POLITECNICO DI TORINO

THESIS

Master’s degree in Computer engineering specialized in Artificial Intelligence and Data Analytics

Automatic classification of healthy / diseased plants using multispectral images

Supervisor:

Prof. Morisio Maurizio

Candidate:

Antoine Wencel

Academic year 2024-2025

Summary

Introduction

Recent advancements in drone technology have significantly increased their applicability across various domains, including search and rescue missions, environmental monitoring, topographic mapping, and general-purpose videography. Within the agricultural sector, drone imagery has emerged as a particularly valuable tool for assessing crop health. However, as agricultural fields expand in scale, the task of monitoring the condition of individual plants becomes increasingly complex. Traditional plant inspection methods, which rely heavily on manual visual assessments and, in some cases, laboratory testing, are often impractical due to their time-consuming and costly nature.

This thesis builds upon the foundation laid by the Dronuts Project, which aims to develop both hardware and software solutions for plant monitoring through remote sensing. The primary objective is to evaluate the health status of individual plants (Hazelnuts trees in Piedmont, Italy) using multispectral images captured by drones. The work presented here focuses on applying a variety of software techniques to analyze drone imagery acquired at different time points, with the goal of autonomously identifying the unhealthy parts of hazelnut trees.

Chapter 1 introduces the concepts of multispectral and hyperspectral imaging, as well as the fundamentals of vegetation reflectance. This chapter explains how plants interact with electromagnetic radiation and how these interactions can be exploited using reflectance-based techniques. It also presents the most relevant Vegetation Indices (VIs) used throughout this study, detailing how they are computed and their role in assessing plant health.

Chapter 2 provides a review of the state of the art in unsupervised learning for vegetation classification and segmentation. It also includes a summary of previous approaches developed within the Dronuts project, which helped guide the development of the methods proposed in this work.

Chapter 3 details the methodology followed during the project, from the data preparation steps to the overall experimental design. It also describes the creation of the dataset used for training and evaluation, including how the multispectral images were processed and transformed into suitable inputs for deep learning models.

Chapter 4 begins with a brief theoretical background on Convolutional Neural Networks (CNNs) and Gaussian Mixture Models (GMMs), the two core components of the proposed approach. It then introduces the architecture implemented in this study, which combines a convolutional autoencoder with GMM-based clustering for semantic segmentation of vegetation.

Chapter 5 discusses the successive stages of the implementation. It starts with the initial approach and its limitations, then presents intermediate results and the improvements introduced to overcome earlier challenges. Finally, it concludes with the most recent results obtained, including a reflection on the model's strengths and weaknesses, and how expert feedback was integrated to refine the pipeline.

Finally, the conclusion summarizes the methodologies explored throughout the project, provides critical reflections, and proposes several strategies for future improvements.

Chapter 1

1.1 Multispectral and Hyperspectral images

Multispectral images are composed of several distinct layers, with each layer capturing data at a specific wavelength band. Multispectral sensors typically operate in the following spectral ranges:

- Blue: 450–520 nm
- Green: 520–600 nm
- Red: 600–690 nm
- Red-Edge (RE): 670–750 nm
- Near-Infrared (NIR): 750–900 nm

These images are widely used in remote sensing applications, enabling the extraction of information from surfaces by analyzing the electromagnetic radiation reflected or emitted, as captured by the sensors. Multispectral imagery can be acquired via various platforms, including drones, aircraft, and satellites, each offering different spatial and spectral resolutions depending on the application. This kind of images is the one used in the project.

In contrast, hyperspectral imaging captures an (almost) continuous spectrum for each pixel in the image. Unlike multispectral images, which are typically limited to 3 to 10 discrete bands, hyperspectral images comprise hundreds of narrowly spaced spectral bands, providing a more detailed and continuous spectral profile. The following figure summarizes the differences between the two:

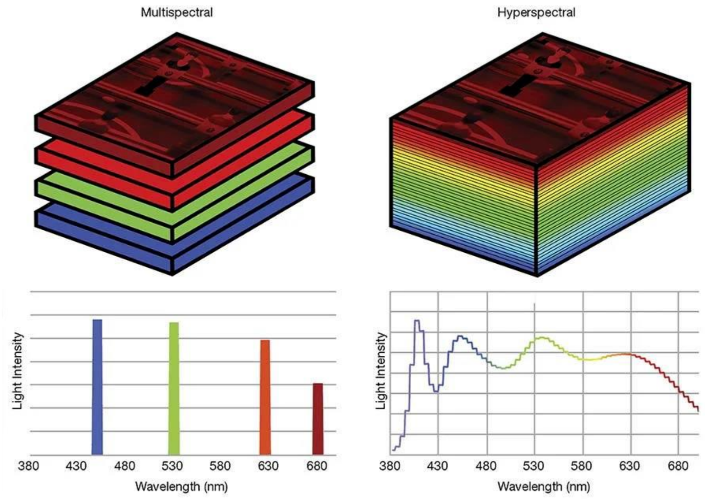

Figure 1.1: Comparison between Multispectral and Hyperspectral imaging

In agricultural applications, hyperspectral imagery can offer advantages over multispectral systems, notably in its ability to detect subtle spectral variations associated with plant diseases, nutrient deficiencies, and soil moisture levels, details often missed by multispectral sensors due to their coarser spectral resolution. However, this enhanced capability comes with drawbacks: hyperspectral data typically require significantly more storage space (ranging from 10 to 1000 times larger than multispectral data) and rely on more expensive and complex sensors.

In general, the remote sensing process involves three main stages:

1. Acquisition of electromagnetic radiation through specialized sensors,
2. Processing of the captured signals and their transformation into digital image data,
3. Interpretation and analysis of the visual information to extract meaningful insights.

1.2 Spectral Reflectance of Vegetation

In addition of the classical wavelength (RGB), the sensors of the drone used to take the pictures of the trees also have 2 additional wavelengths, namely Red-Edge (RE) and Near-Infrared (NIR) that are important when it comes to vegetation:

- The Red-Edge band is situated between the red and the near-infrared ones and in the case of vegetation, rapid changes are usually observed in that very band.
- The Near-Infrared one is above the red, it is very useful when it comes to the analysis of absorbed and reflected radiations by a surface such as trees, soils, etc.

Chlorophyll is a green pigment found in all green plants and plays a fundamental role in photosynthesis by absorbing light energy. It primarily absorbs light within the visible spectrum (400–700 nm). While chlorophyll is responsible for strong absorption in the blue and red portions of the spectrum, the cellular structure of plant leaves reflects a significant portion of light in the near-infrared (NIR) range (700–1100 nm). This characteristic reflectance makes chlorophyll a reliable indicator of a plant’s productivity. Moreover, chlorophyll levels can provide valuable insights into the plant's nutritional status, water stress, disease presence, and other physiological conditions.

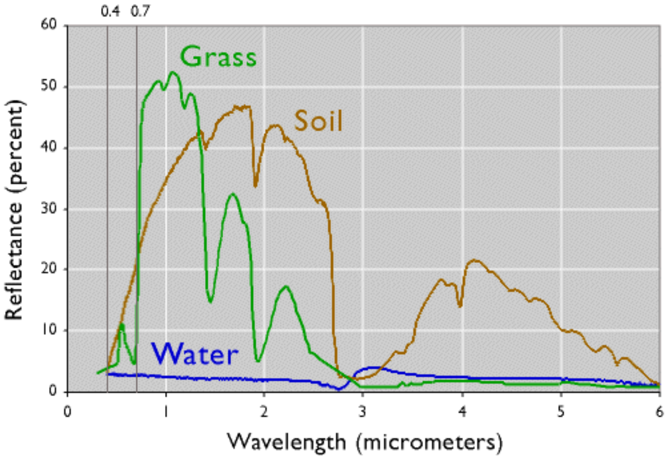

Figure 1.2: Spectral signatures of water, vegetation and soil

As illustrated in Figure 1.2, healthy vegetation typically exhibits low reflectance in the visible range, especially in the blue and red bands, which are most actively involved in photosynthesis. In contrast, reflectance significantly increases beyond the red wavelengths, peaking in the Red-Edge band. This phenomenon occurs because the internal structure of leaf cells has evolved to scatter light in the NIR region, where photon energy is insufficient to drive photosynthetic reactions. The portion of sunlight that plants use for photosynthesis is referred to as Photosynthetically Active Radiation (PAR). Consequently, in imagery captured within the PAR range, vegetation appears darker, whereas in the NIR spectrum, healthy vegetation appears much brighter due to its high reflectance.

Knowing the spectral reflectance profile of healthy vegetation allows for the identification of stressed or diseased plants, which generally exhibit lower reflectance in the NIR region and higher reflectance in the visible spectrum. This deviation from the typical signature helps in diagnosing plant health issues remotely.

In addition to distinguishing between healthy and unhealthy vegetation, spectral analysis can also be used to differentiate vegetation from soil and water. Water bodies, for example, show strong absorption in the infrared spectrum, resulting in very low reflectance values. Bare soil, on the other hand, tends to display a more uniform spectral curve, lacking distinct peaks or valleys across the measured wavelengths.

It is important to note, however, that the spectral signature of soil can vary significantly depending on several factors, including moisture content, texture, surface properties (e.g., rocky, sandy, or clayey composition), the presence of iron oxides, and the amount of organic matter.

1.3 Vegetative Indexes (VIs)

In the previous part, we highlighted the relevance of multispectral and hyperspectral imaging, which offer a wide range of electromagnetic bands essential for assessing the health status of crops. These spectral bands enable the extraction of key information from imagery, allowing for semantic differentiation between various elements in the scene, for example, distinguishing vegetation from soil, or healthy plants from diseased ones. To extract and interpret such spectral information, a set of mathematical tools known as Vegetation Indexes (VIs) is employed.

In essence, Vegetation Indices are mathematical combinations of surface reflectance values at two or more wavelengths, specifically designed to emphasize particular biophysical characteristics of vegetation. They are particularly effective in highlighting photosynthetic activity, making them widely used in remote sensing applications related to agriculture and crop monitoring. Most VIs leverage the inverse relationship between red and near-infrared (NIR) reflectance observed in healthy green vegetation, high NIR reflectance and low red reflectance signal vigorous plant activity.

A wide variety of Vegetation Indices have been proposed in the scientific literature, each tailored to specific features or conditions, and computed from different combinations of spectral bands available in multispectral or hyperspectral images. These indices serve diverse purposes, from detecting chlorophyll concentration to assessing plant water stress or distinguishing vegetation from non-vegetation surfaces.

For the purposes of this project, a selection of Vegetation Indexes was chosen to evaluate plant status based on the types of information most relevant to the available dataset. In order to select those, the results obtained by others students that worked on the same project in the past were re-used (see chapter 2 for more information):

The following were selected:

- Normalized Difference Vegetation Index (NDVI) is one of the most widely adopted vegetation indices in the field of remote sensing. It serves as a fundamental metric for assessing vegetation cover and health by evaluating the normalized contrast between reflectance in the near-infrared (NIR) spectrum, where healthy vegetation reflects strongly and the red spectrum, which is more heavily absorbed by plant chlorophyll. It is obtained via the following formula:

  $$NDVI = \frac{NIR - RED}{NIR + RED}$$

  The resulting values range from -1 to +1. Typically, positive values are indicative of vegetation, with values above 0.3 generally corresponding to dense, healthy vegetation such as crop fields or forested areas. In contrast, values near zero are often associated with bare soils, rocks, or urban surfaces, while negative values tend to signal the presence of water bodies or, in some cases, cloud cover (particularly in satellite imagery).

Furthermore, NDVI is sensitive to plant stress: a significant decrease in NDVI values, particularly below a certain threshold, may signal reduced NIR reflectance, a possible indicator of disease, water stress, or nutrient deficiency in the plant. Note that this indicator was not directly used in the deep learning algorithms, but to create tree masks from the original tree pictures as explained in chapter X.

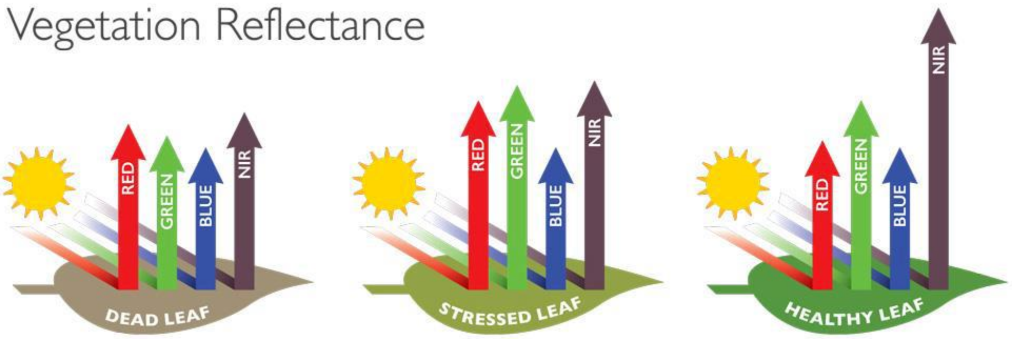

Figure 1.3: Vegetation spectral bands reflectance

For the following five used indexes, a picture of a tree with the considered index will be shown as an example (since the original tree come from the final dataset [discussed in chapter X], the considered grey level is already normalized), for all of them, the following RGB picture will be used:

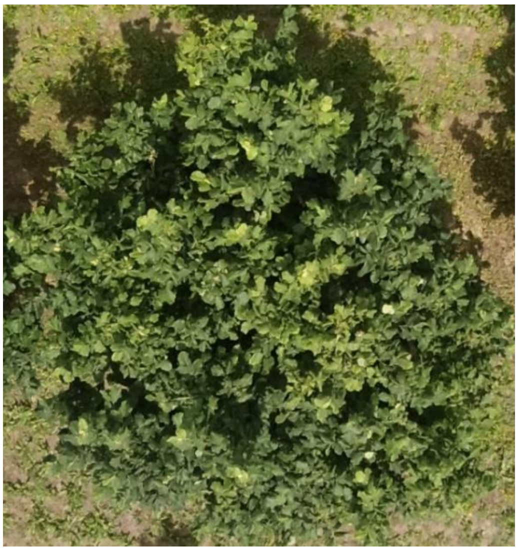

Figure 1.4: Original RGB picture (Carrù Field, plant n°100)

- Green Chlorophyll Index (GCI) is a vegetation index designed to estimate the chlorophyll content within plant canopies, which is a key indicator of crop health and photosynthetic activity. GCI leverages the relationship between reflectance in the near-infrared (NIR) and green spectral bands, two wavelengths particularly relevant for assessing vegetation vigor and greenness. The GCI is calculated using the following formula:

$$GCI = \frac{NIR}{GREEN} - 1$$

The index can take values ranging from -1 to positive infinity. Higher values typically correspond to greater chlorophyll concentrations, and therefore healthier and more productive vegetation. Because chlorophyll strongly influences the green reflectance and is closely related to plant nutrition and stress status, the GCI is especially valuable in precision agriculture applications focused on monitoring nutrient availability, plant growth, and early stress detection.

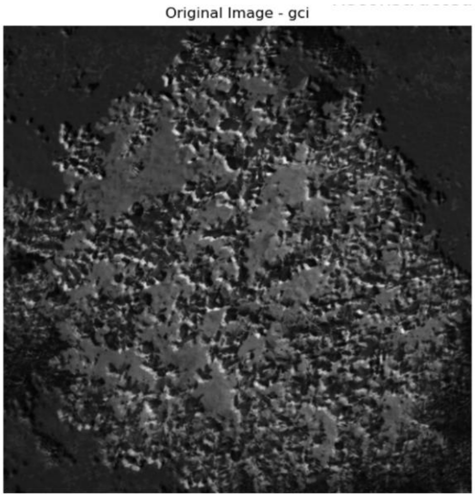

Figure 1.5: GCI grey level of the original tree

- Green Normalized Difference Vegetation Index (GNDVI) is a commonly used vegetation index designed to assess the photosynthetic activity of plants, with a particular emphasis on evaluating nitrogen content and water availability. Structurally similar to NDVI, GNDVI differs by using the green band instead of the red band in its formulation, making it more sensitive to chlorophyll concentration. GNDVI is computed as:

$$GNDVI = \frac{NIR - GREEN}{NIR + GREEN}$$

The resulting values range between -1 and +1, where higher values typically correspond to healthier vegetation. Due to its higher saturation threshold compared to NDVI, GNDVI is especially useful in later growth stages of crops and for dense vegetation canopies, where NDVI might already be saturated. While NDVI is more effective during the early stages of growth for estimating overall vigor, GNDVI provides a more nuanced evaluation of plant health in mature crops, particularly where chlorophyll variability is significant.

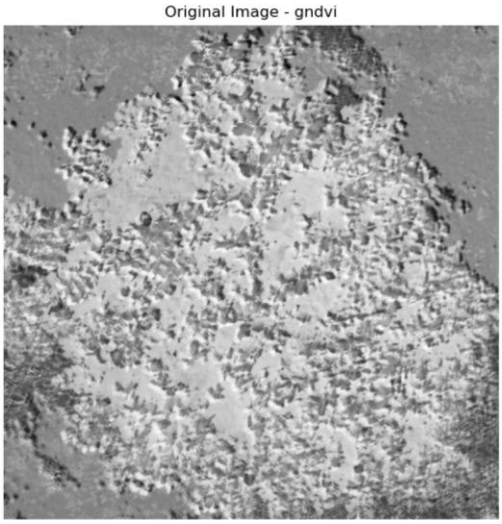

Figure 1.6: GNDVI grey level of the original tree

- Normalized Difference Red-Edge Index (NDREI) is a vegetation index widely employed to estimate chlorophyll concentration in plants, making it a valuable indicator of overall plant health and nutrient status. NDREI leverages reflectance measurements from the near-infrared (NIR) and the red-edge spectral band, a transition region between the red and NIR wavelengths that is particularly responsive to changes in chlorophyll content. The index is computed using the formula:

$$NDREI = \frac{NIR - RED\_EDGE}{NIR + RED\_EDGE}$$

NDREI is especially useful during the mid to late stages of the growing season, when vegetation is maturing and traditional indices like NDVI may begin to saturate. This makes NDREI particularly suitable for portions of the dataset considered in this project, where plant development has progressed beyond early vegetative stages. The use of the red-edge band enables NDREI to capture subtler variations in chlorophyll content, offering a more refined assessment of crop health in later phenological stages.

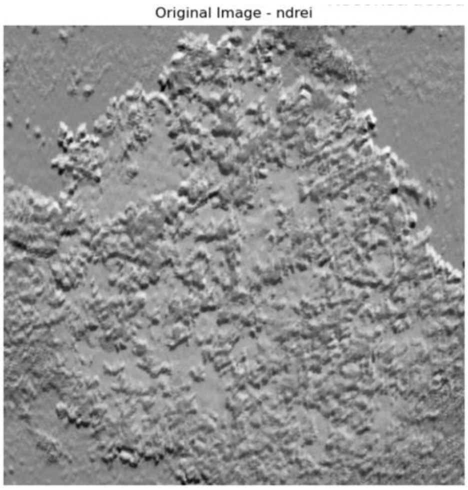

Figure 1.7: NDREI grey level of the original tree

- Nitrogen Reflectance Index (NRI) is a spectral index designed to assess the nitrogen content in vegetation. Nitrogen is a fundamental macronutrient that plays a critical role in plant growth and development, being an essential component of proteins, enzymes, and chlorophyll molecules. Adequate nitrogen levels are directly linked to optimal photosynthetic efficiency and crop productivity.
  The presence of nitrogen in chlorophyll ensures energy availability throughout the plant, contributing to high yields. Conversely, nitrogen deficiency may result in stunted growth, smaller leaves, and a reduction in chlorophyll content, often visible as pale green or yellow foliage.
  NRI detects such deficiencies by analyzing reflectance in the green and red spectral bands. It is calculated as follows:

$$NRI = \frac{GREEN - RED}{GREEN + RED}$$

  In this formulation, lower NRI values correspond to healthier vegetation, as healthy leaves reflect less light in the red spectrum due to active chlorophyll absorption. In contrast, higher NRI values may indicate chlorosis or other stress factors reducing photosynthetic activity.

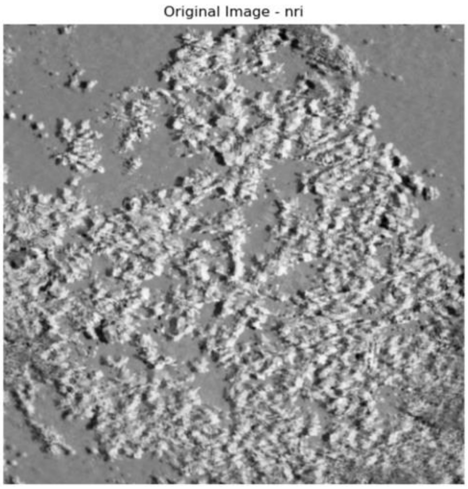

Figure 1.8: NRI grey level of the original tree

- Greenness Index (GI) is a vegetation index used to evaluate the level of greenness in plants, which serves as an indicator of chlorophyll content and, consequently, of plant health. A higher chlorophyll concentration results in stronger green reflectance, and thus, GI is inversely related to vegetation health: the lower the GI value, the healthier the plant. The GI is computed using the following formula:

$$GI = \frac{GREEN}{RED}$$

This ratio captures the contrast between the green and red spectral reflectance, leveraging the fact that healthy vegetation absorbs red light more effectively (for photosynthesis) while reflecting more green light.
Similar to the Nitrogen Reflectance Index (NRI), GI values are lower for healthy vegetation and tend to be higher in stressed or unhealthy plants. Figure 1.9 illustrates this phenomenon, where vegetation areas appear brighter than the surrounding ground, indicating lower GI values and confirming the chlorophyll-rich condition of the plants.

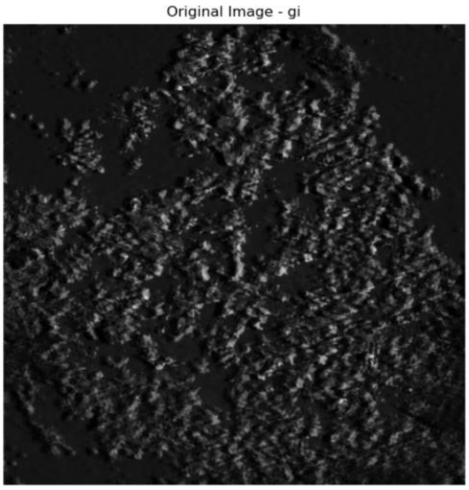

Figure 1.9: GI grey level of the original tree

Chapter 2

2.1 State of the Art

The application of deep learning in plant health monitoring and agricultural image analysis has been the focus of numerous studies in recent years. This section reviews a selection of relevant papers that illustrate key trends, limitations, and opportunities in the domain, particularly with regard to multispectral imagery, aerial data collection, and automated disease detection. Each paper brings complementary insights that have influenced the methodological choices of this project.

Deep Learning in Agriculture: A Survey [1]

This study presents an overview of deep learning applications in agriculture, highlighting the significant benefits of aerial imagery captured via drones. Such methods allow for non-destructive and systematic data collection over large areas. The review emphasizes the prevalence of multispectral imaging, alongside thermal and near-infrared imagery, for vegetation analysis. Convolutional Neural Networks (CNNs), including standard architectures like AlexNet, GoogleNet, and ResNet, are widely employed. The paper also underlines the importance of image preprocessing steps such as resizing, segmentation, background removal, and foreground pixel extraction. While deep learning reduces the need for handcrafted feature extraction, it necessitates larger datasets and data augmentation. Although slightly dated, this paper establishes a solid foundation for understanding the role of deep learning in agricultural image processing.

A Survey on Using Deep Learning Techniques for Plant Disease Diagnosis [2]

This survey investigates the performance of deep learning models in diagnosing plant diseases. It reveals a major limitation: models trained on laboratory images struggle to generalize to real-world conditions. In controlled datasets, deep learning outperforms human experts, achieving accuracy above 94%. However, in practical field conditions, performance significantly drops below 50%. The paper also highlights the lack of a universal standard for assessing disease severity through deep learning, pointing to a gap that hinders reproducibility and model evaluation across studies. These insights underline the importance of adapting models to real-world scenarios in this project.

A Review on the Main Challenges in Automatic Plant Disease Detection Based on Visible Range Images [3]

This review identifies the core challenges in automated disease detection from visible spectrum images. It categorizes the obstacles into two groups. Extrinsic factors include complex backgrounds, variations in lighting and viewing angles, sensor resolution, and reflectance issues. Intrinsic factors involve gradual transitions between healthy and diseased areas, multiple symptoms caused by a single pathogen, and symptom overlap between different diseases. The paper underscores the complexity of disease identification and the necessity for robust models capable of handling real-world variability, considerations that have been carefully integrated into our project pipeline.

Deep Feature-Based Plant Disease Identification Using ML Classification [4]

In this work, the authors propose a hybrid architecture that combines deep CNN feature extraction with traditional machine learning classifiers. This approach seeks to achieve high classification accuracy while minimizing the number of parameters and computational cost. The study shows that the results are highly dependent on dataset quality and diversity. It emphasizes the importance of having a wide range of images to capture various conditions and disease expressions. This hybrid methodology has directly inspired the design of our own classification architecture, as it aligns well with the project's goals of efficiency and generalization.

In addition to these core references, several more general papers concerning land cover classification and multispectral segmentation have also contributed to our understanding. While not specific to plant pathology, these works have informed broader methodological decisions, particularly regarding the handling of multispectral data and image segmentation. (A land cover classification method [5], Applying DL for improved NDVI reconstruction [6], Multispectral vineyard segmentation: a DL study [7])

2.2 Older works on the Dronuts Project

The article Characterization of hazelnut trees in open field through high-resolution UAV-based imagery and vegetation indices [8] proposes a method for monitoring the health status of hazelnut trees using multispectral imagery acquired by drones. The authors collected 4,112 high-resolution images (2 MP each) from 185 hazelnut trees located in two orchards in Italy. To allow a more granular analysis and reduce false negatives, each tree was segmented into nine sub-images. These sub-images were then visually labeled as “healthy” or “non-healthy” by expert agronomists.

Nine vegetation indices (VIs) were evaluated to discriminate between healthy and stressed vegetation. Among them, five indices: GNDVI, GCI, NDREI, NRI, and GI, demonstrated sufficient discriminative power, while the remaining indices (NDVI, SAVI, RECI, and TCARI) were found less effective. This limitation was primarily attributed to the shrubby architecture of hazelnut trees, which renders some conventional indices less informative.

To classify the vegetation health status, three supervised machine learning algorithms such as Random Forest, K-Nearest Neighbors, and Logistic Regression were applied, all yielding similar results. The overall classification accuracy reached approximately 65%, with a false negative rate of 13%, which the authors considered acceptable for practical applications in precision agriculture.

The study also identified several directions for future work, including expanding the dataset to enhance model robustness, incorporating deep learning methods for improved visual pattern recognition, transitioning to multi-class classification to identify specific stress causes (e.g., water stress or pathogen attack), integrating hyperspectral sensors for broader spectral analysis, and establishing regular UAV monitoring to support early and continuous detection.

The present project builds directly upon this study, particularly in the exploration of deep learning approaches, the implementation of data augmentation strategies, and the shift from binary to multi-class classification. These extensions aim to improve the precision and generalizability of vegetation health diagnostics in real-world agricultural contexts.

Chapter 3

3.1 Method Applied

Building upon the insights and limitations identified in the aforementioned study, the present project aims to design a custom deep learning-based pipeline tailored to the task of vegetation health classification from drone imagery. The central idea is to replace the conventional supervised classification step with a hybrid, semi-unsupervised approach that combines a deep autoencoder with a clustering algorithm. Specifically, a convolutional autoencoder is trained to project input images of trees into a compact latent space. The learned latent representations are then clustered using a Gaussian Mixture Model (GMM), which serves to assign each latent feature a label, thus producing a pixel-wise segmentation map of the original image with a fixed number of vegetation health classes.

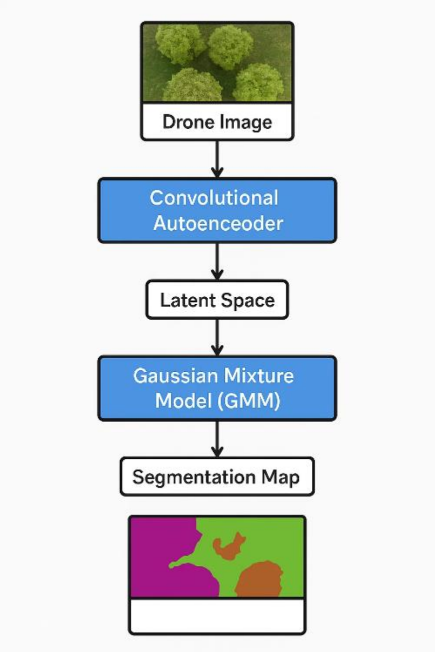

Figure 3.1: Schema of the considered architecture

This transition from classical machine learning to deep learning was motivated by a critical limitation observed in previous approaches. In earlier studies, each tree was subdivided into 1, 4, or 9 regions, and each region was assigned a single label (either healthy or unhealthy) based on expert visual assessment. However, this binary labeling at the region level introduced a major flaw: even if only a small part of a region was affected, the entire region was labeled as unhealthy. Furthermore, the associated input features were aggregated as vegetation index (VI) averages over the whole region, effectively diluting localized signals of plant stress. This averaging masked the presence of small but meaningful anomalies and imposed an upper limit on classification accuracy, which plateaued around 70%.

By leveraging pixel-level reconstruction and unsupervised clustering within a learned latent space, the current deep learning-based method aims to overcome this ceiling and capture finer-grained spatial variations in plant health.

To implement this architecture, whose technical specifications and training methodology are detailed in subsequent sections of this report, significant effort was made to adapt and reorganize fragments of legacy code developed by previous student cohorts. However, reusing this code posed substantial challenges, as it was not consistently documented or structured. Much of the code originated from a disjointed Google Colab notebook that had been split into smaller, isolated components at the end of the earlier project. This fragmentation, combined with limited inline documentation, complicated the integration process and necessitated a thorough reanalysis of the code logic and function dependencies.

The project was implemented using Python (version 3.13.2) within a WSL (Windows Subsystem for Linux) Ubuntu environment. The PyTorch and scikit-learn libraries were the primary frameworks used for deep learning and clustering tasks, respectively. The training and inference steps involving the autoencoder were accelerated using a dedicated NVIDIA GeForce GTX GPU. All the code should be available into the DropBox dedicated to the project. Specifically, an effort was made coding the auxiliary Python functions into dedicated files, with consequent comments in order to make the work easier for the possible future students on that very project. The Semantic Segmentation part was made under Jupyter Notebooks.

From a methodological standpoint, the use of GMM over alternative clustering algorithms such as K-Means is motivated by several factors. GMM, unlike K-Means, does not assume that clusters are spherical and equally sized. Instead, it models the data distribution as a mixture of Gaussian distributions with potentially different shapes, sizes, and orientations. This flexibility is particularly beneficial in the context of latent representations learned by deep networks, which may form complex, anisotropic distributions in feature space. Consequently, GMM is better suited to capture subtle variations in latent features that reflect nuanced differences in vegetation health.

3.2 Creation of the dataset

Although we had access to the original multispectral images of hazelnut trees, the annotated and segmented versions labeled by botanists during a previous phase of the project were unfortunately lost. The botanists involved no longer retained a copy of the data, and it was not possible to re-establish contact with the former students who contributed to the annotation. As a result, all preprocessing work, segmentation, labeling, and organization, had to be redone from scratch.

The dataset itself consists of 4112 multispectral images of 185 hazelnut trees, collected in open fields using a DJI drone equipped with a multispectral camera. These images were originally captured in two orchards in Italy (Carrù and Farigliano) with a resolution of 2 megapixels per image. The imagery includes several spectral bands: red, green, near-infrared (NIR), and red-edge, which are commonly used to compute vegetation indices.

Despite its richness, the dataset presents several challenges that can complicate the image processing pipeline:

- Multiple Trees per Image:

  In many cases, images contain more than one hazelnut tree, making it difficult to isolate the subject of interest. Overlapping branches and foliage further complicate the identification of individual tree boundaries.

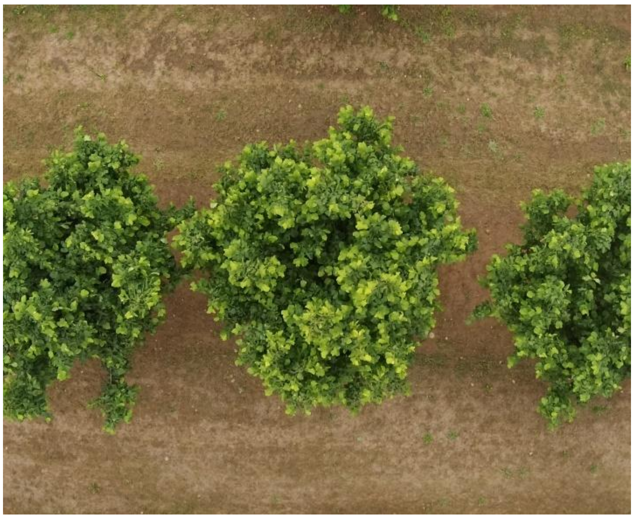

Figure 3.2: Example of image containing multiple trees (Carrù Field, plant n°19)

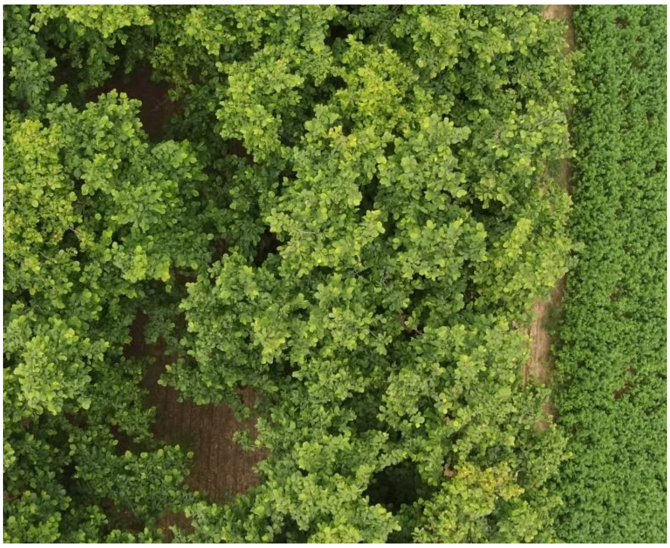

Figure 3.3: Example of overlapping trees (Farigliano Field, plant n°1)

- Variability in Lighting Conditions:

  The lighting conditions under which the images were taken differ considerably across sessions. Variations in sunlight, cloud cover, and sun angle result in inconsistencies in image brightness and shadow distribution across all spectral bands.

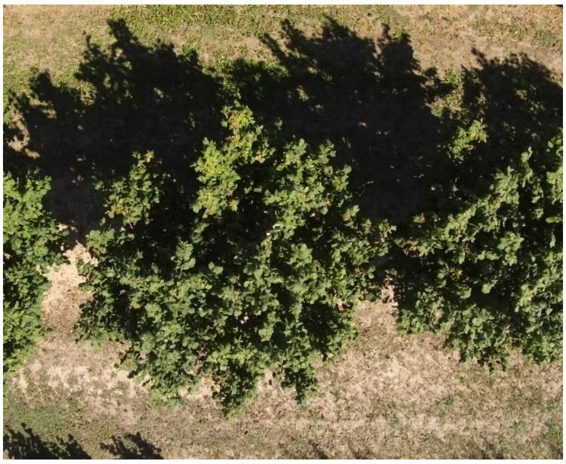

Figure 3.4: Carrù Field, plant n°19 captured under different lighting conditions

As shown, one image displays almost no shadow, while the other shows heavy contrast due to uneven sunlight, potentially causing information loss and unreliable feature extraction.

- Band Misalignment in Multispectral Images:

  Some images show misalignment between spectral bands due to slight drone movements or delays between consecutive captures. This misalignment leads to noticeable artifacts.

  These artifacts appear as low-value outlines along the contours of the tree, which do not correspond to real vegetative patterns. They are the result of spatial mismatches across the bands and must be addressed in preprocessing to avoid introducing bias in the analysis.

Despite these limitations, the dataset remains highly valuable for precision agriculture applications. However, the lack of reliable labels and the inconsistencies in the raw data necessitated a complete re-engineering of the preprocessing pipeline in order to enable unsupervised learning and feature extraction strategies explored later in this report.

In the early stages of this project, the objective was to reuse, with minimal modification, an algorithm developed by a former student (José Doumet Thesis [9]).This script aimed to automatically detect the contours of hazelnut trees directly from NDVI (Normalized Difference Vegetation Index) images. Its dual purpose was to crop the image around the tree of interest and to generate a binary tree/non-tree mask to assist with further analysis.

However, upon implementation, the results proved highly unsatisfactory. While the algorithm could occasionally perform on images containing multiple trees with limited overlap, the detected bounding boxes were frequently either excessively large, capturing a substantial portion of the background, or overly narrow, failing to encompass the full canopy of the tree.

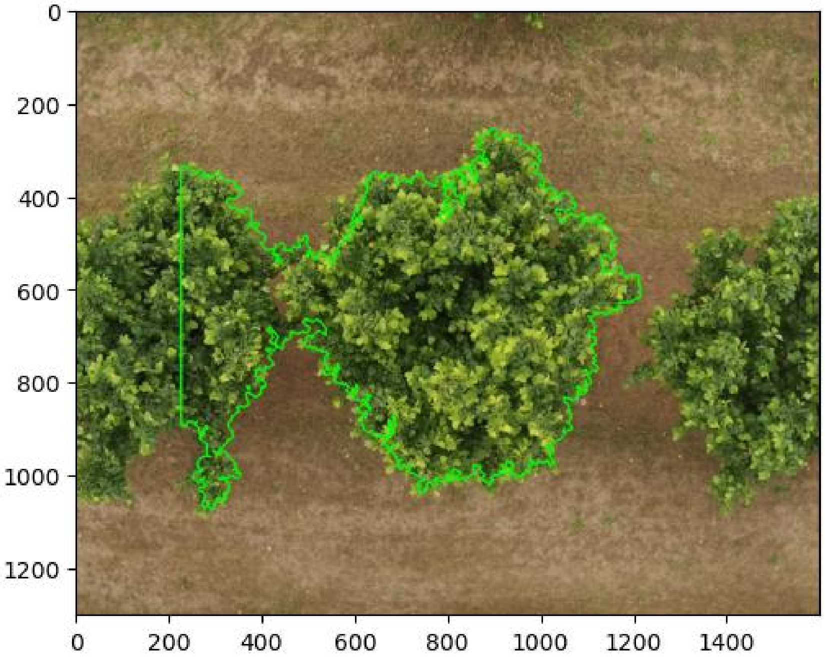

Figure 3.5: Bounding box too wide — overestimation of tree area (Carrù, plant n°19)

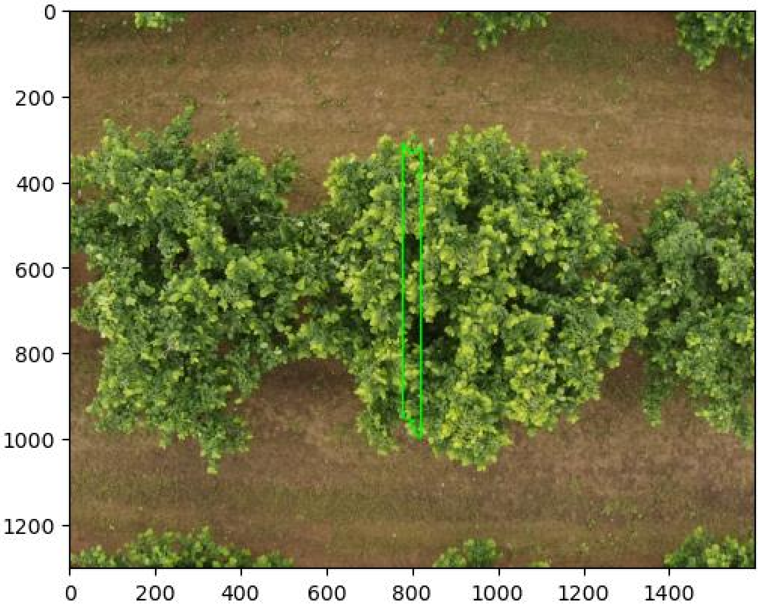

Figure 3.6: Bounding box too narrow — incomplete tree extraction (Carrù, plant n°23)

The situation worsened significantly for images where tree crowns and branches were heavily overlapping. In such cases, the algorithm failed almost entirely to isolate individual trees, resulting in unusable segmentation masks.

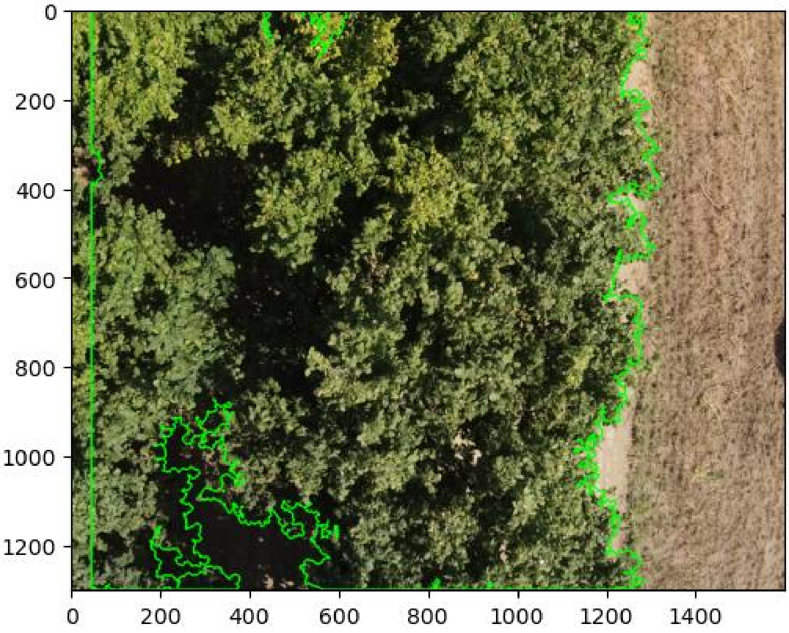

Figure 3.7: Failed segmentation on image with overlapping trees (Farigliano, plant n°1)

Due to the high failure rate, virtually 0% of the dataset yielded acceptable masks, a different strategy had to be adopted. A semi-automatic adaptation of the original code was implemented. In this revised version, the user manually defines the rectangular region enclosing the tree of interest by specifying its four edges. The NDVI filtering is then applied within this manually defined window to generate the binary mask.

Although this approach significantly improves the quality of the extracted masks, especially in challenging cases with multiple or overlapping trees, it suffers from a major drawback: manual intervention. Each image requires individual attention, with the bounding box coordinates input by hand. As a result, the process becomes time-consuming and labor-intensive, limiting the scalability of the pipeline.

Nevertheless, this method has enabled the creation of a subset of high-quality masks and tree-centered crops, which serve as the basis for the unsupervised learning experiments detailed in the subsequent sections. Here's some examples of trees contouring obtained via this method:

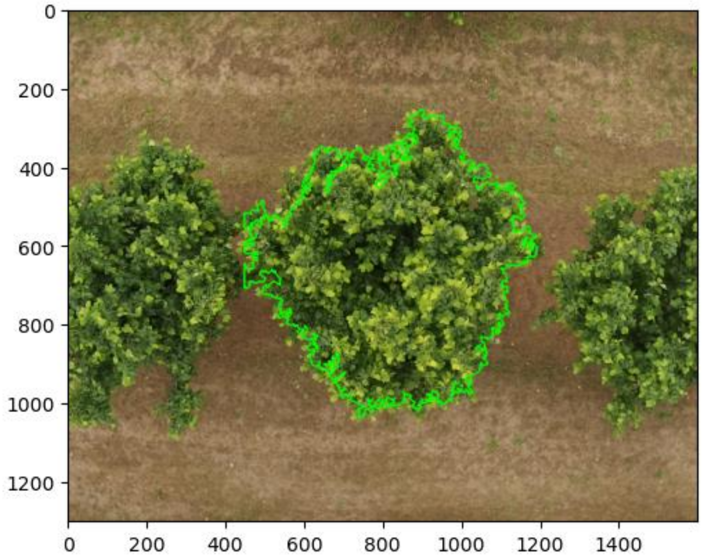

Figure 3.8: Bounding box perfectly shaped — (Carrù, plant n°19)

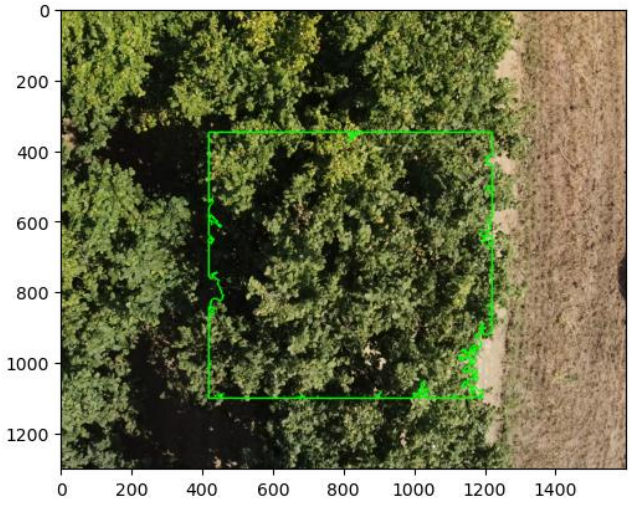

Figure 3.9: Segmentation on image with overlapping trees (Farigliano, plant n°1)

Via this method, we can obtain the following results for the VIs of each contoured tree:

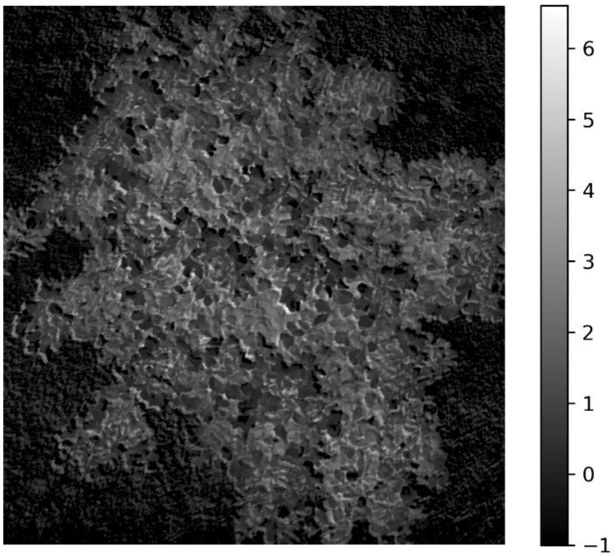

Figure 3.10: Example of a grayscale GCI obtained via this method (Carrù, plant n°97)

Finally, using the NDVI mask, we are able to manually set to -1 the VI value for each pixel outside the tree (there is no VI that can have a value lower than -1):

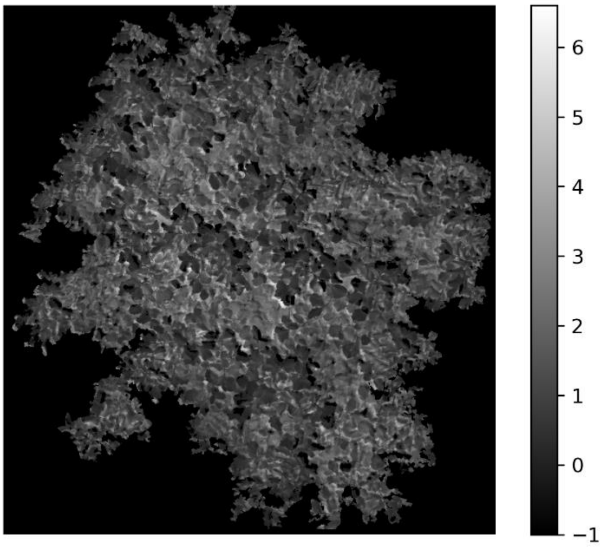

Figure 3.11: Result obtained after applying the binary mask (Carrù, plant n°97)

As we can see on the last picture, there are some Misalignment that cause for instance the presence of pixels that should be considered outside the tree in it (on the left part) and also some parts that are within the tree are considered outside. This remark is going to have some importance in later part of the project, as we will come back to later.

Finally, once all the necessary Vis are computed, they are stacked together into a .npy for each tree for later use.

Chapter 4

4.1 CNN and GMM reminders

In order to design a robust image-based classification pipeline for vegetation health analysis, this project leverages two core machine learning components: Convolutional Neural Networks (CNNs) and Gaussian Mixture Models (GMMs). This section provides an overview of their theoretical foundations and their relevance to the problem at hand.

Convolutional Neural Networks are a class of deep neural networks particularly effective for processing data with a grid-like topology, such as images. Unlike fully connected neural networks, which treat each input feature independently, CNNs are designed to exploit the spatial structure of data by applying learnable filters (also called kernels) that convolve across the input image.

Key components of a CNN architecture include:

- Convolutional layers: These apply multiple filters to the input image to detect local features such as edges, textures, and patterns.
- Activation functions: Typically, ReLU (Rectified Linear Unit), applied after each convolution to introduce non-linearity.
- Pooling layers: These reduce the spatial dimensions of the feature maps while preserving the most salient features, thereby increasing computational efficiency and robustness.
- Fully connected layers (in classification tasks): They aggregate high-level features and produce the final output predictions.

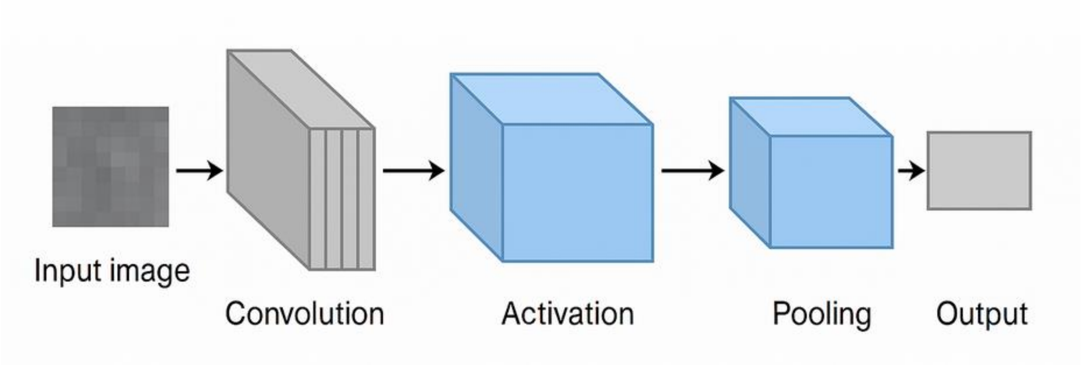

Figure 4.1: Basic schema of a CNN architecture

CNNs have been widely adopted in image classification, segmentation, and object detection tasks due to their ability to learn hierarchical representations directly from pixel data. In this project, they are used as the backbone of an autoencoder architecture.

An autoencoder is a type of neural network designed to learn efficient encodings of input data in an unsupervised manner. It consists of two main parts:

- Encoder: A neural network (typically convolutional when processing images) that maps the input data to a compressed latent space representation.

- Decoder: A symmetrical network that reconstructs the original input from the latent representation.

The network is trained to minimize the reconstruction loss, i.e., the difference between the original and reconstructed input. In doing so, the encoder learns a compressed and structured representation of the input, which can be exploited for tasks such as clustering or anomaly detection.

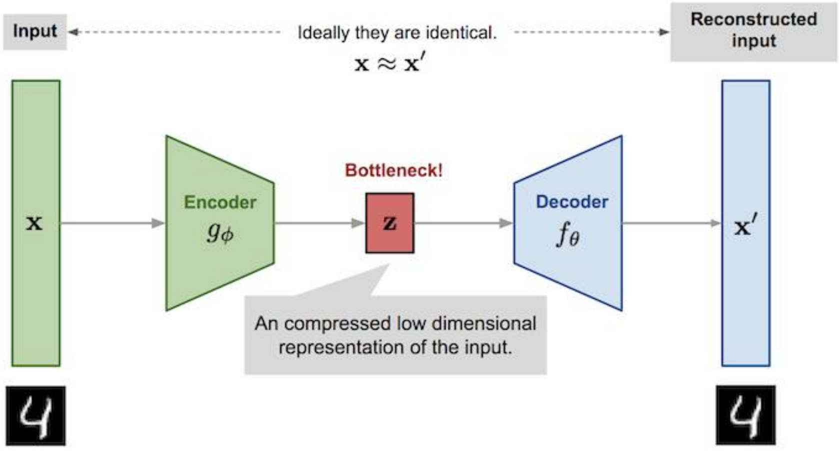
Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
$\mathbf{x}$
Encoder
g_ϕ
Bottleneck!
$\mathbf{z}$
Decoder
f_θ
$\mathbf{x}'$
An compressed low dimensional representation of the input.

Figure 4.2: Basic schema of an AutoEncoder architecture

In this project, the latent features produced by the encoder serve as input for clustering with a GMM, thereby enabling pixel-wise unsupervised classification of vegetation health.

A Gaussian Mixture Model is a probabilistic model that assumes that the data points are generated from a mixture of several Gaussian distributions with unknown parameters. Formally, a GMM is defined as a weighted sum of K Gaussian components:

$$P(x) = \sum_{k=1}^{K} \pi_k \, N(x|\mu_k, \Sigma_k)$$

where π_k are the mixture weights (summing to 1), and each component $N(x|\mu_k, \Sigma_k)$ is a multivariate normal distribution with its own mean μ_k and covariance matrix Σ_k.

GMMs are typically trained using the Expectation-Maximization (EM) algorithm, which iteratively estimates the parameters of each Gaussian component and the posterior probability of each point belonging to a given cluster.

Unlike k-means clustering, which partitions the space using hard assignments and assumes spherical clusters, GMMs provide soft assignments and can model more complex shapes and correlations in the data. This makes GMMs particularly well-suited for clustering in non-linear latent spaces such as those produced by autoencoders.

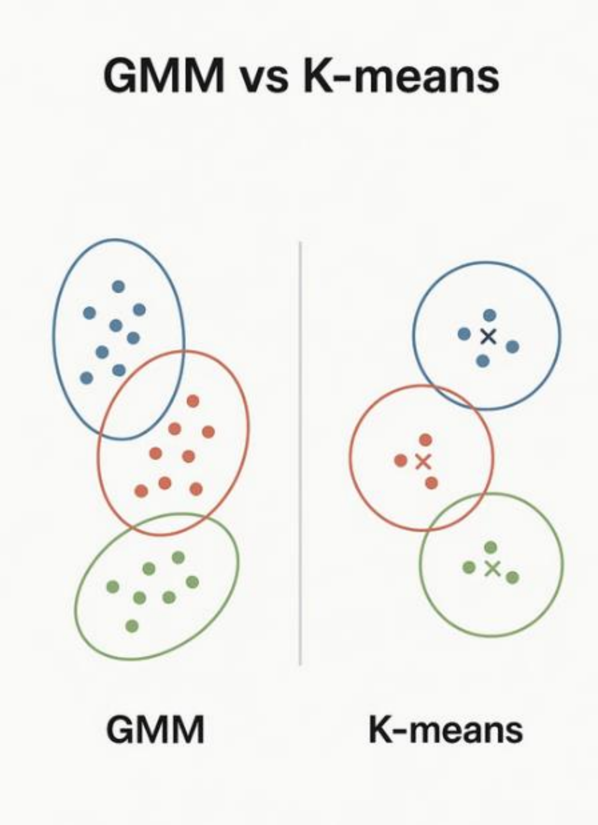

Figure 4.3: Schematic difference between GMM and K-means

4.2 Proposed Architecture: Convolutional Autoencoder + GMM Clustering

To overcome the limitations of traditional vegetation index-based classification, especially the loss of spatial information due to regional aggregation, this study adopts an unsupervised learning approach combining a convolutional autoencoder with a Gaussian Mixture Model. This architecture enables the extraction of compact and meaningful feature representations from multispectral image patches while preserving spatial granularity for fine segmentation.

Convolutional Autoencoder:

The autoencoder is designed to process 144×144-pixels (more on why this specific size on Chapter 5) patches with 5 spectral channels derived from vegetation indices. It is composed of two main parts:

The encoder progressively reduces spatial resolution while increasing feature depth, yielding a compact latent representation. The architecture consists of:

- Conv2D (input: 5 channels → 32 filters), stride 2, padding 1
- ReLU activation
- Conv2D (32 → 64), stride 2, padding 1
- ReLU activation
- Conv2D (64 → latent_dim), stride 2, padding 1
- ReLU activation

The input of shape (5, 144, 144) is encoded into a latent feature map of shape (latent_dim, 18, 18). Once again, for more details about why these size, cf. Chapter 5.

The decoder reconstructs the input patch from the latent representation using transposed convolutions:

- ConvTranspose2D (latent_dim → 64), stride 2, padding 1, output padding 1
- ReLU activation
- ConvTranspose2D (64 → 32), stride 2, padding 1, output padding 1
- ReLU activation
- ConvTranspose2D (32 → 5), stride 2, padding 1, output padding 1

The model is trained using the Mean Squared Error (MSE) loss function and the Adam optimizer. Hyperparameters such as learning rate and number of epochs are subject to optimization in future experiments.

Feature Extraction:

Once trained, the autoencoder is used to extract latent features from the patches, which serve as input for the clustering step. Two extraction modes are considered:

Flattened features: the latent (C, 18, 18) tensor is flattened into a 1D vector, useful for sample-wise clustering.

Spatial features: the full (C, 18, 18) tensor is preserved for pixel-wise clustering and segmentation.

Gaussian Mixture Model Clustering:

A Gaussian Mixture Model is trained on the set of latent features extracted from the full dataset. The GMM assumes that the latent space is composed of a mixture of Gaussian distributions, each representing a potential semantic class (e.g., healthy vegetation, stressed vegetation, non-vegetation).

The GMM is trained on the flattened latent features and later applied to predict a label for each latent pixel using the spatial feature maps. The predicted labels are then reassembled to form a segmentation mask corresponding to the original input image. Note that because the latent space dimension is smaller than the original image patch size, the segmentation mask has overall a smaller resolution than the original image. Thus, the clustering cannot be perfect no matter the result.

Experiments are conducted with different numbers of clusters (n = 3, 4, 5). Post-processing will include merging and reassigning clusters to semantically coherent categories, particularly to isolate stressed or diseased vegetation.

Chapter 5

5.1 First Implementation

The first implementation of the proposed pipeline was constrained by the limited computational power available, specifically the memory capacity of the GPU. As a result, a resizing strategy was initially adopted to reduce the input patch size to 128×128 pixels. Although this allowed the model to process the data without memory overflow, it introduced a major limitation at the output level: after three successive downsampling operations in the encoder (via strided convolutions), the latent representation of each patch was reduced to 8×8. Consequently, the segmentation map generated by clustering the latent space was extremely coarse and lacked the spatial resolution needed to draw any meaningful conclusions, especially in the context of fine-grained vegetation health assessment.

During this early phase, a parallel attempt was made to work directly with RGB imagery instead of the collection of Vegetation Indices (VIs). However, the RGB-based approach yielded poor results. RGB channels alone were insufficient to capture the subtle differences between healthy and stressed vegetation, which are more effectively captured by specific spectral indices.

To overcome the limitations introduced by aggressive resizing and insufficient input modalities, the next step involved a transition to a patch-based processing strategy. The goal was to maintain higher resolution in the encoded features while keeping the memory footprint within manageable bounds. Several attempts were made to define a custom dataset class that could extract same-sized patches from the original images while minimizing the need for resizing. However, this proved difficult due to the variability in original image sizes. Some Implementations overcome this issue but were too heavy to simply process for the data preparation phase (computing the mean, variance on the train dataset for instance)

A compromise was ultimately adopted: all input images were resized to a fixed size of 512×512 pixels. Crucially, this resizing was performed using nearest-neighbor interpolation rather than bilinear resampling, due to the nature of the input data. Pixels with a value of -1, corresponding to areas outside the tree canopy mask, do not represent real values and should not influence the surrounding data. Bilinear interpolation would have blended these placeholder values into valid vegetation regions, introducing artifacts and biasing the downstream model. In that first implementation, all the classicals function to compute data from the VI indexes had to be modified so that it takes into account the presence of -1 pixels that represents a different proportion of the tree data for each of them: for instance, considering them would have greatly influence the mean of the pixels value for the entire dataset and completely dwindle down the variance.

This preprocessing step allowed each image to be split into 16 non-overlapping patches of 128×128 pixels, each of which could be processed independently by the convolutional autoencoder architecture described in the following section. It also allows to artificially multiply by 16 the amount of data for the training process.

Regarding the train-test split, it was decided to do an 80-20 split while keeping all 16 patches of each tree and 3 tree pictures taken at the same date in the same part, to avoid having trees present in the train and the test, because it can cause overfitting issues. At first, it was an idea to have also a validation set to use between the training of the autoencoder of the one of the GMM, but because of the already small amount of data in the training set, the idea was abandoned.

5.2 Firsts results and improvements ideas

The initial results obtained on the test set (with no hyperparameter optimization and number of clusters set to *n* = 3) were encouraging. The clustering process appeared to successfully distinguish between tree-covered areas and non-tree regions. This validated the overall approach, as it showed that the model had learned meaningful representations for separating vegetation from background elements.

However, the clustering within the tree regions was noticeably less consistent. A recurring pattern involved one cluster being associated with the soil–tree border and another with the central part of the canopy. While this segmentation pattern was stable across examples, it did not necessarily reflect a biologically meaningful difference in vegetation health. This behavior is illustrated in the example below:

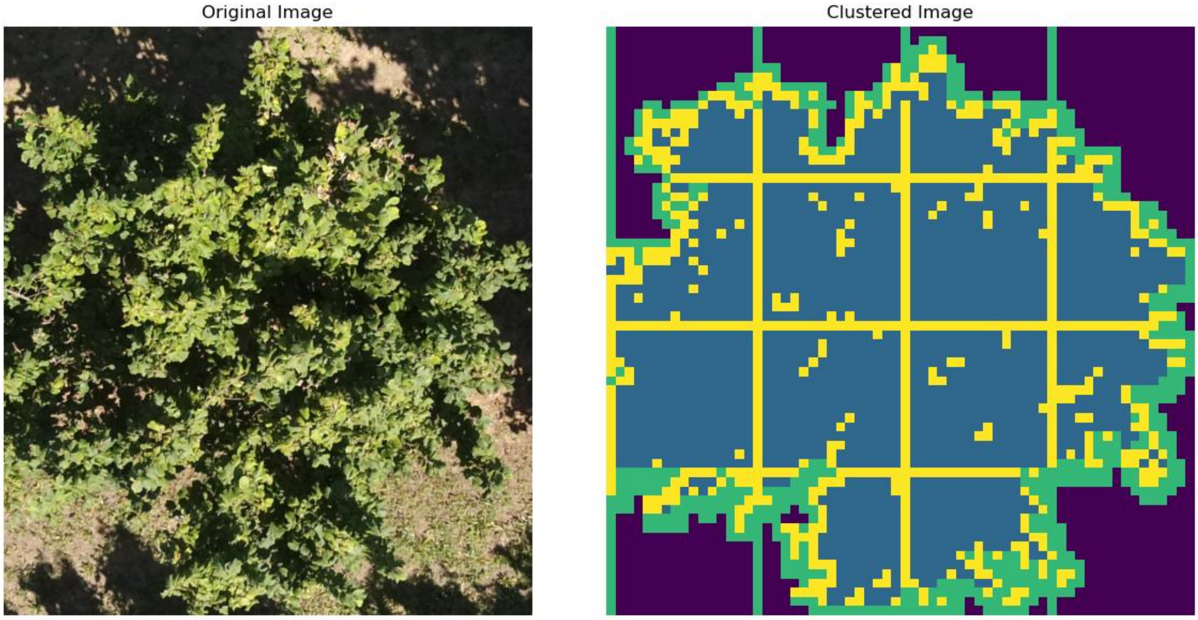

Figure 5.1: Example of a cluster map from a tree of the test dataset

A possible explanation for this inconsistency lies in the imperfect construction of the tree mask. Some pixels that should have been excluded from the analysis were instead included, and vice versa. As previously discussed, the mask might misclassify some non-tree pixels as part of the tree, and the abrupt transition between valid pixel values and those set to -1 outside the mask likely introduced irregularities in the convolutional layers. These sharp edges can interfere with the receptive field of the model and introduce artifacts near the borders of the canopy.

In the same image, a clear grid-like artifact can also be observed, corresponding to the 4×4 patch decomposition initially applied to process the full 512×512 images. These boundary effects seemed to originate from the autoencoder itself, as confirmed by comparing a raw vegetation index (e.g., GCI) to its reconstruction by the decoder:

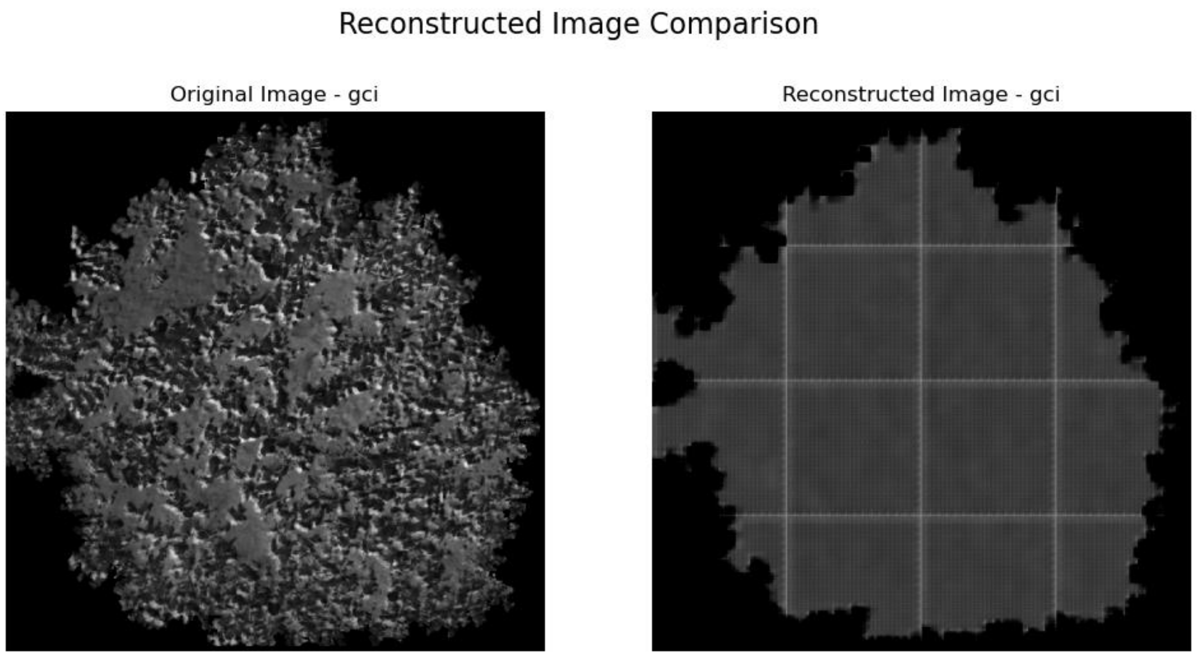

Figure 5.2: Comparison between the original GCI layer and the reconstructed one

Various strategies were explored to mitigate these border artifacts, but none proved fully satisfactory. As a result, the patch size was increased to 144×144 pixels (still within the GPU's memory constraints), with a slight overlap introduced between adjacent patches. Given the original 512×512 image dimensions, this overlap was enough to provide contextual information across patch borders without introducing significant redundancy or risking overfitting. During reconstruction, a small margin was removed from the edges of each patch to eliminate edge effects, before assembling the full image. The result was substantially improved: apart from the outermost top and left margins (where overlapping was not possible), the patch boundary artifacts were largely eliminated.

Following this improvement, another design decision was revisited: the use of the -1 mask for non-tree areas. While theoretically sound, the mask introduced additional complexity and computational overhead during data preprocessing. Moreover, because out-of-tree values were typically very distinct and less variable than those within the canopy, their influence on the clustering and subsequent segmentation was likely minimal. After conducting experiments with and without the mask, and observing comparable results, it was decided to drop the mask altogether. Though conceptually appealing, the masking strategy ultimately added more complexity than benefit in practice. If we just compare the results obtained from the decoder, it even seems that (still without training the hyperparameters) the version without the tree mask is better, as we can see on the following image:

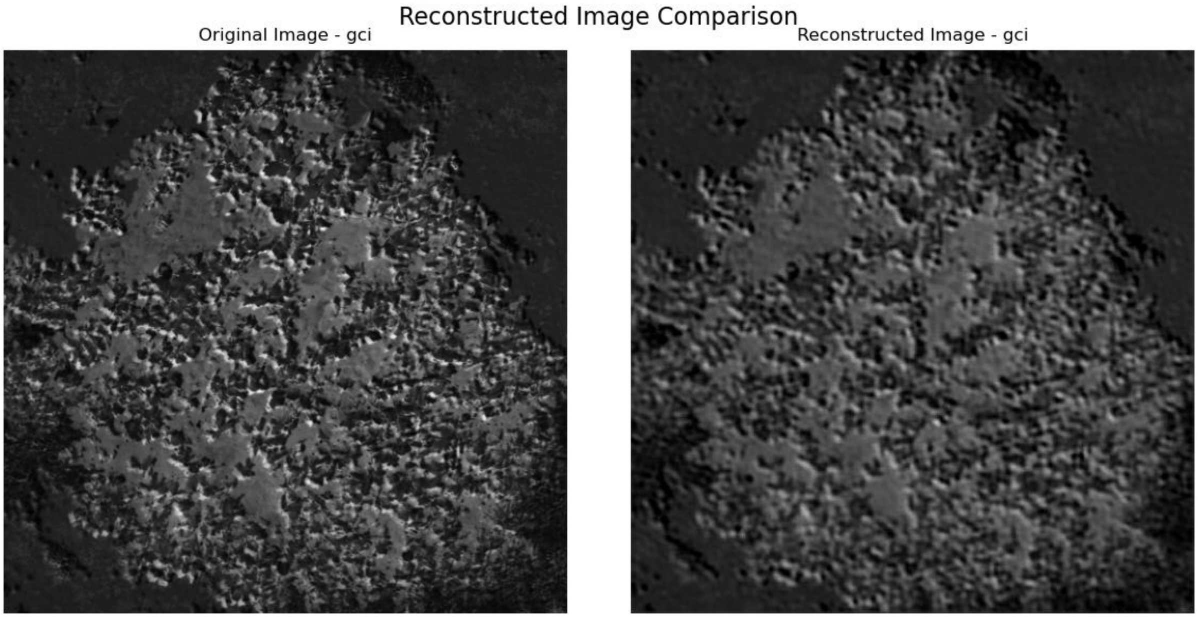

Figure 5.3: Comparison between the original GCI layer and the reconstructed one (without tree mask)

5.3 Final Results

After training an initial version of the semantic segmentation model on unmasked data, it was decided to present preliminary clustering maps to a group of botanists in order to gather expert feedback. As there was no existing ground truth for evaluating the clustering performance, their domain expertise was essential for qualitatively assessing the outputs. However, the feedback received was fairly critical: the clustering maps did not align well with their expectations in identifying diseased regions of the trees. Here’s an example of result obtain for a tree of the test set considering the 3 different dates:

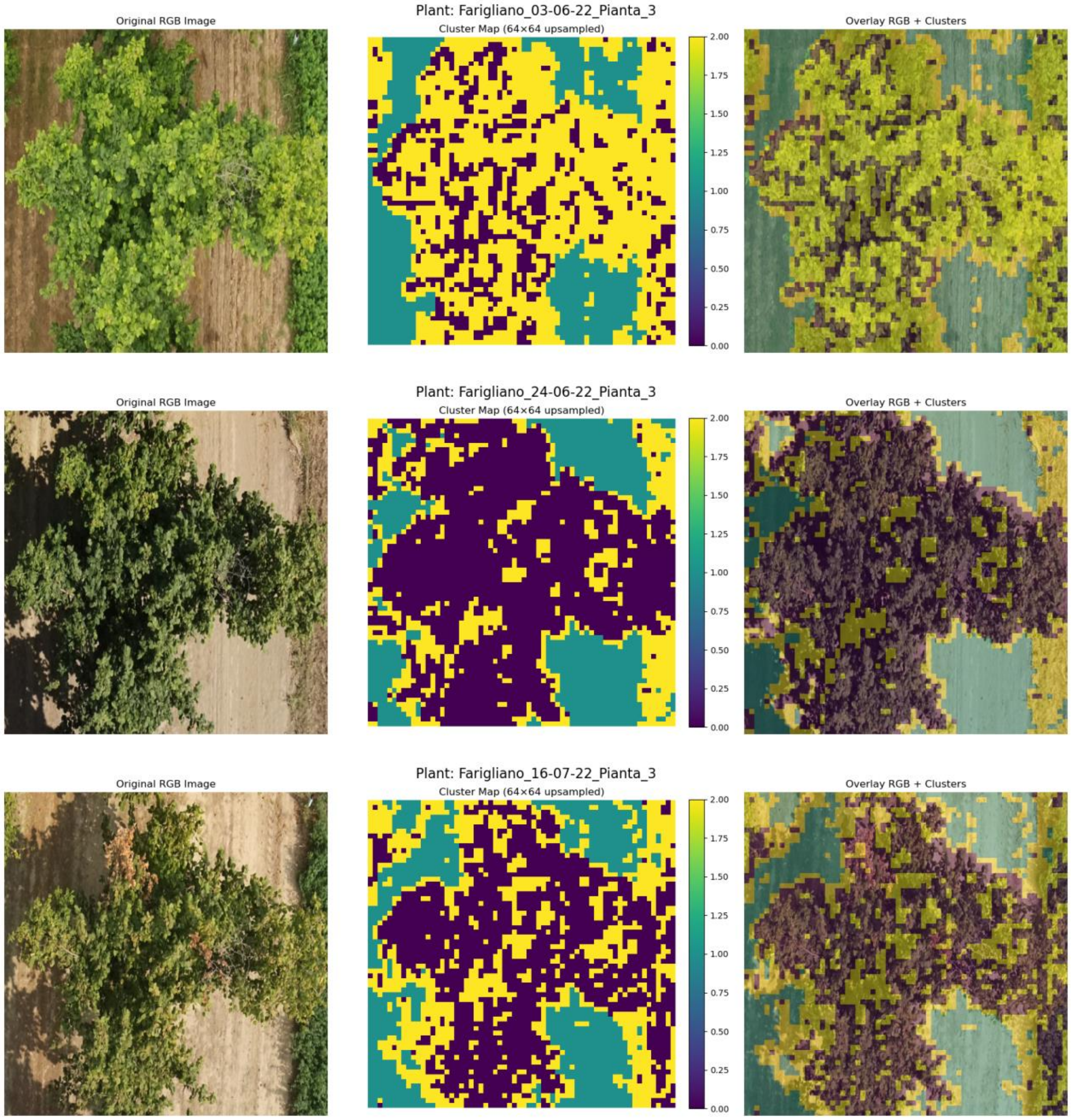

Figure 5.4: Different clustering obtained for a same tree (Farigliano, tree n°3) at different dates

As we can see, depending on the date, the dominant color within the tree changes a lot, and looking at the RGB part, the brown area and the one with less dense foliage on the right (the unhealthy ones) doesn't seem to be colored accordingly.

In response, the botanists manually annotated visibly diseased areas on a subset of approximately ten trees, providing a small but valuable ground truth dataset. These annotations were drawn directly on top of the multispectral images, highlighting regions that should correspond to symptoms of disease. An example annotation is shown below:

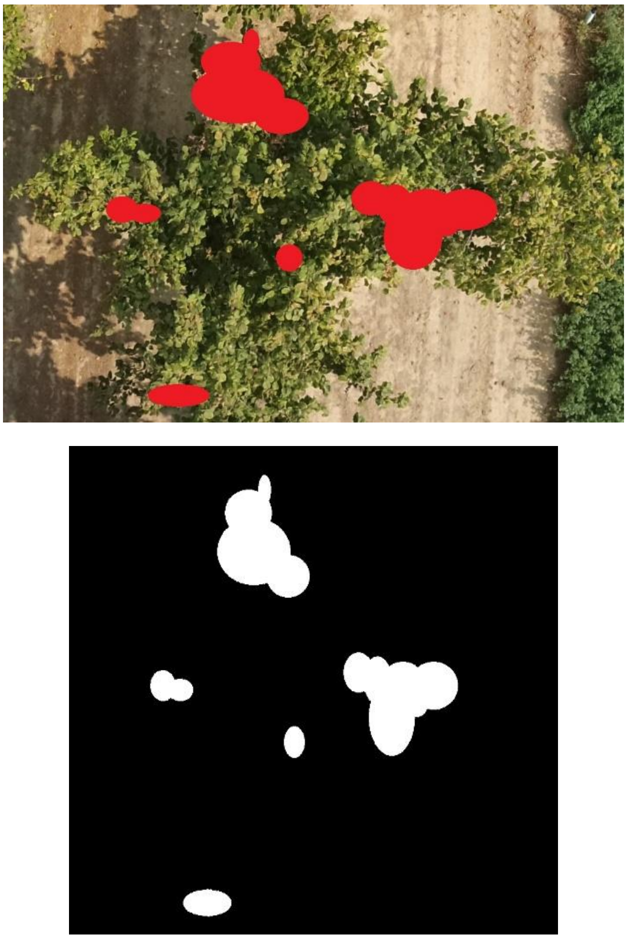

Figure 5.5: Annotations and mask obtained for the tree shown just above

From these expert annotations, binary masks could be generated by isolating the red regions in the images. These masks then allowed the definition of a quantitative metric for evaluation: the Intersection over Union (IoU):

- IoU is a common metric used to evaluate the accuracy of image segmentation algorithms. It is defined as the area of overlap between the predicted segmentation and the ground truth, divided by the area of their union:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

This score ranges from 0 to 1, where 1 indicates perfect alignment between the prediction and the ground truth.

Despite the limited number of annotated trees and the fact that some of them showed no visible signs of disease, the introduction of this evaluation method enabled a more objective assessment of the clustering maps. It also provided guidance for refining the GMM-based clustering, especially in tuning the number of clusters *n* and selecting better hyperparameters.

Furthermore, it became apparent that diseased regions could manifest in different spectral patterns, depending on the type or severity of the disease. Therefore, it was more appropriate not to constrain the model to a fixed number of clusters explicitly representing "healthy," "diseased," and "background." Instead, a more flexible approach was adopted: using a larger number of clusters (e.g., *n* = 4, 5 or 6) and grouping several of them during evaluation to compute the IoU against the manually annotated diseased areas. This aggregation allowed for more nuanced clustering results, accommodating the variability in disease expression while still supporting quantitative performance analysis. After optimizing the different hyperparameters of the 2 algorithms, we obtained the following results:

Best hyperparameters: {'lr': 0.001, 'epochs': 60, 'n_clusters': 6}

With these hyperparameters, we obtain an average IoU of *0.0737* on the few masks that we have access too. While this result is very low, if we compare it to a naïve IoU (that consider all the image as unhealthy) which gives an average IoU of *0.042*, we have a result *75%* better than the naïve one. For instance, the older work (that had a different metric) had a final accuracy of *67%*, which once compared to the naïve classifier that had a *57%* accuracy on the test set only represent a *18%* improvement (Note that while these 2 approaches are not really comparable, our was more complex from beginning)

Conclusion and possible improvements

This project aimed to explore unsupervised methods for the semantic segmentation of trees in aerial imagery, with a particular focus on distinguishing healthy vegetation from potentially diseased areas. In the absence of extensive annotated datasets, the objective was to leverage self-supervised learning and clustering techniques to build a segmentation pipeline capable of detecting relevant patterns using multispectral vegetation indices.

The dataset used in this work originated from drone imagery collected over different time periods and environmental conditions. Each image was processed to compute several vegetation indices (VIs), including NDVI, GI, and NRI, among others. These indices were then spatially aligned and structured into normalized 5-channel arrays, forming the input to the learning pipeline.

The architecture developed consisted of a convolutional autoencoder trained on patches extracted from preprocessed images. After several iterations, the final configuration used 144×144 patches with slight overlaps, reducing boundary artifacts during reconstruction. The autoencoder was able to efficiently compress and reconstruct the vegetation index data, producing a robust latent representation of tree structures.

Clustering was then applied to these latent representations using a Gaussian Mixture Model (GMM) to assign each pixel to a semantic category. While the autoencoder itself performed well, capturing essential structural and spectral features of the trees, the clustering results were more inconsistent. In particular, the segmentation of diseased regions was affected by strong variations in lighting, shadow, and contrast between image acquisition periods. These environmental differences introduced significant noise, limiting the effectiveness of a unified clustering strategy.

One possible solution considered was to train separate models for each acquisition period. However, preliminary results from training period-specific autoencoders were not convincing enough to justify this shift. More importantly, such an approach would have conflicted with the core objective of the project: to develop a generalized, all-in-one segmentation method capable of handling data across different conditions without the need for fine-tuning on a per-epoch basis.

In conclusion, while the autoencoder architecture showed strong potential for learning relevant features from multispectral data in an unsupervised way, the clustering stage remains sensitive to environmental variability. Future work could involve more advanced considerations on how to solve the light/shadow issues on the aerial views that compose the dataset.

Bibliography

[1] Deep learning in agriculture: A survey

Andreas Kamilaris, Francesc X. Prenafeta-Boldú

[2] A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools

Aanis Ahmad, Dharmendra Saraswat, Aly El Gamal

[3] A review on the main challenges in automatic plant disease identification based on visible range images

Jayme Garcia Arnal Barbedo

[4] Deep feature-based plant disease identification using machine learning classifier

Sk Mahmudul Hassan, Arnab Kumar Maji

[5] A Land Cover Classification Method for High-Resolution Remote Sensing Images Based on NDVI Deep Learning Fusion Network

Jingzheng Zhao, Liyuan Wang, Hui Yang, Penghai Wu, Biao Wan, Chengrong Pan and Yanlan Wu

[6] Applying deep-learning enhanced fusion methods for improved NDVI reconstruction and long-term vegetation cover study: A case of the Danjiang River Basin

Shidong Wang, Dunyue Cui, Lu Wang, JinYan Peng

[7] Multispectral vineyard segmentation: A deep learning comparison study

T. Barros, P. Conde, G. Gonçalves, C. Premebida, M. Monteiro, C.S.S. Ferreira, U. J. Nunes

[8] Characterization of hazelnut trees in open field through high-resolution UAV-based imagery and vegetation indices

Maurizio Morisio, Emanuela Noris, Chiara Pagliarani, Stefano Pavone, Amedeo Moine, José Doumet and Luca Ardito

[9] Automatic classification of healthy / diseased plants using multispectral images

José Doumet