



Estimating Leaf Chlorophyll Content of Crops Using Multispectral Remote Sensing and Machine Learning Models

A research report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science (Geographical Information Systems and Remote Sensing) at the School of Geography, Archaeology & Environmental Studies.

Mulivhuweni P. Mphaphuli (1209197)

Supervisor:

Prof. Elhadi Adam

July 2024

DECLARATION

I, Mulivhuweni P. Mphaphuli, solemnly affirm that this research report titled "Estimating Leaf Chlorophyll Content of Crops Using Multispectral Remote Sensing and Machine Learning Models" marks the culmination of my Master of Science degree in the Department of Geographical Information Systems (GIS) and Remote Sensing, within the School of Geography, Archaeology, and Environmental Studies at the University of the Witwatersrand. I declare that this work has not been previously submitted for any degree at this or any other educational institution. I affirm that all content presented in this report is my own work, and I have duly acknowledged all reference materials used herein.

Student: Mulivhuweni P. Mphaphuli

Signature: 

Date: 28-08-2024

Supervisor: Prof. Elhadi Adam

Signature:

Date:

DEDICATION

To my dearest mother, Ellen

And

My beloved brother, Ntanganedzeni,

Who ceaselessly encouraged and strengthened me at times of stress and difficulties.

You've been my biggest fans since day 1.

Always & Forever.

ACKNOWLEDGEMENTS

I express profound appreciation to my Lord and Saviour, Jesus Christ, without whom this paper would not have come to fruition.

I also sincerely appreciate my supervisor, Prof. Elhadi Adam, for his unwavering support and guidance. His patience and immense assistance were instrumental in the success of this study, and I am truly thankful.

My dear mother and brother deserve special thanks for their unwavering love, steadfast support, and understanding throughout this study. Their encouragement has been my fuel through late nights and challenging days. This study bears my name but also the indelible imprint of their sacrifices and belief in me.

To my extended circle of family and friends, especially Sandra, thank you for being my community. Your support has meant the world to me, and I genuinely appreciate each of you.

I express my gratitude to Eugene for consistently offering a helping hand whenever it was needed.

I extend sincere appreciation and gratitude to all those who have assisted in various capacities, whether through significant contributions or small gestures. Your support has profoundly influenced my journey, and I am truly thankful.

ABSTRACT

Tomatoes (*Solanum lycopersicum* L.) are the second most cultivated vegetable in the world, contributing significantly to global production and offering a wide range of health advantages and nutrients. Precise and timely measurement of leaf chlorophyll content (LCC) on a small-scale spatial level is paramount for monitoring a plant's photosynthetic ability and nutritional health. Hence, this study used LCC to monitor the dynamic changes in tomato crops using various machine learning (ML) models.

This study aimed to estimate LCC in tomato crops by leveraging ML models, specifically DTR, RFR, PLSR, and XGB. The models were developed in two phases: Before-Feature-Selection (BFS) and After-Feature-Selection (AFS). In BFS, twenty VIs were used as predictors, while in AFS, the three most relevant VIs for LCC estimation were determined through Pearson's correlation coefficient and the Gini index. Noteworthy findings identified GNDVI, GRVI, and GCI as pivotal VIs, demonstrating a moderate and positive correlation with LCC.

The evaluation and assessment of the machine learning models were conducted utilizing leave-one-out cross-validation (LOOCV) and statistical metrics such as R^2 , MAE, and RMSE. The XGB model revealed the highest R^2 of 0.53 and 0.54, RMSE values of $9.54\mu\text{g}/\text{cm}^2$ and $9.36\mu\text{g}/\text{cm}^2$ and MAE values of $6.39\mu\text{g}/\text{cm}^2$ and $6.04\mu\text{g}/\text{cm}^2$ in both phases. In contrast, the PLSR exhibited the lowest R^2 , with the highest RMSE values of $11.94\mu\text{g}/\text{cm}^2$ and $11.86\mu\text{g}/\text{cm}^2$ and MAE values of $8.79\mu\text{g}/\text{cm}^2$ and $8.55\mu\text{g}/\text{cm}^2$ in both phases. The results revealed that the AFS models demonstrated higher R^2 than the BFS models. The study also generated chlorophyll maps using the ML models, capturing the spatial dynamics of leaf chlorophyll within the tomato field. Therefore, this study highlights the potential of multispectral UAVs combined with ML for LCC estimation, offering valuable insights for decision-making and management in small-scale farms.

TABLE OF CONTENTS

DECLARATION.....	i
DEDICATION	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT.....	iv
LIST OF FIGURES.....	viii
LIST OF TABLES	ix
ACRONYMS & ABBREVIATIONS	x
CHAPTER 1.....	1
INTRODUCTION	1
1.1. Background to the study.....	1
1.2. Problem Statement	3
1.3. AIMS AND OBJECTIVES	3
1.3.1. Aim	3
1.3.2. Research Objectives.....	3
1.4. Significance of the study	3
1.5. Structure of the research report.....	4
CHAPTER 2.....	6
LITERATURE REVIEW	6
2.1. Agriculture in the 4IR.....	6
2.2. The use of UAVs data for Precision Agriculture	7
2.2.1. Imaging Sensor systems.....	8
2.2.1.1. MultiSpectral (MS) Imaging sensors.....	9
2.2.1.2. Hyperspectral (HS) Imaging sensors.....	10
2.3. Agricultural Applications using Satellite Imagery data.....	11
2.4. Machine Learning.....	12
2.4.1. Brief overview	12
2.4.2. Machine Learning Applications in Precision Agriculture	13
2.4.2.1. Crop Yield Prediction.....	14

2.4.2.2. Disease and Weed Detection	15
2.4.2.3. Crop Recognition	16
2.4.2.4. Crop/ Species Quality.....	17
2.5. Machine Learning Modelling in Agriculture	18
2.5.1. Machine Learning modelling with VIs.....	18
2.6. Challenges of ML in Agriculture	19
CHAPTER 3.....	21
METHODOLOGY.....	21
3.1. Study Area	21
3.1.1. Geographic Landscape and Features	21
3.1.2. Soil	21
3.1.2. Climate	23
3.2. Data Acquisition	23
3.2.1. Remote Sensing Data.....	24
3.3.2. Leaf Chlorophyll measurements	25
3.4. Image pre-processing	26
3.5. Vegetation Indices (VIs) calculations.....	26
3.6. Feature Selection	29
3.6.1. Feature importance and selection.....	29
3.7. Modelling Methods	29
3.7.1. Partial-Least Squares Regression (PLSR).....	30
3.7.2. Decision Trees regression (DTR).....	30
3.7.3. Random Forest regression (RFR).....	31
3.7.4. eXtreme Gradient Boosting (XGB).....	31
3.8. Cross Validation	32
3.9. Evaluation of Models.....	32
3.10. Summary.....	33
CHAPTER 4.....	34
RESULTS AND DISCUSSION	34

4.1. Descriptive statistics of the measured Leaf Chlorophyll Content (LCC)	34
4.2. Correlation Analysis between Vegetation Indices and chlorophyll content	35
4.3. Feature Selection/ Importance	37
4.4. Estimation of Tomato Chlorophyll Content by using ML models.....	38
4.5. Mapping the Spatial Distribution of Tomato LCC using ML models	44
4.6. DISCUSSION.....	45
4.6.1. Relationship between VIs and Chlorophyll.....	46
4.6.2. LCC estimation using Vegetation Indices.....	47
4.6.3. Comparison of different ML Estimation models	47
CHAPTER 5.....	49
SUMMARY, CONCLUSIONS AND RECOMMENDATIONS.....	49
5.1. Summary	49
5.2. Conclusions	49
5.3. Limitation of the study	50
5.4. Recommendations for Future Research.....	50
REFERENCES	51
APPENDICES.....	64

LIST OF FIGURES

Figure 2.1: Comparison between MS and HS. The figure on the left represents the MS showing the non-continuous spectral sampling, whereas the figure on the right represents the HS data cube, showing the continuous spectral sampling.....	10
Figure 2.2: A graphical illustration of a typical ML system approach.	13
Figure 2.3: The four pillars of agricultural machine learning as identified by Liakos et al. (2018)	14
Figure 3.1: Soil type/colour	22
Figure 3.2: Map of the study area in Ha-Mphaila irrigation scheme a) Location of the study area where data collection took place, highlighted in red, b) and c) where the study area is located in South Africa.	22
Figure 3.3: General workflow diagram illustrating data collection and analysis methods utilized in this study.	24
Figure 3.4: Field sampling: Tomato plant (left) and the SPAD-502Plus meter during ground measurement (right).....	26
Figure 4.1: Boxplot of the LCC values. The centerline of the box represents the median of the data, the diamond symbol represents the mean, the upper and lower lines represent the 1st and 3rd quantiles, and the two upper and lower whiskers represents the maximum and minimum values of the data, respectively.	35
Figure 4.2: Predictor variable importance score/ contribution in predicting the response (LCC) variable based on the Gini index.	38
Figure 4.3: Scatter plots depicting the correlation between observed and predicted LCC values and the performance level of the ML models ((a)DTR, (b)PLSR, (c)RFR and (d)XGB), accompanied by the best-fit line Before Feature Selection	41
Figure 4.4: Scatter plots depicting the correlation between observed and predicted LCC values and the performance level of the various ML models((a)DTR, (b)PLSR, (c)RFR and (d)XGB), accompanied by the best-fit line After Feature Selection	44
Figure 4.5: Spatial distribution of LCC across a tomato field, employing four distinct ML models, namely DTR, XGB, PLSR and RFR.	45

LIST OF TABLES

Table 1: Pre-Flight settings details.....	25
Table 2: The MS camera's centre wavelength and bandwidth for each spectral band.	25
Table 3: Vegetation Indices (VIs) derived from UAV image.....	27
Table 4: Statistical characteristics of tomato LCC in Mphaila irrigation scheme	34
Table 5: Pearson's Correlation coefficients between VIs and tomato LCC, at 0.05 significance level.	36
Table 6: Estimation results of LCC in tomato plants based on four ML models.	48

ACRONYMS & ABBREVIATIONS

4IR	Fourth Industrial Revolution
AFS	After Feature Selection
BFS	Before Feature Selection
CV	Cross Validation
DTR	Decision Tree regression
GB	Gradient Boosting
GCPs	Ground Control Points
HS	Hyperspectral
IPCC	Intergovernmental Panel on Climate Change
LCC	Leaf Chlorophyll Content
LOOCV	Leave-One-Out Cross Validation
MAE	Mean Absolute Error
ML	Machine Learning
MS	Multispectral
NIR	Near Infra-red
PLSR	Partial Least Squares Regression
RF	Random Forest
RFR	Random Forest regression
RGB	Red Green Blue
RMSE	Root Mean Square Error
RS	Remote Sensing
SPAD	Soil Plant Analysis Development
SVR	Support Vector Regression
UAVs	Unmanned Aerial Vehicles
VIs	Vegetation Indices
XGB	eXtreme Gradient Boost

CHAPTER 1

INTRODUCTION

1.1. Background to the study

Tomatoes, scientifically known as *Solanum lycopersicum* L., are the second most cultivated vegetable globally. They play a significant role in worldwide production and offer various health advantages and nutrients (Ma et al., 2023). According to the Department of Agriculture, Land Reform and Rural Development (2020), the Limpopo province in South Africa is often regarded as the most suitable area for tomato agriculture due to its mild environment. This province spans an area of 3900 hectares. According to the DAFF (2019), tomatoes make up approximately 15.4% of the total vegetative production in South Africa.

Nevertheless, they are often affected by pests, illnesses, and fluctuations in climate. Wang and Liu (2021) and Cardellicchio et al. (2023) stated that precision agriculture has faced difficulties in increasing productivity for this particular crop. As a result, it is crucial to study and evaluate plants' biophysical and biochemical features closely. Chlorophyll exemplifies a biochemical characteristic.

Chlorophyll is a primary indicator of photosynthetic activity in plants. The amount of solar energy a leaf absorbs depends on its pigment content (Arabian, 2015). Hence, it is imperative to evaluate the quality of crops and improve their production and productivity (Cardellicchio et al., 2023). The photosynthesis activity, nitrogen uptake, pests, diseases, crop health, stress, and crop production are all indicators of chlorophyll content (D. Li et al., 2020; Pérez-Patricio et al., 2018). LCC is crucial in controlling energy (Y. Li et al., 2018), carbon absorption and release (Croft et al., 2015), and water flow (Hailemichael et al., 2016) between plants and the atmosphere, primarily due to its significant impact on photosynthetic processes (H. Xu et al., 2022).

Conventional techniques for quantifying plant chlorophyll levels are destructive, arduous, time-consuming, and expensive. Remote sensing has demonstrated its capacity and efficiency in monitoring vegetation patterns across space and time (Khaliq et al., 2019). According to Shah et al. (2019), Remote Sensing (RS) technology presents a viable, non-destructive method for obtaining crop information, managing crops and farms, and estimating yields. As a result, many researchers are interested in using satellite and aerial RS data to study within-field variability and monitor and predict agricultural output (Kayad et al., 2019). Using satellite remote sensing (RS) has demonstrated significant utility in estimating agricultural biophysical or biochemical parameters (Narmilan et al., 2022).

Satellite RS has been widely used to monitor crops. For example, Main et al. (2011) and Nandibewoor, Muddebihal and Hegadi (2017), amongst others, concentrated on monitoring chlorophyll content in crops using a wide range of visible and near-infrared. However, these platforms are limited by a larger ground sampling distance and inadequate revisit period, making them inefficient for field-level analyzes (Sagan et al., 2019). For instance, Sentinel-2 MultiSpectral Instrument (MSI) acquires data on 13 spectral bands, but it is not ideal for monitoring subtle changes in crops and agricultural field management. This limitation arises from the coarse spatial resolution inherent in the majority of satellite RS data, which weakens the correlation between spectral reflectance and leaf chlorophyll content (LCC) observations in the field. Consequently, the accuracy of LCC observations is further reduced (H. Xu et al., 2022). Therefore, satellite RS cannot provide the requisite data resolution for precision agriculture applications, and obtaining timely and reliable data proves challenging (T. Wang et al., 2021).

In contrast, technological advancements in Unmanned Aerial Vehicles (UAVs) present an alternative to data acquisition (Maes & Steppe, 2019), as they are independent of meteorological conditions and variables (Messina et al., 2020). Utilizing UAV remote sensing for monitoring crop growth can effectively address the limitations of satellites. This approach meets the requirements for acquiring real-time image data at a high spatiotemporal resolution on different spatiotemporal scales, demonstrating its practical significance (Ji et al., 2023). UAVs can quickly acquire images with centimetre-scale spatial resolution and revisit periods at the user's will (H. Xu et al., 2022), making them suitable for monitoring the Earth's surface and estimating yields. Additionally, using UAV imagery offers alternative crop monitoring and management strategies in agricultural fields (Yuhao et al., 2020). Most commercial UAVs have digital RGB cameras and are reasonably affordable, however, crop health monitoring and estimation of biophysical and biochemical parameters require multiple spectral bands (Yang et al., 2022) and continuous imagery, which can be costly for UAV observations.

UAVs are increasingly employed alongside machine learning techniques to address diverse and complex agricultural challenges (Zuolkernan et al., 2023). Recently, Machine Learning research has shown considerable promise in constructing plant trait prediction models utilizing spectral reflectance data (Shahi et al., 2022), showcasing enhanced accuracy and robustness (Qian & Liu, 2020). In this study, leaf chlorophyll content was utilized as a proxy and surrogate for crop development, as it has been recognized as one of the most accurate and reliable indicators of productivity and health. In addition, this study explored the efficacy of the Partial least square regression (PLSR), Random Forest regression (RFR), Decision Tree regression (DTR) and eXtreme Gradient Boost (XGB) ML algorithms in estimating leaf chlorophyll content.

1.2. Problem Statement

Crop monitoring in near-real-time serves to efficiently manage resources, reduce inputs, and minimize environmental impact and footprint. The limitations of satellite technology, characterized by coarse ground sampling distance and inadequate revisit periods, hinder the effective monitoring of rapidly changing crop chlorophyll levels. While UAVs present promising features, such as high flexibility, ultra-high spatial resolution (i.e., centimetre-scale), and immunity to cloud conditions (Zhang et al., 2020), the primary emphasis in LCC estimation approaches has been using hyperspectral or sporadically narrow-band reflectance data, as observed by Croft et al. (2014). The potential of multispectral or broadband sensors, particularly in LCC estimation and retrieval, has been largely overlooked (Croft et al., 2015). Therefore, this lack of exploration presents an excellent opportunity to explore and understand the potential of multispectral UAV data and its effectiveness in agricultural monitoring.

1.3. AIMS AND OBJECTIVES

1.3.1. Aim

The aim of this study is to investigate the use of UAV-based multispectral vegetation indices combined with machine learning models to estimate leaf chlorophyll content (LCC) in tomato crops.

1.3.2. Research Objectives

- i. To estimate chlorophyll content variations using the data derived from a multispectral UAV.
- ii. To identify the most effective vegetation indices calculated from UAV multispectral data for predicting chlorophyll content in tomato crops.
- iii. To assess and compare the predictive performance of DTR, RFR, XGB and PLSR models in estimating chlorophyll content.

1.4. Significance of the study

Smallholder farmers, constituting approximately 80% of Africa's agricultural sector, play a crucial role in the continent's economy, as emphasized by Baiyegunhi (n.d). The contribution of these farming systems holds excellent significance in addressing and eradicating hunger, poverty, and unemployment while enhancing trade and investment. According to Kamara et al. (2019), smallholder agricultural systems significantly impact the agricultural output, sustainability of livelihoods, and socioeconomic growth of developing countries.

Precision agriculture advancements in South Africa, primarily driven by UAVs, represent a notable innovation. These technological developments deliver spatially detailed, nearly real-

time data crucial for evaluating crop dynamics, especially beneficial for smallholder farmers, as emphasized by Brewer et al. (2022).

Nevertheless, small-scale agricultural farms frequently struggle with limited resources, resulting in yields that are notably below the land's potential. Given this understanding, it is imperative to implement inventive and cost-effective measures to empower small-scale farmers and enhance their efficiency. These measures are essential for guaranteeing higher and more nutritious crop production, aligning with the overall objective of sustainable agricultural growth (Singh et al., 2020). Furthermore, prompt and precise crop monitoring enables the timely identification and intervention of unhealthy crops, thereby preserving agricultural yield production (Yuhao et al., 2020). Additionally, estimating crop chlorophyll content can contribute to informed decisions regarding nitrogen fertilization (Mamatha & Sreekumar, 2018). Therefore, this study used LCC to monitor and track the changes in the tomato plant, given its recognition as a vital and reliable measure of plant growth, productivity and physiological status (Mamatha & Sreekumar, 2018; Naveed Tahir et al., 2018).

1.5. Structure of the research report

The following outline delineates the logical structure of this text, aiming to provide the reader with a comprehensive overview of the content and substantive themes explored in each chapter.

Chapter 1 serves as the introductory section of the research paper, offering a comprehensive overview of the study. This encompasses an overview of tomato farming, chlorophyll, remote sensing techniques, and machine learning. The chapter outlines the specific research problem aimed at bridging a gap in existing knowledge and elucidates the research's overarching aims and objectives. Furthermore, it delves into the significance of the study within the context of smallholder farmers.

Chapter 2 analyzes scholarly literature on precision agriculture and crop monitoring. The study focuses on studies that primarily used UAV image data. The chapter highlights the importance of using cutting-edge technologies for better crop monitoring. The chapter discusses different applications of ML in precision agriculture and the role of sensors in precision agriculture. Additionally, it provides the challenges of ML in agriculture.

Chapter 3 provides a detailed account of the study's location and an overview of the methodologies employed to achieve the research objectives. It offers insights into the types of data used and the specific techniques involved in data collection and analysis. The research methodology comprises an empirical field survey for data collection and investigation. A vital aspect of the study is exploring the relationship between Vegetation Indices (VIs) and

chlorophyll content. This process is carried out in five stages: feature selection, model learning, cross-validation, model performance evaluation, and mapping of LCC predictions.

Chapter 4 reports the key findings obtained from this study. It aligns these findings with the objectives that shaped and structured this paper. It effectively connects the research efforts to the acquired knowledge as a bridge, providing a coherent link between the study's goals and the knowledge gained.

Chapter 5 presents conclusive remarks, summarizing the paper. It acknowledges the limitations encountered in the current study, and based on the significant findings of this research, provides insightful recommendations for future research.

CHAPTER 2

LITERATURE REVIEW

2.1. Agriculture in the 4IR

Agriculture has undergone some of the most disruptive changes brought about by anthropogenic activity in the previous few decades (Angel & McCabe, 2022). Using herbicides and pesticides is more prevalent in traditional agricultural practices due to the inherent challenge of early detection and prevention of weeds, pests, and diseases (Saranya et al., 2023). These have a number of detrimental repercussions, such as decreased productivity and a bad effect on the health of the environment, the soil, and the crops. In addition, by 2050, the demand for food is anticipated to grow by 70% (FAO, 2009). Conventional farming systems cannot meet the increasing demand for food (Ane & Yasmin, 2019). The Fourth Industrial Revolution (4IR) urges farms to embrace more sustainable and eco-friendly practices.

Agriculture and industry are interdependent and mutually reinforcing. The 4IR encompasses cutting-edge digital technologies that present a transformative chance to revolutionise how humans think and operate. The notion of the 4IR integrates big data and artificial intelligence (AI), which have garnered considerable attention and appeal in precision farming (Torres et al., 2020). These include monitoring crop health status, identifying insect pests, quantifying soil moisture levels, determining when to harvest crops, and reducing the need for labour-intensive human monitoring (Ane & Yasmin, 2019).

The 4IR is a period marked by the swift and profound emergence of innovative and advanced technologies that integrate the digital and physical realms (Mzwakhe & Isaac, 2021). This emerging trend is anchored in several digital technologies, including AI, IoT, big data, robotics, UAVs and nanotechnology (Nkosi & Agholor, 2021; Schwab, 2016). The 4IR in agriculture, cutting-edge digital technology, seeks to increase and revolutionise agricultural outputs. These developments have the potential to boost agricultural profitability and production, particularly in Africa (Aishwarya *et al.*, 2022). The 4IR is extending precision agriculture through robotics and artificial intelligence, and its implementation is smart observation, which extracts real-time information from field level data with minimal human input (David et al., 2022).

Nevertheless, the 4IR risks exacerbating societal inequalities due to insufficient resources for adjusting to novel approaches. This is expected to expand the gap between the affluent and the disadvantaged (Mzwakhe & Isaac, 2021). The limited availability of financial resources is the primary factor contributing to small-scale farmers' low adoption of technology. Njukiet et

al. (2013) also observed that the low rate of technological adoption among smallholder farmers could be related to the insufficient education and technical expertise of these farmers.

Adequate internet access and usage, along with the affordability of data and technological tools, are considered the standards for the digital transformation of the South Africa's agricultural sector in the 4IR (Aguera et al., 2020). Based on the investigation carried out by Research ICT Africa on internet access, South Africa leads among Sub-Saharan African countries, with 53% of the population having internet access (Aguera et al., 2020). The study also highlights disparities in internet access between urban and rural areas, emphasizing the need for increased investments in ICT infrastructure in rural regions to benefit smallholder farmers in the era of the 4IR.

Small-scale farmers face challenges due to inadequate infrastructure, antiquated farming methods, limited agricultural knowledge, and insufficient resources, which prevent them from fully benefiting from the potential available in the agricultural sector. These challenges hinder their ability to compete effectively and significantly contribute to economic expansion, worsened by restricted access to capital, markets, and agricultural resources (Aguera et al., 2020). Given the current circumstances and struggles smallholder farmers face, adopting and implementing the Fourth Industrial Revolution (4IR) may pose significant challenges (Ane & Yasmin, 2019).

As per the 2019 report by StatsSA, nearly one-fifth of South Africa's population faces food insecurity, which could deteriorate without enhancements in food production. The potential solution lies in groundbreaking ideas and cutting-edge technologies arising from the 4IR, as Aguera (2020) highlighted. Furthermore, the 4IR is expected to stimulate South Africa's agricultural sector (Ane & Yasmin, 2019).

2.2. The use of UAVs data for Precision Agriculture

Precision Agriculture (PA) has recently emerged as a prominent field of agricultural research, attracting significant global interest. PA involves a management approach that integrates individual data, spatial, temporal, and other relevant information to support informed decision-making based on estimated variability (IPCC, 2022). The primary objective of PA is to enhance resource utilization sustainability, efficiency, profitability, quality, and productivity within agricultural production systems while minimizing production costs and effects on the environment (Sharma et al., 2021). The concept of PA recognizes variations within the field and aims to identify the spatial location and extent of these variations.

Modern agriculture confronts unprecedented challenges that have not been observed in preceding generations. The first challenge stems from climate change's impact on agricultural

systems, disrupting established farming practices (Mulla et al., 2020). This disruption leads to irregular crop seasons, driven by elevated temperatures and limited water availability in productive agricultural regions (Piao et al., 2019), inevitably leading to new invasive pests or intensification of existing ones. The subsequent challenge entails addressing the food demands of an increasing global population while guaranteeing food security by minimizing agrochemical use and implementing rigorous controls across the entire agricultural supply chain.

Advancements in Unmanned Aerial Vehicles (UAVs) RS technology have played a pivotal role in both remote sensing and PA (Tsouros et al., 2019). Maes and Steppe (2019) highlighted the role of UAVs as a pivotal development in PA research, emphasizing their unique ability to deliver products with exceptional temporal, spatial and spectral temporal resolution. UAVs were developed and optimized to meet the demands of precision farming, serving as a reliable and potent monitoring platform. Equipped with various sensors (RGB, multispectral, hyperspectral or thermal), UAVs offer significant potential for acquiring field data conveniently, efficiently and economically compared to conventional methods. Typically, five- or six-band multispectral cameras are used by UAVs to monitor crop fields. Many studies have provided evidence supporting the efficacy of this method in monitoring crop chlorophyll (Brewer et al., 2022), detecting and mapping early water stress (Brewer, 2021), estimating nitrogen concentration (M. Li et al., 2022), and predicting yield (Nevavuori et al., 2020). For instance, Ban et al. (2022) found significant correlations between VIs such as NDVI and LCC in rice crops using a five-band multispectral camera.

UAVs, equipped with sensors, are significantly transforming the agriculture sector. They serve multiple purposes, including field and crop monitoring, pesticide application and the implementation of drip irrigations. UAVs capture images of crops throughout the crop's lifecycle, which are analyzed utilizing ML and computer vision algorithms to identify diseases and weeds. With increasing concerns among farmers about the unpredictability of weather patterns, UAVs integrated with AI technology offer an efficient solution through the implementation of drip irrigation, effectively addressing water-related challenges faced by farmers (Sharma et al., 2021).

A UAV, which is more flexible, has fewer limitations than airborne and spaceborne systems, and has high spatial and spectral resolution, is one of the instruments used for crop monitoring in PA (Yuhao et al., 2020).

2.2.1. Imaging Sensor systems

Integrating conventional imaging and spectroscopic techniques in spectral imaging enables the generation of spatial and spectral data on a specific scene. Xu and Mishra (2019) proposed

a method that facilitates the concurrent assessment of many physical attributes (e.g., colour, shape, and size) as well as chemical properties (such as water, sugar content, and fat) of the scene(s) under investigation. Omia et al. (2023) demonstrated that the target object's spectral signature can be visualised spatially through distribution maps.

MultiSpectral and Hyperspectral imaging are the two primary imaging methods. These imaging sensors function in the optical segment of the electromagnetic spectrum. Optical sensors typically detect the quantity of reflected radiation within specific wavelength intervals, including visible light, NIR, MIR and thermal radiation (Lustenhouwer, 2009). The distinction among these sensors is their capacity to perceive the spectrum range.

2.2.1.1. MultiSpectral (MS) Imaging sensors

Multispectral (MS) imaging comprises multiple layers of images representing a specific scene, wherein each layer is captured at a distinct portion, sometimes referred to as a band, within the electromagnetic spectrum. According to Omia et al. (2023), the prevailing MS sensors typically employ 3-10 spectral bands for each pixel inside the generated image. These cameras serve the purpose of both colour imaging and spectral monitoring, capturing multiple bands across the invisible and visible spectrum (Akkoyun, 2022). The cameras can encompass specific segments within the infrared and ultraviolet spectrum. The red-edge and NIR wavelengths from the infrared spectrum are the predominant wavelengths utilized by these cameras. According to Ramirez et al. (2022), MS cameras have a sensitive area detector with specified waveband filters or an adjustable light source.

The majority of research in literature showcases the extensive array of MS imaging applications in monitoring field crops, including the determination and retrieval of the photosynthetic pigment, phenotyping of crop biomasses, and evaluation of the variability of nitrogen, amongst other applications (Omia et al., 2023). Various spectral bands accurately represent the distinctive attributes of different plants, effectively differentiating various crops. Brewer et al. (2022) employed the RedEdge multispectral camera by MicaSense to assess maize crop health in their study. Specifically, they focused on chlorophyll content, a reliable indicator of plant productivity. They successfully estimated chlorophyll levels throughout different growth stages, and their study also revealed that VIs derived from the red-edge and NIR wavelength bands were the most influential.

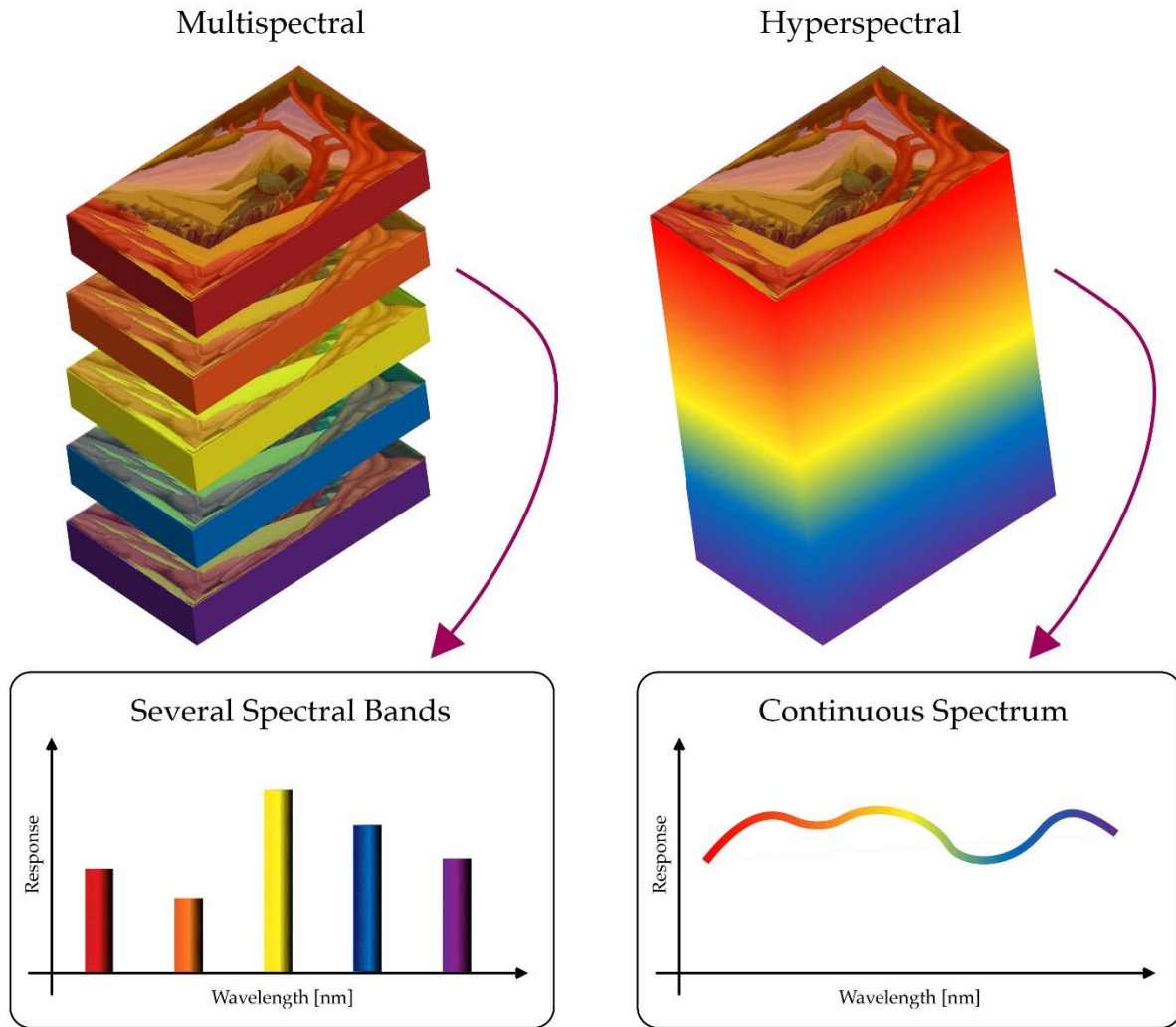


Figure 2.1: Comparison between MS and HS. The figure on the left represents the MS showing the non-continuous spectral sampling, whereas the figure on the right represents the HS data cube, showing the continuous spectral sampling.

2.2.1.2. Hyperspectral (HS) Imaging sensors

Hyperspectral systems have revolutionized data collection by enabling the capture of multiple spectral bands simultaneously, generating spectral data with exceptional detail. These systems operate in the electromagnetic spectrum's visible, NIR, MIR, and TIR portions. Typically, HS sensors gather 200 or more bands, creating nearly continuous reflectance spectra for each pixel in the scene (Lustenhauer, 2009), as shown in Figure 2.1. The increased spectral resolution enhances the ability to effectively distinguish between elements in an image (Govender et al., 2007).

In contrast to MS imaging sensors, which collect several discrete image layers of the identical scene, HS imaging involves capturing spectral images for narrow spectral bands instead of distinct bands (Zhong et al., 2018). HS data significantly benefits monitoring vegetation health,

enabling the capture of diverse physical and chemical responses in plants across different wavelengths. Nevertheless, abundant information and numerous bands in HS data can lead to a high correlation among some bands, resulting in redundancy that may distort prediction results and reduce accuracy (Tuerxun et al., 2023).

Several studies have demonstrated that HS imaging technology can outperform MS imaging technology in various agricultural applications. This superiority can be attributed to its ability to identify and differentiate between distinct object characteristics using multiple narrow contiguous spectral channels. In recent years, HS remote sensing has proven to be a significant advance in vegetation chlorophyll content monitoring, offering high spectral resolution and effective and non-destructive capabilities, as highlighted by Ta et al. (2021). Additionally, it offers a highly effective method for studying the chlorophyll levels in crops and facilitating PA practices.

While HS sensors provide more spectral data than MS sensors (Zhong et al., 2018), their higher cost and lower spatial resolution present limitations. The reduced spatial resolution may result in the loss of detailed information for small targets (T. Wang et al., 2021). Notably, Tao et al. (2020) used HS sensors to investigate the relationship between crop biomass, red-edge parameters, and VIs. Their findings emphasized the importance of using a suitable quantity of spectral characteristics associated with biomass in order to avoid overfitting the model.

2.3. Agricultural Applications using Satellite Imagery data.

Satellite imagery has demonstrated remarkable effectiveness in various applications in agriculture (Karakizi et al., 2016) owing to its significant advantages, such as non-destructive data capture over expansive geographical areas (Liaghat, 2010). Satellite imagery provides data covering hundreds of hectares and areas difficult to access for UAVs. This feature enables rapid data acquisition, facilitating prompt agronomic management decisions to identify potential intervention areas and prevent production losses. According to Alvarez-Vanhard et al. (2021), satellites exhibit limited flexibility in their collection capabilities, subject to cloud cover, view angle, and acquisition time constraints.

Vegetation indices (VIs) calculated from satellite data are widely used for estimating crop chlorophyll content. Recent MS satellite sensors incorporate red-edge bands like RapidEye, WorldView-2 and Sentinel-2. While VIs derived from Sentinel-2 images have been extensively evaluated for estimating canopy chlorophyll content across various crops like winter wheat, potato, maize and soybean, research on RapidEye and WorldView-2 has been relatively limited (Zou et al., 2023). WorldView-2 satellites offer high spatial resolution and exhibit reduced susceptibility to spatial variability. However, these images come at a premium cost and entail lower data processing efficiency.

In a comparative study by Zou et al. (2023), Sentinel-2 images offered more advantages than Planet images, mainly due to their high spectral resolution. Conversely, Croft et al. (2022) employed multispectral Landsat-8 OLI data to analyze the spatial distribution of LCC across various growth stages and nitrogen fertilizer application levels in wheat and corn fields. Their findings revealed a robust correlation between Landsat 8-derived estimates of LCC and directly measured LCC. This correlation was evidenced by an R^2 value of 0.64 and an RMSE of $16.18 \mu\text{g}/\text{cm}^2$, showcasing the potential of RS data for mapping and understanding the spatial fluctuations of LCC in crops.

The literature highlights the effectiveness of satellite RS in mapping extensive areas, particularly in large-scale studies and monitoring. However, limitations persist regarding the spectral, spatial, and temporal resolution required for PA practices (Jafarbiglu & Pourreza, 2022).

2.4. Machine Learning

2.4.1. Brief overview

ML is the subset of computer science that allows computers to learn autonomously without explicitly programming (Arthur Samuel, 1959; Sharma et al., 2021). ML algorithms are primarily employed to tackle intricate problems where human expertise may fall short, including tasks like predicting weather patterns, identifying plant diseases, filtering spam and recognizing patterns. ML techniques encompass a learning phase aimed at gaining knowledge from "experience" (training data) to execute a specific task (Tambakhe et al., 2019). ML algorithms aim to enhance a given activity's performance by leveraging examples or prior knowledge. These algorithms function through two main processes: training and testing, often utilizing characteristic features forming a vector, encompassing binary, numeric, ordinal, or nominal to facilitate the learning process (Khan et al., 2022). The machine can perform the task through experiential learning using training data (Baburao et al., 2023). Following the training phase, the model applies the knowledge gained to classify, predict, or cluster new, unseen data (testing data). Figure 2.2 illustrates a standard ML approach during these phases.

ML algorithms can be grouped into unsupervised, supervised, and reinforcement learning, each employing distinct learning approaches. In supervised learning, the machine acquires knowledge by analyzing input and output data to determine optimal methods for producing desired outputs from given inputs (Baburao et al., 2023). Conversely, unsupervised learning operates without labels, enabling algorithms to independently identify patterns within input data (Benos et al., 2021). Meanwhile, reinforcement learning involves actions within a system

that trigger outcomes to the algorithm, adjusting its behaviour accordingly based on this feedback loop (Tambakhe et al., 2019).

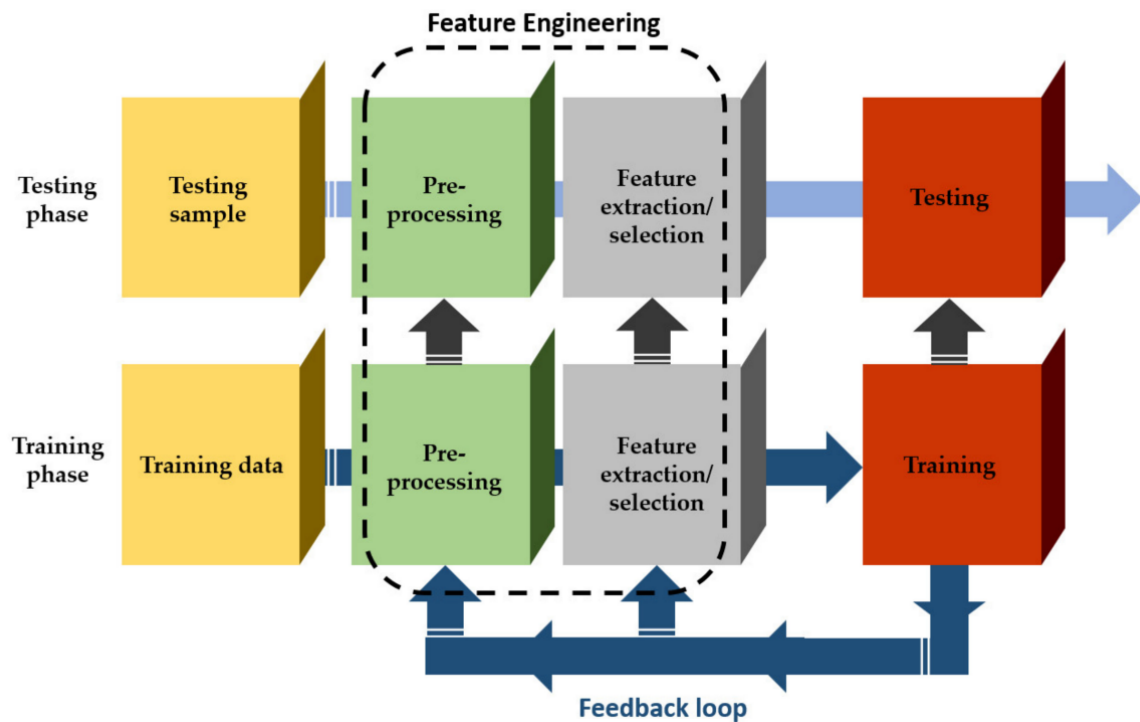


Figure 2.2: A graphical illustration of a typical ML system approach.

2.4.2. Machine Learning Applications in Precision Agriculture

Agriculture is a cornerstone of the nation's economy, providing food for the entire population and interacting with all relevant sectors (Pallathadka et al., 2023). However, modern agriculture faces numerous challenges, such as increasing food demand due to the expansion of the world population, the exhaustion of natural resources, and climate change (Benos et al., 2021). To tackle these issues, there is an increasing inclination towards leveraging advanced technologies, such as Machine Learning (ML), to improve agricultural efficiency while minimizing environmental impact.

Machine Learning has demonstrated significant potential in advancing multiple facets of agriculture 4.0. This discussion seamlessly leads into a comprehensive study conducted by Liakos et al. (2018), reviewing 40 papers focusing on ML's application in agriculture, categorizing them into four main domains, as shown in Figure 2.3. Figure 2.3 elucidates agricultural challenges that can be resolved using ML, as Liakos et al. (2018) outlined. Furthermore, the study, encompassing literature from 2004 to 2018, discovered that approximately 61% of the articles concentrated on crop management. This encompasses predicting crop yield, disease and weed detection, crop quality assessment, and crop recognition.

Furthering the discourse on crop management, applying sophisticated ML techniques in this field has demonstrated substantial benefits. These techniques not only enhance productivity but also bolster financial returns. These elements encapsulate the fundamental objectives of precision agriculture, a modern farming practice that employs data-driven techniques to augment agricultural productivity while concurrently diminishing its environmental footprint. Consequently, ML emerges as a potent instrument in the progression of agriculture, steering it towards a future characterized by enhanced sustainability and productivity (Baburao et al., 2023).

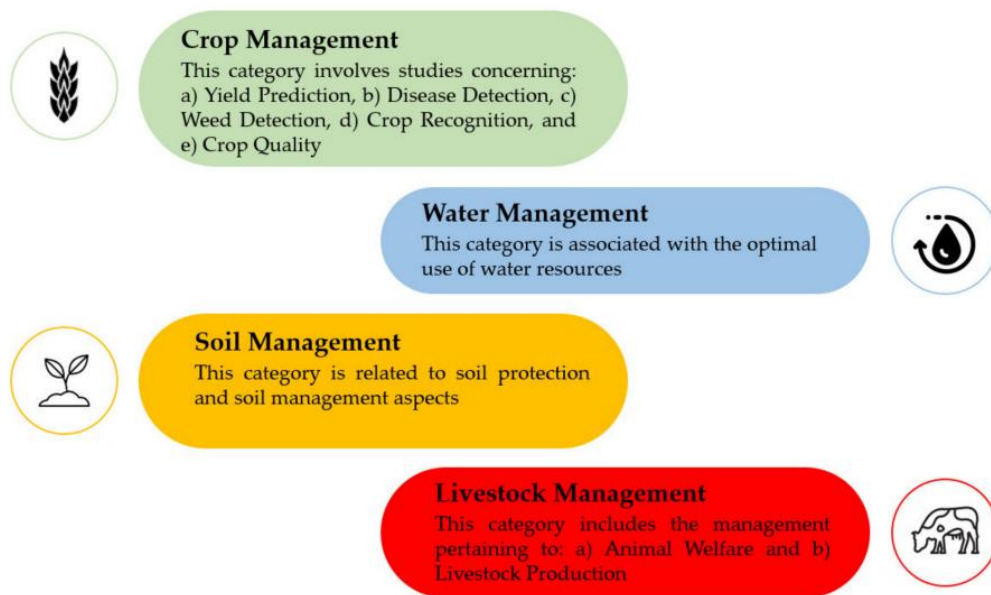


Figure 2.3: The four pillars of agricultural machine learning as identified by Liakos et al. (2018)

Machine learning (ML) has several applications across diverse agricultural domains. In crop monitoring, ML algorithms can examine data derived from RS technologies, such as spectral imaging and satellite images, aiming to forecast and oversee diverse crop attributes, encompassing yield, pests, diseases, and quality. ML algorithms can enhance irrigation and fertilization tactics and optimize the efficiency of PA techniques.

2.4.2.1. Crop Yield Prediction

Predicting crop yield is a complex challenge in precision agriculture, paramount in decision-making at various scales, from international to local contexts (van Klompenburg et al., 2020). ML has emerged as a versatile tool that allows for precise crop yield predictions, enabling strategic decisions about crop cultivation and management throughout the growing season (Iniyan et al., 2023). This recent innovation in agriculture significantly influences the economy by assisting farmers in making well-informed judgements about optimal cultivation and management of crops, considering environmental factors (Ansarifar et al., 2021). In the pursuit

of accurate predictions, advanced models have evolved, incorporating diverse environmental variables such as precipitation, temperature, solar radiation, and soil attributes, among others. Ansarifar et al. (2021) noted that these models go beyond the basic applications of ML and enhance the precision of yield estimation.

Researchers have explored numerous models, with an increasing inclination towards ML in crop yield estimation (Kumar et al., 2021). For instance, Kuradusenge et al. (2023) utilized three ML techniques - Random Forests (RF), Support Vector Regression (SVR), and Polynomial Regression (PR) - to predict the yield of Irish potatoes and maize. Their study, utilizing climate-related data, highlighted the superiority of the RF model, positioning it as the most recommended choice for early crop yield prediction.

Expanding on the application of ML in agriculture, Iniyan and Jebakumar (2022) developed a model using a multilayer stacked ensemble technique to predict the yield of soybean and corn crops. This model, selecting features based on mutual information, exhibited superior performance compared to other ML methods like Decision Tree regression (DTR), Multiple Linear regression (MLR), Random Forest regression (RFR), and Gradient Boosting regression (GBR). This emphasizes the continuous evolution and improvement of ML techniques in addressing the intricacies of agricultural practices.

2.4.2.2. Disease and Weed Detection

Crop diseases substantially threaten agricultural systems, adversely impacting the quality and quantity of yields from production to transportation. Reports of yield losses due to plant diseases are widespread at the farm level, with significant implications for global food security. Recognizing and responding to these challenges, the prompt detection of plant diseases becomes imperative for effective control and management (Verma et al., 2022). Domingues et al. (2022) reported that an estimated 20 to 40% of the annual crop output globally has been lost due to diseases and pests, resulting in economic losses of \$220 billion. Therefore, in crop disease management, image processing plays a pivotal role in advancing the detection and categorization of diseases during the initial stages of crop development. This phase negatively impacts crop productivity and yield. Employing machine learning-based image classifiers is crucial for distinguishing between diseased and healthy crops (Shinde et al., 2022).

ML algorithms have been particularly beneficial for identifying weed, disease, and plague mapping using RS (Lassalle, 2021). Many farmers and growers consider weeds one of the most significant challenges to crop production. Accurate and precise weed detection is a crucial aspect of sustainable agriculture, primarily due to the difficulties in identifying and distinguishing them from crops (Cravero et al., 2022). Integrating sensors and ML algorithms offers a solution, enabling the precise identification and detection of weeds without causing

adverse environmental impacts. These algorithms, trained on extensive datasets, can discern subtle differences in the texture, colour, and shape of weeds and crops. Furthermore, they have paved the way for the development of robots and tools designed to effectively eliminate weeds, thereby reducing the dependence on herbicides and mitigating environmental harm.

Reducing herbicide use is a significant environmental benefit of applying ML algorithms in weed management. Herbicides, which can negatively impact water sources, wildlife, and human health, are minimized, reducing agriculture's environmental footprint. This reduction allows for maintaining high crop yields while minimizing the environmental impact of agricultural practices (Domingues et al., 2022). ML algorithms have significantly enhanced the precise detection and classification of crop quality attributes. These algorithms can identify and differentiate various crop traits, such as size, shape, colour, and texture, helping farmers optimize their production. The use of ML algorithms in agriculture has, therefore, increased the value of products and reduced waste by enabling farmers to produce crops with higher quality and consistency (Mesías-Ruiz et al., 2023).

In summary, integrating sensors and ML algorithms in weed detection has revolutionized sustainable agriculture by enabling farmers to identify and manage weeds more effectively, reduce herbicide use, and optimize crop production (Khanal et al., 2017). Precisely detecting and classifying crop quality characteristics have also enhanced the value of products and minimized waste.

2.4.2.3. Crop Recognition

Crop recognition and mapping pertain to detecting and delineating distinct crop types within agricultural fields (Araújo et al., 2023). It involves leveraging various data sources to identify and categorize various crops and their spatial distribution. Employing ML techniques enables creating precise and detailed crop maps, identifying each crop's unique characteristics. Such information is invaluable for agricultural planning, yield prediction, and resource management (Araújo et al., 2023). Furthermore, accurate crop recognition is pivotal in shaping and formulating future agricultural policies and advancing sustainable development goals (United Nations, 2015).

Satellite remote sensing has played a pivotal role in mapping crop types, benefiting from its extensive coverage, timely observations, and regular monitoring capabilities (Zhong et al., 2019). Broadly, existing approaches for identifying crop types can be categorized into spectral index-based and machine-learning techniques (Pelletier et al., 2019). Spectral indices are meticulously designed by combining reflectance data from various bands, relying on the distinct spectral properties of crops. While effective for certain specific crops, this approach encounters challenges when generalizing to a diverse range of crop types due to the inherent

similarities in their spectral features (Mesías-Ruiz et al., 2023). Machine learning, including supervised and unsupervised methods, is crucial in overcoming these challenges and advancing accurate crop-type mapping.

Feng et al. (2019) studied crop type identification and mapping using Sentinel-2A imagery and ML algorithms, notably RF and support vector machines. Remarkably, their approach achieved an impressive identification accuracy exceeding 95% for each crop type. This research demonstrates the effectiveness of combining remote sensing data and advanced ML techniques for accurate crop classification.

2.4.2.4. Crop/ Species Quality

Crop quality pertains to the attributes of agricultural produce that ascertain its suitability and value. Enhancing crop quality through ML entails monitoring and controlling factors such as crop growth, organoleptic characteristics, nutrient levels, and other relevant parameters (Araújo et al., 2023). The quality of crops is intricately linked to the detection of diseases, which can significantly diminish the quality and quantity of agricultural yield, posing a substantial risk to food safety (Bilal et al., 2023). Consequently, crop quality assumes paramount importance in the agricultural sector, given that the pricing and competitiveness of farmers hinge on the quality of their agricultural products.

Precisely recognizing and classifying characteristics indicative of crop quality can elevate the product's value and reduce wastage, as Liakos et al. (2018) highlighted. Compared to human experts, machines can leverage data and discern relationships that might seem insignificant, thus revealing novel attributes that play a role in determining the overall quality of crops and facilitating their identification. Numerous studies evaluate crop quality through image classification methods, aiming to differentiate between healthy and unhealthy crops. For example, Meghashree et al. (2021) utilized a Convolutional Neural Network (CNN) in their study to appraise crop quality. Their approach involved training the CNN on a dataset of images of healthy and diseased leaves to enable it to identify the differences between the two. The study aimed to enhance the quality and quantity of agricultural production by identifying and addressing crop diseases at an early stage.

Similarly, Chokey and Jain (2019) employed supervised ML techniques to identify plant diseases as part of their crop quality assessment strategy. They used a Quadratic SVM to classify plant images as healthy or unhealthy/diseased. They trained the SVM on a dataset of images of plants with and without diseases to enable it to identify the patterns that distinguish healthy and diseased plants. Their approach yielded an accuracy of 83.3% in disease detection, which is a significant improvement over traditional methods of crop quality assessment.

2.5. Machine Learning Modelling in Agriculture

2.5.1. Machine Learning modelling with VIs

Vegetation Indices (VIs) are metrics derived from remotely sensed data that measure vegetation health or vigour. Calculated using reflectance measurements in various spectral bands, including the NIR and green wavebands, VIs are crucial for understanding variations in LCC. Consequently, identifying optimal VIs for estimating LCC in the NIR and green wavebands underscores the susceptibility of these spectral regions to variations in LCC (Lu et al., 2018).

Various Vegetation Indices (VIs) extracted from remote sensing images offer a versatile means to monitor ground vegetation, enabling quantitative assessments of richness, greenness, and vitality. The use of multiple VIs enhances monitoring accuracy by capturing additional spectral features. Distinct Vegetation Indices possess individual attributes, and employing multiple indices allows for identifying a broader range of spectral features, thereby enhancing monitoring accuracy significantly. VIs are mathematical formulas combining reflectance measurements across numerous spectral bands, providing valuable information for assessing crop growth, vitality, and various vegetation characteristics such as chlorophyll and biomass. Mapping these indices enables a more thorough comprehension of the vegetation conditions' temporal and spatial fluctuations (Ferchichi et al., 2022).

Numerous studies have explored the estimation of LCC using VIs derived from UAV images, revealing a strong correlation between VIs and LCC, with each VI performing optimally for different crop types. For instance, in a study by Angel and McCabe (2022), the comparison between LCC estimation using spectral bands and VIs predictors demonstrated that ML models performed better with VIs as predictor variables. Similarly, Brewer et al. (2022) evaluated thirteen VIs for estimating maize chlorophyll content across various phenotyping stages. They identified NDRE, RECI and NDVI as the most influential variables during the early vegetative growth stage. Other VIs, including MCARI and the EVI, have proven effective for LCC estimation, with MCARI particularly suitable for crops with high canopy density, like rice and wheat (Li et al., 2017), and EVI found effective for cotton crops (Wang et al., 2019).

The integration of MS images captured by UAVs and VIs derived from those images offers a reliable and efficient strategy for estimating LCC (Ban et al., 2022). This method provides valuable insights into crop health and shows significant promise in precision agriculture, facilitating site-specific management techniques and increasing agricultural productivity. Therefore, as technology advances, UAV-based LCC estimation will become vital for managing crops sustainably and effectively.

In PA, ML has advanced accurate, nondestructive sensing and high throughput and data processing technologies. Machine learning has demonstrated significant promise in providing innovative findings in estimating leaf chlorophyll levels. Nevertheless, the successful implementation of applications can be impeded by various challenges, such as the selection of appropriate algorithms, the availability and quality of training data, the choice of effective learning strategies, and the selection of suitable predictors (Verrelst *et al.*, 2019). Researchers such as An *et al.* (2020), Angel and McCabe (2022) and Ban *et al.* (2022) have also made extensive use of regression algorithms to monitor and retrieve crop chlorophyll using spectral data collected from UAV photos. For example, Angel and McCabe (2022) tested 60 VIs formulated from a hyperspectral imaging system over six regression algorithms, and their findings revealed that the models performed better when using VIs as predictors compared to spectral bands. However, a study by Yang *et al.* (2022) revealed that stacking ML algorithms had more significant advantages in estimating chlorophyll content than single machine learning algorithms.

2.6. Challenges of ML in Agriculture

Despite the encouraging and substantial outcomes of incorporating ML into the agricultural sector, some hurdles remain. Some of these challenges arise during the implementation of ML algorithms in agriculture (Meshram *et al.*, 2021).

The foundation for constructing ML models rests on data. Numerous researchers encounter data-related challenges, including issues such as insufficient and unavailability of data in the desired format, poor data quality, and the potential inclusion of extraneous features in the dataset (Meshram *et al.*, 2021). In ML, the quality and quantity of data significantly influence the success and accuracy of the model's prediction. Data scarcity can occur when the amount of data collected is insufficient to represent the problem being solved. This leads to overfitting, in which the model exhibits an excessive resemblance to training data, resulting in poor performance on new data.

According to Oikonomidis *et al.* (2023), another significant challenge lies in the limited availability of extensive training datasets, which increases the risk of overfitting and ultimately impacts the practical performance of the models. Constructing an accurate model requires substantial training data, along with essential testing and validation to assess its accuracy prior to deployment (Meshram *et al.*, 2021). Developing a model from scratch for optimal outcomes involves lengthy training and repeated testing, which can be time-consuming. Additionally, it necessitates high-performance hardware resources, skilled programmers, and practical testing tools. Challenges such as overfitting and underfitting commonly arise during the model-building process.

Another significant unresolved issue is that most farmers lack expertise in machine learning, hindering their ability to grasp the patterns generated by ML algorithms entirely. Cravero et al. (2022) and Meshram et al. (2021) noted that there is a need for the development of more user-friendly systems.

CHAPTER 3

METHODOLOGY

3.1. Study Area

3.1.1. Geographic Landscape and Features

The study area is located within the rural area of Ha-Mphaila village in Limpopo province, South Africa (22°53'58" S, 30°08'55" E), as shown in Figure 3.2. This irrigation scheme, reliant on the perennial Mutshedzi River, covers approximately 71 hectares and supports about 74 farmers. However, this study was directed solely towards a plot encompassing approximately 1.6 hectares dedicated to tomato cultivation, utilizing a drip irrigation system. It is noteworthy to mention that Limpopo province is a significant hub for agricultural schemes, housing a substantial portion of the country's Smallholder Irrigation Schemes (SIS), with 170 out of 302 schemes within its bounds (van Averbek et al., 2011).

3.1.2. Soil

Reddish-brown loamy soils dominate the study area, as shown in Figure 3.1 and support diverse agricultural cultivation, fostering the growth of various vegetables and crops such as, maize, cabbage and tomatoes, to name a few (Jiyane & Simalenga, 2019). Iron oxides, particularly hematite, are primarily responsible for the reddish colour of red loamy soils (Ittianath, 2024). This soil type is optimal for agriculture due to its well-draining properties while retaining adequate moisture and nutrients necessary for robust crop development (Jiyane & Simalenga, 2019).

The smallholder farmers practice manual planting, tending and harvesting of crops. Additionally, herbicide backpack sprayers are utilized for weed and grass control. These farming practices are crucial in guaranteeing food security and maintaining sustainable livelihoods. Furthermore, the locally cultivated crop is directed towards local markets, bolstering the community's economic well-being.



Figure 3.1: Soil type and colour in the study area.

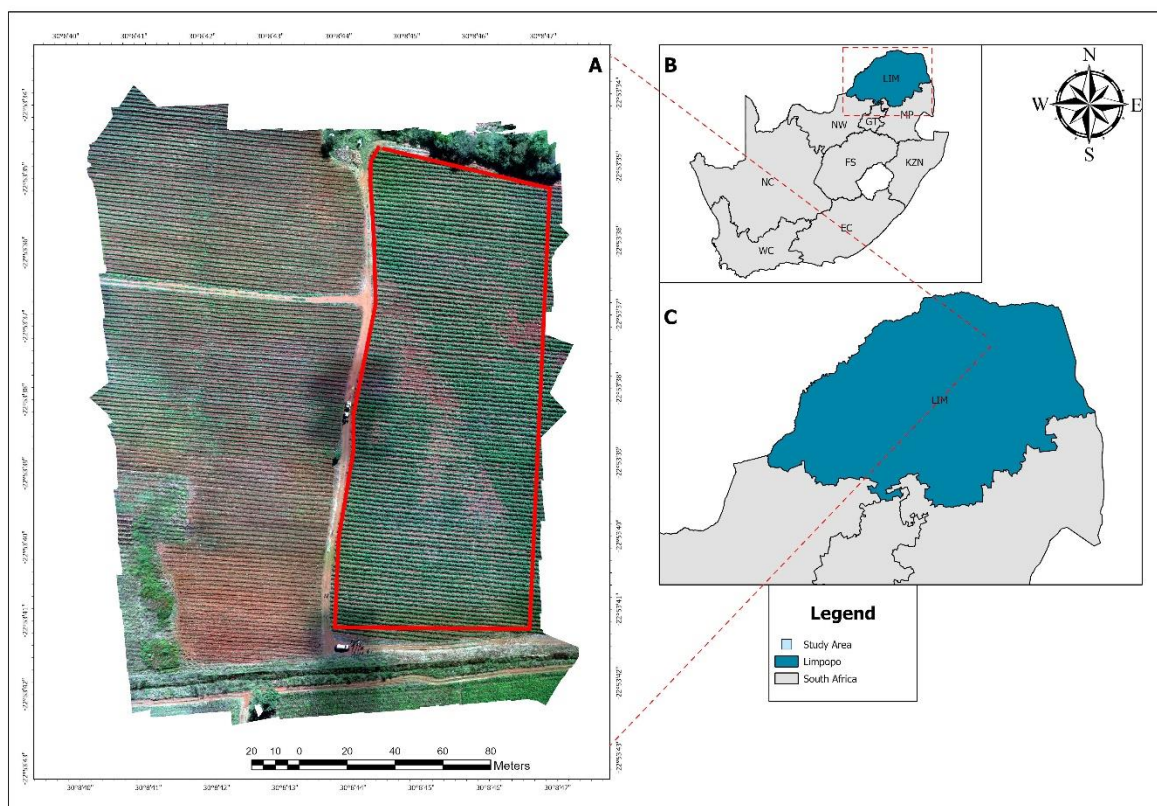


Figure 3.2: Map of the study area in Ha-Mphaila irrigation scheme a) Location of the study area where data collection took place, highlighted in red, b) and c) where the study area is located in South Africa.

3.1.2. Climate

The region experiences a mean annual rainfall of 400 to 900 mm (Moeletsi et al., 2013), a crucial water source for crop growth. This precipitation level is a significant resource essential for crop growth and development. The area experiences a diverse array of temperatures over the year, with average yearly temperatures fluctuating between 9° and 17°C in winter and 22° and 37°C in summer (Rusere et al., 2019). These temperature ranges create an environment conducive to the growth of various crops. For instance, crops like maize, cabbage, and tomatoes thrive in these warm conditions.

Moreover, the region is renowned for cultivating subtropical fruits like mangoes, litchis, bananas, citrus, avocados, pecan, and macadamia nuts (Oni et al., 2011). These fruits thrive in warm temperatures, and the region's climate provides the optimal conditions for their cultivation. Overall, the region's combination of rainfall and temperature creates a favourable environment for diverse agricultural production. The climate supports the growth of both vegetable crops and subtropical fruits, contributing to the region's agricultural diversity and productivity (Materechera & Scholes, 2022).

3.2. Data Acquisition

Figure 3.3 presents a detailed workflow diagram illustrating the systematic approach adopted for data acquisition and analysis in this study. It visually depicts the methods employed, emphasizing each critical step and its interconnections.

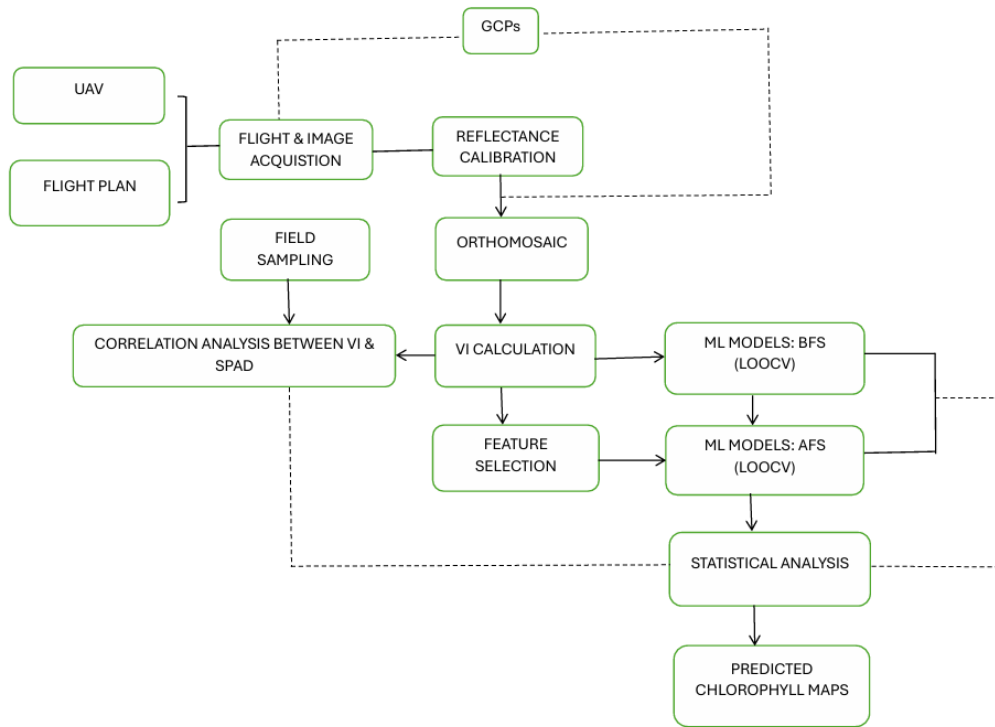


Figure 3.3: General workflow diagram illustrating data collection and analysis methods utilized in this study.

Notes: GCPs (Ground Control Points), UAV (Unmanned Aerial Vehicle), BFS (Before Feature Selection), AFS (After Feature Selection) & LOOCV (Leave-One-Out cross validation)

3.2.1. Remote Sensing Data

UAV photos were systematically collected alongside ground truth data between 12:15 pm and 01:00 pm local time, a period recognized for optimal crop photosynthetic activity. Clear, sunny weather prevailed during this window, reducing image distortions and mitigating wind interference (Wan et al., 2018). The DJI Matrice-300 drone was deployed with a MicaSense RedEdge-MX MS camera capable of capturing images across five spectral bands, including visible, red-edge and NIR. Table 2 presents information about the bandwidth and wavelength for each band, as specified by MicaSense (2017).

To establish flight parameters, Pix4DCapture software guided the drone's trajectory at a 60-meter above ground, positioning the sensor angle at 90°, resulting in the ground sampling distance (GSD) of ~ 2.96 cm/pixel. Table 1 clearly details all the pre-flight settings and conditions. To ensure comprehensive coverage, flight settings featured 85% frontal and 75% side overlap percentages. Reflectance calibration occurred prior to and subsequent to each flight using the MicaSense calibration reflectance panel (CRP) for necessary adjustments.

Human-operated piloting facilitated the capture of spectral photographs at an altitude above the controlled release point.

Table 1: Pre-Flight settings details

SETTING	SPECIFICATION
Time	12:15 – 13:00
Weather condition	Clear and sunny
Image Overlap	Front overlap: 75% Side overlap: 75%
Altitude	60m
Camera angle	90°
GSD	~ 2.96cm/pixel

Ground control points (GCPs) were strategically positioned along the flight route for precise geometric calibration, resulting in five collected GCPs, one at each corner of the field and the last at the centre. These GCPs were marked using a Trimble handheld GPS device with an impressive precision of 2 cm. This meticulous process ensured accurate geometric alignment and calibration of the captured photographs.

Table 2: The MS camera's centre wavelength and bandwidth for each spectral band.

Spectral band	Wavelength (nm)	Bandwidth (nm)
Blue	475	20
Green	560	20
Red	668	10
Red Edge	717	10
Near Infrared	840	40

3.3.2. Leaf Chlorophyll measurements

The total area of the tomato field was 1.60 hectares. However, the field was subdivided into subplots of 10 x 10 m, from which 22 random sample points were collected. This method was deliberately chosen for its ability to offer a well-rounded and representative sample of the entire study area.

A handheld SPAD-502 chlorophyll meter was utilized to quantify the chlorophyll concentration of the spectrally sampled leaves, as depicted in Figure 3.4. The SPAD-502Plus detected leaf absorbances in the red and near-infrared spectrums. Plants were randomly selected. The

chlorophyll content of a leaf was assessed by calculating the average LCC value across five distinct locations on the leaf. All leaf chlorophyll concentrations within each plot were averaged to calculate the mean chlorophyll concentration in leaves at the field level. Simultaneously, GPS points were recorded at the spots where LCC data was collected.



Figure 3.4: Field sampling: Tomato plant (left) and the SPAD-502Plus meter during ground measurement (right).

3.4. Image pre-processing

Radiometric calibration was performed on all the captured images, incorporating the before and after-flight CRP images to transform the original MS images into reflectance images. According to Brewer et al. (2022), the radiometric calibration panel is a white balance card that offers information regarding its reflectance characteristics across the whole range of light detected by the camera. This functionality allows the software to accurately calibrate and rectify the images' reflectance, accounting for prevailing atmospheric conditions at the time of image collection. Subsequently, the preprocessed reflectance images were merged to generate an orthomosaic, which was saved as a GeoTIFF file. To enhance the geolocation accuracy for subsequent image processing, the orthomosaic was georeferenced using five GCPs. Finally, the georeferenced orthomosaic was utilized to compute vegetation indices (VIs).

3.5. Vegetation Indices (VIs) calculations

Vegetation indices (VIs) are calculated by combining a minimum of two distinct spectral regions within the electromagnetic spectrum. The imaging platform employed in this study utilizes UAV technology to detect reflectance in the visible, red-edge, and NIR portions of the spectrum. Consequently, all potential combinations of UAV spectral bands were evaluated to

forecast crop chlorophyll levels precisely. Using the Google Earth Engine (GEE) platform, twenty commonly utilized VIs, crucial in estimating crop chlorophyll, were computed. Table 3 showcases the respective abbreviations and calculation formulas for each VI. These VIs were selected based on the reported correlation by other researchers with chlorophyll content, such as Brewer et al. (2022), Khoshrou et al. (2021), Nandibewoor et al. (2017) and Pérez-Patricio et al. (2018), among others. The sampling points were used to extract the values of the VIs for the estimation of LCC.

Table 3: Vegetation Indices (VIs) derived from UAV image.

VEGETATION INDEX	FORMULA	REFERENCES
NDVI (Normalized Difference Vegetation Index)	$\frac{NIR - R}{NIR + R}$	Rouse (1974); Xue and Su, (2017) ; Iizuka et al. (2019)
NGRDI (Normalized Green-Red Difference Index)	$\frac{G - R}{G + R}$	Gitelson et al. (2002) ; Wan et al. (2018)
VARI (Visible Atmospherically Resistant Index)	$\frac{G - R}{G + R - B}$	Gitelson et al. (2004) ; Wan et al. (2018)
GRVI (Green Ratio Vegetation Index)	$\frac{NIR}{G}$	Ban et al., (2022)
GNDVI (Green Normalized Difference Vegetation Index)	$\frac{NIR - G}{NIR + G}$	Gitelson et al. (1996); Wahab et al. (2018) ; Iizuka et al. (2019) ; Themistocleous (2019)
NDRE (Normalized Difference Red Edge Index)	$\frac{NIR - RE}{NIR + RE}$	Gitelson and Merzlak (1997); Barnes <i>et al.</i> (2000)
ARVI (Atmospherically Resistant Vegetation Index)	$\frac{NIR - (R - 2 * (B - R))}{NIR + (R - 2 * (B - R))}$	Gitelson <i>et al.</i> , (2002)
RVI (Ratio Vegetation Index)	$\frac{NIR}{R}$	Huete <i>et al.</i> , (2002)

GCI (Green Chlorophyll Index)	$\frac{NIR}{G} - 1$	Gitelson <i>et al.</i> , (2005); Zhang and Zhou (2015)
RECI (Red-Edge Chlorophyll Index)	$\frac{NIR}{RE} - 1$	Gitelson <i>et al.</i> , (2005); Zhang and Zhou (2015)
SAVI (Soil Adjusted Vegetation Index)	$\frac{1.5 (NIR - R)}{NIR + R + L}$ L is a constant between 0 and 1.	Huete (1988); Xue and Su (2017)
EVI (Enhanced Vegetation Index)	$\frac{2.5 (NIR - R)}{NIR + 6R - 7.5B + 1}$	Martina <i>et al.</i> , (2018)
MCARI (Modified Chlorophyll Absorption in Reflectance Index)	$\frac{1.5 [2.5(NIR - R) - 1.3(NIR - G)]}{\sqrt{(2NIR + 1)^2 - (6NIR - 5\sqrt{R})} - 0.5}$	Daughtry <i>et al.</i> , (2000); Wu <i>et.</i> (2008)
LCI (Leaf Chlorophyll Index)	$\frac{NIR - RE}{NIR + RE}$	Eitel <i>et al.</i> (2011)
NPCI (Normalized Pigment/ Chlorophyll Index)	$\frac{RE - B}{RE + B}$	Peñuelas <i>et al.</i> (1994); Albetis <i>et al.</i> (2019)
DVI (Difference Vegetation Index)	$NIR - R$	Imran <i>et al.</i> (2020)
ExR (Excess Red)	$\frac{1.4R - G}{R + G + B}$	Narmilan <i>et al.</i> (2022)
ExG (Excess Green)	$\frac{2G - R - B}{R + G + B}$	Narmilan <i>et al.</i> , (2022)
ExGR (Excess Green Red)	$ExG - ExR$	Narmilan <i>et al.</i> (2022)
TVI (Triangular Vegetation Index)	$60(NIR - G) - 100(R - G)$	Eitel <i>et al.</i> (2011)
OSAVI (Optimized Soil-Adjusted Vegetation Index)	$\frac{NIR - R}{NIR + R + 0.16}$	Xue and Su (2017)

Note: B refers to the blue band, G to the green band, R to the red band, NIR to the near-infrared, and RE to the red edge.

3.6. Feature Selection

Feature Selection is a powerful technique that eliminates irrelevant variables, thereby reducing the dimensionality of input data (Otchere et al., 2022). It selects a subset from the entire feature set, streamlining the dimensionality and enhancing the performance and interpretability of regression models by eliminating unnecessary and redundant features (Kvaktun et al., 2022).

This study used Correlation Analysis to assess the linear association between Vegetation Indices (VIs) and LCC. After data collection and preprocessing, the correlation between the VIs and LCC was computed. Specifically, Pearson's correlation coefficient was employed to quantify the strength of linear relationships and select features that highly correlate with the LCC. It is important to note that correlation indicates the simultaneous change in both variables but does not necessarily imply a cause-and-effect relationship (Asuero et al., 2006; Joshi et al., 2023). Angel and McCabe (2022) employed Pearson's correlation coefficient to identify potential multicollinearity within predictor sets, while Narmilan et al. (2022) utilized Pearson's correlation for feature selection.

3.6.1. Feature importance and selection

The Gini index is a widely used method for selecting features most relevant to the predictions, proving particularly effective in Classification and Regression Trees (CART), especially when dealing with high-dimensional features (Bouke et al., 2023). In the RF model, determining the importance of features involves constructing a classification and regression tree to create out-of-bag (OOB) sample data. Using this OOB data, the RF model assesses the significance of input features and assigns an importance score to each feature, represented by the mean decrease in accuracy (Zhao et al., 2022). Therefore, the Gini index integrated into the RF model was utilized in this study to identify potentially prognostic predictor variables for estimating tomato LCC (Menze et al., 2009). As a metric, the Gini index quantifies the significance of a variable for predictive accuracy. It is derived during the model training process, with higher Gini index values indicating increased significance for the model's predictive performance (Simpson et al., 2022).

3.7. Modelling Methods

In this study, Machine Learning (ML) models, including PLSR, DTR, RFR and XGB, were tested to estimate LCC in tomato crops. These models were built using the Leave-One-Out cross-validation (LOOCV) technique. In LOOCV, each data point in the dataset is left out once for validation, while the remaining data is utilized for training, creating a separate model for each data point.

The study also investigated which VIs are most influential in the estimation of LCC in tomato plants. The process entails five steps: selecting features, learning the model, cross-validation and rating the performance of the model and mapping the LCC predictions. Using Pearson correlation matrix, any multicollinearity between individual VIs and LCC values were visualized. LCC data was used as response variable (y) and VIs were used as predictor variables (X).

3.7.1. Partial-Least Squares Regression (PLSR)

PLSR is a multivariate statistical method frequently employed in constructing predictive models, especially when dealing with a substantial number of highly correlated or collinear explanatory variables (Maimaitijiang et al., 2020). It integrates various statistical techniques, including correlation analysis, canonical correlations and principal component analysis, to effectively tackle data collinearity issues (Tao et al., 2020; Ji et al., 2023). It can simplify a complex set of collinear factors into a smaller orthogonal principal component. These components are uncorrelated variables that capture the maximum variance in the data. This transformation to orthogonal principal components helps prevent overfitting the model to the data and eliminates redundant information.

Essentially, PLSR transforms the original data space into a new space where the variables (principal components) are uncorrelated. This makes the modelling process more stable and reliable. Selecting the number of principal components to retain in the model is a crucial step in PLSR. This number can significantly impact the model's predictive performance and interpretability. Therefore, the LOOCV technique was used. Several Studies have argued that PLSR outperforms alternative regression methods such as the MLR (Jin & Wang, 2019).

3.7.2. Decision Trees regression (DTR)

A Decision Tree (DT) is a prediction model representing the mapping relationship between the qualities of objects and their respective values (Gao et al., 2023). Each node represents an object within the tree structure, while the branches delineate various properties. This hierarchical arrangement organizes conditions sequentially, starting from the root of the tree and progressing to its leaf (Breiman et al., 2017). The DT algorithm's fundamental premise is to partition a complex problem into numerous smaller ones, leading to a more comprehensible solution (Xu et al., 2005).

In practical terms, building a DT model involves recursively partitioning training set data and conducting multiple regressions at each node. This iterative process progressively splits the data until a predefined stopping point is reached (Rodriguez-Galiano et al., 2015). At the leaf nodes of the trained tree, simple regression models are employed exclusively for that particular segment (Rodriguez-Galiano et al., 2015).

3.7.3. Random Forest regression (RFR)

Random Forest (RF) is an ensemble learning technique that extends from DT and integrates Breiman's "bagging" concept and random feature selection (Pei et al., 2017; Singhal et al., 2019). It generates multiple samples from the original dataset, each being used to construct a decision tree. These trees are then aggregated to produce predictions, with the final prediction determined through a majority voting process.

Renowned for its robust resistance to overfitting and its ability to retain information, RF consistently achieves convergence of generalization error. This has led to its widespread use in chlorophyll studies (Singhal et al., 2019; An et al., 2020). The regression variant of this algorithm, RFR, has shown remarkable efficiency in estimating rice chlorophyll levels. An et al. (2020) highlighted the superior performance of RFR, which outperformed SVR, Gaussian Process Regression, and Gradient Boosting Regression Tree models, achieving a strong R^2 of 0.80 for the validation set.

The performance of RF is enhanced by the randomization of splitting rules (Ishwaran, 2015). To achieve optimal results, three parameters: *ntree* (number of trees), *node size* (the minimum size of terminal nodes), and *mtry* (the number of variables randomly sampled as candidates at each split) - were optimized using the LOOCV technique. This process ensures the RF model is fine-tuned to provide accurate and reliable predictions.

3.7.4. eXtreme Gradient Boosting (XGB)

XGB, introduced by Chen and Guestrin in 2016, revolutionized the machine-learning landscape with its exceptional efficiency and adaptability (Chen & Guestrin, 2016). This technique, rooted in Quinlan's Decision Tree algorithm, continuously refines its predictions by sequentially combining weak predictive tree models (Hastie et al., 2009). Its outstanding performance in many regression tasks solidifies its reputation as a versatile and robust tool (Kganyago et al., 2022; Narmilan et al., 2022; Joshi et al., 2023).

However, XGB's strength extends beyond its sequential model-building approach. Employing gradient-boosted decision trees and a refined formulation avoids overfitting while efficiently managing missing or sparse data (Kganyago et al., 2022). Moreover, its adeptness in utilizing parallel and distributed computing streamlines the handling of large models, showcasing its adaptability and computational efficiency (Kganyago et al., 2022; Narmilan et al., 2022; Joshi et al., 2023). Additionally, it harnesses parallel and distributed computing for swift tree construction and handling of large models (Kganyago et al., 2022).

In practice, the model initiates with an initial guess, often the mean of the training set's measured values and introduces a learning rate known as the shrinkage parameter (Kganyago

et al., 2022). This parameterization ensures a smooth progression in model improvement by integrating new trees, effectively preventing overfitting, a common hurdle in machine-learning.

3.8. Cross Validation

Cross-validation (CV) is a statistical analytic technique for evaluating the effectiveness of a regression model (An et al., 2020). CV approach is employed to choose models and optimise parameters to prevent the model from overfitting. This study chose the LOOCV technique to evaluate machine learning (ML) models, which is particularly advantageous when working with smaller datasets. Veiga et al. (2021) and Cao et al. (2023) highlighted that ML-based prediction models tended to overfit the data. Hence, employing the CV technique is imperative to minimize model inaccuracies. Additionally, when the sample size is below 30, Xu et al. (2023) recommend LOOCV as the preferred evaluation method.

LOOCV divides observations into two distinct parts, as James et al. (2023) highlighted. LOOCV designates a single observation (X_1, y_1) to the validation set, while the remaining observations $\{(X_2, y_2), \dots, (X_n, y_n)\}$ constitute the training set. The statistical learning method is trained on the $n - 1$ training observations, predicting the excluded observation's value $(\hat{y}_1)$ by utilizing its X_1 value. This iterative process recurs n times, aligning with the number of observations within the dataset.

The resultant LOOCV estimate for the Mean Squared Error (MSE) test is derived by averaging these n -test error estimates, providing a robust measure of the model's predictive performance. Employing LOOCV eliminates overfitting issues, enabling the model to exhibit a higher potential for generalization to unseen data. Equation 3.1 delineates the calculation methodology adopted by LOOCV to determine the MSE.

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i \quad (3.1)$$

3.9. Evaluation of Models

The performance evaluation of ML models (PLSR, RFR, DTR, and XGB) was assessed and compared using statistical metrics, namely mean absolute error (MAE), root mean square error (RMSE), and the determination of coefficient (R^2). These metrics were computed at a significance level of 0.05.

The RMSE is a metric that quantifies the average magnitude of errors between measured (y_i) and predicted/estimated ($\hat{y}_i$) values, providing an overall measure of prediction accuracy. A lower RMSE indicates a smaller deviation between the actual (measured) and predicted values, reflecting higher precision in the model's estimations (T. Wang et al., 2022). In contrast,

MAE calculates the average absolute variances between the measured and predicted values, representing the typical error in the model's predictions.

Furthermore, R^2 quantifies the proportion of variability in the dependent variable explained by the independent variables (Wu et al., 2023). It measures how well the model aligns with the measured data. As the R^2 value increases, the model's ability to make accurate predictions strengthens, indicating a higher level of explanatory power in the model. The equations for calculating the RMSE, MAE and R^2 are provided below.

$$RSME = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3.2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3.3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.4)$$

'n' represents the quantity of samples, while ' $\hat{y}_i$ ' and ' y_i ' are the predicted and measured parameters. $\bar{y}$ is the mean values of the measured LCC values.

3.10. Summary

The chapter outlines the methodology employed in this study. The study adopted a Stratified Random Sampling technique to ensure comprehensive data collection of the reference data. The study utilized UAV imaging platforms and ground control points for data acquisition, and the captured images underwent radiometric correction and orthomosaic creation. Additionally, the study involved the calculation of twenty VIs using the Google Earth Engine (GEE) platform. The chapter further details the use of ML models, such as PLSR, DTR, RFR and XGB, for estimating leaf chlorophyll content in tomato crops, along with the process of feature selection, model learning, cross-validation, and performance evaluation. Additionally, the geographic landscape and features of the study area, along with specific techniques involved in data collection and analysis, were also described in the chapter.

CHAPTER 4

RESULTS AND DISCUSSION

4.1. Descriptive statistics of the measured Leaf Chlorophyll Content (LCC)

Figure 4.1 presents a boxplot that depicts the distribution of the LCC data. The boxplot provides a five-number summary, which includes the minimum, first quartile (Q1), median, third quartile (Q3), and maximum. The centerline of the box represents the median, which is the dataset's middle value when sorted in ascending order. The upper and lower edges of the box represent the Q1 and Q3, respectively. The whiskers of the boxplot extend from the box to the maximum and minimum values within the dataset. These whiskers provide a comprehensive view of the spread of the LCC data, providing a comprehensive view of the spread and central tendency of the LCC data.

Table 4 shows the descriptive statistics summary of the measured tomato LCC, providing more details to complement Figure 4.1. In this study, 22 samples were analyzed, and the results, detailed in Table 4, revealed that LCC data ranged from a minimum of $14.30\mu\text{g}/\text{cm}^2$ to a maximum of $55.60\mu\text{g}/\text{cm}^2$. The calculated mean value, obtained by dividing the sum of all the values by the total number of samples, was $40.28\mu\text{g}/\text{cm}^2$. This gives us a general idea of the central location of the data. However, the median, representing the central value separating the upper and lower halves of the dataset, exhibited a significantly higher value of $46.10\mu\text{g}/\text{cm}^2$.

The significant difference between the mean and median suggests a lack of symmetry in the data, indicating that the LCC data does not follow a normal distribution. This asymmetry is further illustrated in the boxplot, where the distance from the first quartile (Q1) to the median is larger than from the median to the third quartile (Q3). This pattern signifies a distribution tilted to the right, with the right tail being longer or thicker than the left.

The standard deviation, which quantifies the degree of variability or spread within the dataset, was calculated as 14.20. It reveals how spread out the values are from the mean. The interquartile range (IQR), another measure of statistical dispersion, was found to be 21.58. The IQR, calculated as $Q3 - Q1$, gives us the range within which the central 50% of the values fall. In this case, Q1 and Q3 were calculated as 28.60 and 50.18, respectively. This further emphasizes the spread and skewness of the LCC data.

Table 4: Statistical characteristics of tomato LCC in Mphaila irrigation scheme

Analysis Variable : LCC

N	Minimum	Mean	Std Dev	Q1	Median	Q3	IQR	Maximum
22	14.30	40.28	14.20	28.60	46.10	50.18	21.58	55.60

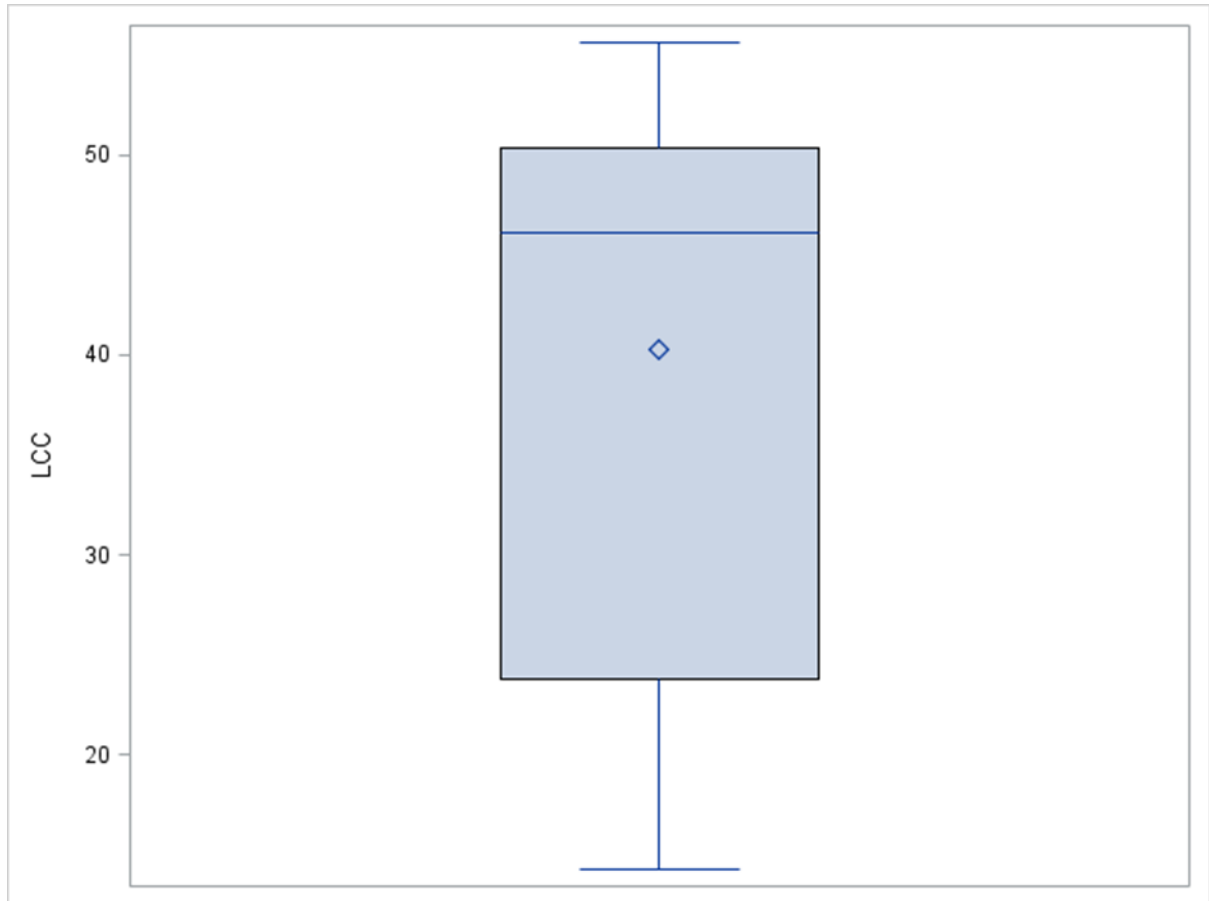


Figure 4.1: Boxplot of the LCC values. The centerline of the box represents the median of the data, the diamond symbol represents the mean, the upper and lower lines represent the 1st and 3rd quantiles, and the two upper and lower whiskers represents the maximum and minimum values of the data, respectively.

4.2. Correlation Analysis between Vegetation Indices and chlorophyll content

One of the objectives of this study was to determine which VIs were most sensitive and responsive to changes and variations in the LCC of tomatoes. A Pearson's correlation coefficient test was performed to achieve this, and the results are presented in *Table 5* and Figure A1.

Table 5 reveals that GNDVI, GRVI and GCI responded more favourably to LCC. GNDVI and GRVI exhibited the most robust positive correlation among the various VIs, with an r value of 0.67. The GCI is closely behind, with a correlation coefficient of 0.66. In contrast, several other VIs, such as NDVI, SAVI, RVI, DVI, OSAVI, ExGR, ExG, TVI, and NGRDI, displayed positive correlation coefficients below the 0.50 threshold, suggesting a comparatively weaker relationship. For instance, NDVI, OSAVI, SAVI, and NPCI followed a similar pattern, presenting r values around 0.42. RVI has an r value of 0.41. Meanwhile, the correlation coefficients for DVI and MCARI were 0.36, while ExGR and ExG demonstrated a similar r value of 0.32. TVI, on the other hand, had an r value of 0.29.

Interestingly, certain VIs, namely the NDRE, RECI, LCI, ARVI, VARI and EVI, displayed negative correlation coefficients. This indicates an inverse relationship with LCC, meaning that as the LCC level increases, the values of these VIs decrease.

The study's results offer significant insights into identifying the most sensitive VIs for estimating tomato LCC, which is pivotal for accurately estimating LCC and facilitating plant health monitoring. Consequently, these findings can improve agricultural practices, making them more effective and efficient.

Table 5: Pearson's Correlation coefficients between VIs and tomato LCC, at 0.05 significance level.

VEGETATION INDEX	CORRELATION COEFFICIENT
NGRDI	0.053
OSAVI	0.42
EXGR	0.32
EXG	0.32
EXR	-0.18
DVI	0.36
TVI	0.29
MCARI	0.36
RECI	-0.53
NPCI	0.42
GRVI	0.67
VARI	-0.029
LCI	-0.49
GNDVI	0.67
NDVI	0.42

EVI	0.0019
NDRE	-0.52
RVI	0.41
SAVI	0.42
GCI	0.66
ARVI	0.066

4.3. Feature Selection/ Importance

The RF and GB models incorporate feature selection mechanisms to assess the relevance of VIs predictors in predicting the LCC. Figure 4.2, a horizontal bar graph, illustrates the importance of each predictor variable (VIs) in the models on the y-axis against their significance on the x-axis. The x-axis depicts the significance of the variables based on the Gini index, ranging from 0 to 0.40. The graph (Figure 4.2) aids in identifying variables crucial for predicting LCC.

Both models identified GNDVI, GCI and GRVI as the most important predictor variables, albeit with varying ranks. Specifically, RF prioritized GNDVI, although the difference is insignificant, while GB favoured GCI, showing a significant difference from the other predictor variables. Conversely, MCARI was consistently the least impactful across both techniques.

During the phase of OOB calculation, the average reduction in accuracy caused by each predictor is determined. This process involves using a single variable to divide a node, calculating the Gini uncertainty for the resulting child nodes and comparing it with the output factor node.

In ML, feature importance assigns scores to input features based on their contribution to predicting the response variable. These scores elucidate each feature's impact on the model's predictive power, with higher scores denoting greater significance in achieving accurate predictions. Consequently, these scores guide the selection of relevant features, augmenting the model's performance and interpretability. Figure 4.2 reinforces the significance of GNDVI, GCI and GRVI in the models.

In this study, the selection of predictor variables for the models was guided by two criteria: VIs with a strong correlation with LCC and a feature importance threshold based on the mean decrease in Gini uncertainty exceeding 0.15, which shows how much each predictor helps to improve the model's predictions. The correlation analysis highlighted a substantial relationship between GRVI, GNDVI, and GCI with LCC. These indices proved highly impactful and substantially relevant to predicting LCC, as shown in *Table 5* and Figure 4.2. Subsequently,

these variables were employed for training and testing the models following the feature selection process.

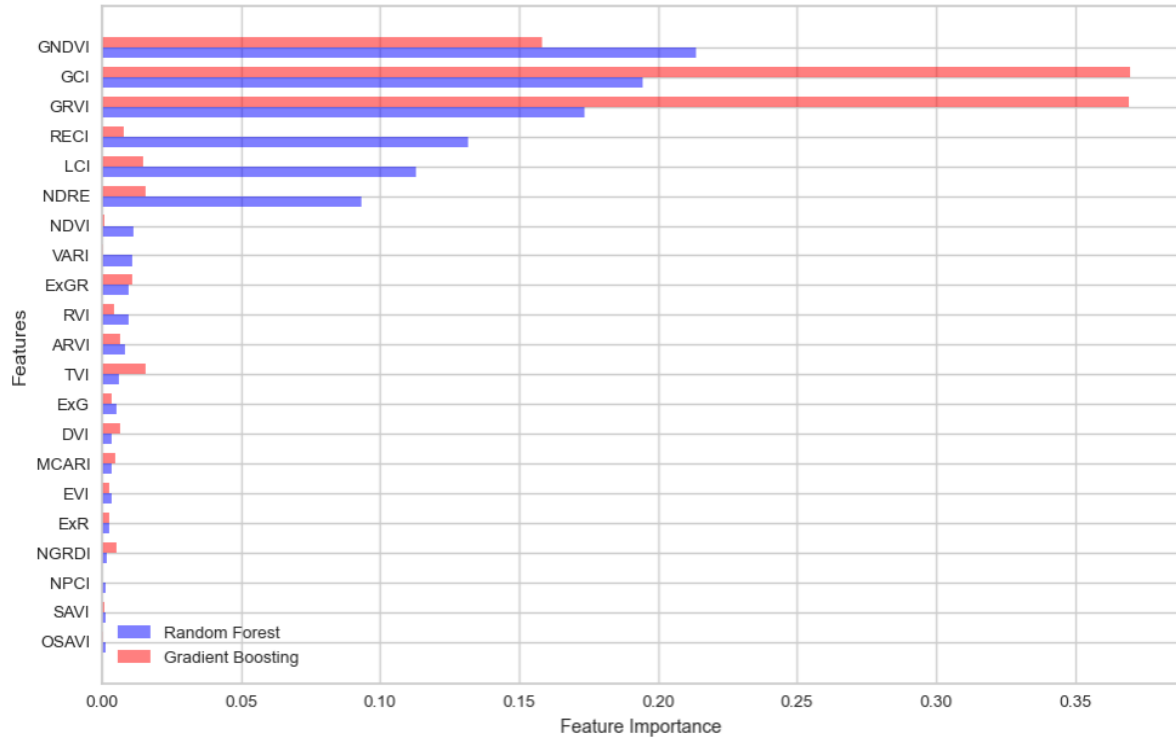


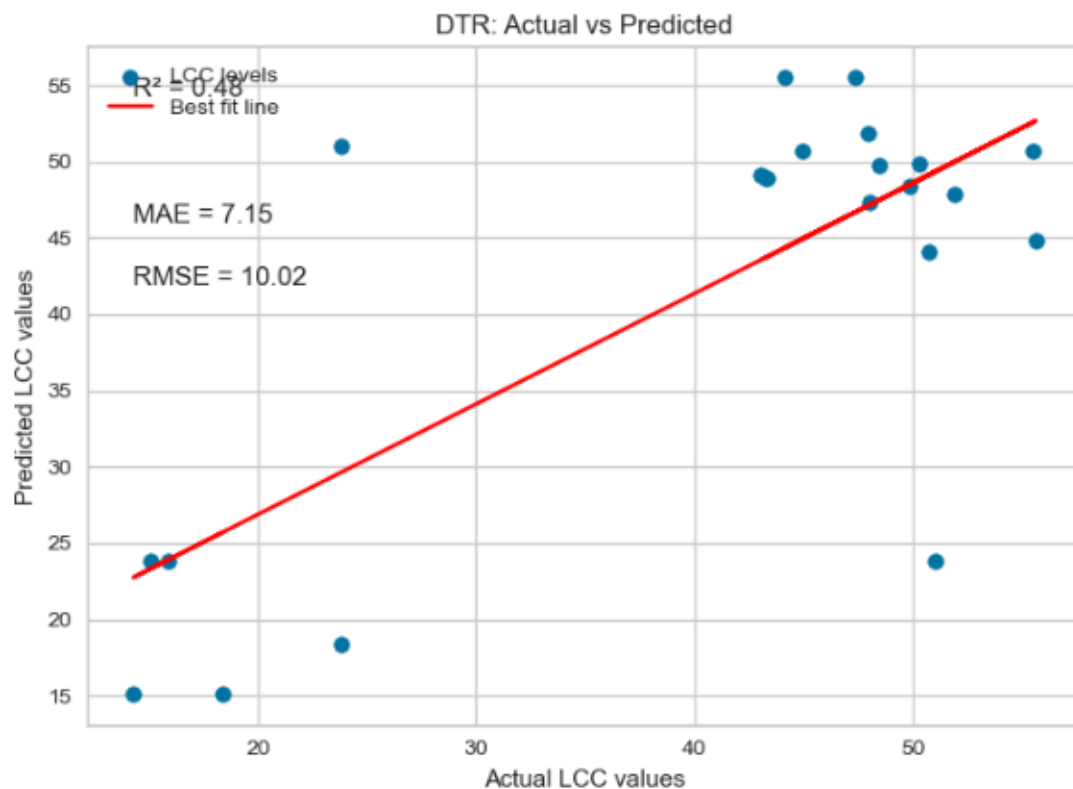
Figure 4.2: Predictor variable importance score/ contribution in predicting the response (LCC) variable based on the Gini index.

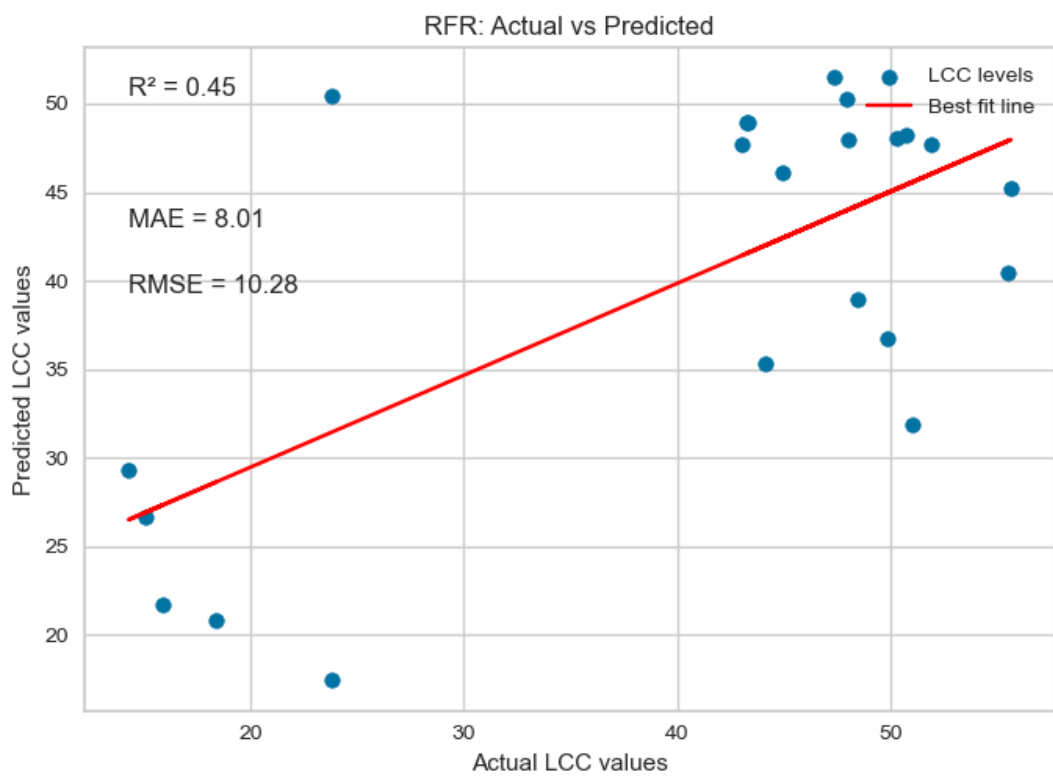
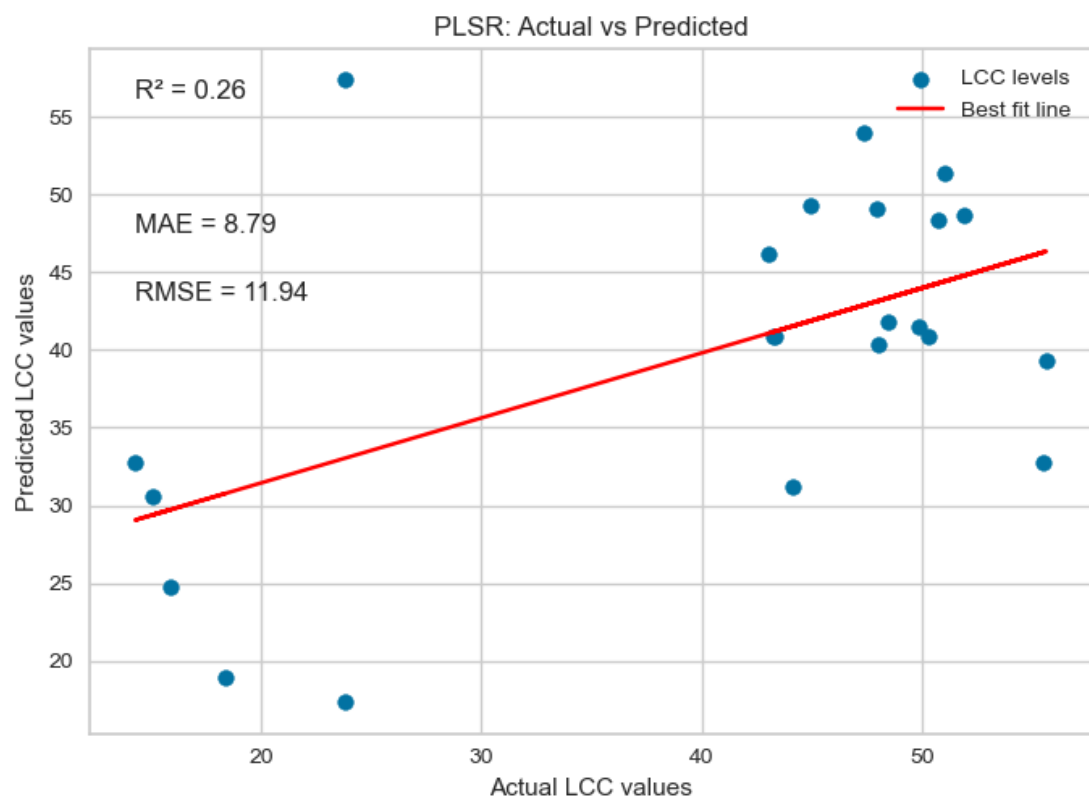
4.4. Estimation of Tomato Chlorophyll Content by using ML models

In this study, four ML models, including PLSR, DTR, RFR and XGB, were developed and evaluated. The development and evaluation of these models were characterized into two phases: Before Feature Selection (BFS) and After Feature Selection (AFS). During the BFS phase, models were trained using all twenty vegetation indices (VIs). In contrast, in the AFS phase, models were trained using only the most influential VIs for tomato LCC based on Pearson's correlation coefficient (as detailed in *Table 5*) and the feature importance derived from RF and GB techniques (as depicted in Figure 4.2). Under both BFS and AFS approaches, the performance of these models was compared based on statistical metrics such as R^2 , MAE and RMSE, as evidenced in Figure 4.3 and Figure 4.4.

Figure 4.3 and Figure 4.4 present best-fit line plots, a visual representation demonstrating the comparison of all ML models employing both BFS and AFS approaches. These lines accurately depict the relationship between data points and accurately represent the relationship between the data points. Each red line represents the optimal regression line for the respective model.

Figure 4.3 highlights that the XGB model exhibited an R^2 value of 0.53, making it the only model that exceeded an R^2 threshold of 0.50 during the BFS phase. Following closely are the DTR and RFR, which achieved R^2 values of 0.48 and 0.45, respectively. Conversely, the PLSR demonstrated an R^2 of 0.26, making it the lowest performing model. The MAE for all models, reflecting their average absolute prediction errors, ranged between $6.39\mu\text{g}/\text{cm}^2$ and $8.79\mu\text{g}/\text{cm}^2$. The RMSE of these models varied between $9.54\mu\text{g}/\text{cm}^2$ and $11.94\mu\text{g}/\text{cm}^2$.





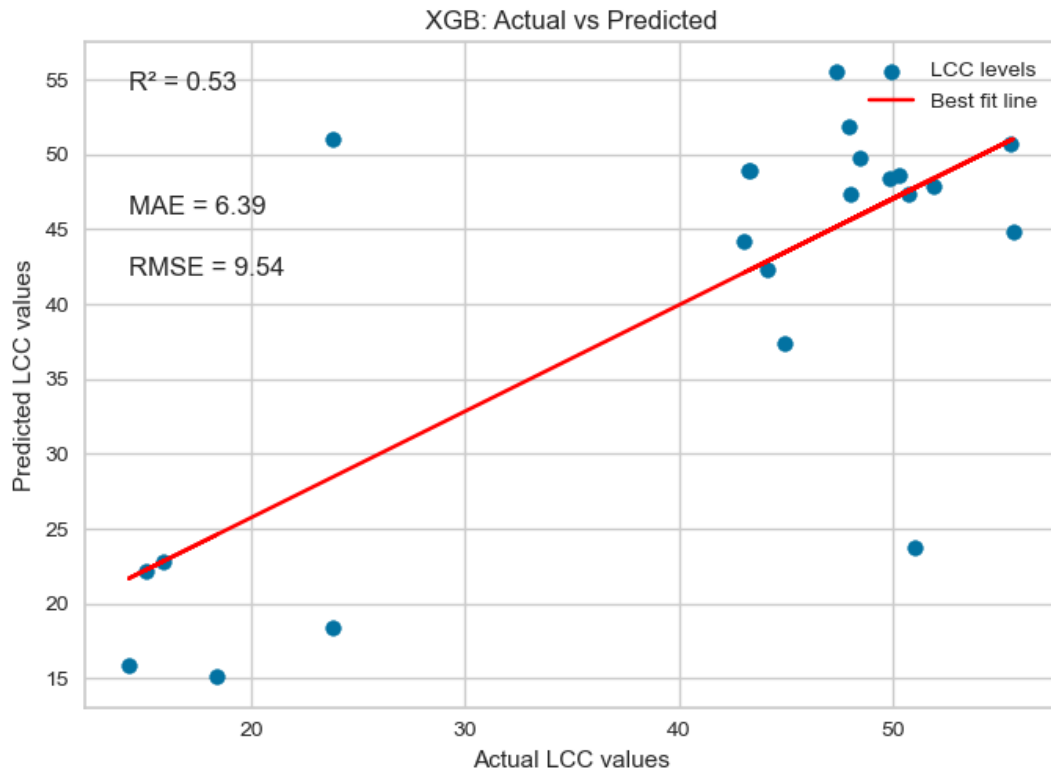


Figure 4.3: Scatter plots depicting the correlation between observed and predicted LCC values and the performance level of the ML models ((a)DTR, (b)PLSR, (c)RFR and (d)XGB), accompanied by the best-fit line **Before Feature Selection**.

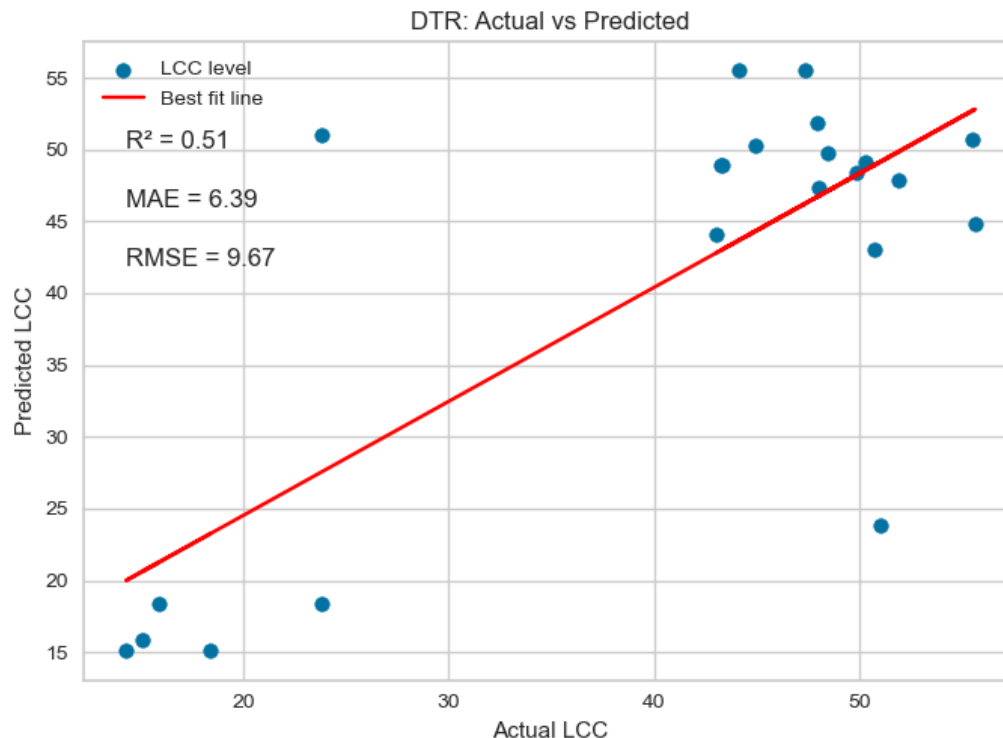
This study identified three VIs, GRVI, GNDVI and GCI, as the most influential predictor variables in predicting LCC. These VIs were selected based on Pearson's Correlation Coefficients, Random Forest, and gradient-boosting screening techniques. These selected VIs were then utilized as predictor variables to establish regression models for tomato LCC estimation, marking the second phase of the study, referred to as After Feature Selection (AFS).

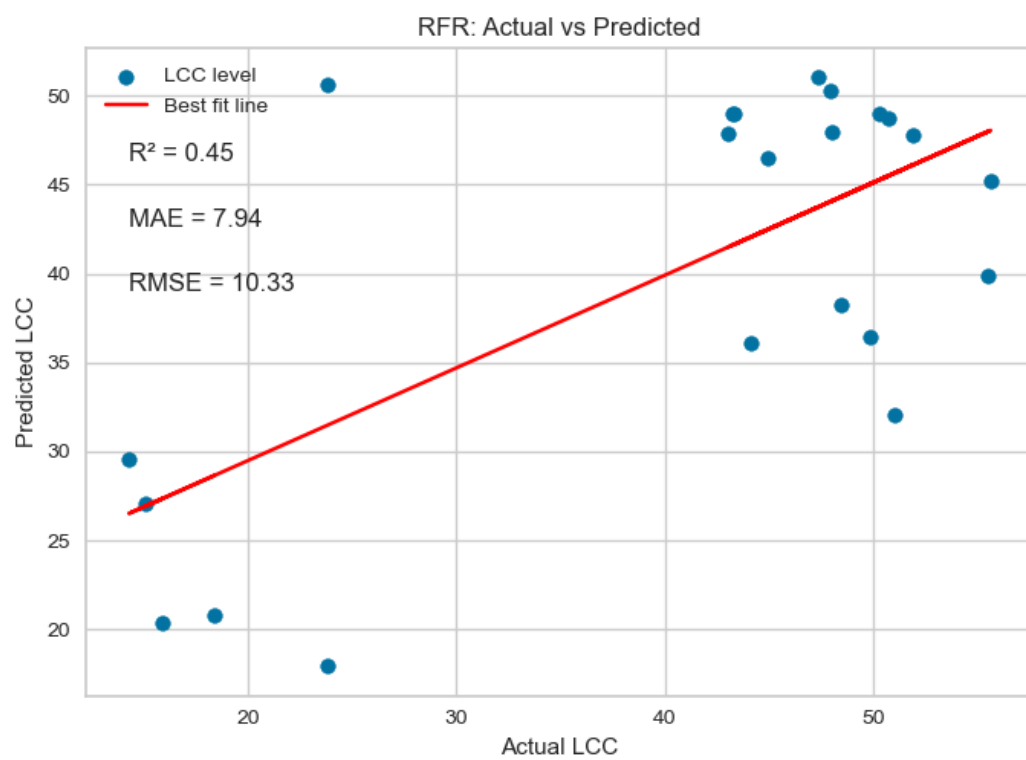
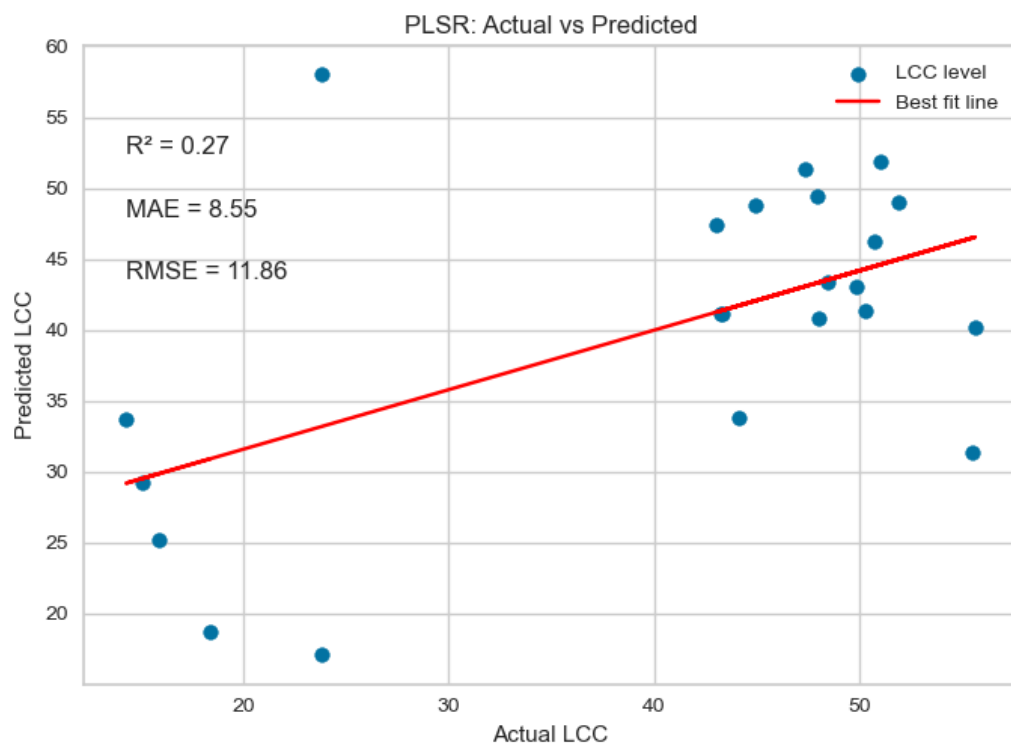
Figure 4.4 illustrates the models' goodness of fit, as indicated by the coefficient of determination (R^2). Among the models, XGB emerged as a top performer, showcasing a relatively superior R^2 of 0.54. Additionally, it demonstrated an MAE of 6.04 and an RMSE of $9.39\mu\text{g}/\text{cm}^2$. Following closely, the DTR exhibited R^2 values of 0.51, with MAE and RMSE values of $6.39\mu\text{g}/\text{cm}^2$ and $9.67\mu\text{g}/\text{cm}^2$, respectively. Interestingly, the R^2 of the RFR model remained unchanged across both phases (BFS and AFS).

The R^2 value of the XGB model implies that the model explains approximately 54% of the variability in the target variable (LCC), indicating a moderate level of predictive power. The MAE of $6.04\mu\text{g}/\text{cm}^2$ denotes the average absolute disparity between the actual and predicted LCC values, indicating that, on average, predictions deviate by approximately 6.04 units from the actual values. Meanwhile, the RMSE of $9.36\mu\text{g}/\text{cm}^2$ represents a measure of discrepancies between actual and predicted values, emphasizing more significant errors due to the squared differences.

All the models improved during the AFS phase except for the RFR model, which showed the same R^2 value in both phases. However, other statistical metrics, such as RMSE and MAE, showed slight variations between the two phases. For instance, the RMSE value increased from 10.28 to $10.33\mu\text{g}/\text{cm}^2$, showing a difference of 0.05, whereas the MAE value decreased from 8.01 to $7.94\mu\text{g}/\text{cm}^2$, showing a decrease of 0.07.

XGB consistently outperformed in the BFS and AFS methods, maintaining an R^2 above 0.50. It showcased its exceptional and reliable predictive capabilities across diverse data configurations, solidifying its status as the consistently superior-performing model in the analysis.





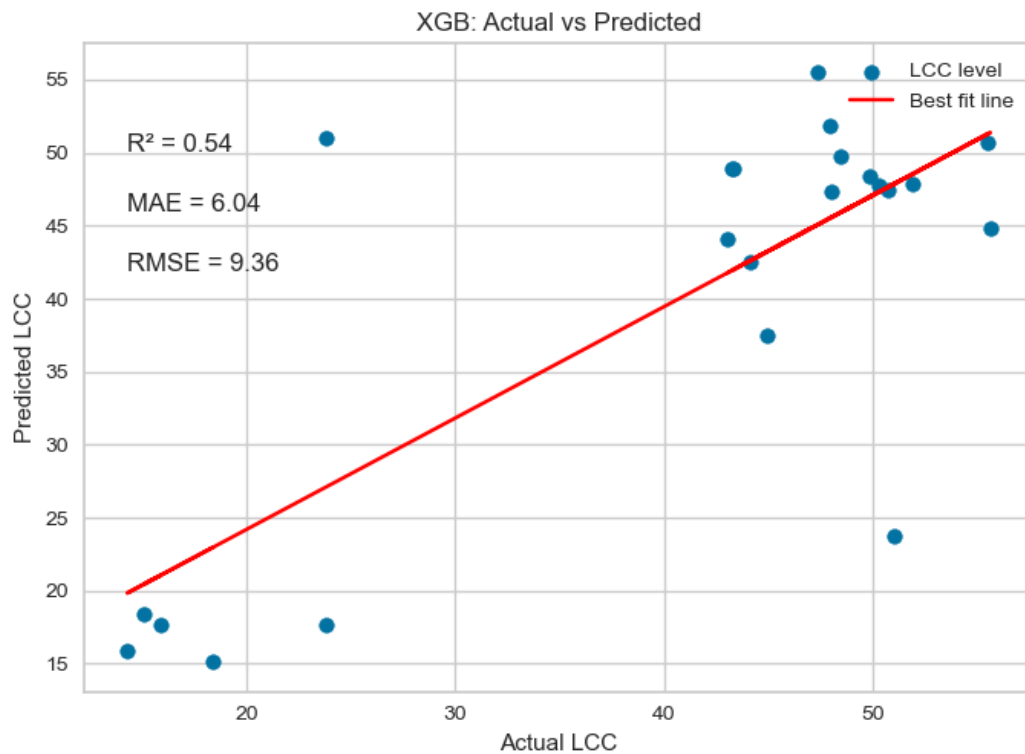


Figure 4.4: Scatter plots depicting the correlation between observed and predicted LCC values and the performance level of the various ML models((a)DTR, (b)PLSR, (c)RFR and (d)XGB), accompanied by the best-fit line **After Feature Selection**.

4.5. Mapping the Spatial Distribution of Tomato LCC using ML models

Spatial distribution maps were created using equations derived from the selected features following the feature selection process. Figure 4.5 illustrates the modelled LCC utilizing four distinct machine learning models: DTR, RFR, PLSR, and XGB.

Examining Figure 4.5 reveals a subtle variation between the minimum and maximum LCC values across all four models. Specifically, the LCC values derived from the DTR model spanned from $8.30\mu\text{g}/\text{cm}^2$ to $12.55\mu\text{g}/\text{cm}^2$, while those generated by the XGB model ranged from $8.59\mu\text{g}/\text{cm}^2$ to $12.68\mu\text{g}/\text{cm}^2$. Notably, these two models exhibited nearly identical ranges of values, emphasizing their comparable performance in predicting LCC.

In contrast, the LCC values obtained from the PLSR model showed a range between $23\mu\text{g}/\text{cm}^2$ and $25.26\mu\text{g}/\text{cm}^2$. Conversely, for the RFR model, the LCC values varied from $18.83\mu\text{g}/\text{cm}^2$ to $21.62\mu\text{g}/\text{cm}^2$. This divergence in range highlights distinct characteristics in the predictions made by PLSR and RFR, underscoring the nuanced differences in their respective modelling approaches and outcomes.

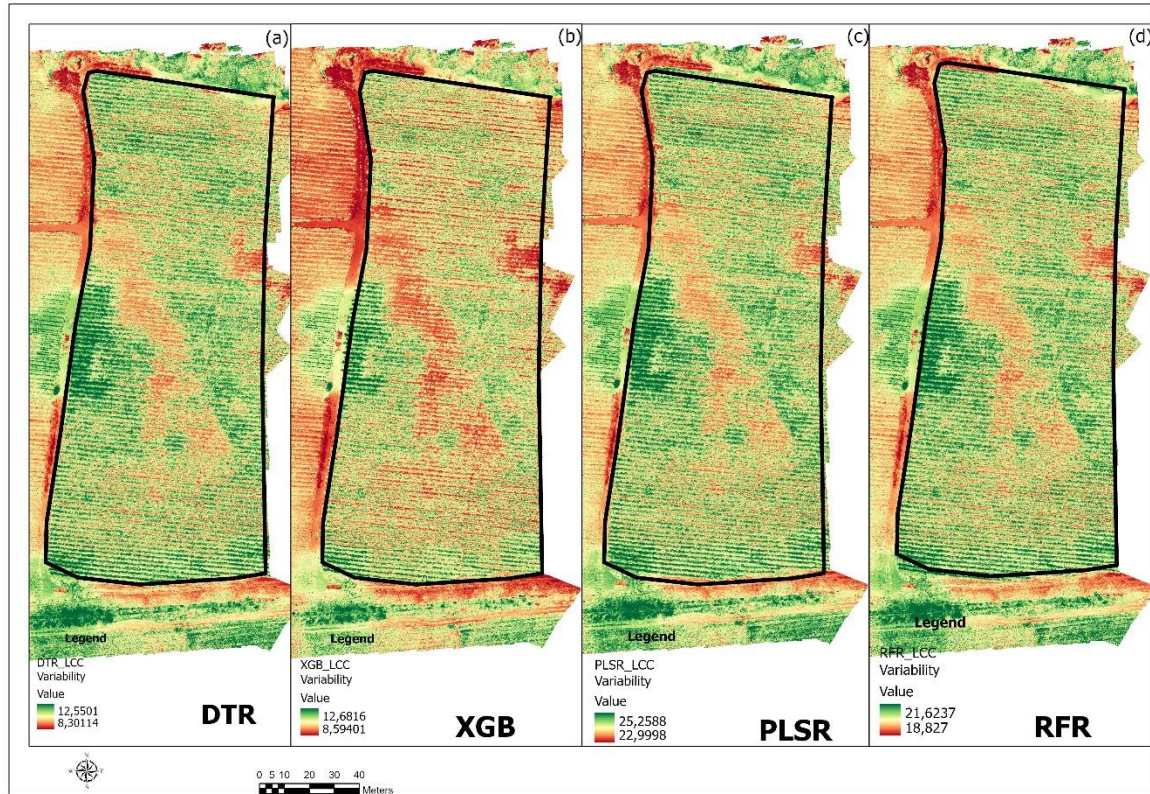


Figure 4.5: Spatial distribution of LCC across a tomato field, employing four distinct ML models, namely DTR, XGB, PLSR and RFR.

4.6. DISCUSSION

The LCC plays a pivotal role in a plant's health. Variations in the LCC can indicate various aspects of a plant's condition, including its photosynthetic capacity, growth phase, nutritional status, and environmental stress levels (Sonobe & Wang, 2017). Moreover, LCC changes can also indicate diseases and pest infestations (Hasan et al., 2023). Multispectral (MS) RS has emerged as a non-invasive technique for estimating LCC. This method has the potential to provide detailed data on the unique spectral reflectance properties of plants, effectively distinguishing them from other elements on the ground, such as water and soil (Dian et al., 2016). Rapid access to information about crop growth conditions in farmland is a crucial component of smart agriculture. It facilitates fast sensing and informed decision-making, enhancing agricultural efficiency and productivity (Wu et al., 2023).

This study computed VIs from orthomosaic derived from UAV to construct LCC estimation models using ML algorithms. Four ML modelling techniques were tested: RFR, DTR, PLSR and XGB. These models were constructed using a complete set of predictor variables (i.e., before feature selection) and a subset of those predictor variables (after feature selection).

4.6.1. Relationship between VIs and Chlorophyll

Pearson correlation coefficients were used in this study to compare various VIs with the LCC data collected in the study area. VIs, known to correlate with agronomic features, have been recognized as significant substitutes for conventional agronomic indicators (Wu et al., 2023). This study demonstrated a moderate correlation between the LCC of tomatoes and VIs such as GRVI, GNDVI and GCI. These VIs, as depicted in *Table 5*, were identified as the most optimal predictors for estimating tomato LCC in this study. Notably, the variables derived from NIR and Green wavebands consistently scored highest in feature importance screening (see Figure 4.2). Existing Literature notes that during the mid-vegetative growth phase, plants exhibit reduced sensitivity to the red band and increased sensitivity to the green band (Brewer et al., 2022). This observation elucidates why VIs derived from the green band were highly significant in estimating LCC.

The relationship observed in this study is consistent with prior research, which identified a direct linear correlation between VIs and LCC data. For instance, Ban et al. (2022) reported a linear relationship between GNDVI and GRVI with LCC readings in a rice crop in Ningxia, China. However, they reported a much higher positive linear relationship (0.61 and 0.59, at $p < 0.001$) than observed in this study. The sensitivity of GCI to chlorophyll contributes to accurate predictions of chlorophyll deficiencies and the detection of stress in tomato plants.

The strong relationship between LCC and VIs allows for the inference of LCC from these indices. The absorption of light by chlorophyll and other pigments in the visible spectrum and the scattering of light from leaves in the NIR band generate a distinct reflective feature. This feature, particularly the robust link between the vegetation index from the red edge and LCC, enables effective capture and estimation of chlorophyll content through RS methods (Dian et al., 2016). Wu et al. (2023) further contributed to this understanding by noting that the correlation between LCC values and VIs gradually decreases as the reproductive process progresses. They proposed that this phenomenon could be attributed to the crop's nutritional growth phase, where a significant portion of absorbed nitrogen is deposited in nutritional organs like leaves. Consequently, this results in heightened sensitivity of spectral reflectance.

However, the quest for a universally applicable index for all plant species remains elusive, as highlighted by Sonobe and Wang (2017). In their research on beech trees, they reported a strong linear relationship ($r = 0.64$, $p < 0.001$) between MCARI and the LCC, notably surpassing the correlation reported in the present study ($r = 0.36$). This underscores the variability in the linear relationship between LCC and different VIs across diverse plant species.

4.6.2. LCC estimation using Vegetation Indices

In this paper, various VIs were tested to explore their potential in estimating the LCC. The study identified GNDVI, GRVI, and GCI as the most optimal predictor variables. The significance of a feature for a model's predictive accuracy is quantified by the Gini index, with higher values indicating greater importance (Simpson et al., 2022). Therefore, the exclusion of these variables during the training process results in a decrease in the predictive accuracy of the model (Boyko et al., 2022).

4.6.3. Comparison of different ML Estimation models

The integration of ML models with remote sensing data has become extensive with applications ranging from crop growth monitoring to yield estimation and disease detection and identification (Fu et al., 2020; García-Martínez et al., 2020; Guo et al., 2021; Li et al., 2019; Ma et al., 2022; Zhang et al., 2021; Zheng et al., 2019). This study centred on tomato plants and employed four machine learning algorithms—RFR, DTR, XGB, and PLSR—to construct LCC estimation models. The performance of these models was evaluated and compared, as detailed in *Table 6*, revealing XGB as the top-performing model, providing a moderate degree of explanation for LCC variation.

XGB consistently outperformed in the BFS and AFS phases, maintaining an R^2 of 0.50 and above. It showcased its exceptional and reliable predictive capabilities across diverse data configurations, solidifying its status as the consistently superior-performing model in the analysis. These findings align with the observations by Narmilan et al. (2022), emphasizing XGB's consistently strong performance in similar contexts. Compared with other models, XGB exhibited higher R^2 and lower RMSE and MAE for LCC estimates in both phases, indicating that the XGB algorithm can provide accurate LCC estimations.

Surprisingly, this study unveiled weak predictive power in models that usually have superiority in predictions. For example, the XGB had extremely high predictive power higher than 0.8 ($R^2 > 0.8$) in a study by Narmilan et al. (2022) and Hasan et al. (2023), estimating chlorophyll content of sugarcane and litchi crops, respectively. Furthermore, the findings suggested that while feature selection techniques can enhance the model's predictive power and accuracy, the differences between the two phases were minimal.

Despite rigorous optimization efforts to mitigate overfitting and underfitting, the models struggled with a general lack of predictive power. This highlights challenges stemming from factors such as the quality or quantity of the training data, model parameter choices, or the inherent complexity of the addressed problem. Even though the models were meticulously calibrated, accurate predictions on new data remained elusive, underscoring the multifaceted challenges in applying machine learning to complex agricultural scenarios.

Table 6: Estimation results of LCC in tomato plants based on four ML models.

ML MODELS	BEFORE FEATURE SELECTION			AFTER FEATURE SELECTION		
	R ²	RMSE	MAE	R ²	RMSE	MAE
PLSR	0.26	11.94	8.79	0.27	11.86	8.55
DTR	0.48	10.02	7.15	0.51	9.67	6.39
RFR	0.45	10.28	8.01	0.45	10.33	7.94
XGB	0.53	9.54	6.39	0.54	9.36	6.04

Prior research has also demonstrated that the estimation of chlorophyll content is influenced by using a variable significance plot in RF, which reduces the number of features. The RF model's contribution to each spectral index reveals how each index can anticipate fluctuations in chlorophyll content (see Figure 4.2) As evident in Figure 4.3 and Figure 4.4, the R² of the RFR model remains consistent across both phases. This consistency can be attributed to the incorporation of the Gini index into the RF model and the model's implicit feature selection process, which utilizes a small subset of potent variables for predictions (Menze et al., 2009; Boyko et al., 2022). The feature importance score, a by-product of the RFR training, provides a comparative rating of the input features.

UAV RS platforms have emerged as an efficient means of swiftly obtaining ground data, particularly in crop growth monitoring. Given the intricacy of the farmland environment, it is essential to modify the flight height and velocity according to real-time conditions. Conversely, lowering the flight speed and height is imperative to capture comprehensive crop information (G. Yang et al., 2017).

CHAPTER 5

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1. Summary

The study's aim was to estimate LCC in tomato crops through the application of ML models, specifically DTR, RFR, PLSR and XGB. In pursuit of the study's aims and objectives, images of UAVs were captured using the DJI Matrice 300 drone, equipped with a MicaSense RedEdge MX-MS camera. Leaf chlorophyll measurements were obtained using a SPAD-502 Plus chlorophyll meter, with 22 samples collected within a 1.6-hectare tomato field.

The study comprised two phases: Before Feature Selection (BFS) and After Feature Selection (AFS). During the BFS phase, ML models were developed using twenty VIs as predictors. In the AFS phase, models were constructed using only the most relevant VIs for estimating tomato LCC, which was determined through the Gini index incorporated into the RF model. Feature selection aims to identify the most compact subset of features from the initial set to maximize generalizability. The results revealed that GNDVI, GRVI and GCI were most influential in predicting LCC, exhibiting a strong and positive correlation with LCC.

The LOOCV method was employed to evaluate ML model performance. This technique is particularly advantageous when dealing with small datasets, as it optimally utilizes available data. Scatterplots were utilized to visualize the linear relationship between measured and predicted LCC values. Subsequently, regression model equations from the scatterplots were applied to generate spatial distribution maps of LCC within the field. Overall, four maps were produced, highlighting the variabilities across different models.

5.2. Conclusions

Chlorophyll is a green pigment in plant cells, and it is crucial for assessing crop health and making preliminary predictions. It captures sunlight and converts it into chemical energy through photosynthesis, indicating crop productivity and well-being and enabling farmers in making well-informed choices regarding crop management practices.

The research aimed to demonstrate the practical application of multispectral UAV data and machine learning for crop monitoring across a tomato field. Specifically, the study aimed to estimate LCC using ML models, such as PLSR, RFR, DTR and XGB. Vegetation indices derived from multispectral UAV system served as predictor variables, while LCC data collected through a SPAD-502 Plus chlorophyll meter was used as response variable.

Several noteworthy aspects emerged from this research, aligning with the predefined research objectives explored:

- Three VIs, namely the GNDVI, GRVI and GCI, demonstrated a moderate and positive correlation with the LCC of tomato crops.
- The implementation of feature selection improved the predictive accuracy of the ML models.
- The XGB model consistently outperformed the RFR, PLSR and DTR models in both phases (i.e., BFS and AFS), while the PLSR underperformed in both phases.

In light of these findings, it is evident that UAV Multispectral imagery exhibits significant potential for accurate, rapid and economic LCC estimation. Spectral imaging is crucial in crop monitoring, facilitating informed decision-making for implementing agronomic procedures and inputs that vary across space. Applying vegetation indices derived from UAV multispectral images, coupled with ML, offers an effective means of assessing chlorophyll content in crops, contributing to yield estimation, diagnosis of plant nutrition, and enhance comprehension of plant-environment interactions.

5.3. Limitation of the study

This study's small sample size (22 samples) is its primary limitation, which influenced the efficacy of the ML models. Although RF is well-suited for handling small quantities of sampled data, the performance of the RFR model is contingent upon the sample size. Generally, the greater the sample points, the more precise the estimations become (Lie et al. (2019); Narmilan et al. (2022)). This highlights the importance of data as the cornerstone of an ML model's performance, as emphasized by Xu et al. (2023).

5.4. Recommendations for Future Research

In subsequent research endeavours, it is imperative to collect and analyze chlorophyll measurements throughout the entire growth season, from establishment to pre-harvest, to enhance the generalizability of the findings. Given the focus of this study on regression models using ML, future studies should explore the generation of chlorophyll prediction models using Deep Learning techniques. Additionally, since this study lacked the resources to conduct experiments under various environmental conditions, future research must prioritise addressing this limitation. Furthermore, considering the less-than-impressive prediction accuracy results obtained in this study, future studies should explore stacking ensemble methods to enhance the accuracy of the predictions.

REFERENCES

- Albetis, J., Jacquin, A., Michel, G., Poilvé, H., Rousseau, J., Clenet, H., Dedieu, G., Duthoit, S., Goulard, M., & Dedieu, G. (2019). On the potentiality of UAV multispectral imagery to detect flavescence dorée and grapevine trunk diseases remote sensing On the Potentiality of UAV Multispectral Imagery to Detect Flavescence dorée and Grapevine Trunk Diseases. 11(1), 23. <https://doi.org/10.3390/rs11010023>
- Alvarez-Vanhard, E., Corpetti, T., & Houet, T. (2021). UAV & satellite synergies for optical remote sensing applications: A literature review. *Science of Remote Sensing*, 3, 100019. <https://doi.org/10.1016/j.srs.2021.100019>
- An, G., Xing, M., He, B., Liao, C., Huang, X., Shang, J., & Kang, H. (2020). Using machine learning for estimating rice chlorophyll content from in situ hyperspectral data. *Remote Sensing*, 12(18). <https://doi.org/10.3390/RS12183104>
- Ane, T., & Yasmin, S. (2019). *Agriculture in the Fourth Industrial Revolution* (Vol. 23, Issue 2). Online. www.seekmomentum.com/blog/manufacturing/the-evolution-of-industry-from-1-to-4
- Angel, Y., & McCabe, M. F. (2022). Machine Learning Strategies for the Retrieval of Leaf-Chlorophyll Dynamics: Model Choice, Sequential Versus Retraining Learning, and Hyperspectral Predictors. *Frontiers in Plant Science*, 13. <https://doi.org/10.3389/fpls.2022.722442>
- Ansarifar, J., Wang, L., & Archontoulis, S. V. (2021). An interaction regression model for crop yield prediction. *Scientific Reports*, 11(1). <https://doi.org/10.1038/s41598-021-97221-7>
- Arabian, J. (2015). Retrieving leaf chlorophyll content in wheat and corn using Landsat-8 imagery Retrieving leaf chlorophyll in crops using Landsat-8 images. MSc. University of Toronto.
- Araújo, S. O., Peres, R. S., Ramalho, J. C., Lidon, F., & Barata, J. (2023). Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives. *Agronomy*, 13(12), 2976. <https://doi.org/10.3390/agronomy13122976>
- Asuero, A. G., Sayago, A., & González, A. G. (2006). The correlation coefficient: An overview. In *Critical Reviews in Analytical Chemistry* (Vol. 36, Issue 1, pp. 41–59). <https://doi.org/10.1080/10408340500526766>

- Baburao, P. S., Kulkarni, R. B., Kharade, P. A., & Patil, S. S. (2023). Review of Machine Learning Model Applications in Precision Agriculture (pp. 916–930). https://doi.org/10.2991/978-94-6463-136-4_81
- Ban, S., Liu, W., Tian, M., Wang, Q., Yuan, T., Chang, Q., & Li, L. (2022). Rice Leaf Chlorophyll Content Estimation Using UAV-Based Spectral Images in Different Regions. *Agronomy*, 12(11). <https://doi.org/10.3390/agronomy12112832>
- Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. In *Sensors* (Vol. 21, Issue 11). MDPI AG. <https://doi.org/10.3390/s21113758>
- Bilal, A., Liu, X., Long, H., Shafiq, M., & Waqar, M. (2023). Increasing Crop Quality and Yield with a Machine Learning-Based Crop Monitoring System. *Computers, Materials and Continua*, 76(2), 2401–2426. <https://doi.org/10.32604/cmc.2023.037857>
- Bouke, M. A., Abdullah, A., ALshatebi, S. H., Abdullah, M. T., & Atigh, H. El. (2023). An intelligent DDoS attack detection tree-based model using Gini index feature selection method. *Microprocessors and Microsystems*, 98. <https://doi.org/10.1016/j.micpro.2023.104823>
- Boyko, N., Omeliukh, R., & Duliaba, N. (2022). The Random Forest Algorithm as an Element of Statistical Learning for Disease Prediction. In: *CITRisk'2022: 3rd International Workshop on Computational & Information Technologies for Risk-Informed Systems*, Neubiberg, Germany, 12 January 2023.
- Brewer, K. (2021). Assessment of maize crop health and water stress based on multispectral and thermal infrared unmanned aerial vehicle phenotyping in smallholder farms. MSc dissertation, University of KwaZulu-Natal, Pietermaritzburg, South Africa.
- Brewer, K., Clulow, A., Sibanda, M., Gokool, S., Naiken, V., & Mabhaudhi, T. (2022). Predicting the Chlorophyll Content of Maize over Phenotyping as a Proxy for Crop Health in Smallholder Farming Systems. *Remote Sensing*, 14(3), 518. <https://doi.org/10.3390/rs14030518>
- Cao, Y., Liu, G., Sun, J., Bavirisetti, D. P., & Xiao, G. (2023). PSO-Stacking improved ensemble model for campus building energy consumption forecasting based on priority feature selection. *Journal of Building Engineering*, 72. <https://doi.org/10.1016/j.jobbe.2023.106589>
- Cardellicchio, A., Solimani, F., Dimauro, G., Petrozza, A., Summerer, S., Cellini, F., & Renò, V. (2023). Detection of tomato plant phenotyping traits using YOLOv5-based single stage

- detectors. *Computers and Electronics in Agriculture*, 207. <https://doi.org/10.1016/j.compag.2023.107757>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. <https://doi.org/10.1145/2939672.2939785>
- Cravero, A., Pardo, S., Sepúlveda, S., & Muñoz, L. (2022). Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review. In *Agronomy* (Vol. 12, Issue 3). MDPI. <https://doi.org/10.3390/agronomy12030748>
- Croft, H., Arabian, J., Chen, J. M., Shang, J., & Liu, J. (2020). Mapping within-field leaf chlorophyll content in agricultural crops for nitrogen management using Landsat-8 imagery. *Precision Agriculture*, 21(4), 856–880. <https://doi.org/10.1007/s11119-019-09698-y>
- Croft, H., Chen, J. M., Froelich, N. J., Chen, B., & Staebler, R. M. (2015). Seasonal controls of canopy chlorophyll content on forest carbon uptake: Implications for GPP modeling. *Journal of Geophysical Research: Biogeosciences*, 120(8), 1576–1586. <https://doi.org/10.1002/2015JG002980>
- David, L. O., Nwulu, N. I., Aigbavboa, C. O., & Adepoju, O. O. (2022). Integrating fourth industrial revolution (4IR) technologies into the water, energy & food nexus for sustainable security: A bibliometric analysis. *Journal of Cleaner Production*, 363. <https://doi.org/10.1016/j.jclepro.2022.132522>
- Department of Agriculture, Land Reform and Rural Development. (2020). A Profile of the South African Tomato Market Value Chain 2020. www.dalrrd.gov.za
- Dian, Y., Le, Y., Fang, S., Xu, Y., Yao, C., & Liu, G. (2016). Influence of Spectral Bandwidth and Position on Chlorophyll Content Retrieval at Leaf and Canopy Levels. *Journal of the Indian Society of Remote Sensing*, 44(4), 583–593. <https://doi.org/10.1007/s12524-015-0537-2>
- Domingues, T., Brandão, T., & Ferreira, J. C. (2022). Machine Learning for Detection and Prediction of Crop Diseases and Pests: A Comprehensive Survey. In *Agriculture (Switzerland)* (Vol. 12, Issue 9). MDPI. <https://doi.org/10.3390/agriculture12091350>
- Ferchichi, A., Abbes, A. Ben, Barra, V., & Farah, I. R. (2022). Forecasting vegetation indices from spatio-temporal remotely sensed data using deep learning-based approaches: A systematic literature review. In *Ecological Informatics* (Vol. 68). Elsevier B.V. <https://doi.org/10.1016/j.ecoinf.2022.101552>

- Gao, G., Zhang, L., Wu, L., & Yuan, D. (2023). Estimation of Chlorophyll Content in Wheat Based on Optimal Spectral Index. <https://doi.org/10.20944/preprints202311.1022.v1>
- Gitelson, A. A., Kaufman, Y. J., Stark, R., & Rundquist, D. (2002). Novel algorithms for remote estimation of vegetation fraction. www.elsevier.com/locate/rse
- Hailemichael, G., Catalina, A., González, M. R., & Martin, P. (2016). Relationships between Water Status, Leaf Chlorophyll Content and Photosynthetic Performance in Tempranillo Vineyards. *South African Journal of Ecology and Viticulture*, 37(2).
- Hasan, U., Jia, K., Wang, L., Wang, C., Shen, Z., Yu, W., Sun, Y., Jiang, H., Zhang, Z., Guo, J., Wang, J., & Li, D. (2023). Retrieval of Leaf Chlorophyll Contents (LCCs) in Litchi Based on Fractional Order Derivatives and VCPA-GA-ML Algorithms. *Plants*, 12(3), 501. <https://doi.org/10.3390/plants12030501>
- Iizuka, K., Kato, T., Silsigia, S., Soufiningrum, A. Y., & Kozan, O. (2019). Estimating and examining the sensitivity of different vegetation indices to fractions of vegetation cover at different scaling Grids for Early Stage Acacia Plantation Forests Using a Fixed-Wing UAS. *Remote Sensing*, 11(15). <https://doi.org/10.3390/rs11151816>
- Iniyar, S., Akhil Varma, V., & Teja Naidu, C. (2023). Crop yield prediction using machine learning techniques. *Advances in Engineering Software*, 175. <https://doi.org/10.1016/j.advengsoft.2022.103326>
- Iniyar, S., & Jebakumar, R. (2022). Mutual Information Feature Selection (MIFS) Based Crop Yield Prediction on Corn and Soybean Crops Using Multilayer Stacked Ensemble Regression (MSER). *Wireless Personal Communications*, 126(3), 1935–1964. <https://doi.org/10.1007/s11277-021-08712-9>
- Jafarbiglu, H., & Pourreza, A. (2022). A comprehensive review of remote sensing platforms, sensors, and applications in nut crops. In *Computers and Electronics in Agriculture* (Vol. 197). Elsevier B.V. <https://doi.org/10.1016/j.compag.2022.106844>
- James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *Statistics An Introduction to Statistical Learning with Applications in Python*. Switzerland: Springer. ISBN 978-3-031-38747-0 (eBook). <https://doi.org/10.1007/978-3-031-38747-0>
- Ji, J., Li, N., Cui, H., Li, Y., Zhao, X., Zhang, H., & Ma, H. (2023). Study on Monitoring SPAD Values for Multispatial Spatial Vertical Scales of Summer Maize Based on UAV Multispectral Remote Sensing. *Agriculture*, 13(5), 1004. <https://doi.org/10.3390/agriculture13051004>

- Jin, J., & Wang, Q. (2019). Evaluation of informative bands used in different PLS regressions for estimating leaf biochemical contents from hyperspectral reflectance. *Remote Sensing*, 11(2). <https://doi.org/10.3390/rs11020197>
- Jiyane, J., & Simalenga, T. E. (2019). Factors Influencing Under-Utilisation of Smallholder Irrigation Schemes and Opportunities to Improve the Schemes' Productivity in Limpopo province, South Africa Report to the Water Research Commission. www.wrc.org.za
- Joshi, A., Pradhan, B., Chakraborty, S., & Behera, M. D. (2023). Winter wheat yield prediction in the conterminous United States using solar-induced chlorophyll fluorescence data and XGBoost and random forest algorithm. *Ecological Informatics*, 77. <https://doi.org/10.1016/j.ecoinf.2023.102194>
- Kamara, A., Conteh, A., Rhodes, E. R., & Cooke, R. A. (2019). The relevance of smallholder farming to african agricultural growth and development. *African Journal of Food, Agriculture, Nutrition and Development*, 19(1), 14043–14065. <https://doi.org/10.18697/AJFAND.84.BLFB1010>
- Kayad, A., Sozzi, M., Gatto, S., Marinello, F., & Pirotti, F. (2019). Monitoring within-field variability of corn yield using sentinel-2 and machine learning techniques. *Remote Sensing*, 11(23). <https://doi.org/10.3390/rs11232873>
- Kganyago, M., Adjorlolo, C., Sibanda, M., Mhangara, P., Laneve, G., & Alexandridis, T. (2022). Testing Sentinel-2 spectral configurations for estimating relevant crop biophysical and biochemical parameters for precision agriculture using tree-based and kernel-based algorithms. *Geocarto International*. <https://doi.org/10.1080/10106049.2022.2146764>
- Khaliq, A., Comba, L., Biglia, A., Ricauda Aimonino, D., Chiaberge, M., & Gay, P. (2019). Comparison of satellite and UAV-based multispectral imagery for vineyard variability assessment. *Remote Sensing*, 11(4). <https://doi.org/10.3390/rs11040436>
- Khan, M. A., Khan, R., & Ansari, M. A. (2022). *Application of Machine Learning in Agriculture*. London: Academic Press. ISBN 978-0-323-90550-3.
- Khoshrou, M. I., Zarafshan, P., Dehghani, M., Chegini, G., Arabhosseini, A., & Zakeri, B. (2021). Deep Learning Prediction of Chlorophyll Content in Tomato Leaves. 9th RSI International Conference on Robotics and Mechatronics, ICRoM 2021, 580–585. <https://doi.org/10.1109/ICRoM54204.2021.9663468>
- Kumar, T., Jyoti, K., & Singla, S. K. (2021). Design and Development of Machine Learning Model for Crop Yield Prediction. In *Proceedings of the International Conference on*

- Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., Uwitonze, C., Ngabonziza, J., & Uwamahoro, A. (2023). Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *Agriculture (Switzerland)*, 13(1). <https://doi.org/10.3390/agriculture13010225>
- Kvaktun, D., Hoffmann, A., & Schiffers, R. (2022). Analysis of feature extraction algorithms for quality prediction using machine learning in injection molding. *Procedia CIRP*, 112, 590–595. <https://doi.org/10.1016/j.procir.2022.09.059>
- Li, D., Chen, J. M., Zhang, X., Yan, Y., Zhu, J., Zheng, H., Zhou, K., Yao, X., Tian, Y., Zhu, Y., Cheng, T., & Cao, W. (2020). Improved estimation of leaf chlorophyll content of row crops from canopy reflectance spectra through minimizing canopy structural effects and optimizing off-noon observation time. *Remote Sensing of Environment*, 248. <https://doi.org/10.1016/j.rse.2020.111985>
- Li, M., Shamshiri, R. R., Weltzien, C., & Schirrmann, M. (2022). Crop Monitoring Using Sentinel-2 and UAV Multispectral Imagery: A Comparison Case Study in Northeastern Germany. *Remote Sensing*, 14(17). <https://doi.org/10.3390/rs14174426>
- Li, Y., He, N., Hou, J., Xu, L., Liu, C., Zhang, J., Wang, Q., Zhang, X., & Wu, X. (2018). Factors influencing leaf chlorophyll content in natural forests at the biome scale. *Frontiers in Ecology and Evolution*, 6(JUN). <https://doi.org/10.3389/fevo.2018.00064>
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. In *Sensors (Switzerland)* (Vol. 18, Issue 8). MDPI AG. <https://doi.org/10.3390/s18082674>
- Lu, S., Lu, F., You, W., Wang, Z., Liu, Y., & Omasa, K. (2018). A robust vegetation index for remotely assessing chlorophyll content of dorsiventral leaves across several species in different seasons. *Plant Methods*, 14(1). <https://doi.org/10.1186/s13007-018-0281-z>
- Lustenhouwer, N. (2009). The application of remote sensing in vegetation science. http://earthobservatory.nasa.gov/GlobalMaps/view.php?d1=MOD13A2_M_NDVI#
- Ma, M., Taylor, P. W. J., Chen, D., Vaghefi, N., & He, J. Z. (2023). Major Soilborne Pathogens of Field Processing Tomatoes and Management Strategies. In *Microorganisms* (Vol. 11, Issue 2). MDPI. <https://doi.org/10.3390/microorganisms11020263>

- Maes, W. H., & Steppe, K. (2019). Perspectives for Remote Sensing with Unmanned Aerial Vehicles in Precision Agriculture. In *Trends in Plant Science* (Vol. 24, Issue 2, pp. 152–164). Elsevier Ltd. <https://doi.org/10.1016/j.tplants.2018.11.007>
- Maimaitijiang, M., Sagan, V., Sidike, P., Daloye, A. M., Erkbol, H., & Fritsch, F. B. (2020). Crop monitoring using satellite/UAV data fusion and machine learning. *Remote Sensing*, 12(9). <https://doi.org/10.3390/RS12091357>
- Main, R., Cho, M. A., Mathieu, R., O'kenedy, M. M., Ramoelo, A., & Koch, S. (2011). An investigation into robust spectral indices for leaf chlorophyll estimation. Pretoria: Council for Scientific and Industrial Research, Natural Resources and Environment.
- Mamatha, V., & Sreekumar, K. (2018). A Review on Image Analysis Approaches for Estimating Chlorophyll Content from Leaf Images. In *International Journal of Computer Applications* (Vol. 179, Issue 36). www.ijcaonline.org
- Materechera, F., & Scholes, M. C. (2022). Understanding the Drivers of Production in South African Farming Systems: A Case Study of the Vhembe District, Limpopo South Africa. In *Frontiers in Sustainable Food Systems* (Vol. 6). Frontiers Media S.A. <https://doi.org/10.3389/fsufs.2022.722344>
- Meghashree, Mendonca, A. E., & Shetty, A. S. (2021). Automated Quality Assessment of Crops Using CNN - Keras. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 641–645. <https://doi.org/10.32628/cseit2173186>
- Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10. <https://doi.org/10.1186/1471-2105-10-213>
- Meshram, V., Patil, K., Meshram, V., Hanchate, D., & Ramkteke, S. D. (2021). Machine learning in agriculture domain: A state-of-art survey. In *Artificial Intelligence in the Life Sciences* (Vol. 1). Elsevier B.V. <https://doi.org/10.1016/j.ailsci.2021.100010>
- Mesías-Ruiz, G. A., Pérez-Ortiz, M., Dorado, J., de Castro, A. I., & Peña, J. M. (2023). Boosting precision crop protection towards agriculture 5.0 via machine learning and emerging technologies: A contextual review. In *Frontiers in Plant Science* (Vol. 14). Frontiers Media S.A. <https://doi.org/10.3389/fpls.2023.1143326>

- Messina, G., Peña, J. M., Vizzari, M., & Modica, G. (2020). A comparison of UAV and satellites multispectral imagery in monitoring onion crop. An application in the 'Cipolla Rossa di Tropea' (Italy). *Remote Sensing*, 12(20), 1–27. <https://doi.org/10.3390/rs12203424>
- Nandibewoor, A., Muddebihal, S., & Hegadi, R. (2017). Estimation Of Chlorophyll Content in Crop Using Multispectral Satellite Imagery. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, 6, 8. www.ijstr.org
- Narmilan, A., Gonzalez, F., Salgadoe, A. S. A., Kumarasiri, U. W. L. M., Weerasinghe, H. A. S., & Kulasekara, B. R. (2022). Predicting Canopy Chlorophyll Content in Sugarcane Crops Using Machine Learning Algorithms and Spectral Vegetation Indices Derived from UAV Multispectral Imagery. *Remote Sensing*, 14(5). <https://doi.org/10.3390/rs14051140>
- Naveed Tahir, M., Zaigham Abbas Naqvi, S., Lan, Y., Zhang, Y., Wang, Y., Afzal, M., Jehanzeb Masud Cheema, M., & Amir, S. (2018). Real time estimation of chlorophyll content based on vegetation indices derived from multispectral UAV in the kinnow orchard. *International Journal of Precision Agricultural Aviation*, 1(1), 24–31. <https://doi.org/10.33440/j.ijpaa.20180101.0001>
- Nevavuori, P., Narra, N., Linna, P., & Lipping, T. (2020). Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models. *Remote Sensing*, 12(23), 1–18. <https://doi.org/10.3390/rs12234000>
- Nkosi, M. & Agholor, A.I., 2021. The Fourth Industrial Revolution and its Implication for Agricultural Advisory Services in South Africa: A review. *Turkish Online Journal of Qualitative Inquiry (TOJQI)*, 12(7), pp.8691-8707.
- Omia, E., Bae, H., Park, E., Kim, M. S., Baek, I., Kabenge, I., & Cho, B. K. (2023). Remote Sensing in Field Crop Monitoring: A Comprehensive Review of Sensor Systems, Data Analyses and Recent Advances. In *Remote Sensing* (Vol. 15, Issue 2). MDPI. <https://doi.org/10.3390/rs15020354>
- Otchere, D. A., Ganat, T. O. A., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, 208. <https://doi.org/10.1016/j.petrol.2021.109244>
- Pallathadka, H., Mustafa, M., Sanchez, D. T., Sekhar Sajja, G., Gour, S., & Naved, M. (2023). IMPACT OF MACHINE learning ON Management, healthcare AND AGRICULTURE. *Materials Today: Proceedings*, 80, 2803–2806. <https://doi.org/10.1016/j.matpr.2021.07.042>

- Pei, H., Li, C., Feng, H., Yang, G., Liu, M., & Wu, Z. (2017). Hyperspectral Estimation Methods for Chlorophyll Content of Apple Based on Random Forest. *Beijing Research Center for Information Technology in Agriculture*.
- Pérez-Patricio, M., Camas-Anzueto, J. L., Sanchez-Alegría, A., Aguilar-González, A., Gutiérrez-Miceli, F., Escobar-Gómez, E., Voisin, Y., Rios-Rojas, C., & Grajales-Coutiño, R. (2018). Optical method for estimating the chlorophyll contents in plant leaves. *Sensors (Switzerland)*, 18(2). <https://doi.org/10.3390/s18020650>
- Qian, X., & Liu, L. (2020). Retrieving crop leaf chlorophyll content using an improved look-up-table approach by combining multiple canopy structures and soil backgrounds. *Remote Sensing*, 12(13). <https://doi.org/10.3390/rs12132139>
- Sagan, V., Maimaitijiang, M., Sidike, P., Maimaitiyiming, M., Erkbol, H., Hartling, S., Peterson, K. T., Peterson, J., Burken, J., & Fritsch, F. (2019). Uav/satellite multiscale data fusion for crop monitoring and early stress detection. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 42(2/W13), 715–722. <https://doi.org/10.5194/isprs-archives-XLII-2-W13-715-2019>
- Saranya, T., Deisy, C., Sridevi, S., & Anbananthen, K. S. M. (2023). A comparative study of deep learning and Internet of Things for precision agriculture. *Engineering Applications of Artificial Intelligence*, 122, 106034. <https://doi.org/10.1016/j.engappai.2023.106034>
- Shah, S. H., Angel, Y., Houborg, R., Ali, S., & McCabe, M. F. (2019). A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat. *Remote Sensing*, 11(8). <https://doi.org/10.3390/rs11080920>
- Shahi, T. B., Xu, C. Y., Neupane, A., & Guo, W. (2022). Machine learning methods for precision agriculture with UAV imagery: a review. *Electronic Research Archive*, 30(12), 4277–4317. <https://doi.org/10.3934/era.2022218>
- Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. In *IEEE Access (Vol. 9, pp. 4843–4873)*. Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ACCESS.2020.3048415>
- Shinde, A., Raykar, T., Patil, P., Mali, H., Gurav, Y., & Students, B. E. (2022). Plant Disease Detection and Identification using Leaf Images using deep learning. *International Research Journal of Engineering and Technology (IRJET)*, 9(5), pp.288-291.
- Simpson, G., Jin, W., Spieler, B., Portelance, L., Mellon, E., Kwon, D., Ford, J. C., & Dogan, N. (2022). Predictive Value of Delta-Radiomics Texture Features in 0.35 Tesla Magnetic

- Resonance Setup Images Acquired During Stereotactic Ablative Radiotherapy of Pancreatic Cancer. *Frontiers in Oncology*, 12. <https://doi.org/10.3389/fonc.2022.807725>
- Singh, P., Pandey, P. C., Petropoulos, G. P., Pavlides, A., Srivastava, P. K., Koutsias, N., Deng, K. A. K., & Bao, Y. (2020). Hyperspectral remote sensing in precision agriculture_ present status, challenges and future trends. *In* Prasad S., Thenkabail, P. and Joshi, A., (eds). *Hyperspectral Remote Sensing of Agriculture and Vegetation*. Elsevier, pp. 121-156. DOI: 10.1016/B978-0-08-102894-0.00009-7.
- Singhal, G., Bansod, B., Mathew, L., Goswami, J., Choudhury, B. U., & Raju, P. L. N. (2019). Chlorophyll estimation using multi-spectral unmanned aerial system based on machine learning techniques. *Remote Sensing Applications: Society and Environment*, 15. <https://doi.org/10.1016/j.rsase.2019.100235>
- Sonobe, R., & Wang, Q. (2017). Hyperspectral indices for quantifying leaf chlorophyll concentrations performed differently with different leaf types in deciduous forests. *Ecological Informatics*, 37, 1–9. <https://doi.org/10.1016/j.ecoinf.2016.11.007>
- Ta, N., Chang, Q., & Zhang, Y. (2021). Estimation of apple tree leaf chlorophyll content based on machine learning methods. *Remote Sensing*, 13(19). <https://doi.org/10.3390/rs13193902>
- Tambakhe, Gulhane, V. S., & Karnewar, S. (2019). Machine Learning Application in the field of Agriculture: A Review. *International Journal of Research in Advent Technology*, 7(4), 5. www.ijrat.org
- Tao, H., Feng, H., Xu, L., Miao, M., Long, H., Yue, J., Li, Z., Yang, G., Yang, X., & Fan, L. (2020). Estimation of crop growth parameters using UAV- based hyperspectral remote sensing data. *Sensors (Switzerland)*, 20(5). <https://doi.org/10.3390/s20051296>
- Themistocleous, K. (2019). DEM modeling using RGB-based vegetation indices from UAV images. 21. <https://doi.org/10.1117/12.2532748>
- Torres, A. B. B., da Rocha, A. R., Coelho da Silva, T. L., de Souza, J. N., & Gondim, R. S. (2020). Multilevel data fusion for the internet of things in smart agriculture. *Computers and Electronics in Agriculture*, 171. <https://doi.org/10.1016/j.compag.2020.105309>
- Tsouros, D. C., Bibi, S., & Sarigiannidis, P. G. (2019). A review on UAV-based applications for precision agriculture. *In* *Information (Switzerland)* (Vol. 10, Issue 11). MDPI AG. <https://doi.org/10.3390/info10110349>

- Tuerxun, N., Zheng, J., Wang, R., Wang, L., & Liu, L. (2023). Hyperspectral estimation of chlorophyll content in jujube leaves: integration of derivative processing techniques and dimensionality reduction algorithms. *Frontiers in Plant Science*, 14. <https://doi.org/10.3389/fpls.2023.1260772>
- van Averbek, W., Denison, J., & Mkeni, P. N. S. (2011). Smallholder irrigation schemes in South Africa: A review of knowledge generated by the Water Research Commission. In *Water SA* (Vol. 37, Issue 5, pp. 797–808). South African Water Research Commission. <https://doi.org/10.4314/wsa.v37i5.17>
- van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. In *Computers and Electronics in Agriculture* (Vol. 177). Elsevier B.V. <https://doi.org/10.1016/j.compag.2020.105709>
- Veiga, R. K., Veloso, A. C., Melo, A. P., & Lamberts, R. (2021). Application of machine learning to estimate building energy use intensities. *Energy and Buildings*, 249, 111219. <https://doi.org/10.1016/j.enbuild.2021.111219>
- Verma, S., Choudhary, P. K., Kumar, S., & Gunjan, Prof. Dr. R. (2022). Plant Disease Detection Using Deep Learning. *International Journal for Research in Applied Science and Engineering Technology*, 10(6), 1009–1013. <https://doi.org/10.22214/ijraset.2022.43700>
- Wahab, I., Hall, O., & Jirstrom, M. (2018). Remote sensing of yields: Application of UAV imagery-derived ndvi for estimating maize vigor and yields in complex farming systems in Sub-Saharan Africa. *Drones*, 2(3), 1–16. <https://doi.org/10.3390/drones2030028>
- Wan, L., Li, Y., Cen, H., Zhu, J., Yin, W., Wu, W., Zhu, H., Sun, D., Zhou, W., & He, Y. (2018a). Combining UAV-based vegetation indices and image classification to estimate flower number in oilseed rape. *Remote Sensing*, 10(9). <https://doi.org/10.3390/rs10091484>
- Wan, L., Li, Y., Cen, H., Zhu, J., Yin, W., Wu, W., Zhu, H., Sun, D., Zhou, W., & He, Y. (2018b). Combining UAV-based vegetation indices and image classification to estimate flower number in oilseed rape. *Remote Sensing*, 10(9). <https://doi.org/10.3390/rs10091484>
- Wang, T., Gao, M., Cao, C., You, J., Zhang, X., & Shen, L. (2022). Winter wheat chlorophyll content retrieval based on machine learning using in situ hyperspectral data. *Computers and Electronics in Agriculture*, 193. <https://doi.org/10.1016/j.compag.2022.106728>
- Wang, T., Liu, Y., Wang, M., Fan, Q., Tian, H., Qiao, X., & Li, Y. (2021). Applications of UAS in Crop Biomass Monitoring: A Review. In *Frontiers in Plant Science* (Vol. 12). Frontiers Media S.A. <https://doi.org/10.3389/fpls.2021.616689>

- Wang, X., & Liu, J. (2021). Tomato Anomalies Detection in Greenhouse Scenarios Based on YOLO-Dense. *Frontiers in Plant Science*, 12. <https://doi.org/10.3389/fpls.2021.634103>
- Wu, Q., Zhang, Y., Zhao, Z., Xie, M., & Hou, D. (2023). Estimation of Relative Chlorophyll Content in Spring Wheat Based on Multi-Temporal UAV Remote Sensing. *Agronomy*, 13(1), 211. <https://doi.org/10.3390/agronomy13010211>
- Xu, H., Wang, J., Qu, Y., Hu, L., Tang, Y., Zhou, Z., Xu, X., & Zhou, Y. (2022). Estimating Leaf Chlorophyll Content of Moso Bamboo Based on Unmanned Aerial Vehicle Visible Images. *Remote Sensing*, 14(12). <https://doi.org/10.3390/rs14122864>
- Xu, P., Ji, X., Li, M., & Lu, W. (2023). Small data machine learning in materials science. In *npj Computational Materials* (Vol. 9, Issue 1). Nature Research. <https://doi.org/10.1038/s41524-023-01000-z>
- Xue, J., & Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. In *Journal of Sensors* (Vol. 2017). Hindawi Limited. <https://doi.org/10.1155/2017/1353691>
- Yang, G., Liu, J., Zhao, C., Li, Z., Huang, Y., Yu, H., Xu, B., Yang, X., Zhu, D., Zhang, X., Zhang, R., Feng, H., Zhao, X., Li, Z., Li, H., & Yang, H. (2017). Unmanned aerial vehicle remote sensing for field-based crop phenotyping: Current status and perspectives. In *Frontiers in Plant Science* (Vol. 8). Frontiers Media S.A. <https://doi.org/10.3389/fpls.2017.01111>
- Yang, H., Hu, Y., Zheng, Z., Qiao, Y., Zhang, K., Guo, T., & Chen, J. (2022a). Estimation of Potato Chlorophyll Content from UAV Multispectral Images with Stacking Ensemble Algorithm. *Agronomy*, 12(10). <https://doi.org/10.3390/agronomy12102318>
- Yang, H., Hu, Y., Zheng, Z., Qiao, Y., Zhang, K., Guo, T., & Chen, J. (2022b). Estimation of Potato Chlorophyll Content from UAV Multispectral Images with Stacking Ensemble Algorithm. *Agronomy*, 12(10). <https://doi.org/10.3390/agronomy12102318>
- Yuhao, A., Norasma, N., Ya, C., Roslin, A., & Ismail, M. R. (2020). Rice Chlorophyll Content Monitoring using Vegetation Indices from Multispectral Aerial Imagery. *Pertanika J. Sci. & Technol*, 28(3), 779–795.
- Zhang, M., Zhou, J., Sudduth, K. A., & Kitchen, N. R. (2020). Estimation of maize yield and effects of variable-rate nitrogen application using UAV-based RGB imagery. *Biosystems Engineering*, 189, 24–35. <https://doi.org/10.1016/j.biosystemseng.2019.11.001>

- Zhao, Y., Zhu, W., Wei, P., Fang, P., Zhang, X., Yan, N., Liu, W., Zhao, H., & Wu, Q. (2022). Classification of Zambian grasslands using random forest feature importance selection during the optimal phenological period. *Ecological Indicators*, 135. <https://doi.org/10.1016/j.ecolind.2021.108529>
- Zou, X., Jin, J., & Möttus, M. (2023). Potential of Satellite Spectral Resolution Vegetation Indices for Estimation of Canopy Chlorophyll Content of Field Crops: Mitigating Effects of Leaf Angle Distribution. *Remote Sensing*, 15(5). <https://doi.org/10.3390/rs15051234>
- Zuolkernan, I., Abuhani, D. A., Hussain, M. H., Khan, J., & ElMohandes, M. (2023). Machine Learning for Precision Agriculture Using Imagery from Unmanned Aerial Vehicles (UAVs): A Survey. *Drones*, 7(6), 382. <https://doi.org/10.3390/drones7060382>

APPENDICES

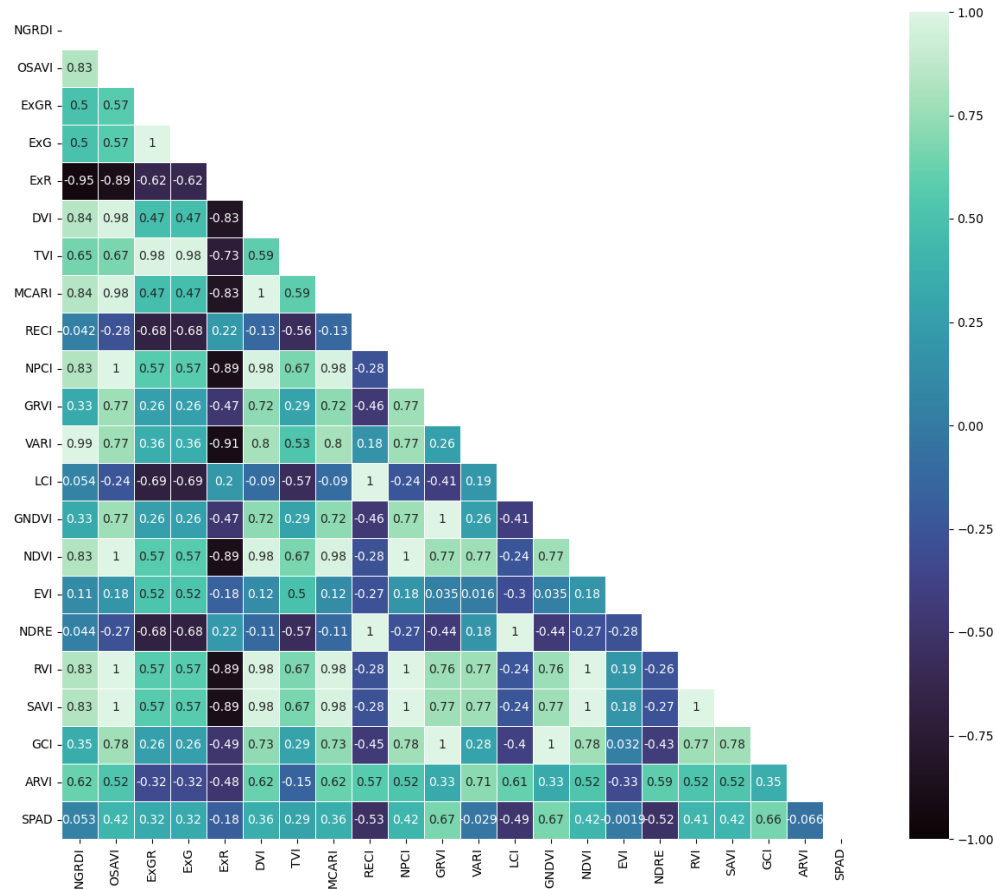


Figure A1: Pearson's Correlation matrix between VIs and tomato LCC.