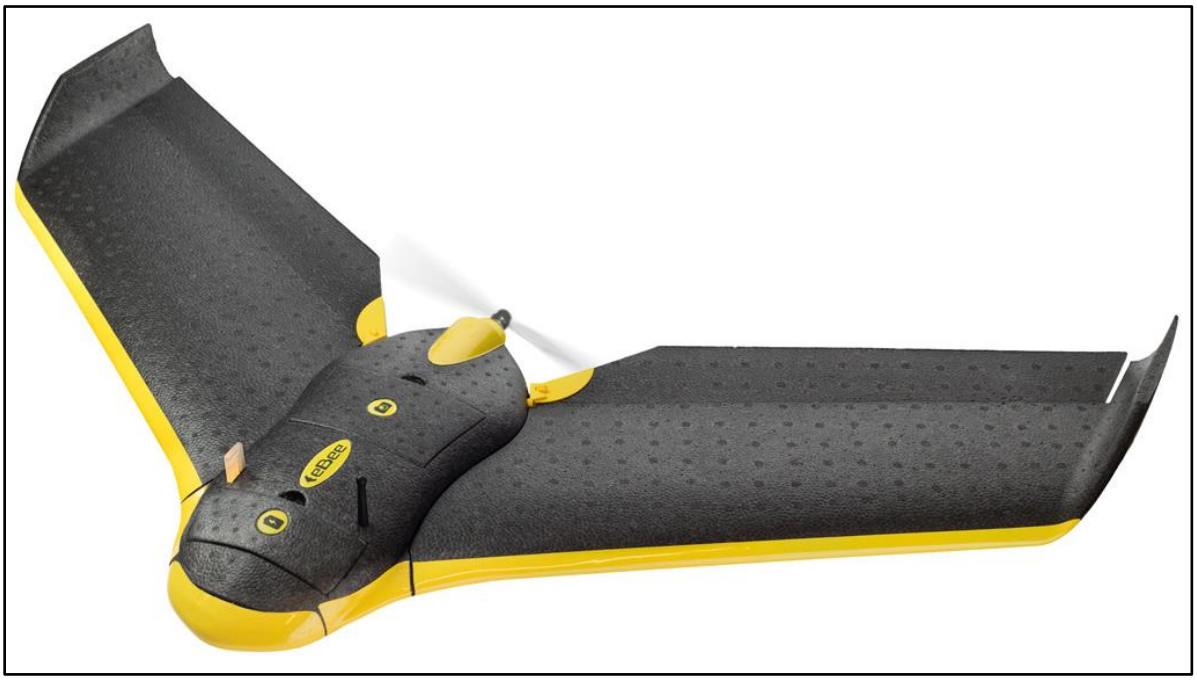


Figure 1. Example of a fixed-wing UAV, the senseFly eBee (SenseFly).

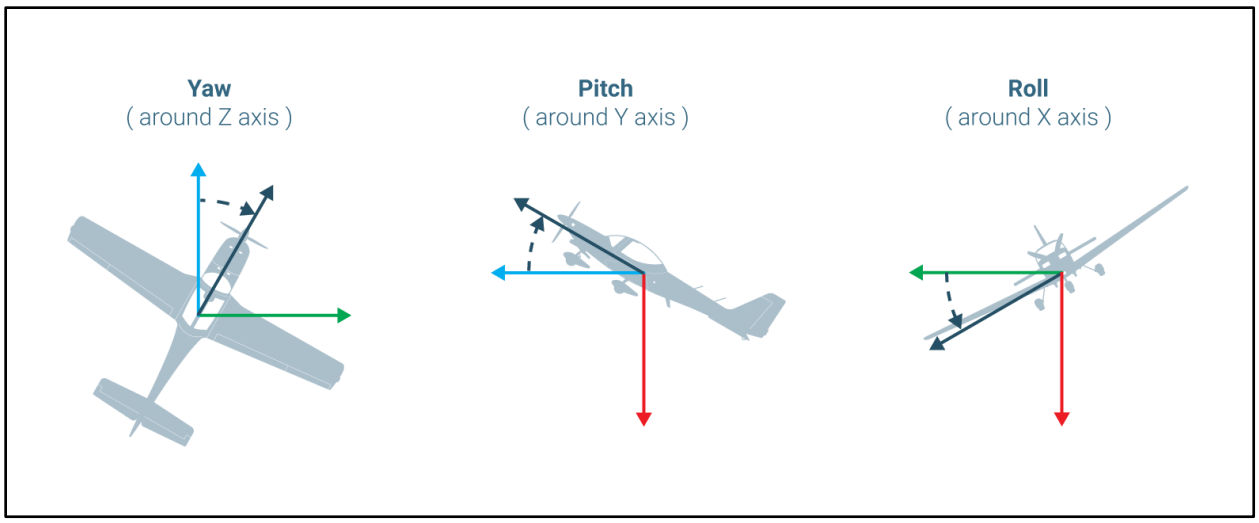


Figure 2. Visualisation of the manned-aircraft-borrowed terms Yaw, Pitch and Roll (source: Pix4D.com[1]).

The second category, rotary-winged UAVs can be further divided into subclasses based on the number of their rotor blades (Figure 3). Their common characteristic is that they can hover in place, which is usually their default action when they do not execute or receive any orders. They are agile and highly maneuverable aircrafts, capable of executing very demanding flight missions, as they can hover or move along a target in close quarters. Furthermore, they can vertically take-off and land in place, and they do not require airflow over the blades to move, as the required airflow is generated by the rotation of the blades. The main advantage of rotary-wing UAVs is that they can fly in lower altitudes with lower speeds, which results in a very high resolution collected data. As it is expected, however, they can cover smaller areas compared to fixed-wing UAVs and require a longer flight time over the same area to cover it properly. Increasing the number of rotors results in lower crash risk, higher stability during cruising and hovering. This results in lower vibrations, translated to lower levels of blur in collected images, even without the presence of gimbals and stabilisers and under higher wind speeds or even in the presence of gusts.

[1] https://support.pix4d.com/hc/en-us/articles/202558969-Yaw-Pitch-Roll-and-Omega-Phi-Kappa-angles

Figure 3. Example of a quadcopter UAV, the DJI Inspire 2 (SZ DJI Technology Co.).

Finally, efforts have been made to combine the fixed-wing and rotary-winged UAVs into hybrid models in an attempt to develop a system that maintains their respective advantages while overcoming their limitations (Bapst et al., 2015). Such aircrafts are called Vertical Take-Off and Landing - VTOL, and combine the cruising flight efficiency of fixed-wing aircraft with the convenient vertical landing of their rotary counterparts (Figure 4). The main flight strategy of VTOLs, however, remains similar to the fixed-wing UAVs, and thus the inability to perform low airspeed flights still exists. To properly address this, a VTOL aircraft also requires a specific system to allow the transition and reposition of the rotors after take-off, something that increases both complexity and manufacturing costs. Simple VTOL aircrafts have found applications for missions where large fields in steep and uneven terrains (i.e. mountainous orchards or vineyards) should be covered, as they do not require areas with no obstacles to take-off and land, something that is often difficult to locate in these terrains. Moreover, the transition between rotary and fixed-wing modes during the flight heavily impacts the aerodynamics of the aircraft, something that often leads to very challenging aircraft control situations, demanding high skills from the VTOL operator. Complete comparisons of fixed-wing, rotary-wing, and hybrid UAVs are presented in Table 2. In general, the characteristics and configurations of the UAV depend exclusively on each different application.

Figure 4. Example of a VTOL UAV, the Wingtra-One (Wingtra AG).

Table 2. The characteristics of different UAV types (Kanellakis and Nikolakopoulos, 2017).

	Advantages	Disadvantages
Multi-Rotor	• VTOL flight • Hover flight • Maneuverability • Indoors/outdoors • Small and cluttered areas • Simple design	• Area coverage • Limited payload • Short flight time
Fixed-Wing	• Long endurance • Large coverage • Fast flight speed • Heavy Payload	• Launch-Landing specific space • No hover flight • Constant forward velocity to fly
Hybrid	• Long endurance • Large coverage • VTOL flight	• Transistion between hovering and forward flight

In the selection of the UAV alone, multiple different characteristics should be considered, such as the system's performance, autonomy and load capacity, which are all critical for agricultural applications. To this end, a detailed analysis should be conducted, taking into consideration the unique aspects and requirements of each foreseen application, and then deciding on the optimal UAV and sensor

configuration to achieve these objectives. Moreover, a proper risk management analysis should be included in the analysis, as environmental factors such as turbulence, gusts and extreme temperatures drastically affect the components, performance and safety of the aircraft and nearby people. Therefore, careful consideration of both proper aircraft components and mission planning is the only way to effectively address the aforementioned factors, which is crucial for agricultural UAV applications.

1.4.3 Sensor Specifications

There are several types of sensing systems that can be mounted on UAVs for agricultural applications. The most commonly used type of sensor is a camera, collecting imagery data on the reflectance from the crops in specific spectral bands. As flight planning is based around the characteristics of the sensing system carried by the UAV, there are several factors that should be considered when selecting the sensing system in the first place. For each particular application, a set of basic sensor parameters, namely spatial resolution, spectral resolution and radiometric resolution are what determine whether the sensing system is appropriate for this application.

Spatial resolution refers to the size of the smallest object that can be detected in an image, or simply the size that a pixel, the smallest unit in an image, covers in the real world. A spatial resolution of 50 cm means that each pixel represents an area of 50x50 cm^2. The smaller an area represented by one pixel, the higher the resolution of the image, and therefore more distinguishable and detectable are smaller objects.

Spectral resolution describes the ability of the sensor to define wavelength intervals. For each spectral band that the sensor can sense and generate data, shorter wavelength widths can be distinguished in higher spectral resolution images on this band. An example of finer spectral resolution is a hyperspectral sensor capable of measuring energy in narrower bands (often 10-20nm) compared to a multi-spectral sensor (which normally ranges around 50nm). The narrow bands of hyperspectral imagery are more sensitive to variations in measured spectral bands and therefore have a greater deviation detection capacity compared to multi-spectral imagery.

Radiometric resolution refers to the sensitivity of a remote sensor to variations in the reflectance levels of a single spectral band. The higher the radiometric resolution of the sensor, the more sensitive it is to small differences in reflectance values, allowing for more precise detections on a specific portion of the electromagnetic spectrum.

1.4.4 Mission Flight Methodologies

As described in the previous Chapter, UAVs are small unpiloted aircrafts with on-board positioning system devices, stabilizing on-board 3-axis gyro sensors, allowing them to fly autonomously given a certain set of predefined commands from the Ground Control Station (GSC). This set of commands is usually a sequence of points in the real world (waypoints), that the aircraft is tasked to navigate towards using its internal positioning and orientation systems. Each waypoint, which is perceived by the UAV as a set of orders, includes not only the coordinates of the waypoint, but also other flight parameters, such as cruising speed and altitude. These parameters can be either fixed for the entire duration of the

mission flight, or change from point to point. Moreover, certain actions are also provided in the form of orders, such as take-off and landing, as well as the signal to trigger certain components (i.e. the sensing device to capture an image) whenever there is an established link with the UAV (Figure 5).

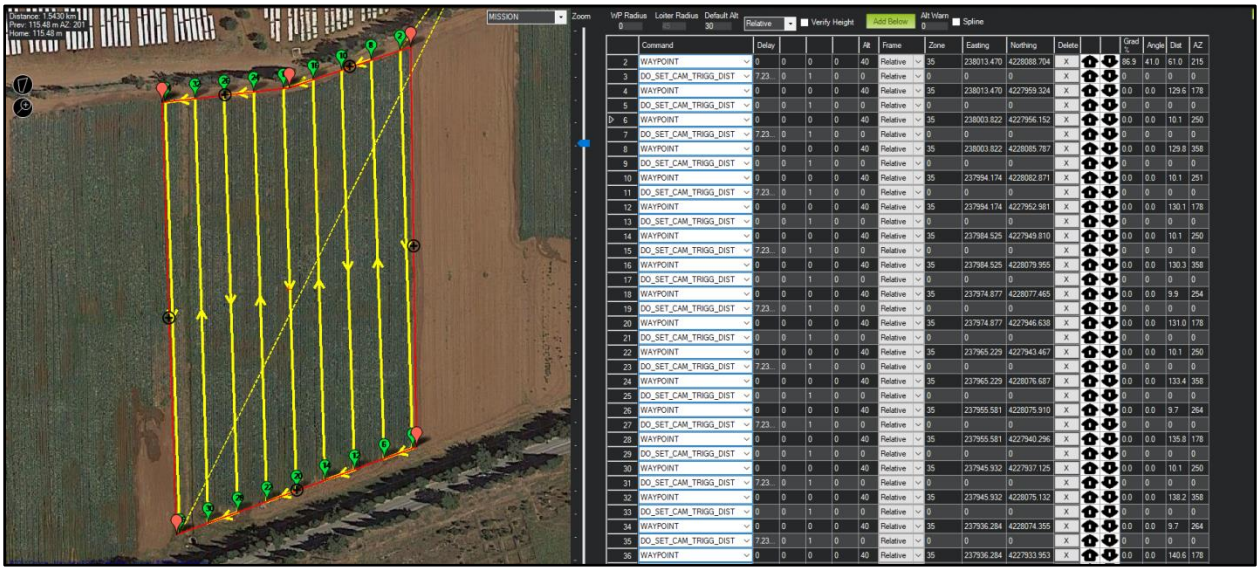


Figure 5. A predefined waypoint flightplan (left) and the commands received by the UAV (right).

Based on the existence of an established link between UAV and sensing device(s), two methods of flight missions exist. Prior to that, however, a number of other flight parameters should first be presented. The objective of most agricultural UAV missions is to create a 2D map of the surveyed field. 3D modelling of agricultural crops is another critical aspect of UAV surveying in agriculture, demanding its own methods of mission flight design and execution, but will not be analysed as it falls outside the scope of this thesis. The UAV 2D mapping process involves the collection of individual images to create orthophotogrammetric (or orthomosaic) maps. The photogrammetric process essentially “stitches” each individual image with its neighbours, based on geo-reference metadata and their extracted common features (the common points between neighbouring images). To achieve this, the generated dataset should guarantee that the sequence of aerial images can meet the photogrammetric requirements and there are enough corresponding feature points to complete image mosaicking. To this end, a number of flight parameters related to the function of both the aircraft and the sensing device should be determined and combined in a delicate way to ensure that the dataset is eligible for mosaicking. The first parameter that is often considered when designing a UAV flight mission is the sensing instruments field of view. It is expressed in two (2) angle values, representing the vertical and horizontal field of view respectively. Through these values, we can easily calculate the area covered by the sensing instrument in each image capture, by calculating the triangular values of these angles and the perpendicular side of the aircraft’s altitude (Figure 6).

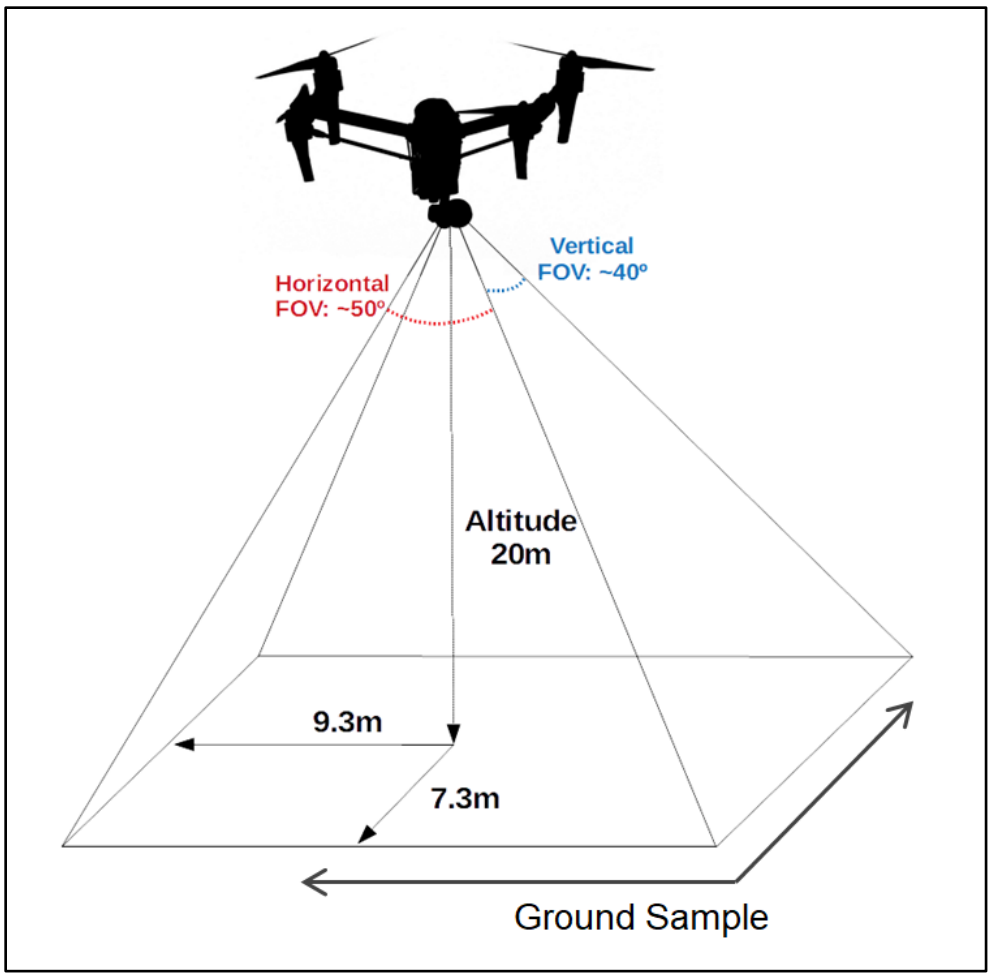


Figure 6. Example of an aerial sensing system's FOVs and sampling area values.

Image overlap is the parameter that ensures that sufficient common features between the dataset's images exist, and therefore high-quality mosaicking is achievable (Haala and Rothermel, 2012). In aerial imagery, there are two overlap parameters: front overlap (or frontlap) and side overlap (or sidelap) (Figure 7). In agricultural surveys, UAVs normally move in straight lines, called flight lines. Frontal overlap represents the overlapping percentage of consecutive images captured on the same flight line. Sidelap refers to the percentage of overlap between different flight lines, or simply the common area scanned between neighbouring images of consecutive flight lines (Figure 7). Overlaps are the only parameter that are affected by all other flight parameters, both related to the aircraft (i.e. speed and altitude) and the sensing device (i.e. camera field of view and capture interval) (Xing et al., 2010, Dandois et al., 2015; Fraser et al., 2018; Torres-Sánchez et al., 2018). As a result, they are considered a standard requirement, and are the first parameter to be taken into consideration. A common practice for nadir flights (the most common type of flights in agricultural surveys, where the flightlines are parallel to the earth's surface and data is collected vertically, with the camera's field of view being perpendicular to the ground), a standard practice is to design flightplans with 70% and 80% frontal and side overlaps respectively (Ni et al., 2018), as higher values might result in unnecessary large volumes of data and higher processing times at later stages.

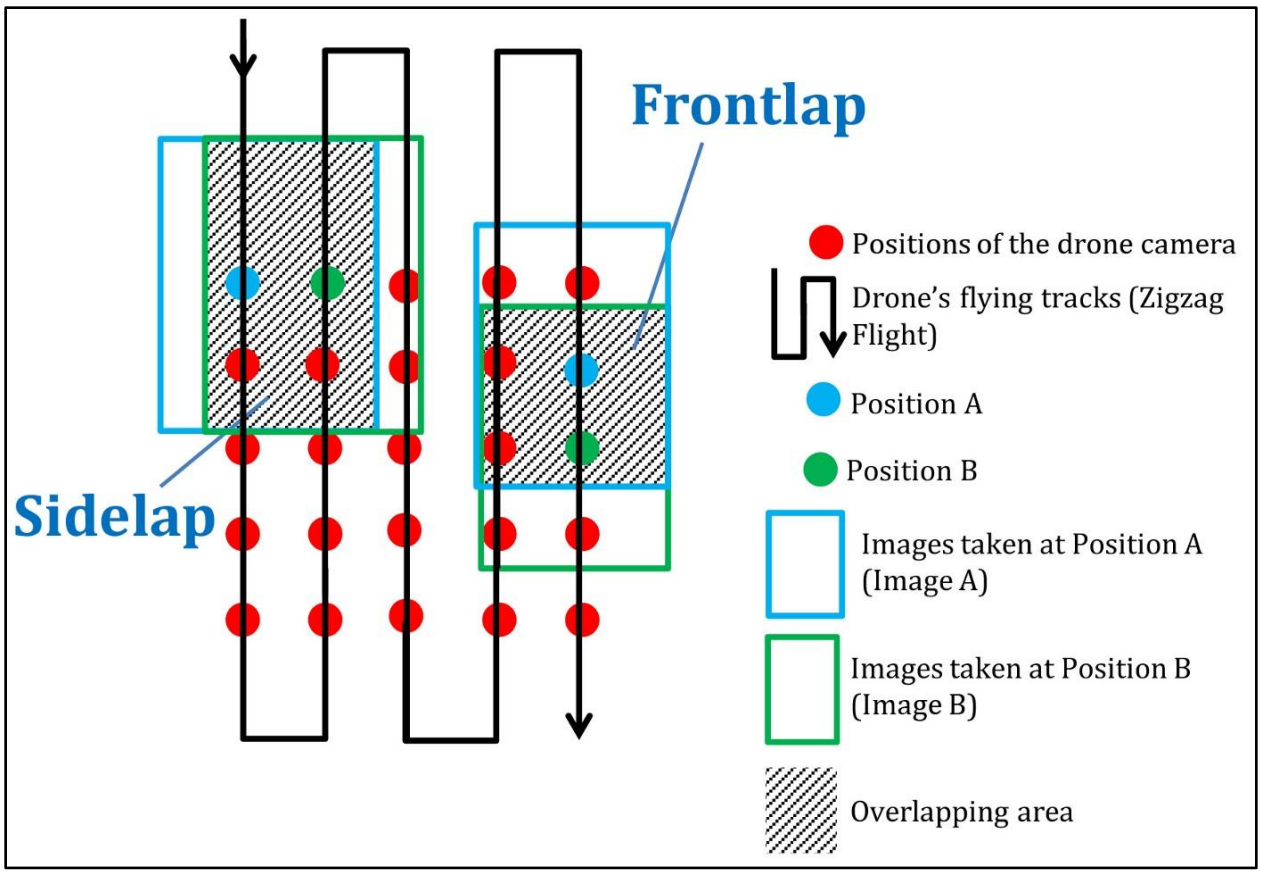


Figure 7. Representation of the front and side overlap parameters (image source: Medium.com[2]).

Ground Sampling Distance (GSD) is a critical parameter in all drone mapping and surveying projects. GSD is defined as the distance between the centres of two adjacent pixels measured in the real world (the ground, in nadir flights). Essentially, GSD translates distances in the final orthomosaic to actual distances on the ground (Figure 8). It is described in side length per pixel (cm/px), and is related to the camera's focal length and resolution (number and arrangement of pixels), and of course the camera's distance from the surveyed area (altitude). The further the camera is from the target, the less 'space' the target will occupy on the image. GSD scales linearly with the drone's altitude if it is carrying a camera with a fixed focal length lens, and linearly with focal length if the camera has a variable lens. Naturally, lower altitude flights improve the GSD and smaller objects become more distinguishable in the final orthomosaic. GSD is highly dependent on each specific survey and application. For simple crop monitoring tasks, a higher GSD is acceptable (i.e 5-10 cm/px). For tasks where smaller objects should be clearly visible to the final image, however, GSD should be considerably higher (i.e. lower than 2 cm/px), and the entire flight plan should be designed to ensure this value.

[2] https://medium.com/@kitty.j/what-does-overlap-really-mean-in-3d-modelling-cf8d321bf25

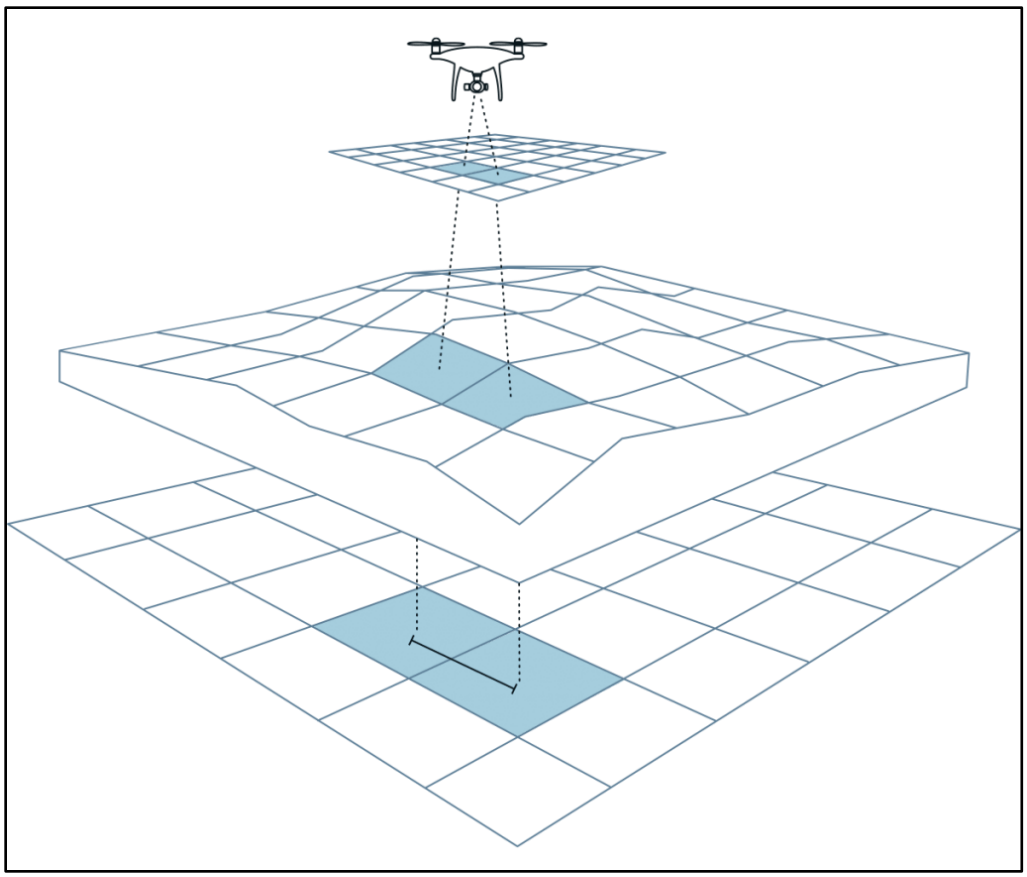


Figure 8. Representation of GSD, based on aircraft altitude (image source: Wingtra.com[3]).

Altitude and speed are the two aircraft-related flight parameters that must be properly balanced to produce high-resolution imagery with sufficient overlap values. Lower altitude flights generate data of higher resolution (GSD), however, the aircraft must travel at a lower speed to avoid distortion and motion blur, as well as to maintain sufficient frontal overlap. Moreover, the flightlines should become more dense, to ensure sufficient side overlap, as the camera moves closer to the ground and the area scanned with its default field of view is reduced. As a result, flight duration increases significantly, and often the operators are met with challenging situations where the data quality and flight efficiency trade-off should be optimised (Psiroukis et al., 2021).

The final parameter of flight planning is the data capturing method of the UAV system, and two different approaches exist based on whether the UAV has an established link with the data capturing device (camera sensor). The first method is the most common one, when there is no established link between the sensing device and the aircraft. In this case, the sensing instrument is set to capture data at a fixed interval, and the entire flight mission is designed around this parameter. This case perfectly represents one of the very first definitions given to aerial photography, by DeLatil in 1961, referring to it as "a means of fixing time within the framework of space" (Weng, 2012). The second method involves a set communication between the aircraft and the camera. In this case, the entire flight plan is designed to

[3] https://wingtra.com/mapping-drone-wingtraone/image-quality/

achieve the desired parameters (overlaps and GSD), and the capturing interval is calculated last, as the system can signal the sensor on when to capture each image, essentially adjusting the capture speed in real time. The signal is either triggered when the UAV has travelled a certain distance since the start of each flight line or the previous capture, or a certain time between captures is calculated at the start of each flight line (according to the predefined overlaps) and the sensor captures in this interval.

Finally, another parameter that should be considered when designing a UAV flight mission is aircraft stability. This parameter cannot be quantified by a certain metric, but should be always optimised in an attempt to maximise the quality of the collected data. Drones have known operational limits, set by the manufacturer, corresponding to the environmental conditions under which the aircraft can operate safely. High wind speeds and the presence of gusts and turbulence are the most common factors for flight cancelations. Even within the safe operational limits, however, the UAV might be able to successfully complete the mission, but the quality of data might not be as high as expected. This is because aircraft stability is a crucial factor for collecting high-quality images, and it is heavily impacted by wind parameters. Collision with gusts can tilt the aircraft and the sensing system in certain angles, deviating from its nadir position and resulting in data captured from an oblique angle. This causes distortion in the final orthomosaic map, which reduces accuracy and limits the potential for accurate measurements and reliable surveys. Moreover, most drones have been assigned with their operational limits when flying by themselves. In the case of customisations, the modifications alter the aircrafts aerodynamics, while the integration of additional payloads (i.e. the sensing devices) also affects the airborne response capacity and agility of the UAV. To address these issues, two (2) common methods have been devised to ensure UAV stability: 1) the use of controllers that keep the aircraft level when hit by gusts, allowing it to recover quickly and 2) the use of gimbals for the mounted cameras, to maintain the nadir capture angle even when the UAV is not level.

1.5 Artificial Intelligence

1.5.1 Computer Vision and Machine Learning

Humans have the ability to instantly recognize what objects are demonstrated in the image with a single glance. The human visual system is fast and when combined with the human knowledge, it enables the accurate judgment about the nature of the object, something that allows us to perform complex tasks with little to no conscious effort. For machines, however, this is a very challenging and complex task, composing the single fundamental problem of Computer Vision: "How can we create a system that imitates the ability of the human visual system to detect objects?”.

Intelligence is defined as the ability to learn and understand, to solve problems and to make decisions. Artificial intelligence (AI) aims to allow machines to perform tasks that would require intelligence if performed by humans (Boden, 1977). In AI, a model replicates a decision process to enable automation. AI models are mathematical algorithms that are “trained” using data and human expert input to replicate a decision an expert would make when presented to the same situation, and provided that same information. Ideally, the model should also reveal the rationale behind its decision to help interpret the decision process. Most often, the training step involves the processing of large volumes of

data through an algorithm to maximize likelihood or minimize cost, ultimately yielding a trained model. By analyzing data from many sources, collected under different conditions, the model learns to detect all the types of patterns and distinguish these when presented to them anew. A model attempts to replicate a specific decision process that a human expert would make if they were presented new available data.

Machine learning uses two techniques: supervised and unsupervised learning. In supervised learning, there are the input variables (x) and an output variable (y) and the algorithm is tasked to figure out the mapping function that best describes the translation of the inputs to the output ($y=f_{(x)}$). When using supervised machine learning, the learning algorithm is provided with known quantities to support future predictions. Supervised learning is usually implemented for classification problems where the association between inputs and output labels is sought or for regression problems where the aim is to map input to a continuous output. In classification, the goal is to create a mapping function (f) from input variables (x) to discrete output variables (y). Classification algorithms are used when the output is a discrete label such as 'tomato' or 'apple', 'green' or 'red'. In regression the aim is to create a function (f) that maps input variables (x) to a continuous output variable (y). Such outputs could be the 'crop_height' or 'fruit_weight'. Unsupervised learning is a technique where the machine tries to discover patterns in the presented data by itself. When using unsupervised learning only the input data are available. The goal is to model the structure or distribution of the input data in order to learn more about them. Unsupervised learning is used to infer patterns from a dataset without labeled outcomes. This present thesis revolves around a detection and classification problem in UAV images, and therefore, supervised learning is exclusively used.

An artificial neural network is an information processing paradigm inspired by the way biological nervous systems process information. It comprises a large number of highly interconnected processing elements (neurons) working together in parallel to solve specific problems such as pattern recognition or data classification through a learning process. Neural networks learn by example (training phase) similarly to humans. They are not specifically programmed to perform a task, but they learn by themselves how to solve each problem. There are two architectures of neural networks: first the feed-forward networks that allow one-way travel, from input to output. They are straight forward networks that associate inputs with outputs. They are usually used in pattern recognition. Second, the feedback networks where the information can travel in both directions by using loops of forward and backward propagations. This type of network is very powerful and can become very complex. Feedback networks are dynamic, meaning that they are constantly changing until they reach an equilibrium point where they remain until the input changes.

A neural network consists of three types of layers: the input layer, the hidden layers and the output layer (Figure 9). The input layer is connected to the first hidden layer, which are connected with each other, leading to the last hidden layer that is connected to the output layer. The input layer consists of the information that is fed into the network; the hidden layers are determined by the activities of the input layer and the weights of the connections between the input and the hidden layers. The output layer depends on the activity of the hidden layers and the weights between the hidden and output layer. The weights determine when each hidden unit is active. The weights are adjusted during the training

phase in such a way that the error between the generated output (predictions) and actual output (labels) is minimised. This means that the neural network has to calculate how the error changes as each weight increases or decreases. The most widely used method is the back-propagation algorithm (Stergiou and Siganos, 2010). When the neural network finds an error, the algorithm calculates the gradient of the error function, which is adjusted according to the network's various weights. The gradient of the last layer is calculated first, while the gradient of weights for the first layer is calculated last. The calculations of the gradient from one layer are used again to determine the gradient for the previous layer (Goodfellow et al., 2016).

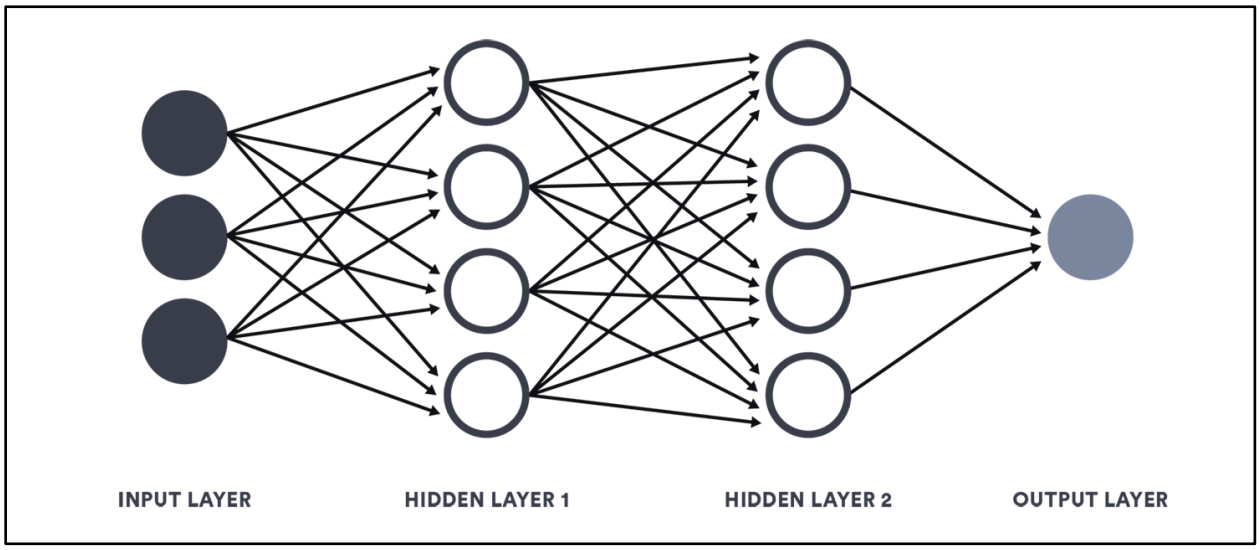


Figure 9. Schematic representation of a Neural Network.

The Universal Approximation Theorem (UAT) declares that only one hidden layer, with a finite number of neurons, is enough for approximating any looked-for function (Hornik et al., 1989), summarising the immense capacity of neural networks in problem solving. However, although a feed-forward network with a single layer may theoretically be sufficient to represent any function, it may not only be infeasibly large, but it would also clearly overfit to the training examples and therefore fail to learn and generalize correctly (Goodfellow et al., 2016). For this reason, instead of attempting to achieve extremely large layers, the focus has shifted towards deeper models that reduce the number of units required to represent any function while reducing the amount of generalization error.

1.5.2 Deep Learning

Deep Learning is a subclass of Machine Learning algorithms whose peculiarity is its generally higher level of complexity (Figure 10). The fundamental advantage of Deep Learning algorithms is that these models can be trained on unstructured data, with unlimited access to information, and this powerful condition provides them the opportunity to obtain more profitable learning. Deep learning is a form of machine learning that allows computers to learn from experience and understand the world in terms of a hierarchy of concepts (Goodfellow et al., 2016). In deep learning the computer operator (human) does not need to specify the knowledge. Deep learning has gained popularity in recent years mainly because

there are more data available that can be fed to learning algorithms and because only recently the computational power of computing units has allowed the training of complex neural networks that are big enough to make use of all the data available. The right choice of the Machine Learning technique is based on the amount of data available and on the performance. When the amount of data is limited, using large/deep neural networks makes little to no sense, as the gain would be small or even null compared to smaller architectures and or other traditional Machine Learning techniques (Ng, 2017). The main limiting factors when training a deep neural network is the computational power available. Therefore, in some cases, it may be beneficial to use a small network as it will require less computational power and is trained faster.

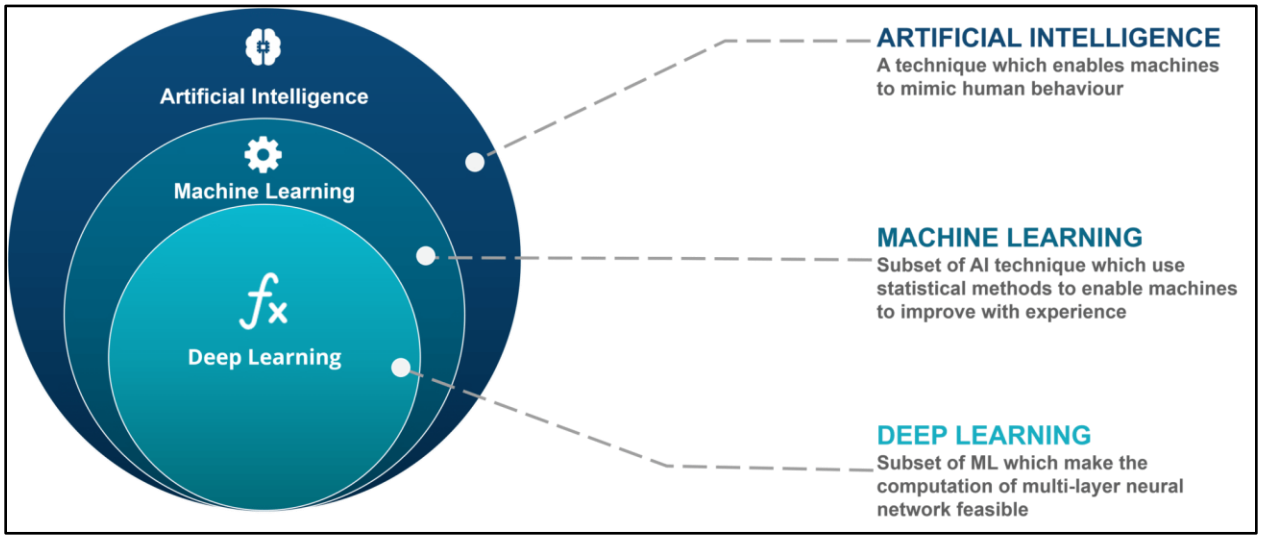


Figure 10. Artificial Intelligence and the hierarchy of Machine Learning and Deep Learning. (image source: Ospreydata.com[4])

Convolutional neural networks (CNN) are a specific type of neural networks used to process data that has a known grid-like topology or a combination of multiple arrays (Goodfellow et al., 2016). Convolutional neural networks usually follow a typical architecture, which deviates from typical neural network architectures due to their "depth" or simply complexity (Figure 11). The computer understands images as an array of pixels, where the horizontal dimension is given by the number of input channels. A grayscale image will be two-dimensional, while an RGB image will have three dimensions. Pixel absolute values can naturally range to the depth (bits) of the image's channels. When comparing two images, all values have to be equal for the pictures to be classified the same. Convolutional neural networks compare images piece (user specified, usually rectangular, bounding boxes) by piece. The pieces that the CNN is looking for are called features. Each feature can be seen as a mini image and each feature represents a common part between the images. When a new image is presented to the CNN, the CNN does not know where the features are located so it searches in every possible position. A map is created, which actually is a filtered version of the original images and shows the matches between the images. The same procedure takes place for every feature.

[4] https://ospreydata.com/getting-to-know-ai-and-ml-models/

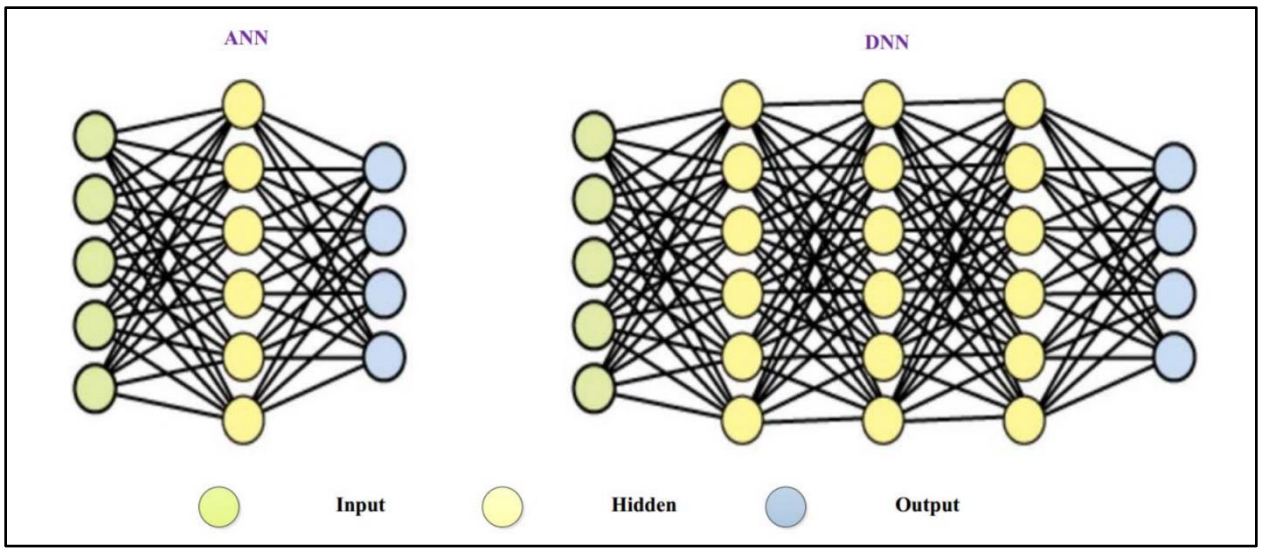


Figure 11. Typical architectures of artificial neural networks (left) and deep neural networks (right) (Aslam et al., 2019).

Pooling allows the CNN to use large images and reduce their size while preserving the most important information in them. If the data size is excessive the model will take a long time for training. During pooling, a small rectangle (window) is slid across the image, keeping the maximum pixel value inside each rectangle. This allows the CNN to locate features in an image without being concerned about its location. By doing so it deals with the problem of computers being too literal (looking for an absolute match), as otherwise the network would be useless with images that do not exactly match the training data. ReLU or Rectified Linear Unit, which is a specific type of activation function, is another important part of a CNN. The images are classified based on the probability of the output. Since the probability ranges between 0 and 1 all negative values have to be changed to zero, which is what the ReLU is responsible for. This allows the CNN to keep working as it prevents it from getting stuck close to zero or growing towards infinity. The product of the convolution and the pooling are reduced feature-filtered images. Those images can be filtered and shrunken in deeper layers. An example of a very deep convolutional neural network that consists of 16 convolutional layers, 5 maxpool layers and 3 fully connected layers has been described by Simonyan and Zisserman (2014). With every step the features become larger and more complex while the images become more compact. As a result, lower layers represent simple features of the image, edges and bright spots, while higher layers represent the more complex aspects of the image, shapes and patterns. At the end of a CNN, there is always at least one layer called the fully connected layer. This layer considers the high-level filtered images and decides in which class they belong. More than one fully connected layer can be part of a CNN. When classifying an image not every feature has the same importance in the classification process. Some values are more important when classifying an image as type A or type B. How important each value is, is expressed by its weight which is determined during the training phase. The weights are chosen using back-propagation. Back-propagation uses images of which their class is already known. These images are presented to an untrained CNN. This results in every feature and weight having a random value. As the process continues, the features and weights are evaluated by calculating the prediction error. The values

are then adjusted, and the error is calculated again. The values with lower error are kept. This process is repeated for every feature and weight and for every image in the network.

The final step in setting up a CNN is to set the hyperparameters, which are parameters whose values are set before the learning phase. Some examples of hyperparameters are the number of the features, the rectangle (window) size of each pooling layer, the stride (how many pixels the window shifts each time), and the number of layers. The values of other parameters, such as the weights, are derived via training. The hyperparameter values are decisions made by the designer. The possibilities are endless and there is no correct or wrong way. The hyperparameters are usually chosen based on values tested by other users that have shown to work well (LeCun and Bengio, 1995).

Summing up, the convolutional layer detects conjunctions of features from previous layers and creates a feature map. The pooling layer is then used to merge similar features inside the feature map into one, resulting in a pooled feature map. Pooling units reduce the dimensions of the representation and create a spatial invariance to small shifts and distortions. A typical convolutional network architecture often consists of two or three sets of convolution, a function to increase non-linearity (ReLU) and pooling. Those are followed by more convolutional layers, a flattening step where the pooled feature map is flattened into a vector in order to become the input for the last layer, which is the fully connected layer whose purpose is to combine the features and classify the images (LeCun et al., 1995).

1.5.3 Object Detection – Yolo Algorithm

Object detection is one of the primary tasks in computer vision which consists of determining the location on the image where certain objects are present, as well as classifying those objects. However, object detection systems should not only classify objects in images, but also categorise them. Initially, the first methods used to address this problem consisted of two stages: (1) Feature Extraction stage, where different areas in the image are identified using sliding windows of different sizes and (2) Classification stage, where the classes of the objects detected are estimated. A common method for the implementation of the classifier across the image is a sliding window approach, with the classifier running at evenly spaced locations over the entire image (Felzenszwalb et al., 2010). These approaches were segmented to multiple stages and are very demanding in computational resources. Speed and accuracy are the two prerequisites for which an object detection algorithm is examined, while these initial systems are fundamentally very difficult to be optimized in both speed and efficiency.

Prior detection systems repurpose classifiers or localizers to perform detection, and then apply the model to an image at multiple locations and scales. Finally, the high-scoring regions of the image are considered detections. Recent approaches like the R-CNN and its variations use region proposal methods to initially generate a number of potential bounding boxes across the image and then run the classifier on these proposed boxes. During training with this approach, after every classification, post-processing steps are used to refine the bounding boxes, usually by increasing the score of the best performed bounding boxes and decreasing the worse ones, ultimately eliminating potential duplicate detections (Girshick et al., 2014). Faster R-CNN is one of the most widely used two-stage object detectors. In Faster R-CNN, detection is performed in two stages. The first stage uses a region proposal

network, an attention mechanism developed as an alternative to the earlier sliding window based approaches. In the second stage, bounding box regression and object classification are performed (Figure 12). Faster R-CNN is fairly recognized as a successful architecture for object detection, but it is not the only meta-architecture (Huang et al., 2017) able to reach state-of-the-art results.

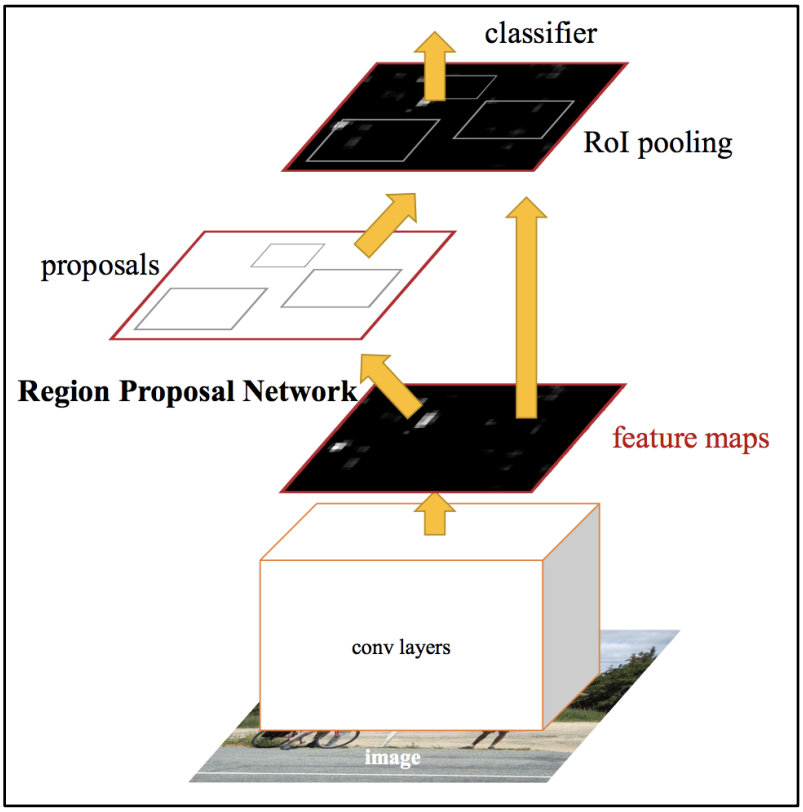


Figure 12. The Faster R-CNN detection pipeline (image source: Towardsdatascience.com[5])

The YOLO model (You Only Look Once) uses a different, more efficient approach. It integrates the entire object detection process into a single neural network, using features from an entire image to generate each bounding box prediction. Therefore, it can perform all bounding box predictions, for all classes and across the entire image, with a single network evaluation. Systems like YOLO that handle object detection as a single regression problem straight from image pixels to bounding box coordinates and class probabilities have constantly gained momentum since their release, mainly due to their performance in both speed and accuracy for detecting and determining object coordinates (Redmon et al., 2016). Initially, Yolo begins every inference by dividing the input image into an assigned SxS grid. Each of these grid cells predicts an also assigned number of bounding boxes (B) and calculates a respective confidence score for each predicted box (Figure 13). The confidence score represents the model's level of confidence on the likelihood of that box containing an object, as well as the expected accuracy of the prediction. The models confidence is calculated as: Pr(Object) * IOU(truth-pred), in order to get the confidence scores for each individual box. If a grid cell contains no objects within it, the confidence scores should naturally be zero. In the case that an object does exist within the grid cell, then

[5] https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e

the optimal confidence score equals the intersection over union (IOU) (Figure 14) between the predicted box and the ground truth, meaning that the model estimated correctly how close it came to the assigned human-generated ground truth box.

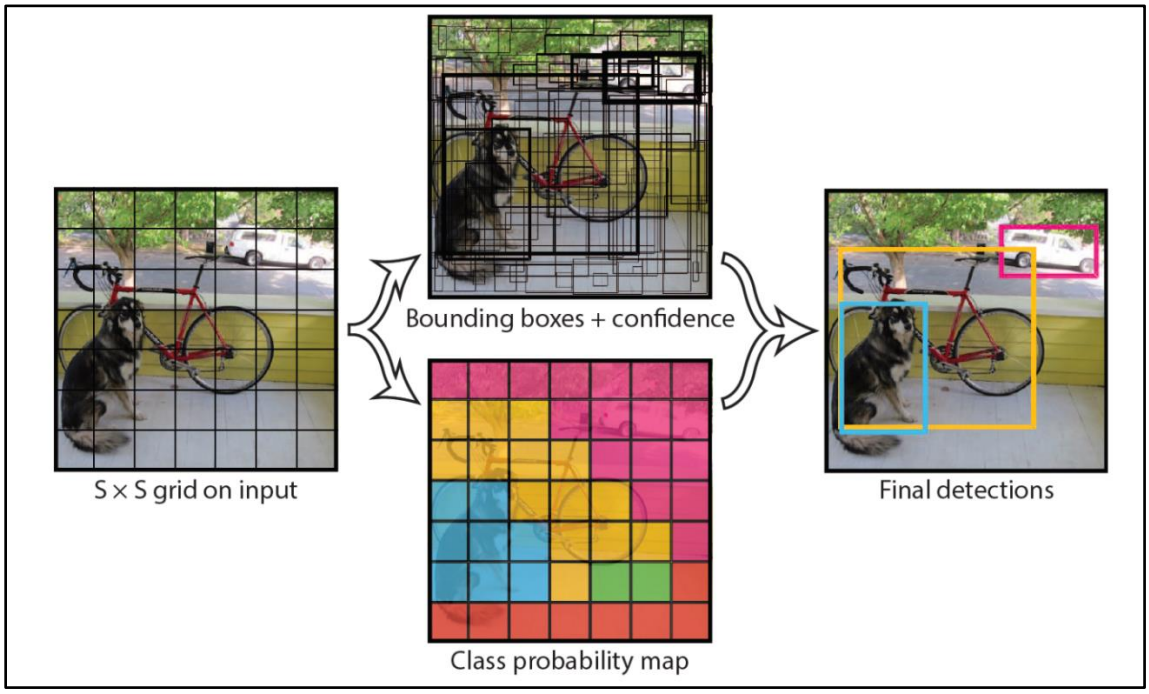


Figure 13. The YOLO model detection pipeline (image source: Redmon et al., 2016).

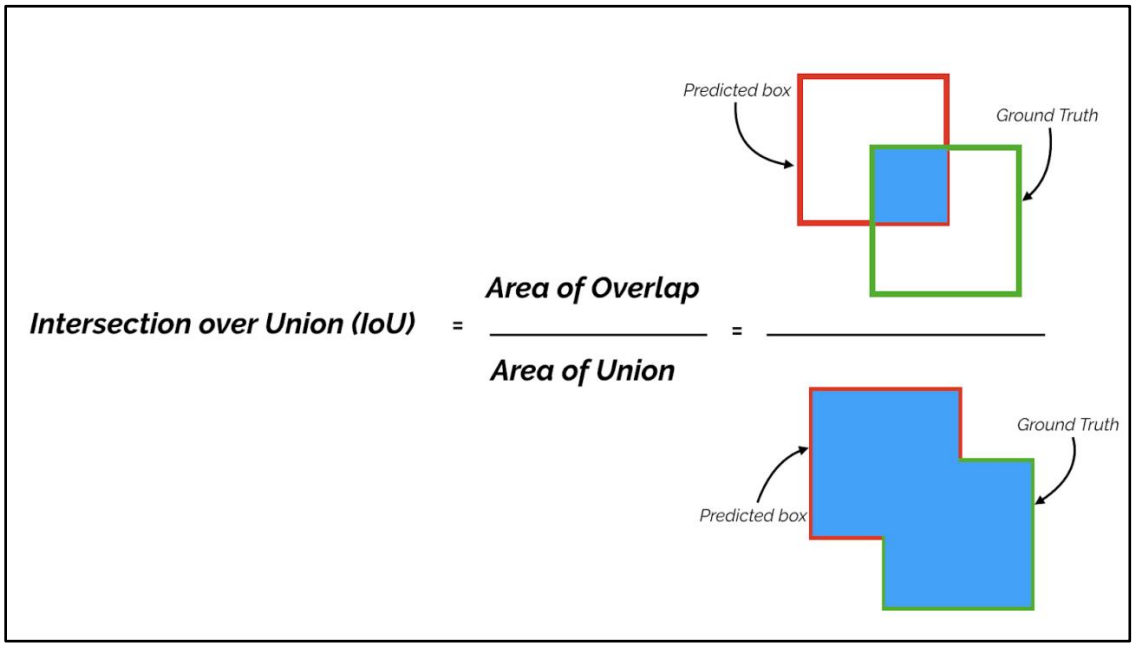


Figure 14. Visualisation of the Intersection over Union (IoU) (image source: Towardsdatascience.com[6]).

[6] https://towardsdatascience.com/map-mean-average-precision-might-confuse-you-5956f1bfa9e2

Each generated bounding box consists of 5 predictions: x, y, w, h, and confidence. The (x, y) refer to the coordinates of the center of the predicted bounding box (relative to the bounds of the grid cell), while the (w, h) represent the width and height (relative to the entire image). The confidence prediction represents the similarity (IOU) between the predicted box and any ground truth box, and is calculated as explained above. Additionally, each grid cell predicts a number of conditional class probabilities, Pr(Classi|Object), representing the probabilities of the grid cell containing an object of a certain class. An entire set of class probabilities (equal to the number of classes used) is uniform for each grid cell, regardless of the number of bounding boxes predicted. These conditional class probabilities are then multiplied with the individual box confidence predictions, providing class-specific confidence scores for each bounding box as:

$$Pr(Class|Object) * Pr(Object) * IOU(truth\text{-}pred) = Pr(Class) * IOU(truth\text{-}pred)$$

These scores are valuable metrics, since they incorporate both the probability of that class appearing in that specific box, but also the detection's overall quality (how well the predicted box fits the object boundaries). This approach allows the model to perform training on "full images", something that optimizes detection performance, while also making the model significantly faster and more consistent with background errors (Redmon et al., 2016). The latter is a parameter of extreme importance, especially in the context of agricultural aerial imagery, where the majority of images share background pixels (usually soil) of similar patterns. Fast R-CNN is known for encountering difficulties in distinguishing background and foreground pixels, often leading to mistakes, classifying background patches in images as objects due to its inability to run on a larger context - a full image. As a reference, even the initial version of YOLO makes less than half the number of background errors compared to Fast R-CNN (Girshick, 2015; Redmon et al., 2016). Moreover, when YOLO is trained on natural images, it normally outperforms top detection methods like Fast R-CNN and its variations, as it learns generalizable representations of objects (Redmon et al., 2016). This, naturally, makes the entire model more generalizable, gaining an edge when applied to new domains or, more importantly for agricultural applications, unexpected inputs.

After the release of the original YOLO model, the algorithm has been upgraded five times within the following five years, integrating several innovative ideas from international computer vision research. The first three versions are researched and developed by the author of the YOLO algorithm, Joseph Redmon. However, he announced to discontinue his research in the computer vision field after the release of YOLOv3, after recognising the multitude of applications computer vision had in military applications. Nevertheless, he did not dispute the continuation of research by any individual or organization based on the early ideas of the YOLO algorithm. As a result, in 2020, a new update on the YOLO algorithm, YOLOv4, was published by Alexey Bochkovskiy et al. (2020). Finally, after only a single month after the launch of YOLOv4, researcher Glenn Jocher and the Ultralytics LLC research department (who previously built YOLO algorithms on the Pytorch framework) published YOLOv5, which impressed the research community with its outstanding performance compared to all four previous versions (Jocher, 2020). In the present thesis, the YOLOv5 architecture has been selected for the object detection tasks, implemented in Pytorch.

2. Literature Review

Automation in agriculture presents a more challenging situation compared to industrial automation due to field conditions and the outdoor environment in general (Oetomo et al., 2009; Kirkpatrick, 2019). Fundamentally, most tasks demand high accuracy of crop detection and localization, as they are both critical components for any automated task in agriculture (Duckett et al., 2018). The fact that there is a constant downwards trend of available agricultural labour force (Roser, 2019) also adds to this problem and sets automation of several production aspects as a necessity. Accurate crop detection and classification are essential for several applications, including crop/fruit counting and yield estimation. Crop detection is often the preliminary step, followed by the classification operation, such as the infestation level through identification of disease symptoms (Barbedo, 2019), or as per the subject of the present thesis, maturity detection for the automation of harvest surveying. At the same time, it is the single most crucial component on automated real-time actuation tasks, such as automated targeted spraying applications or robotic harvesting.

Focusing on horticultural crops, automation in growth stage identification has been an open challenge for multiple decades due to the very nature of the crops, which in their majority are high-value and demand timely interventions to maintain top-cut yield quality. Moreover, as harvesting is one of the most laborious operations throughout the growing season, especially for open-field vegetables, and at the same time heavily affects the final quality of the produce, numerous different approaches have been implemented to achieve automated mapping of crop growth across larger fields and assist harvesting. One of the first attempts was performed by Wilhoit et al. (1990), who correlated the grey-level run lengths of broccoli plant images captured in an indoor controlled environment with broccoli head sizes. Their findings indicated an exponential relationship between the run functions of the images and the area of the broccoli heads. Qui and Shearer (1992) were able to differentiate between various levels of maturity for different varieties of broccoli by using the average response values from various frequency bands. Shearer et al. continued their research, and in 1994 they used a similar approach to determine the maturity of broccoli, this time using a single grey-scale line sampled from each broccoli flower head image of their dataset, and then using a co-occurrence texture analysis method. Ramirez (2006) collected an RGB imagery dataset and a set of simple image processing techniques to identify broccoli heads. Tu et al. (2007) was the first to combine image analysis techniques and neural networks to identify broccoli quality parameters, although their initial dataset consisted of only broccoli heads imaged on a white light stable background, isolated from the leaves.

As developments in computer vision allowed the research to move from simple image analysis frameworks to more complex and automated pipelines, the interest shifted towards Artificial Intelligence during the start of the current century. Commercial RGB cameras and machine learning algorithms can provide affordable and versatile solutions for crop detection. Computer vision systems based on deep CNN (LeCun et al., 2015) are immune to variations in illumination and large inter-class variability (He et al., 2016), both of which have posed challenges in agricultural imaging in the past, thus achieving robust recognition of the targets in open-field conditions. Recent researches (Bargoti and Underwood, 2017, Sa et al., 2016) have shown that the Faster R-CNN (region-based convolutional neural

network) architecture (Ren et al., 2015) can produce accurate results for a large set of horticultural crops and fruit orchards. Bargoti and Underwood (2017) also used a Faster R-CNN network for crop, trained on orchard images captured by a robotic ground vehicle, creating a tree dataset of apples and mangoes, as well as another separate one for almonds, captured by a Digital Single-Lens Reflex camera (DSLR). They achieved an F1-score of 0.9 for the robot-captured mango-apple dataset, outperforming their almond DSLR dataset, which achieved a 0.77 F1-score. This specific finding might indicate that the sensing device might in several cases not be as important as the crop features and how easily they can be captured in the training images. However, it should be noted that in both tests, the authors used several data augmentation techniques, something that might have cancelled some advantages of the DSLR imagery. Sa et al. (2016) used a VGG16, the perceptual backbone in the Faster R-CNN architecture (Simonyan and Zisserman, 2015) trained on ImageNet (Deng et al., 2009) to detect greenhouse peppers and melons, achieving an F1-score of 0.83. Their dataset consisted of 4-channel arrays, generated from the fusion of an RGB and NIR sensor, which is quite similar to the approach used in this thesis. Madeleine et al. (2016) integrated a Faster R-CNN to a multi-sensor framework consisting of an imaging sensor and a LiDAR to detect mangos. Kusumam et al. (2017) used an RGB-D camera mounted on a terrestrial vehicle and a 3D image analysis pipeline consisting of a histogram-based feature extraction and an automated classifier (SVM) for the detection of high-maturity broccoli heads. Birrel et al. (2020) deployed a YOLOv3 object detector and a Darknet object classification network to investigates their potential to initially detect and then classify iceberg lettuce based on their maturity in real-time and under open-field conditions, in order to automate the harvesting operations with an autonomous terrestrial harvester. Chen et at. (2019) utilized UAV RGB orthoimages the Faster R-CNN and ResNet-50 networks to develop strawberry detection system for yield estimation. Furthermore, crop detection results have been integrated by data association approaches, by employing object tracking or mapping to perform fruit/head counting for row-based crops (Liu et al., 2019). Santos et al. (2020) compared YOLO and Mask R-CNN on their ability to detect and cluster wine grapes, and also evaluated their performance with a Mask R-CNN instance segmentation architecture approach. Koirala et al. (2020) used a set of YOLO and RCNN variations to detect mango tree panicles. Junos et al. (2021) used a modified YOLOv3 model to detect loose oil palm fruits on the ground, using an RGB image dataset collected from a very small UAV (DJI Tello). Mutha et al. (2021) used YOLOv3 to detect and classify the maturity of the tomatoes using an RGB image dataset. García-Manso et al. (2021) used a Faster R-CNN modification, using only square base Regions Of Interest (ROIs) (Lin et al., 2017) to detect and categorise broccoli heads on proximal images based on their maturity level, and how close they are to optimal harvesting.

In recent horticultural research literature, several studies have also focused on the localization of broccoli heads, without any evaluation on their maturity. Blok et al. (2016), used a methodology based on texture filters, applied on data of 228 broccoli heads. Though it presented a high precision, the recall was relatively low due to overfitting and generalization problems of the algorithm. However, in 2021, the same authors introduced deep learning techniques in their pipeline, achieving better broccoli location and segmentation results, while succeeding in an overall more robust solution with significantly improved generalization than in their initial study (Blok et al., 2021). Le Louedec et al. (2020), used a system for the localisation of broccoli heads, based on 3D information obtained from RGB-D sensors,

along with a CNN auto-encoder trained for the task of semantic segmentation using these 3D information. Bender et al. (2020) used a Faster Region-Based architecture to locate broccoli and cauliflower plants (not the harvestable heads) using hyperspectral data collected by a terrestrial robot (Ladybird). Finally, Zhou et al. (2020) compared 3 different deep architectures (GoogleNet, VGG16, and ResNet 50) for the segmentation and localization of broccoli heads, followed by a grading based on the head quality (e.g. presence of yellow or black spots) and a regression to estimate yield (head weight).

Broccoli is an example of a high-value crop that requires delicate handling throughout the growing season and during its post-harvesting handling. As the broccoli head can be easily damaged, resulting in visible stains, it is thus still harvested by hand using handheld knives. On top of that, it allows for a very strict time window of "optimal maturity" when the high-end quality broccoli heads should be harvested, before they remain exposed for too long in high humidity conditions and become susceptible to fungal infections and quality degradation. Even slight delays from this time window can result in major losses in final production (Figure 15). However, manual harvesting is a very laborious task, not only for the process of harvesting itself, but for the scouting required to initially identify the field segments where several broccoli plants have reached this maturity level. This scouting process is performed on foot, as agricultural vehicles cruising across the fields result in soil compaction, something highly undesirable in horticulture, especially in the case of organic systems (Soane and van Ouwerkerk, 1994).

Figure 15. Examples of broccoli fields at full bloom, representing losses due to errors in harvest planning.

This case creates a very interesting challenge. First, the scouting process can be automated using machine learning, drastically increasing the overall efficiency and reducing human effort required. At the same time, UAVs can act as a double-benefit factor. They can easily supervise large areas rapidly, while diminishing any potential soil compaction problems. There is a growing need for automated horticultural operations due to increasing uncertainty in the reliability of labor and to allow for more targeted, data-driven harvesting (Bechar & Vigneault, 2016). The scope of this thesis is to deploy a state-of-the-art object detection CNN model, YOLOv5, to assess its potential in maturity classification of open-field broccoli plants, by using a multispectral image dataset collected from low altitude UAV flights.

3. Materials and Methods

3.1 Experimental Overview

The experiment took place in Marathon region, Greece, in a commercial organic vegetable production unit. This region is specifically known for its horticultural production, being the main vegetable provider for Athens, the capital of Greece. The aim of the experiment was to collect low-altitude multispectral imagery of broccoli crops, and to this end, the timing of the data acquisition flights was specifically designed to be performed a few hours prior to the first wave of selective harvesting. This was desired for two (2) reasons: 1) to ensure that the entire field was intact (no broccoli heads were harvested), maximising the number of sample density in every image and the generated field orthomosaic and 2) to make sure that individual plants of different maturity levels were present across the field, as it was at the very start of the harvesting season. The selected experimental parcel was located in the south-west part of the production unit, expanded on an area of 1 ha. The segment on which the ground truth targets were deployed covered slightly more than half of it (Figures 16 and 17).

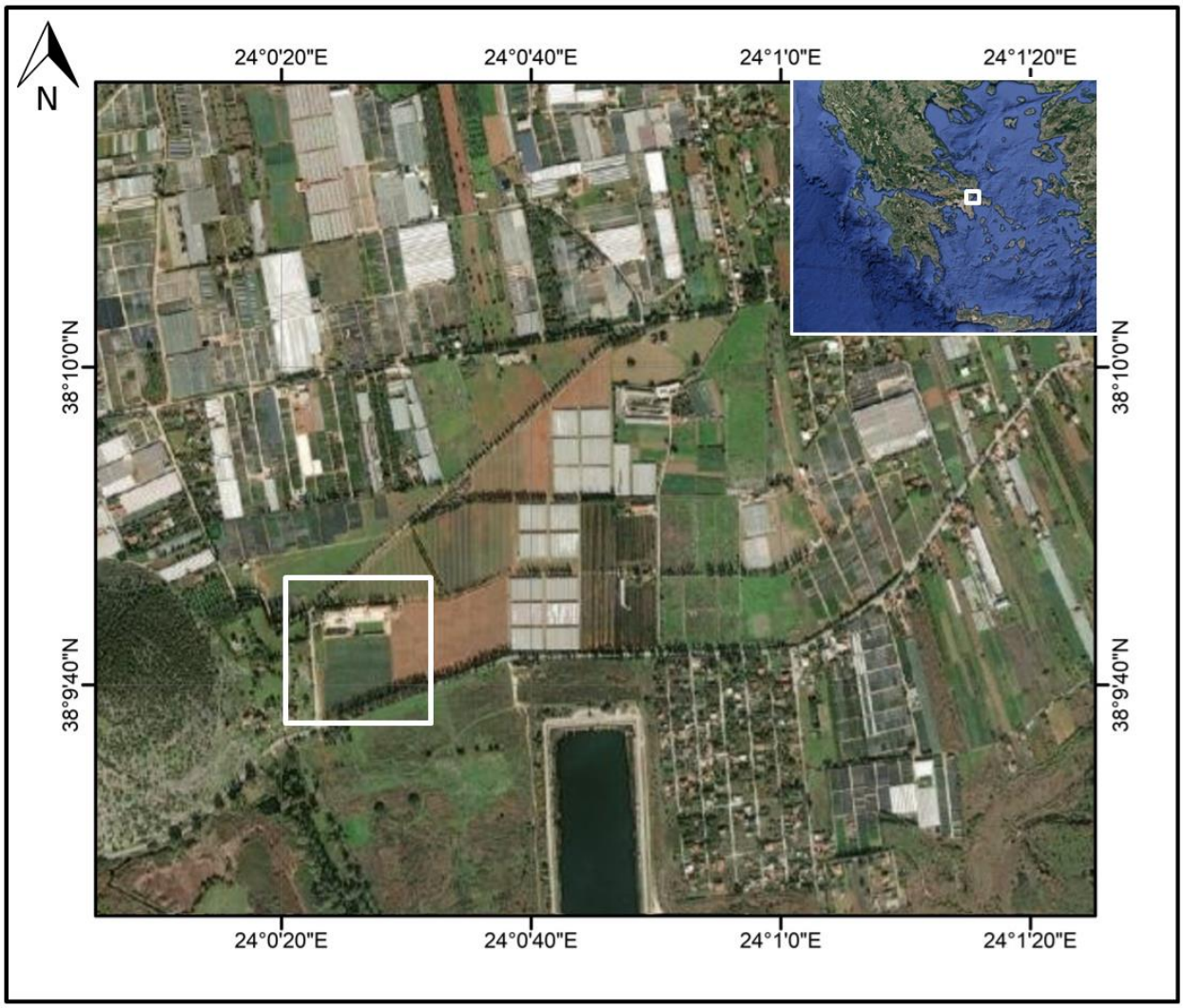


Figure 16. The location of the selected experimental parcel.

Figure 17. The segment of the experimental parcel that was selected as our area of interest.

Data collection was performed by two quadcopter drones, a DJI Phantom 4 Pro (SZ DJI Technology Co.) and a custom quadcopter, each mounted with identical multispectral cameras (Parrot Sequoia+) (Figure 18). The UAVs generated low-altitude aerial imagery datasets of four (4) spectral bands using the respective monochromatic global shutter sensors of the camera, along with RGB imagery using its rolling shutter sensor. The central wavelengths of each single-band channel are 530-570 nm, 640-680 nm, 730-740 nm and 770-810 nm respectively for the monochromatic sensors, with a ± 40nm spectral band width. At the same time, the integrated RGB camera of the DJI Phantom 4 UAV was also used to execute flights and generate individual datasets at the same altitude and overlap values as the Sequoia RGB flights. The specifications of each sensing system are presented below (Table 3). Moreover, both sensing arrays were equipped with the default Parrot Sequoia irradiance sensor to ensure that the captured imagery were radiometrically accurate, while an additional manual radiometric calibration was performed prior to each flight using a reflectance 4-band calibration panel (Airinov Aircalib). The individual raw images were saved on the internal memory of the sensor, to eliminate the potential of communication disruption with the SD cards located inside the irradiance sensor. The monochromatic images were saved as RAW 10-bit TIFF files and the RGB imagery in JPG format, accompanied by the georeferencing metadata of each image.

Table 3. The technical specifications of the mounted sensors.

Parameter	Monochromatic sensors	Sequoia RGB	Phantom RGB
Pixel size	3.75 µm	1.34 µm	2.41 µm
Focal length	3.98 mm	4.88 mm	24mm
Resolution	1280 x 960	4608 × 3456	4096×2160
HFOV	62°	64°	74°
VFOV	49°	50°	51°
DFOV	74°	74°	84°

Figure 18. The two UAVs deployed for the experiment and their preparation.

3.2 Data Acquisition Methodology

The flight missions were executed during the time window between 11:00 and 13:00 (when the solar elevation angle was greater than 45°), to avoid drastic deviations in solar illumination between flights, with the two (2) UAVs performing flights consecutively in rotation. The flights were executed with similar flight parameters for both UAVs, operating at a fixed altitude of 10m AGL, which is considered consistent as the field was leveled by a flattening cultivation roller approximately 4 months prior to the measurements. Both sensors were set to capture images at a fixed interval of 2sec/capture, and

therefore the flightplans were designed around this parameter. The side and front overlaps were both selected to be 80%, resulting in a cruising speed of approximately 0.9 m/s and 1.1 m/s for the multispectral and RGB flights respectively. Finally, the orientation of the flight lines was selected to be parallel to the orientation of the planting rows. The flightplan parameters for each generated dataset are presented in the following table (Table 4). The generated flightplans executed by the UAVs are also presented below (Figure 19). Throughout the data collection day, a total of five (5) flights were executed, resulting in the creation of three (3) multispectral and two (2) RGB datasets.

Table 4. Flightplan parameters for each dataset.

Parameter	Multispectral	RGB
Frontal overlap	80%	80%
Side overlap	80%	80%
Flying speed	0.9 m/s	1.1 m/s
Photo interval	2 sec	2 sec
Altitude AGL	10 m	10 m
GSD	0.95 cm	0.25 cm

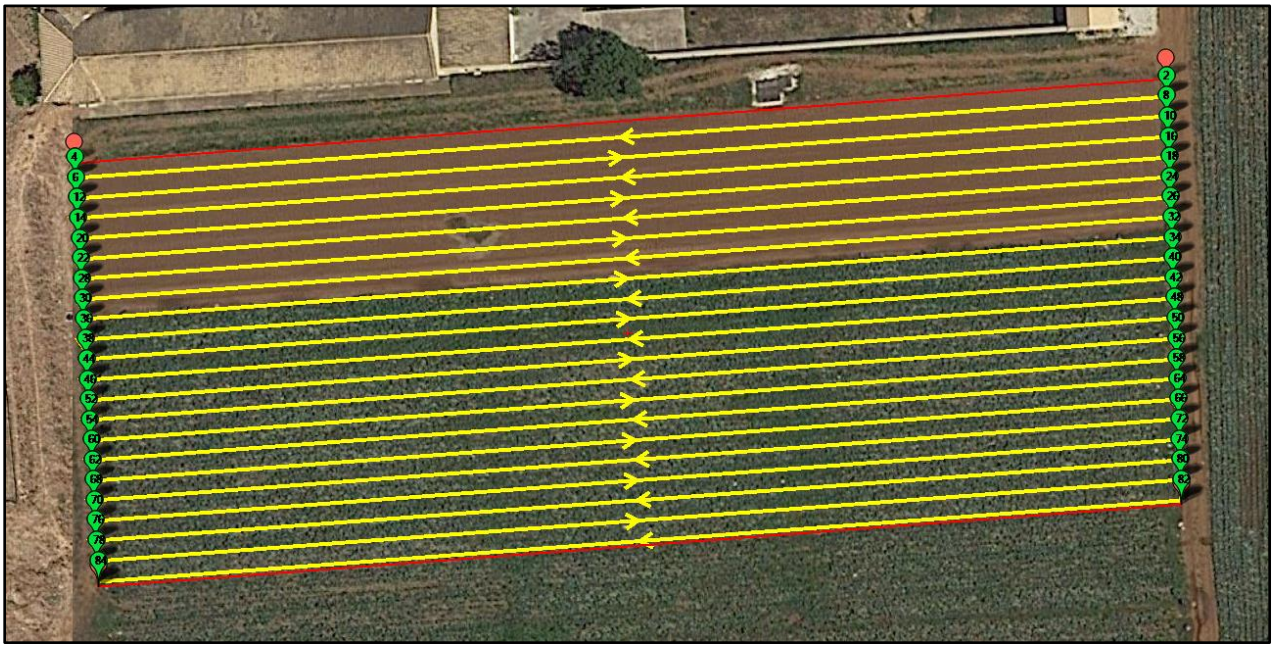


Figure 19. Example of the generated flightplan, for the multispectral data collection.

Before the start of the data collection flights, a total of 45 ground truth targets were deployed and stabilised across the field, to support the annotation stage (explained in next chapter). The targets indicated the maturity level of selected broccoli crops, as they were categorised by an expert agronomist who participated in the process. The human expert would indicate a total of 15 broccoli heads of 3 different broccoli maturity classes. These classes ranged from 1 to 3, with class 1 representing immature crops that will not be harvested for at least during the following 15 days, class 2 representing heads that are estimated to reach harvesting level within a week, and finally, class 3 which contained

exclusively “ready to harvest” heads. Due to the high humidity of the air near the surface and the constantly wet soil, the targets were enveloped inside transparent plastic cases, to protect them from decomposing as they would remain on the field for approximately two hours. The cases had been tested in the university campus to verify that the labels (numbering) remained visible from the UAV imagery. In case a ground truth label was covered heavily by surrounding leaves, a bright yellow point-like object was also placed nearby as a pointer. Examples of the deployed targets are presented below (Figure 20).

Figure 20. The deployed maturity ground truth targets and the bright “index” pointers.

3.3 Dataset Pre-processing

The first step of the data processing phase was to generate the orthomosaic maps from each flight. For this process, the photogrammetric software Pix4D Mapper (Pix4D SA) was used, generating a total of two (2) RGB orthomosaics and three (3) multispectral ones. The RGB orthomaps were exported as 3-band georeferenced TIFF files, while the multispectral one were exported as individual single-band tiffs and merged with each other using a GIS (QGIS), also resulting in georeferenced (5-band) TIFF files. In the following step, the best orthomosaics were selected for each dataset. This was done to make sure that the final dataset(s) that are going to be presented to the machine do not contain duplicates of the same

crops, as this would increase the initial bias of the experiment. For both the multispectral and the RGB dataset, after close inspection, the mosaic of the second flight with the DJI Phantom was selected as it produced the best result (less blurry spots and zero holes in the mosaic), potentially indicating that the flight conditions were better during that time window and enabled the UAV to perform its flight in a more optimal way with less disruptions (further explained in the Discussion section).

Once the mosaics were selected, the next step was to create the two datasets that were initially aimed to be fed to the model. As the following step involved image labeling on individual images, the annotations would be best drawn on top of a high resolution layer, to avoid errors and mistakes. The RGB mosaic has four times higher GSD and thus sixteen times higher resolution compared to the monochromatic ones. For this reason, the decided strategy was to initially rectify the two raster layers (orthomosaics) using ground control points, and then once they were aligned, the low resolution layers (multispectral) would be resampled to the resolution of the RGB layer. This was done, not only to increase the annotation accuracy, but also for another critical reason. The annotating format for the YOLO model requires a pixel-positioning encoding, meaning that once drawn, each annotation file is exported in the form of a text file containing the sequence of the four (4) coordinates of each bounding box within the image. Therefore, to make sure that the same annotations are used in each experiment regardless of dataset, all layers should initially be aligned, and resampled to the same resolution.

The result of the previous step was 2 aligned orthomosaics with the same resolution, one containing only the RGB layers and the other one the RGB layers but also including the additional monochromatic layers (Red-edge and NIR) of the multispectral camera. As both mosaics were georeferenced, an initial crop with a rectangular vector layer was performed in QGIS to eliminate the majority of the black, zero-valued pixels that were created during the mosaicking process (the exported TIFF mosaic is written in a minimum-bounding-box method, surrounding the mosaic map with black pixels to create a rectangle, were all bands are assigned a zero value for the pixels not containing any data). This step however served another purpose, as the next phase involved “cutting” the mosaics into smaller images so that they can be used as input to the CNN. This was easily performed using Python, with a simple script that iterated an entire rectangular mosaic and then copied the first X number of pixels in one direction and Y number of pixels in the other direction, for all bands of the initial mosaic, and then saving them as a new image. In our case the desired image dimensions were 500 x 500 pixels, and therefore the step for each loop (one for each axis, as the mosaic gets scanned) was set to 500. This resulted in two (2) datasets of images with a resolution of 500 x 500. The datasets were then reviewed manually, to eliminate any potential instances with high blur or poor mosaicking areas.

The final step of the preprocessing was the labeling. In this phase, a single dataset was imported to the Computer Vision Annotation Tool (CVAT[7]), and the images were annotated using the ground truth labels as a basis (Figure 21). The annotations consisted of rectangular boxes assigned with the respective maturity class of each broccoli head they contained. The annotations naturally applied for both datasets thanks to the preprocessing strategy described in the previous steps. Each of the final datasets contained a total of 288 images with over 700 annotations (Figure 22).

[7] https://cvat.org/

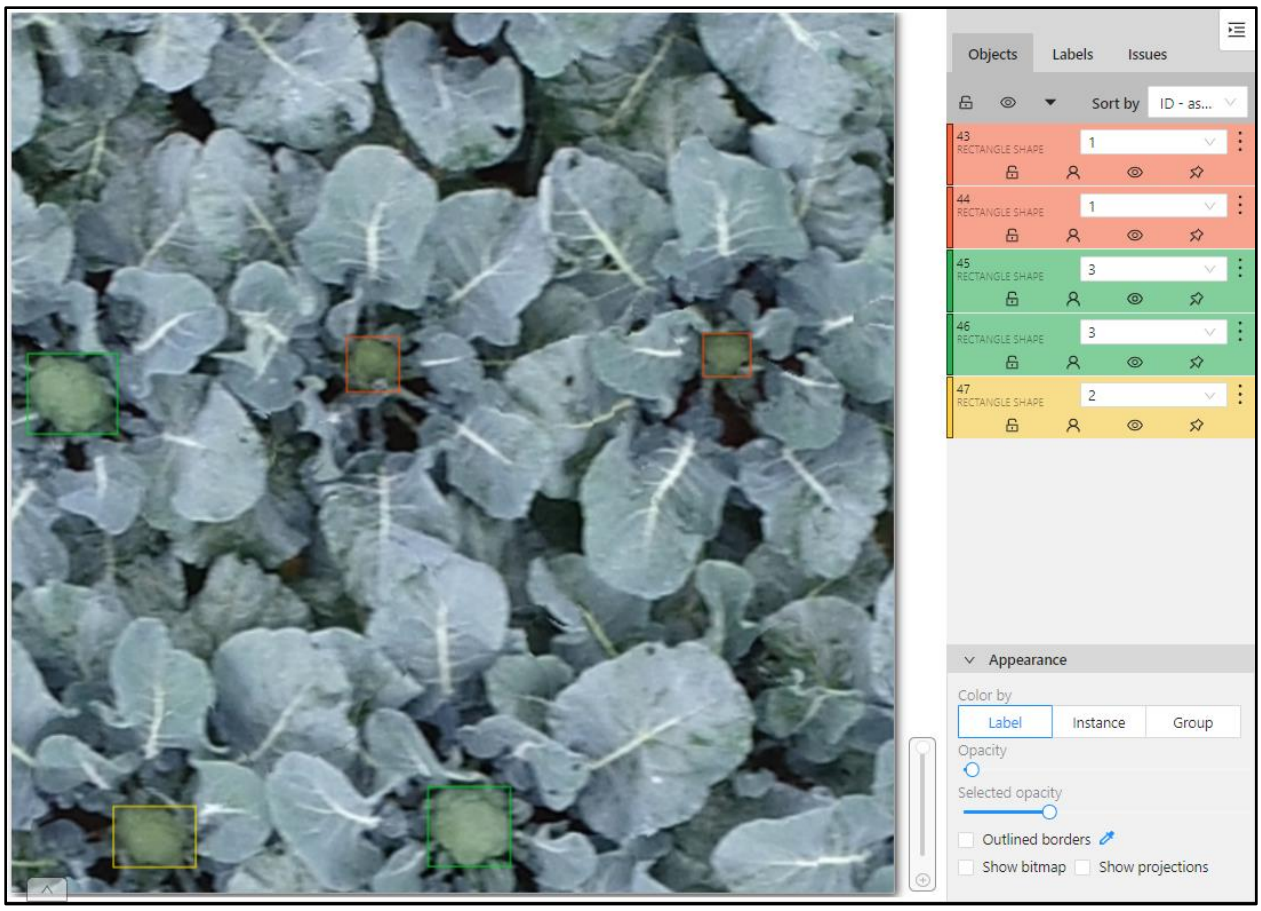


Figure 21. Example of the annotation process on CVAT.

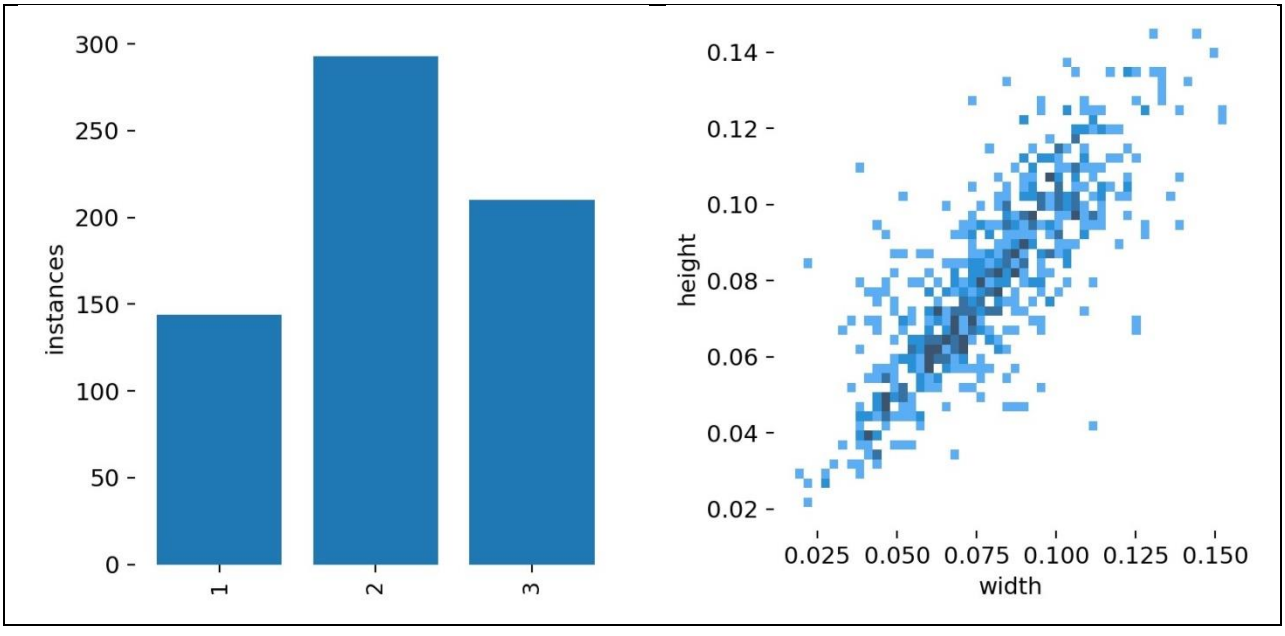


Figure 22. The annotation instances' representation and size plot.

3.4 Model Training

The final step of the analysis involved the training of the YOLOv5 model with the two generated datasets. The data was split 70/30 between training and validation and the model was trained by adjusting YOLO's native resolution to that of our images. The selected weights of the network were the YOLOv5L, which were trained from scratch. Moreover, data augmentation was applied during the training phase to improve network generalization. Two types of image augmentation were applied, image flips (all directions) and image mosaicking (merge-in-one), to assist the network in effectively identifying smaller objects, such as class 1 broccoli heads or larger heads that got "cut" into different images during the mosaic tiling phase, now only covering a few pixels each. The number of epochs was set to 300 for all repetitions, while batch size was a parameter that was further explored with separate tests, as it can generally affect the way that the model generalises, but at the same time, requires high computational power to achieve greater batch size training. The overview of the selected training hyperparameters is presented below (Table 5). The training took place exclusively in a desktop Windows computer with a GeForce RTX 2080 Ti (NVIDIA Corporation) Cuda GPU and a 16-core AMD Ryzen Threadripper 1950X (Advanced Micro Devices, Inc.) CPU.

Table 5. Overview of the hyperparameters used for the model training.

Hyperparameter	Value
Learning rate (start)	0.00658
Momentum	0.934
Weight Decay	0.00047
Warmup Epochs	5
Warmup Momentum	0.90
Intersection Over Union threshold	0.2
Augmentation	Image flipping (0.5), Image mosaicking (0.5)
Epochs	300
Image size	500 x 500
Batch size	2, 4 & 6

4. Results

For the evaluation of the model's performance, a set of widely used metrics was selected. Precision and Recall is a set of fundamental machine learning evaluation metrics. Precision is the fraction of relevant instances among the retrieved instances, while recall is defined as the fraction of samples from a class which are correctly predicted by the model. As these two parameters have an equilibrium dynamic, a common way to visualise them is with a Precision-Recall curve that shows the tradeoff between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both Precision and Recall simultaneously show that the classifier is returning accurate results (high Precision), as well as returning a majority of all positive results (high Recall). Another very popular evaluation metric which combines the two aforementioned metrics is the F1-score, which is essentially the harmonic mean of Precision and Recall. The aforementioned metrics are therefore calculated as:

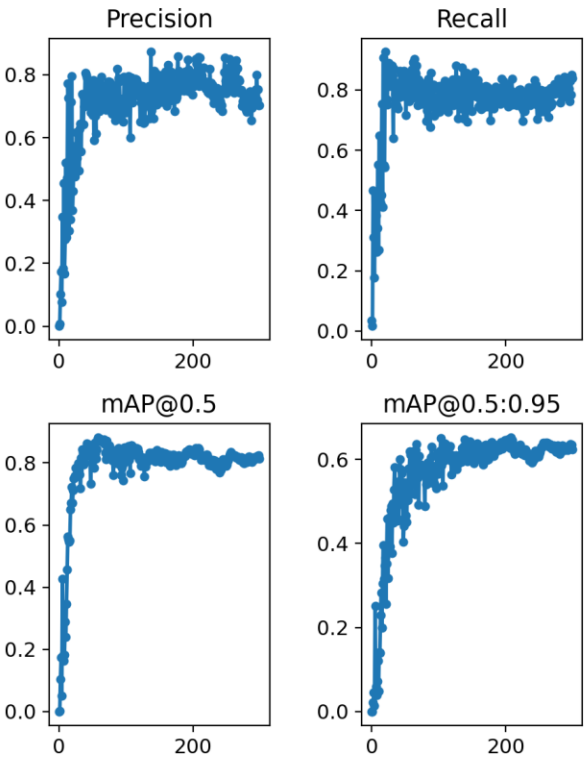
$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN} \qquad \text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Average precision (AP) is the area under the recall-precision curve, and is considered the standard performance measure for object detection. AP[0.5:0.95] corresponds to the average AP for IoU from 0.5 to 0.95 with a step size of 0.05. In the case of AP at IoU threshold 0.5, the confidence score threshold is set to 0.5 and the IoU threshold is also set to 0.5. The results of the training iterations of each dataset are presented in the following table (Table 6). The "progress" plots of specific metrics throughout each training iteration, along with the associated Precision-Recall curves, are presented in order (Figure 23 - 29). Finally, an example of a trained model's detections is presented (Figure 30).

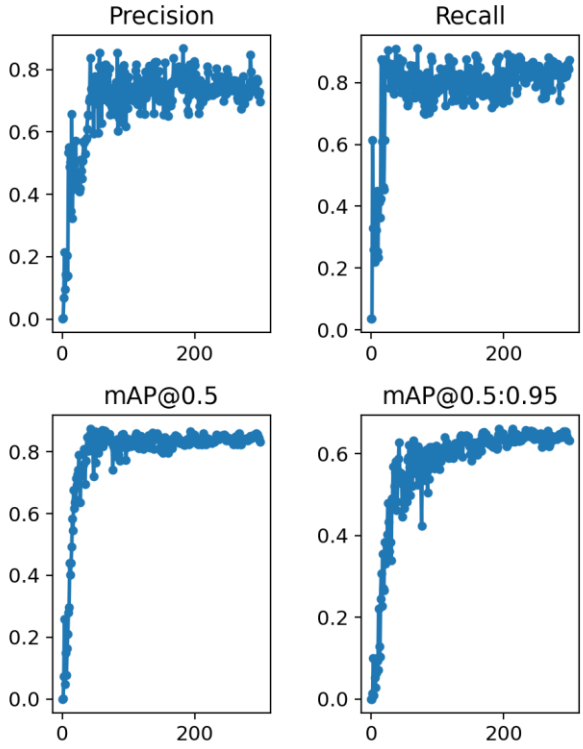
Table 6. The overview of the experimental results and the best metrics achieved in each iteration.

Batch, epochs dataset	Precision	Recall	F-1 score	Map@0.5	Map@0.95
2, 300, RGB	0.87	0.93	0.812	0.882	0.652
2, 300, Multi	0.87	0.91	0.826	0.874	0.661
4, 300, RGB	0.86	0.88	0.818	0.865	0.663
4, 300, Multi	**0.89**	**0.94**	0.817	0.88	**0.671**
6, 300, RGB	0.86	0.912	0.821	0.88	0.651
6, 300, Multi	0.88	0.884	**0.829**	0.873	0.666

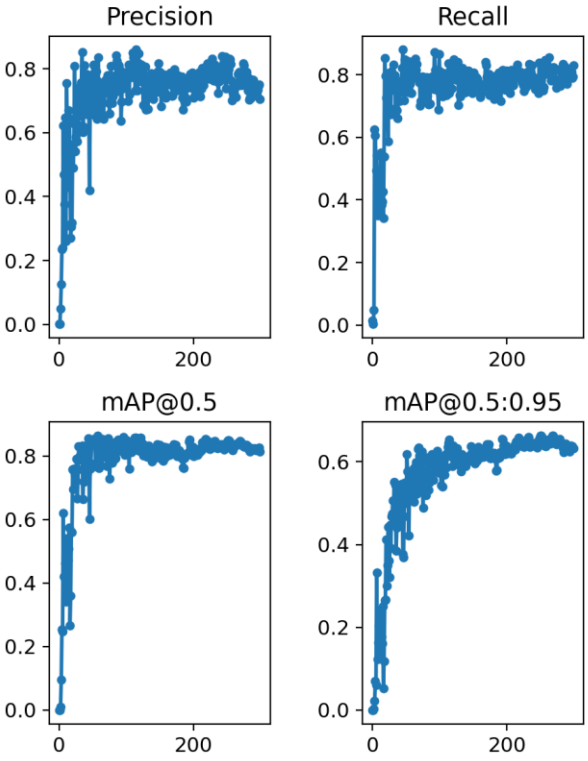
2, 300, RGB



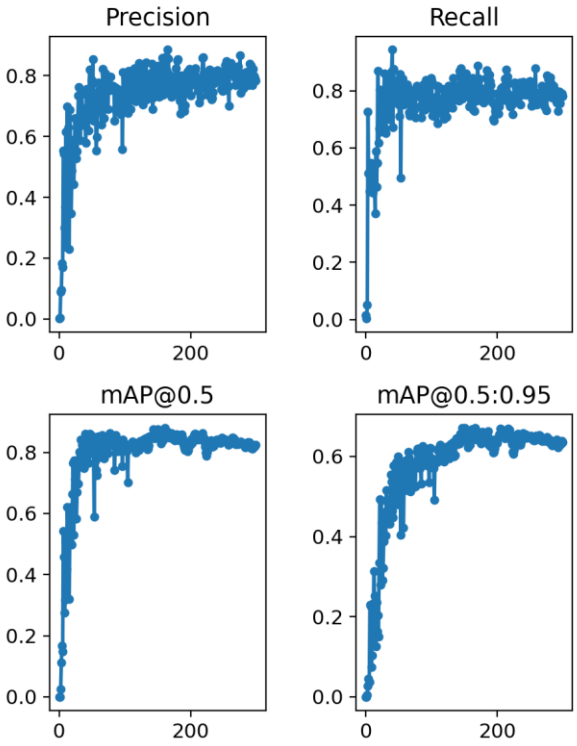
2, 300, Multispectral



4, 300, RGB



4, 300, Multispectral



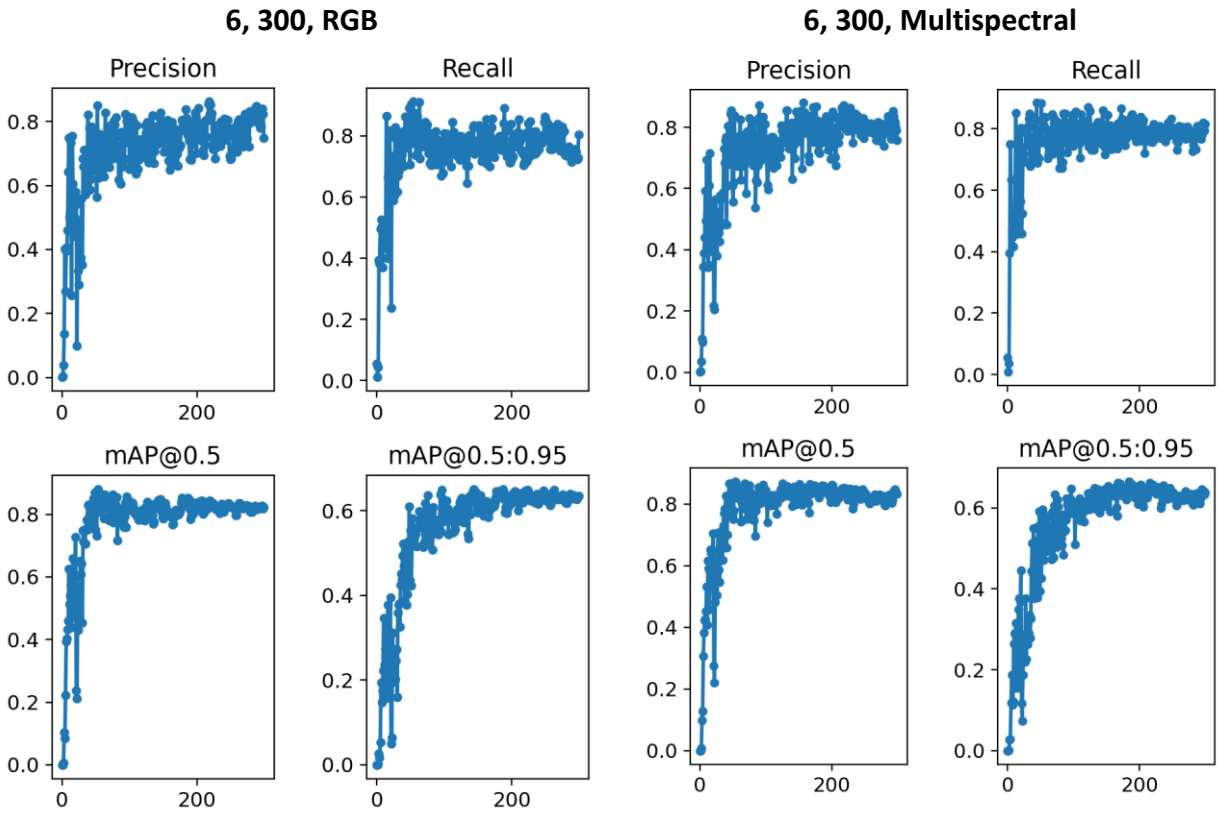


Figure 23. The progress of Precision, Recall and MaP at the two thresholds.

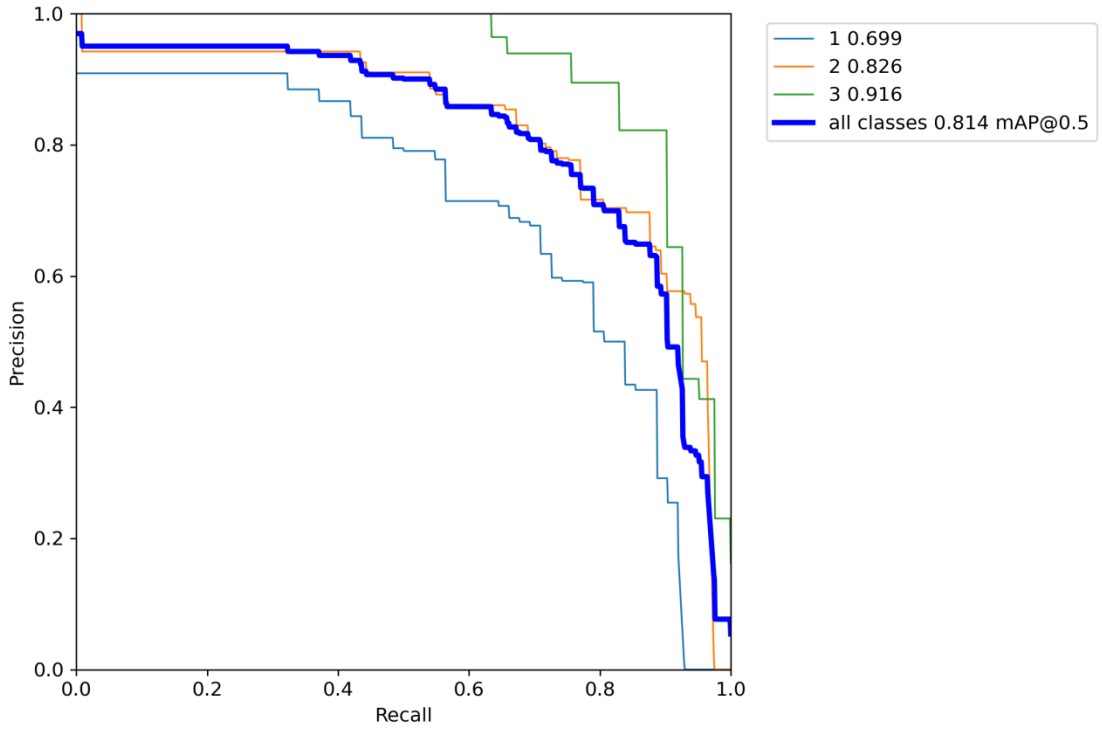


Figure 24. The PR curve of the iteration (3, 300, RGB)

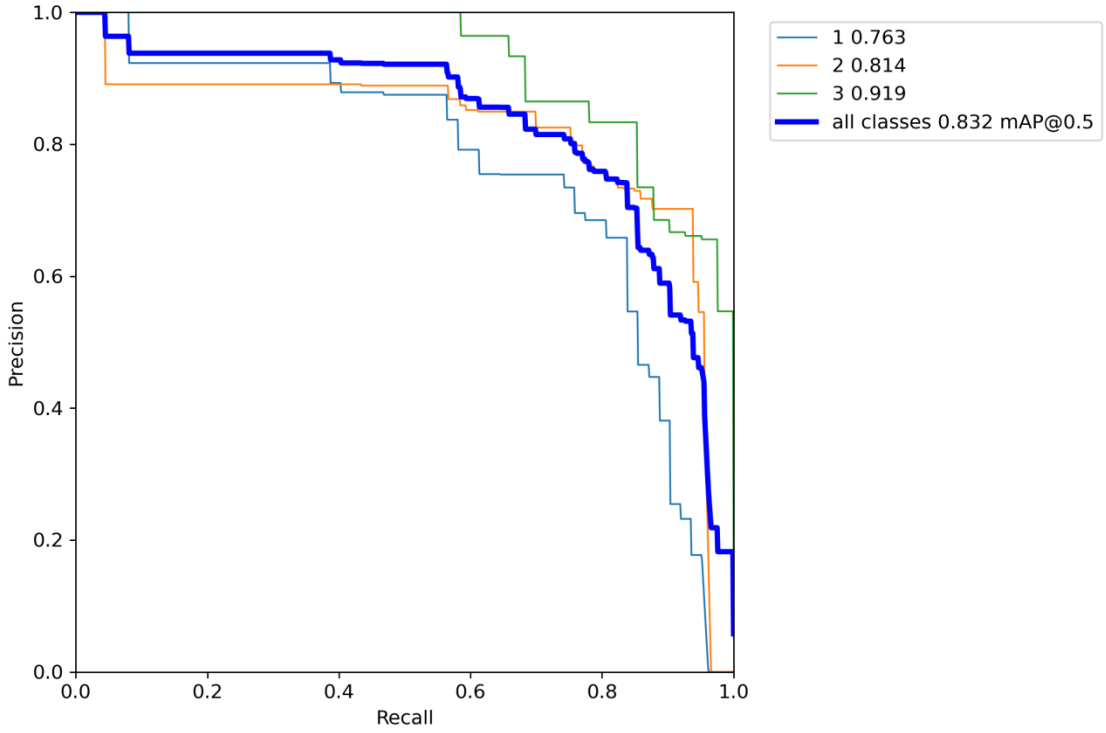


Figure 25. The PR curve of the iteration (3, 300, Multispectral)

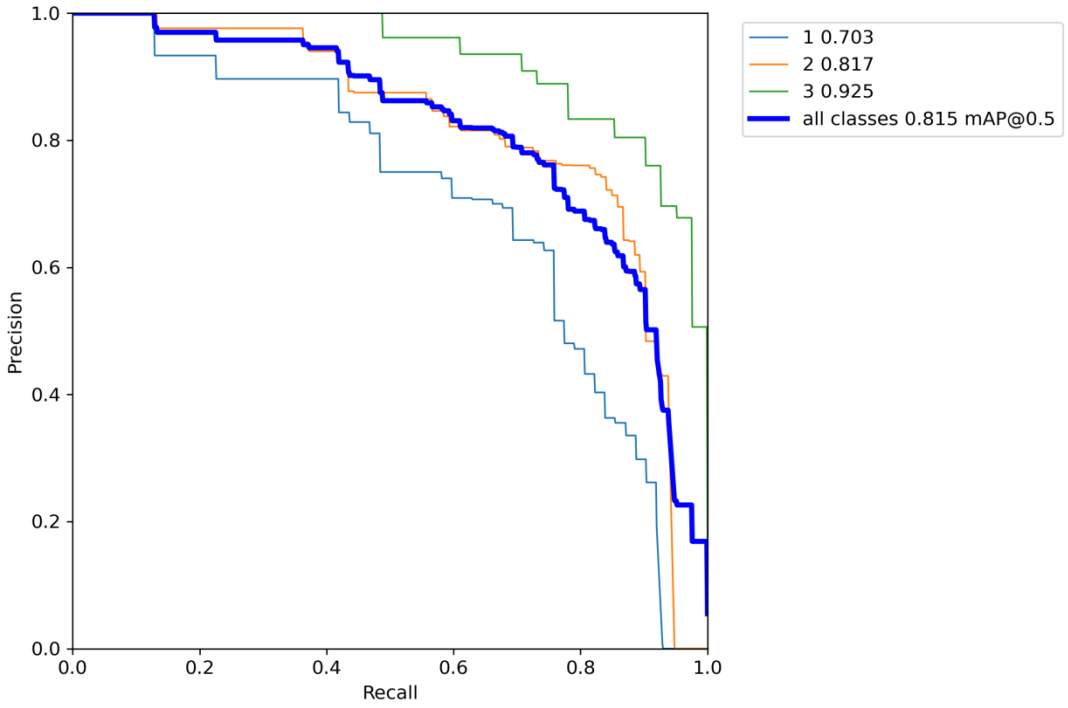


Figure 26. The PR curve of the iteration (4, 300, RGB).

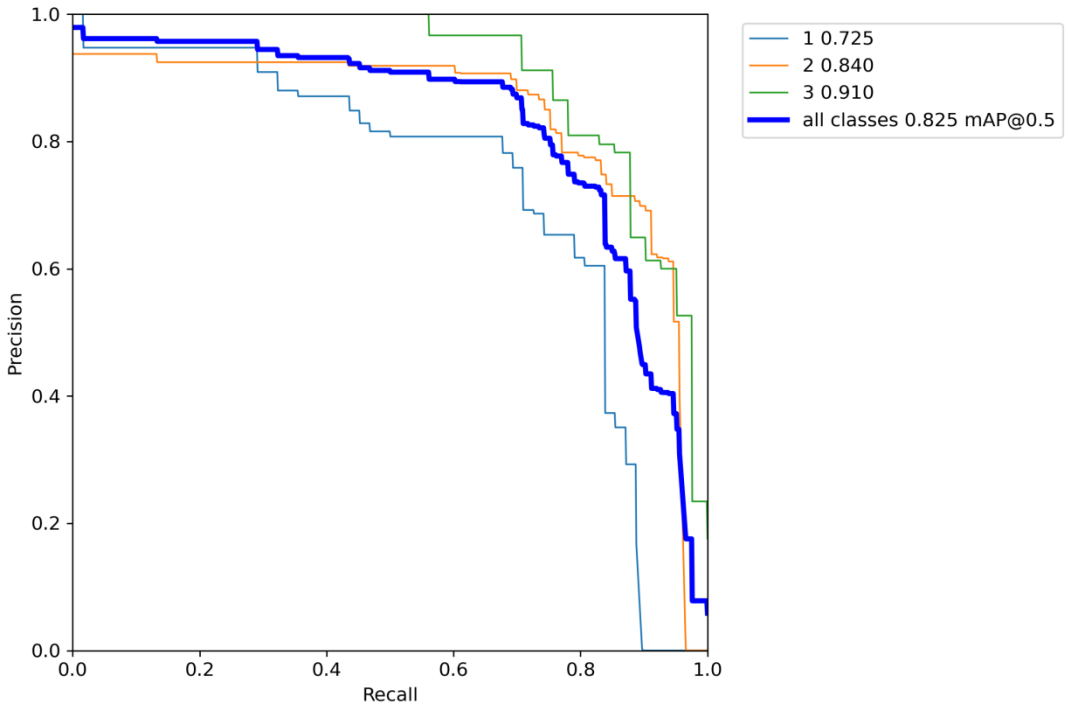


Figure 27. The PR curve of the iteration (4, 300, Multispectral).

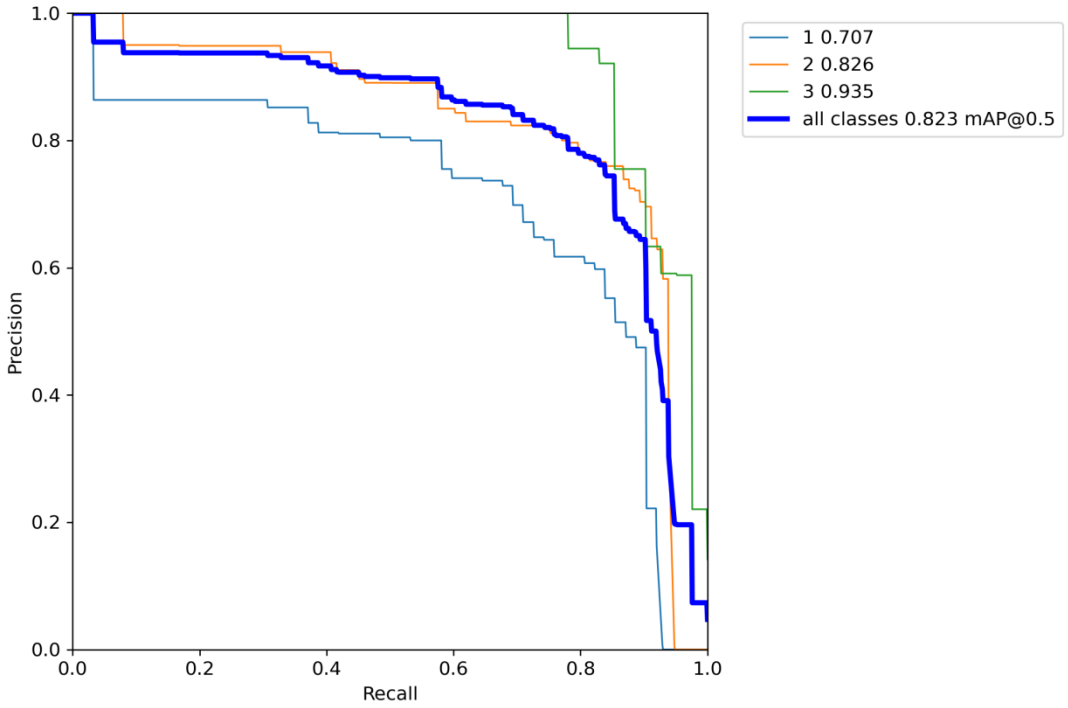


Figure 28. The PR curve of the iteration (6, 300, RGB).

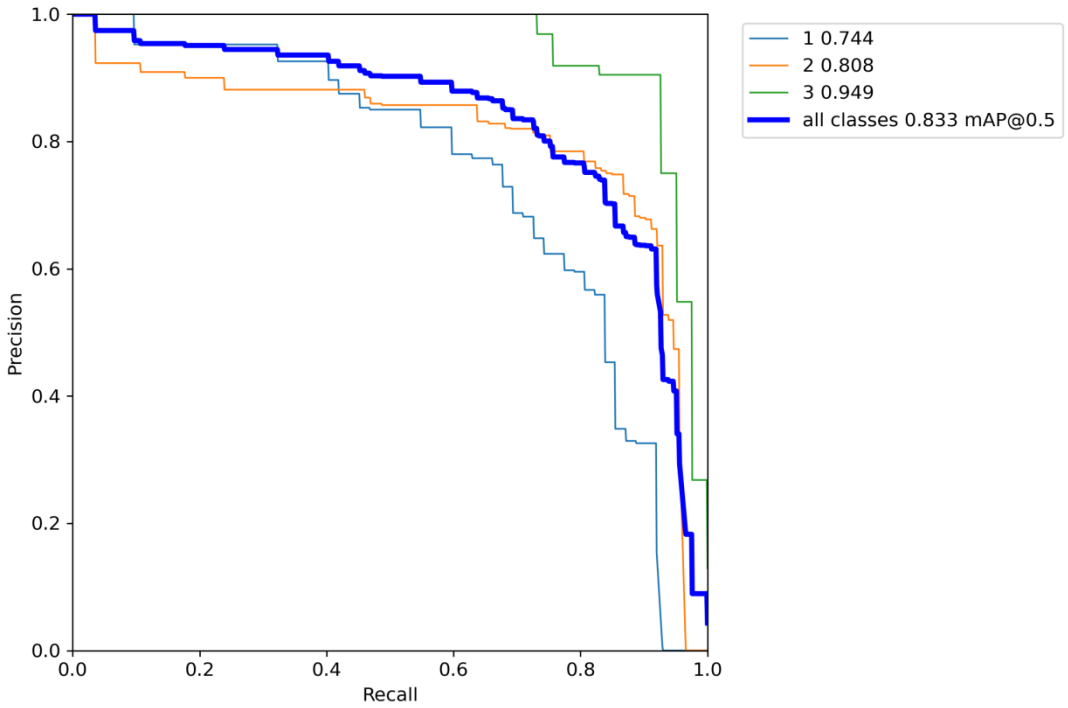


Figure 29. The PR curve of the iteration (6, 300, Multispectral).

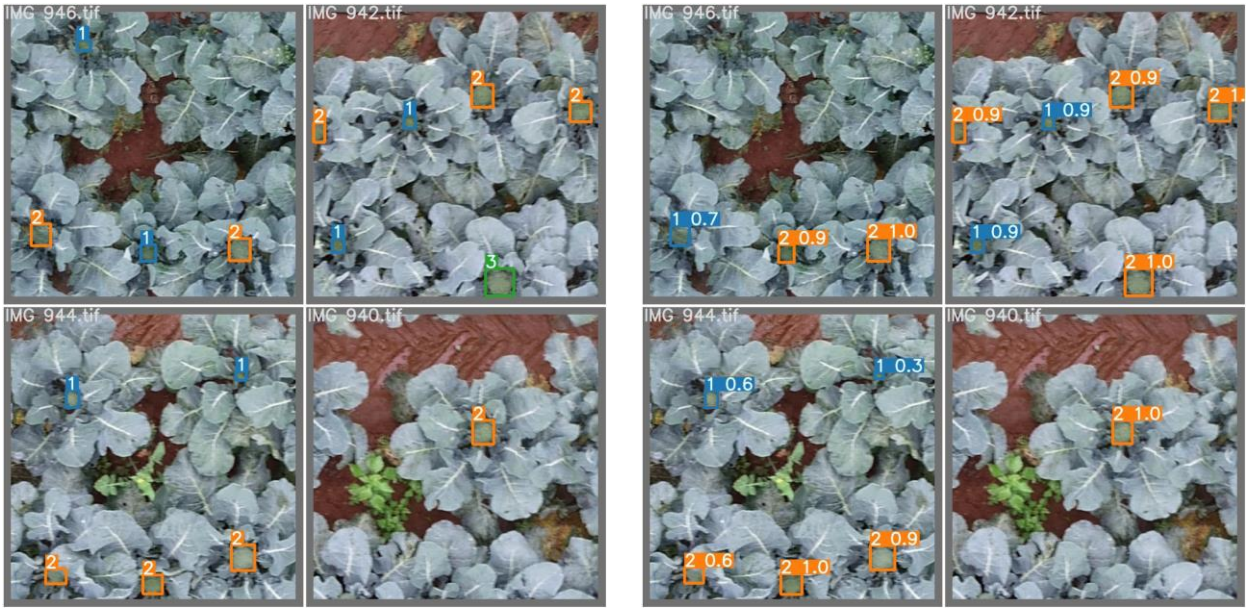


Figure 30. Examples of labelled images (left) and model detections (right) with their respective confidence scores.

5. Discussion

The results of the training and validation experiments indicate that the model is clearly able to perform very well for the task of automated maturity detection. All experimental iterations maintained an F-1 score higher than 0.81 and a MAP@0.5 higher than 0.865, which are solid performances considering the open-field nature of the datasets. Across all iterations, the best possible Precision, Recall, F-1 score and MaP at IoU 0.95 were achieved by the multispectral dataset, while the MaP at IoU 0.5 demonstrated no major differences between the best performing iterations for both datasets. An interesting finding is that no major differences in performance were observed between these two datasets. This might indicate that the RGB sensors, which are both significantly easier to purchase and operate (regarding data processing), can yield strong results, without the need for multispectral instruments. Naturally, this is only a hypothesis and multiple iterations with different flights across different cultivation seasons should be performed before reaching this conclusion. From a physical point of view, however, it is reasonable that the wavelengths in the visible spectra and especially the Green and Blue bands can have a greater impact in maturity assessment, as significant changes occur in these wavelengths when broccoli crops shift towards later maturity stages. Moreover, broccoli heads have one of the highest transpiration rates across all horticultural crops (Kader and Saltveit, 2003), and as a result, might not exhibit the same biochemical thermoregulatory operations, that are directly connected to NIR reflectance rates, as other crops (Lillesaeter, 1982).

Regarding the performance of the YOLOV5 model, the best overall performance was achieved by the multispectral dataset at a batch size of 4. This is logical as medium batch size values tend to provide sufficient regularisation, without taking a significant coil on the computation resources. In the present experiment, a batch size of 6 was the greatest value that could be achieved without the GPU overloading. The most noticeable performance is this iteration's Recall of 0.93, which despite its maximum Precision of 0.89 (also the highest Precision score achieved) displayed a relatively low maximum F-1 score (0.817) compared to other iterations. This can be easily explained as the maximum Precision and Recall values reported simply did not occur during the same epoch, indicating a more "dynamic" tradeoff between these two metrics, something that can also be observed in its PR curve (Figure 31).

Regarding the performance on each individual maturity class, the highest performing class among all iterations was maturity class 3, followed by class 2 and finally class 1. This is also expected due to a variety of reasons. Initially, larger broccoli heads naturally cover more space in the image, and are therefore more distinguishable in the first place. The initial stage of our pipeline, similarly to all object detection and classification tasks, involves the detection of the broccoli heads, and then, once detected, their classification. A simple reason that class 3 outperforms the other two, while being the second most represented class, and therefore bias from over-representation are excluded, is because the machine can more easily detect them in the first place. At the same time, most class 3 detections were performed with a high confidence score, as demonstrated below (Figure 33).

Figure 31. Another example on the detection confidence for the different maturity classes.

From the above image we can observe that all class 3 broccoli heads were detected with a confidence score of 0.8-1. At the same time, two existing maturity class 1 broccoli heads were not detected as an object in the first place, due to their smaller size. This image can also give another interesting indication, as we can see that a class 1 broccoli has been detected successfully, but classified as class 2 instead. This is another logical outcome considering that the "boundaries" between classes was an approximation method for field labeling, and not an actual measurable metric. As a result, some level of confusion might occur even for a human observer tasked to separate instances of these two classes. Therefore, taking into consideration that the model had a relative difficulty in separating broccoli heads of class 1 and 2, while performed exceptionally well on broccoli heads of later maturity stages (class 3), a speculation could be made that, in future research, the model would potentially yield even better results and demonstrate a stronger performance, if the problem presented was a simplified "ready to harvest" and "not ready yet" 2-class problem.

On a more technical aspect, the present study has been an excellent opportunity to obtain knowledge regarding the proper planning and deployment of such experiments. First of all, during the orthomosaicing process, it is a common phenomenon that the flight lines around the perimeter of the flight plan, which often also reflect the boundaries of the experimental field, are the ones with the lowest values of overlap. This is because this area is only scanned once throughout the entire flight, and thus the surrounding areas are only captured in the images of a single flight line (Figure 32). This can easily result in poor mosaicking quality in these areas, and if the majority of the targets are placed there, the entire mission is at risk of being fruitless. However, these areas around the perimeter of the field are the easiest ones to deploy ground truth targets on, as they are easily accessible, the targets are not covered by vegetation, and the lower humidity level can potentially increase the time the targets can

“survive” before getting soggy because of the humidity, if not protected properly (as described in the Materials and Methods section). This can create a “trap”, which should always be considered, and the ground truth targets should be properly scattered towards the middle of the experimental area and ideally across a larger area.

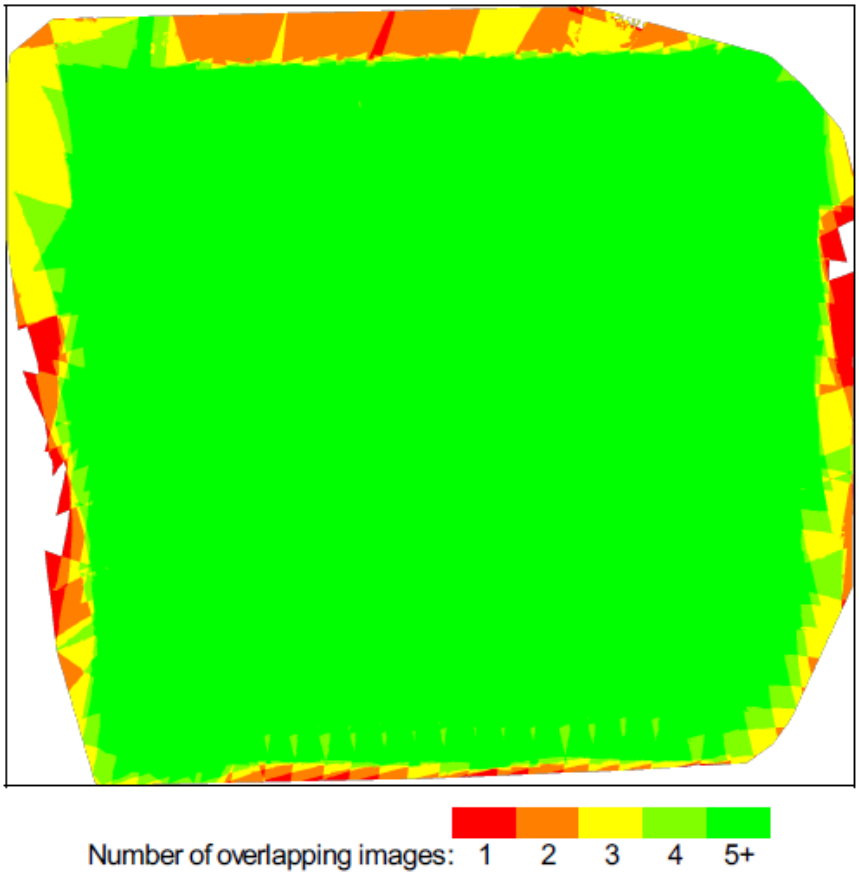


Figure 32. Number of overlapping images computed for each pixel of the orthomosaic.

Finally, as the timeframe during which data collection can be performed for this specific type of experiments is very strict, utmost attention should be given in minimising as many risks as possible before committing to the field visit. One of the most unpredictable factors for all UAV missions is the weather. Certain UAVs have the capacity to perform flights under harsher conditions, while known thresholds are always a safety switch for the pilots to potentially cancel high-risk missions in time if they judge that the conditions do not allow for a safe flight. Weather forecasts can mitigate this risk to a certain extent by giving the pilots enough time to adjust/select the appropriate fleet for each mission based on the conditions they expect to encounter, however, drastic changes, especially in wind speed and direction, are anything but rare. The reason two UAVs of different operation capacity were deployed for this experiment was because there was a forecast indication of potential strong gusts (of over 55km/h) during the day of the data collection. As described in the Experimental Layout section, this experiment only allows for a window of a few hours, as harvesting operation cannot be further delayed in commercial farms, and flights should be performed in a very efficient manner regardless of unpredicted conditions.

6. Conclusions

With the completion of the present thesis some of the aspirations that hopefully have been achieved is that not only new knowledge on the research subject of agricultural computer vision and artificial intelligence in horticulture has been created, but also that thought for future research on this critical topic has been nurtured. I am confident that the methodology described in this thesis has proven to be both efficient, based on the overall framework that was implemented across all steps and phases, as well as effective, based on the results that it yielded. At the same time, the documentation on aspects of optimal data acquisition preparation for challenging UAV flights, such as the ones presented here, as well as technical difficulties similar to the ones encountered during this experiment, will hopefully serve potential future researchers that will continue this research in the same domain, or transfer this knowledge to their respective field. Finally, as the focal point of this research is an open agricultural problem, it is of utmost importance that future research will uphold the principles that constitute the essence of modern Agriculture: to ensure food security for the human population and to safeguard the environment through sustainable development.

References

1. Anastasiou, E., Balafoutis, A., Darra, N., Psiroukis, V., Biniari, A., Xanthopoulos, G., & Fountas, S. (2018). Satellite and proximal sensing to estimate the yield and quality of table grapes. Agriculture, 8(7), 94.
2. Aslam, S., Herodotou, H., Ayub, N., & Mohsin, S. M. (2019). Deep Learning Based Techniques to Enhance the Performance of Microgrids: A Review. 2019 International Conference on Frontiers of Information Technology (FIT). doi:10.1109/fit47737.2019.00031.
3. Bapst, R., Ritz, R., Meier, L., Pollefeys, M. (2015, September). Design and implementation of an unmanned tail-sitter. In 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 1885-1890).
4. Barbedo, J.G.A. (2019) Plant disease identification from individual lesions and spots using deep learning. Biosyst. Eng., 180 (2019), pp. 96-107, 10.1016/j.biosystemseng.2019.02.002
5. Bargoti, S., Underwood, J., 2017a. Deep fruit detection in orchards. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 3626–3633. doi: 10.1109/ICRA.2017.7989417.
6. Bechar, A., & Vigneault, C. (2016). Agricultural robots for field operations: Concepts and components. Biosystems Engineering, 149, 94–111.
7. Bender, A., Whelan, B., & Sukkarieh, S. (2020). A high-resolution, multimodal data set for agricultural robotics: A Ladybird's-eye view of Brassica. Journal of Field Robotics, 37(1), 73-96.
8. Berni, J.; Zarco-Tejada, P.J.; Suarez, L.; Fereres, E. (2009). Thermal and Narrowband Multispectral Remote Sensing for Vegetation Monitoring from an Unmanned Aerial Vehicle. IEEE Trans. Geosci. Remote Sens. 47, 722–738.
9. Birrell, S., Hughes, J., Cai, J. Y., & Iida, F. (2020). A field-tested robotic harvesting system for iceberg lettuce. Journal of Field Robotics, 37(2), 225-245.
10. Blok, P. M., Barth, R., & Van Den Berg, W. (2016). Machine vision for a selective broccoli harvesting robot. IFAC-PapersOnLine, 49(16), 66-71.
11. Blok, P. M., van Evert, F. K., Tielen, A. P., van Henten, E. J., & Kootstra, G. (2021). The effect of data augmentation and network simplification on the image-based detection of broccoli heads with Mask R-CNN. Journal of Field Robotics, 38(1), 85-104.
12. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934..
13. Boden, M. (1977). Artificial intelligence and natural man. MIT Press. ISBN 978-0-262-52123-9.
14. Burkart, A.; Aasen, H.; Alonso, L.; Menz, G.; Bareth, G.; Rascher, U. (2015). Angular Dependency of Hyperspectral Measurements over Wheat Characterized by a Novel UAV Based Goniometer. Remote Sens. 7, 725–746.
15. Chen, Y., Lee, W.S., Gan, H., Peres, N., Fraisse, C., Zhang, Y., He, Y. (2019). Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages. Remote Sens. 11, 1–21.
16. Colomina, I.; Molina, P. (2014). Unmanned Aerial Systems for Photogrammetry and Remote Sensing: A Review. ISPRS J. Photogramm. Remote Sens. 92, 79–97.

17. D'sa, R.; Jenson, D.; Henderson, T.; Kilian, J.; Schulz, B.; Calvert, M.; Heller, T.; Papanikolopoulos, N. (2016). SUAV: Q—An Improved Design for a Transformable Solar-Powered UAV. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); IEEE: New York, NY, USA.
18. Dandois, J.; Olano, M.; Ellis, E. (2015). Optimal Altitude, Overlap, and Weather Conditions for Computer Vision UAV Estimates of Forest Structure. Remote Sens. 7, 13895–13920, doi:10.3390/rs71013895.
19. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L. (2009) Imagenet: a large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition 2009.
20. Domingo, Ørka, Næsset, Kachamba, & Gobakken. (2019). Effects of UAV Image Resolution, Camera Type, and Image Overlap on Accuracy of Biomass Predictions in a Tropical Woodland. Remote Sensing, 11(8), 948. doi:10.3390/rs11080948.
21. Duckett, T., Pearson, S., Blackmore, S., Grieve, B., 2018. Agricultural robotics: The future of robotic agriculture. CoRR, abs/1806.06762. http://arxiv.org/abs/1806.06762. arXiv:1806.06762.
22. FAO (2017). The future of food and agriculture. Trends and challenges (http://www.fao.org/3/i6583e/i6583e.pdf)
23. Felzenszwalb, P. F., Girshick, R. B., McAllester, D., Ramanan. D. (2010) Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645.
24. Fraser, B.; Congalton, R.; Fraser, B.T.; Congalton, R.G. (2018). Issues in Unmanned Aerial Systems (UAS) Data Collection of Complex Forest Environments. Remote Sens. 10, 908, doi:10.3390/rs10060908.
25. García-Manso, A., Gallardo-Caballero, R., García-Orellana, C. J., González-Velasco, H. M., & Macías-Macías, M. (2021). Towards selective and automatic harvesting of broccoli for agri-food industry. Computers and Electronics in Agriculture, 188, 106263.
26. Gatti, M.; Giulietti, F. (2013). Preliminary Design Analysis Methodology for Electric Multirotor. IFAC Proc. 46, 58–63.
27. Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 580–587.
28. Girshick., R. B. (2015) Fast R-CNN. doi:10.1109/ICCV.2015.169.
29. Goodfellow, I., Bengio, Y., Courville A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org (accessed on October, 2021).
30. Haala, N., Rothermel, M. (2012). Dense multiple stereo matching of highly overlapping UAV imagery. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.
31. Hardin, P.J.; Hardin, T.J. (2010). Small-Scale Remotely Piloted Vehicles in Environmental Research: Remotely Piloted Vehicles in Environmental Research. Geogr. Compass. 4, 1297–1311.
32. He, K., Zhang, X., Ren, S., Sun, J. (2016) Deep Residual Learning for Image Recognition. Conference on Computer Vision and Pattern Recognition (CVPR), IEEE 778. doi: 10.1109/cvpr.2016.90.

33. Holmes, R. (2013). Falling upwards : how we took to the air. London: HarperPress. ISBN 978-0-00-738692-5.
34. Hornik, K., Tinchcombe, M., White, H. (1989). Multilayer Feedforward Networks are Universal Approximators. Neural Networks. 2. Pergamon Press. https://cognitivemedium.com/magic_paper/assets/Hornik.pdf (accessed on October, 2021).
35. House of Representatives 725 — 105th Congress: Precision Agriculture Research, Education, and Information Dissemination Act of 1997." https://www.govinfo.gov/content/pkg/BILLS-105hr725ih/pdf/BILLS-105hr725ih.pdf (accessed: October 2021).
36. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fiscer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7310-7311).
37. Hunt, E.R., Jr.; Cavigelli, M.; Daughtry, C.S.T.; Mcmurtrey, J.E., III; Walthall, C.L. (2005). Evaluation of Digital Photography from Model Aircraft for Remote Sensing of Crop Biomass and Nitrogen Status. Precis. Agric. 2005, 6, 359–378.
38. Hunt, E.R., Jr.; Hively, W.D.; Fujikawa, S.; Linden, D.; Daughtry, C.S.; McCarty, G. (2010). Acquisition of NIR-Green-Blue Digital Photographs from Unmanned Aircraft for Crop Monitoring. Remote Sens. 2, 290–305.
39. Jocher, G., Nishimura, K., Mineeva, T., Vilariño, R.: YOLOv5 (2020). Link: https://github.com/ultralytics/yolov5. (Accessed August, 2021).
40. Jones, G.P., IV; Pearlstine, L.G.; Percival, H.F. (2006). An Assessment of Small Unmanned Aerial Vehicles forWildlife Research. Wildl. Soc. Bull. 34, 750–758.
41. Junos, M. H., Khairuddin, A. S. M., Thannirmalai, S., & Dahari, M. (2021). Automatic detection of oil palm fruits from UAV images using an improved YOLO model. The Visual Computer, 1-15.
42. Kader, A. A. & Saltveit, M. E. (2003). Respiration and gas exchange. In: Bartz, J. A. & Brecht, J. K. (eds.), Postharvest physiology and pathology of vegetables, 2nd ed. New York, U.S.A.: Marcel Dekker Inc., pp. 7-29.
43. Kanellakis, C, Nikolakopoulos, G. (2017). Survey on computer vision for UAVs: current developments and trends. J Intell Robot Syst 87: 141–168.
44. Kicherer, A., Herzog, K., Bendel, N., Klück, H.-C., Backhaus, A., M. Wieland, J.C. Rose, L. Klingbeil, T. Läbe, C. Hohl, W. Petry, H. Kuhlmann, U. Seiffert, R. Töpfer. (2017) Phenoliner: a new field phenotyping platform for grapevine research. Sensors 17. doi: 10.3390/s17071625.
45. Kirkpatrick, K. (2019) Technologizing agriculture Communications of the ACM 62, pp. 14-16, doi: 10.1145/3297805
46. Koirala, A., Walsh, K.B., Wang, Z., Anderson, N. (2020). Deep learning for mango (Mangifera indica) panicle stage classification. Agronomy, 10(1), 143.
47. Kusumam, K., Krajník, T., Pearson, S., Cielniak, G., Duckett, T. (2016). Can you pick a broccoli? 3D-vision based detection and localisation of broccoli heads in the field. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 646-651, doi: 10.1109/IROS.2016.7759121.

48. Laliberte, A.S.; Rango, A.; Herrick, J. (2007). Unmanned Aerial Vehicles for Rangeland Mapping and Monitoring: A Comparison of Two Systems. In Proceedings of the ASPRS Annual Conference, Tampa, FL, USA, 7–11 May.
49. Le Louedec, J., Montes, H. A., Duckett, T., & Cielniak, G. (2020). Segmentation and detection from organised 3D point clouds: a case study in broccoli head detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 64-65).
50. LeCun, Y., Bengio, Y., Hinton, G. (2015) Deep learning. Nature 521. doi: 10.1038/nature14539.
51. LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10).
52. Liknes, G.C.; Perry, C.H.; Meneguzzo, D.M. (2010) Assessing tree cover in agricultural landscapes using high-resolution aerial imagery. Journal of Terrestrial Observation. 2(1): 38-55.
53. Lillesaeter, O. (1982). Spectral reflectance of partly transmitting leaves: laboratory measurements and mathematical modeling. Remote sensing of Environment, 12(3), 247-254.
54. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).
55. Liu, X., Chen, S. W., Aditya, S., Sivakumar, N., Dcunha, S., Qu, C., Camillo, J.T., Kumar, V. (2018, October). Robust fruit counting: Combining deep learning, tracking, and structure from motion. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1045-1052). IEEE.
56. Lucieer, A.; Malenovský, Z.; Veness, T.; Wallace, L. (2014). HyperUAS-Imaging Spectroscopy from a Multirotor Unmanned Aircraft System: HyperUAS-Imaging Spectroscopy from a Multirotor Unmanned. J. Field Robot 31, 571–590.
57. Madeleine, S., Bargoti, S., Underwood, J. (2016). Image based mango fruit detection, localisation and yield estimation using multiple view geometry. Sensors 16(11).
58. Matese, A., Toscano, P., Di Gennaro, S.F., Genesio, L., Vaccari, F.P., Primicerio, J., Belli, C., Zaldei, A., Bianconi, R., Gioli, B. (2015). Intercomparison of UAV, Aircraft and Satellite Remote Sensing Platforms for Precision Viticulture. Remote Sens. 7:2971–2990. doi: 10.3390/rs70302971.
59. Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (1983) Machine Learning - An Artificial Intelligence Approach. ISBN 978-3-662-12405-5.
60. Mogili, U.M.R.; Deepak, B.B.V.L. (2018). Review on Application of Drone Systems in Precision Agriculture. Procedia Comput. Sci. 133, 502–509.
61. Mutha, S.A., Shah, A.M., Ahmed, M.Z. (2021). Maturity Detection of Tomatoes Using Deep Learning. SN Computer Science, 2(6), 1-7.
62. Ni, W.; Sun, G.; Pang, Y.; Zhang, Z.; Liu, J.; Yang, A.; Wang, Y.; Zhang, D. (2018). Mapping Three-Dimensional Structures of Forest Canopy Using UAV Stereo Imagery: Evaluating Impacts of Forward Overlaps and Image Resolutions With LiDAR Data as Reference. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 3578–3589, doi:10.1109/JSTARS.2018.2867945.
63. Oetomo, D., Billingsley, J., & Reid, J. F. (2009). Agricultural robotics. Journal of Field Robotics, 26(6-7), 501-503.
64. Pádua, L.; Vanko, J.; Hruška, J.; Adão, T.; Sousa, J.J.; Peres, E.; Morais, R. (2017). UAS, Sensors, and Data Processing in Agroforestry: A Review towards Practical Applications. Int. J. Remote Sens. 2017, 38, 2349–2391.

65. Pajares, G. (2015). Overview and Current Status of Remote Sensing Applications Based on Unmanned Aerial Vehicles (UAVs). Photogramm. Eng. Remote Sens. 81, 281–330.
66. Psiroukis, V., Malounas, I., Mylonas, N., Grivakis, K. E., Fountas, S., & Hadjigeorgiou, I. (2021). Monitoring of free-range rabbits using aerial thermal imaging. Smart Agricultural Technology, 1, 100002.
67. Qihao Weng, (2012). An Introduction to Contemporary Remote Sensing. ISBN-13: 978-0071740111.
68. Qui, W., Shearer, S.A. (1992). Maturity assessment of broccoli using the discrete Fourier transform. Transactions of the ASAE. 35(6):2057-2062.
69. Ramirez, R. A. (2006). Computer vision based analysis of broccoli for application in a selective autonomous harvester (Doctoral dissertation, Virginia Tech).
70. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
71. Ren, S., He, K., Girshick, R., Sun J. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems 28, Microsoft Research.
72. Roser, M., 2019. Employment in agriculture. Our World in Data https://ourworldindata.org/employment-in-agriculture (accessed: October 2021)
73. Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C. (2016) DeepFruits: a fruit detection system using deep neural networks. Sensors, 16. doi: 10.3390/s16081222.
74. Santos, T.T., de Souza, L.L., dos Santos, A.A., Avila, S. (2020). Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association, Computers and Electronics in Agriculture 170. doi: 10.1016/j.compag.2020.105247.
75. Segarra, J., Buchaillot, M. L., Araus, J. L., & Kefauver, S. C. (2020). Remote sensing for precision agriculture: Sentinel-2 improved features and applications. Agronomy, 10(5), 641.
76. Shakhatreh, H.; Sawalmeh, A.H.; Al-Fuqaha, A.; Dou, Z.; Almaita, E.; Khalil, I.; Othman, N.S.; Khreishah, A.; Guizani, M. (2019) Unmanned Aerial Vehicles (UAVs): A Survey on Civil Applications and Key Research Challenges. IEEE Access 7, 48572– 48634.
77. Shearer, S.A., Burks, T.F., Jones P.T., Qui W. (1994). “One-dimensional image texture analysis for maturity assessment of broccoli.” Presented at the 1994 International Summer Meeting, ASAE Paper No. 94-3017. ASAE, 2950 Niles Rd., St. Joseph, MI 49085-9659 USA.
78. Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
79. Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In ICLR. arxiv.org/abs/1409.1556.
80. Soane, B.D., van Ouwerkerk, C. (1994). Developments in Agricultural Engineering Chapter 1 - Soil Compaction Problems in World Agriculture. ISSN 0167-4137, ISBN 9780444882868, Doi: 10.1016/B978-0-444-88286-8.50009-X.
81. Torres-Sánchez, J.; López-Granados, F.; Borra-Serrano, I.; Peña, J.M. (2018). Assessing UAV-collected image overlap influence on computation time and digital surface model accuracy in olive orchards. Precis. Agric. 19, 115–133, doi:10.1007/s11119-017-9502-0.

82. Tsouros, D.C, Bibi, S., Sarigiannidis, P.G. (2019). A Review on UAV-Based Applications for Precision Agriculture. Information 10, p.349. doi:10.3390/info10110349.
83. Tu, K., Ren, K., Pan, L., & Li, H. (2007). A study of broccoli grading system based on machine vision and neural networks. In 2007 International Conference on Mechatronics and Automation (pp. 2332-2336). IEEE.
84. Wilhoit, J.H., Koslav, M.B., Byler, R.K., Vaughan. D.H. (1990). Broccoli head sizing using image texture analysis. Transactions of the ASAE 33(5):1736-1740.
85. Xing, C., Wang, J., & Xu, Y. (2010). Overlap Analysis of the Images from Unmanned Aerial Vehicles. 2010 International Conference on Electrical and Control Engineering. doi:10.1109/icece.2010.360.
86. Zhang, C.; Kovacs, J.M. (2012). The Application of Small Unmanned Aerial Systems for Precision Agriculture: A Review. Precis. Agric. 13, 693–712.
87. Zheng, Y., Wu, B., Zhang, M. (2017) Estimating the above ground biomass of winter wheat using the Sentinel-2A data. J. Remote Sens. 21:318–328. doi: 10.11834/jrs.20176269.
88. Zhou, C., Hu, J., Xu, Z., Yue, J., Ye, H., & Yang, G. (2020). A monitoring system for the segmentation and grading of broccoli head based on deep learning and neural networks. Frontiers in plant science, 11, 402.
89. Zhou, N., Siegel, Z.D., Zarecor, S., Lee, N., Campbell, D.A., Andorf, C.M., et al. (2018). Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning. PLoS Comput Biol 14. doi: 10.1371/journal.pcbi.1006337.