MACHINE LEARNING BASED RICE YIELD PREDICTION AND RICE GROWTH STAGES IDENTIFICATION USING SENTINEL-1 SAR AND AUXILIARY DATA IN MWEA IRRIGATION SCHEME, KENYA

KEVIN AARON KARUGU

MASTER OF SCIENCE IN

SOIL AND WATER ENGINEERING

JOMO KENYATTA UNIVERSITY

OF

AGRICULTURE AND TECHNOLOGY

2026

Machine Learning Based Rice Yield Prediction and Rice Growth Stages Identification Using Sentinel-1 SAR and Auxiliary Data in Mwea Irrigation Scheme, Kenya

Kevin Aaron Karugu

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Soil and Water Engineering of the Jomo Kenyatta University of Agriculture and Technology

2026

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature……………………………… Date.......................................

Kevin Aaron Karugu

This thesis has been submitted for examination with our approval as the University Supervisors:

Signature……………………………… Date.......................................

Dr. Joseph Sang, PhD

JKUAT, Kenya

Signature……………………………… Date.......................................

Dr. Eunice Nduati, PhD

JKUAT, Kenya

DEDICATION

This thesis is dedicated to my loving family for the cheering and support you have given me throughout my life. This has always given me a cause to keep going.

ACKNOWLEDGEMENT

Principally, I thank God Almighty for His, grace, mercies, courage, strength and the gift of life that He has bestowed upon me. The culmination of this work benefited from various people with whom I express my most sincere gratitude. I wish to recognize my supervisors Dr. (Eng). Joseph Sang and Dr. Eunice Nduati, for their steadfast support and guidance in providing intuitive knowledge and guidance that has moulded this research work to its best present form. I attribute the level of my Master's degree to their encouragement and effort without which this research work would not have been completed. I would also like to express my gratitude to staff at the Department of Soil, Water and Environmental Engineering (SWEED) and Department of Geomatic Engineering and Geospatial Information Systems (GEGIS) of Jomo Kenyatta University of Agriculture and Technology (JKUAT) for guidance through the research work. My gratitude also goes to the DAAD for granting me a scholarship to pursue my Master's degree in Soil and Water Engineering. I also thank National Irrigation Authority (NIA), Mwea office for granting me permission to conduct my research at the scheme. Further, I thank Mr. Joseph Thuku of NIA for providing research support during data collection. I thank Mr. Benjamin K. Ndung'u for his support and efforts during the data analysis phase that helped shape this study. Finally, I am thankful to my beloved family, friends, and classmates for their unwavering support and encouragement throughout my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATION AND ACRONYMS

FAO	Food Agriculture Organization
GEE	Google Earth Engine
GRD	Ground Range Detected
IRRI	International Rice Research Institute
IW	Interferometric Wide Swath
LAI	Leaf Area Index
LR	Linear Regression
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MIS	Mwea Irrigation Scheme
ML	Machine Learning
MSE	Mean Squared Error
NDRE	Normalized Difference Red Edge
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NIR	Near Infrared
RF	Random Forest
RMSE	Root Mean Squared Error
ROI	Region of Interest
S1	Sentinel-1
S2	Sentinel-2
SAR	Synthetic Aperture Radar
SVM	Support Vector Machine
SWIR	Short Wave Infrared
VH	Vertical transmit/ Horizontal receive
VV	Vertical transmit/ Vertical receive
XGBoost	Extreme Gradient Boosting

ABSTRACT

Rice is considered the third most important staple food in Kenya after maize and wheat making its productivity of significant concern. Additionally, conventional preharvest crop yield estimation techniques are dependent on data collection from ground-based field visits which are often subjective, costly and prone to huge errors resulting to poor crop estimates. Satellite remote sensing has been widely accepted as a solution to overcoming these challenges. Therefore, the primary objective of this research was to identify rice growth stages and develop a rice yield-prediction model using Sentinel-1 SAR data and climatic auxiliary data with machine learning (ML) in Mwea Irrigation Scheme (MIS), Kenya. Hence, this study focused on the characterization of rice growth stages in MIS using spectral data derived from Sentinel satellites on Google Earth Engine (GEE) cloud. Thereafter, the SAR backscatter data, coupled with rainfall, and temperature data covering the period 2021-2024 were used to predict rice yield applying four ML models including Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM) and Linear Regression (LR) using python programming language on the web-based Jupyter Notebook platform. Finally, evaluation of the performance of RF, XGBoost, SVM models in predicting rice yield for the 2024 season was done using LR as a benchmark. The proposed method produced a rice extent map having an overall accuracy of 83% with a kappa coefficient of 0.53 and effectively identified the stages of rice growth during the 2024 main rice growing season in the study area. From the identified rice clusters, it can be shown that transplanting takes place in July and August while the maturity phase is attained in October and November, giving an indication that the main cropping season was approximately 120 days/ 4 months. The findings from this study illustrate that the RF model outperforms the other models with lower error values, achieving a RMSE of 246.91 kg/acre, MAE of 202.90 kg/acre and MAPE of 7.67% for rice yield predictions utilizing both SAR backscatter and auxiliary climatic datasets. On the other hand, LR model shows suboptimal performance of the four models as indicated by its higher error values attaining a RMSE of 384.09 kg/acre, MAE of 290.86 kg/acre, and MAPE of 11.06%. Furthermore, in predicting yields for the upcoming 2024 season using datasets from previous seasons, the RF model emerged with the highest accuracy obtaining the lowest RMSE of 288.52 kg/acre, MAE of 195.13 kg/acre and MAPE of 8.43%, followed by XGBoost with a RMSE of 290.44 kg/acre, MAE of 201.45 kg/acre and MAPE of 8.56% and finally the SVM model with a RMSE of 493.66 kg/acre, MAE of 395.21 kg/acre and MAPE of 16.73%. The findings show that integration of Sentinel SAR backscatter and optical time series data applying a phenology-driven framework on the GEE platform provides an effective approach for rice mapping and growth stage identification. The study demonstrates that ML models calibrated using previous seasons' datasets can reasonably predict rice yield for subsequent growing seasons. RF provided the most accurate predictions one month prior to harvesting, followed by XGBoost, and SVM models. The LR model failed to generalize to an unseen growing season, indicating that linear relationships are inadequate for capturing seasonal variability in rice production. Machine learning-based rice yield prediction, particularly using the RF model, should be incorporated into agricultural monitoring frameworks to provide early estimates of rice production before harvest. Reliable predictions would support government

and state agencies in evidence-based decision-making on grain procurement, import and export planning, food reserve management and market stabilization. Availability of quality weed control, fertilization and pest management data for rice production was a major limitation in this study which future studies could incorporate for rice yield predictions.

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Rice is considered the third most important staple food in Kenya, after maize and wheat making it an alternative cereal to supplement maize (Ng'endo et al., 2022). In 2018, the National Rice Development Strategy-2, 2019-2030 estimated the annual national rice consumption in Kenya to be approximately 949,000 metric tons while the annual production stood at about 180,000 metric tons (Republic of Kenya, 2020). With a projected population growth rate of 2.7% per year, the estimated annual national need may reach up to 1,292,000 metric tons by the year 2030. The Government of Kenya has identified rice as one of the priority value chains in the National Agriculture Investment Plan (NAIP 2018-2028) which aims to accelerate the country's agricultural transformation towards a commercial and modern sector that sustainably supports food and nutrition security as well as socio-economic development (Republic of Kenya, 2020).

In 2018, rice imports in Sub-Saharan Africa cost the region approximately USD 6.4 billion/year and is expected to reach USD 11 billion by the year 2030 unless serious interventions are taken (International Rice Research Institute, 2020). It is estimated that the African continent has the highest reserves for unexploited natural resources of food production, with well over 230 million ha suitable for rice production, but only about 12 million ha being used for the crop. In recent years, yields on the continent have improved, but they still remain low, with production at about half the global average (International Rice Research Institute, 2020).

Accurate rice yield estimates before the actual harvest could help address the gap between demand and production through improved decision-making enabling timely estimation of production deficits, determination of import requirements, help logistics providers optimize distribution, target extension services and agricultural subsidies (Chu & Yu, 2020; De Clercq & Mahdi, 2024). Moreover, crop yield estimates before harvesting are essential for farmers and land managers in determining and evaluating crop yield variability because of various treatment factors such as fertilizer use, pest control, water use among others (Lobell et al., 2015). Therefore, accurate crop yield prediction is critical for the growth and stability of the agricultural sector (Mokesh Rayalu & Saruk BS, 2025). Additionally, the agricultural field has benefited immensely from recent advancement in sensor technology, data science and machine learning (ML) (Cravero et al., 2026; Mokesh Rayalu & Saruk BS, 2025). These technologies allow the integration of large-scale climatic, soil, genetic, and socioeconomic datasets, permitting the development of predictive models that support informed agronomic decision-making and improve agricultural management (Cravero et al., 2026).

As consequence, digital technologies such as artificial intelligence (AI), machine learning and Big Data have opened avenues for the processing of multisource datasets including climatic, satellite and Internet of things (IoT) sensor data for precision agriculture and monitoring of key production chains (Ajith et al., 2025; Cravero et al., 2026; Rejeb et al., 2021). Specifically, remote sensing spectral data from can be utilized to monitor crop growth and development through optical and microwave bands (Yu et al., 2023). Even though optical vegetation indices such as are often used in crop yield prediction they are often cloud cover which is popular in rice growing areas (Yu et al., 2023). On the other hand, satellite SAR backscatter data unlike optical data remain unaffected with clouds or haze and can also be deployed in crop yield prediction (Yu et al., 2023). Therefore, this research study attempts to provide a solution to rice value chain monitoring through rice yield prediction using satellite SAR data, and auxiliary climatic data using ML models.

1.2 Statement of the Problem

Conventional preharvest crop yield estimation techniques are dependent on data collection from ground-based field visits which are often subjective, costly and prone to huge errors resulting in poor crop estimates (Ali et al., 2022). These limitations hinder the ability to capture intra-seasonal variability while restricting spatial coverage, temporal resolution, and operational flexibility, particularly for large agricultural regions, data-scarce areas, and rapidly changing climatic environments (Radočaj et al., 2026). Satellite remote sensing has been widely accepted as a solution to overcoming these challenges as it allows for constant tracking of crop development and conditions covering expansive areas (Meghraoui et al., 2024; Radočaj et al., 2026). Therefore, the development of remote sensing technologies and data derived model creation has revolutionized crop yield prediction due to ML capabilities in modelling non-linear relationships between yield and different environmental variables (Radočaj et al., 2026; Sánchez et al., 2025). Despite these developments crop yield prediction based solely on satellite data may fail to attain high accuracy of yield variance since yield is influenced only in part by vegetation characteristics detected through satellite imagery hence necessitating the incorporation of meteorological data, soil data and management variables to depict crop yield dynamics comprehensively (Radočaj et al., 2026). In light of this, several studies have combined climatic data and satellite datasets for enhanced yield prediction performance (Cao et al., 2021; Yu et al., 2023).

According to National Irrigation Authority (NIA) Mwea office, rice yield estimations for Mwea Irrigation Scheme are determined through rice production surveys undertaken on sample farms after harvesting. Furthermore, the obtained data may be obtained too late for informed decision making to mitigate food shortages. Rice has become a crucial staple food in Kenya, making its productivity of significant concern (Nyangau et al., 2014). Synthetic Aperture Radar (SAR) satellite data has shown potential for estimating rice yields at field scale (Wang et al., 2019). However, integration of SAR data with other data sources such as climatic, soil data and management data utilizing machine learning has not been extensively explored for rice yield prediction in the Kenyan context. This study therefore, sought to predict rice yield using Sentinel-1 SAR backscatter data (VV and VH polarization) and auxiliary climatic data (rainfall, minimum and maximum temperature) with ML models in Mwea Irrigation Scheme, in Kenya. It is envisaged that rice yield predictions could possibly, provide key data to policy makers in making strategic decisions, such as organizing import and export activities, formulating crop-pricing strategies, and forecasting potential famine.

1.3 Objectives

1.3.1 Main Objective

The primary objective of this research was to identify rice growth stages and develop a rice yield-prediction model using Sentinel-1 SAR data and climatic auxiliary data with machine learning in Mwea Irrigation Scheme, Kenya.

1.3.2 Specific Objectives

The specific objectives of this study were:

i. to characterize rice growth stages using SAR and Optical data from Sentinel satellites in Mwea Irrigation Scheme.
ii. to predict rice yield using machine-learning models based on Sentinel-1 SAR data and climatic auxiliary data.
iii. to evaluate the performance of machine learning models in predicting rice yield using linear regression model as a benchmark.

1.4 Research Questions

i. How can SAR and optical data from Sentinel satellites be effectively utilized to distinguish between different rice growth stages in Mwea Irrigation Scheme?
ii. How do different machine learning algorithms compare in their ability to predict rice yield based on Sentinel-1 SAR data and climatic auxiliary datasets?
iii. How does the performance of different machine learning models in predicting rice yield compare to that of linear regression model when applied to the same dataset?

1.5 Justification

Rice is a vital staple food in Kenya contributing to the country's food security and economic development. However, the rice value chain faces major obstacles such as unfavourable weather conditions, inadequate water for irrigation, low and declining land productivity as well as high cost of inputs (Atera et al., 2018). This study sought to utilize advanced technologies such as SAR backscatter data and ML algorithms to identify rice growth stages and develop rice yield prediction model that captures the dynamic interaction between environmental factors (i.e. temperature and precipitation) and yield. This study entailed integration of multiple data sources including precipitation, minimum and maximum temperature alongside Sentinel-1 SAR datasets. This holistic approach enabled a comprehensive analysis of the factors influencing rice yield resulting in higher accuracy predictions. A more precise and dependable prediction system would allow decision-makers to track crop growth and overall conditions throughout the growing season. This, in turn, would support informed decision-making for farm level management, insurance evaluation, national grain reserve planning as well as international trade forecasting (Radočaj et al., 2026). Additionally, rice phenology information is critical for its practical significance in undertaking timely field management activities such as irrigation activities, fertilization, weeding and pest control for precision agriculture (Wang et al., 2022). Insights gained from implementing this study can be applied to other agricultural contexts in Kenya and beyond, promoting innovation and sustainability in food production systems.

1.6 Scope of Study

This study focused on the characterization of rice growth stages in Mwea Irrigation Scheme using spectral data derived from Sentinel satellites on Google Earth Engine (GEE) cloud. Huge amounts of data may be extracted from the Sentinel satellites, however, this study worked with SAR backscatter data i.e. Vertical transmit/ Vertical receive (VV) and Vertical transmit/ Horizontal receive (VH) polarization data from Sentinel-1 (S1) and vegetation indices i.e. Normalized Difference Vegetation Index (NDVI), Normalized

Difference Water Index (NDWI) and Normalized Difference Red Edge (NDRE) from Sentinel-2 (S2) satellite. The SAR data was selected for its capability in predicting yield utilizing backscattering coefficients and its invulnerability to the effects of clouds or haze. Rice cropping calendar was determined based on representative profiles of the VH polarization, NDVI, NDWI and NDRE time series data covering the 2024 main rice growing season. The specific rice growth stages that this study sought to identify included transplanting, vegetative, reproductive and maturity stages. Thereafter, the SAR backscatter data (VV and VH polarization), coupled with auxiliary climatic data (rainfall data, minimum and maximum temperature data) covering the period 2021-2024 were used to predict rice yield utilizing four ML models including random forest (RF), extreme gradient boosting (XGBoost), support vector machine (SVM) and linear regression (LR) using python programing language on the web-based Jupyter Notebook platform. Finally, evaluation of the performance of RF, XGBoost, SVM models in predicting rice yield for the 2024 season was done using LR as a benchmark. Three statistical indicators for accuracy were utilized in evaluating the performance of the ML models including the root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE).

1.7 Limitations of the Study

Availability of quality rice management data on weed control, fertilization and pest management data for rice production in the scheme covering the period 2021-2024 was a major constraint and hence the study did not include rice management data as one of the variables in rice yield prediction. As such the predictions in this study were undertaken using SAR satellite data with rainfall and temperature datasets without management data which is likely to have an effect on rice yield.

CHAPTER TWO

LITERATURE REVIEW

2.1 Using Satellite Data for Crop Yield Estimation

Remote sensing technologies have manifold applications in agriculture, including crop discrimination, monitoring crop growth and stress, inventorying crops, estimating soil moisture, monitoring and managing nutrient levels, irrigation and pest management, estimating crop acreage, and predicting yields (Meena et al., 2018). One of the most widely utilized methods for crop yield forecasting is the use of Vegetation Indices (VIs) that is based on vegetation reflectance values in the visible and near-infrared spectrum (Singh Boori et al., 2023). The VIs from remote sensing datasets can be modelled with crop biophysical and biochemical measurements, and with proper calibration, these parameters can be used to estimate biomass and yield (Huang et al., 2018).

Methods of forecasting spatial crop yield employing satellite images can be categorized into process-based modelling and empirical regression (Jeong et al., 2022). Process based models simulate crop development, growth and yield mechanically using various inputs such as weather, soil, and cultivar parameters, which may be unavailable in certain regions (Gao et al., 2020). Conversely, empirical regression models utilize relatively simple input dimensions based on single and multiple regression techniques thereby forecasting crop yield based on the empirical relationships of Leaf Area Index (LAI), vegetation indices and environmental factors with actual field measurements making them relatively easy to apply (Jeong et al., 2022). Table 2.1 outlines details of some of the common VIs used for agricultural production monitoring (Nakalembe et al., 2021).

Presently, high-density geospatial data is assembled using sensor equipment of varying capabilities continuously estimating crop canopies and crop yields. Therefore, the collection of these large and more detailed datasets necessitates the advancement in data analytics for more accurate and consistent findings and results to aid land managers, farmers, policy makers and other key stakeholders in making management decisions in agricultural landscapes (Santos et al., 2024). Further, Desloires et al. (2023) premised that enriching remotely sensed data with explanatory variables related to genotype, soil characteristics, agronomic factors and field management practices could aid in enhancing the accuracy of crop yield variation.

Table 2.1: Satellite Based Indices for Agricultural Monitoring

Category	Indices	Definition
Crop Conditions Indicators	NDVI	Normalized Difference Vegetation Index is the ratio between the red (R) and near infrared (NIR) which is used to quantify vegetation greenness and is given by (NIR-R)/ (NIR+R)
	EVI	Enhanced Vegetation Index is ratio between the R and NIR values while reducing the background noise, atmospheric noise and saturation in most cases.
	VCI	Vegetation Condition Index compares the current NDVI to the range of values observed in the same period in previous years.
	NDRE	Normalized Difference Red Edge is an index used for measuring chlorophyll content in plants.
Drought indicators	NDWI	Normalized Difference Water Index is a ratio between the NIR and short-wave near-infrared (SWIR) which shows the moisture content in plants and soil and is given by (NIR-SWIR)/ (NIR+SWIR)
	WSI	Water satisfaction index is an indicator of crop or rangeland performance based on the availability of water to the crop during the growing season.
	SPI	Standardized precipitation index is used for estimating wet or dry conditions based on precipitation variables.
	SPEI	Standardized Precipitation Evapotranspiration Index used for estimating the severity, duration, frequency, and spatial extent of drought.

Category	Indices	Definition
	LST	Land Surface Temperature is how hot the surface of the earth would feel to the touch in a particular location.
	RFE	Rainfall estimates is the estimated precipitation.

With specific reference to predicting rice yield through the incorporation of ML with satellite data, various combinations of factors have been utilized. For instance, Jeong et al. (2022) used VIs, transplanting dates, solar radiation, temperature data, yearly paddy maps as well as county and province information to forecast rice yield at pixel scale utilizing deep leaning models with satellite data. In a recent study, Clarke et al. (2024) utilized crop production and climatic datasets to investigate the effect of dataset construction and data pre-processing on the XGBoost algorithm when employed for head rice yield forecasting and recommended the incorporation of data on irrigation management, remotely sensed VIs and crop phenology to improve accuracy of models generated. In another recent work, Brinkhoff et al. (2024) utilized weather time series and remote sensing datasets with ML models to generate rice yield predictors for their study on analysis and prediction of rice yield using phenology-based aggregation of satellite and weather data in Australia.

Furthermore, it has been established that SAR backscatter datasets can be employed in agricultural studies to monitor key factors linked to crop growth (Bogdanovski et al., 2023). The SAR backscatter data is superior to optical sensors in that they are unaffected by atmospheric conditions and have the ability to penetrate vegetation canopies (Wessels et al., 2023). Moreover, these sensors do not suffer from saturation when crops attain medium to high canopy cover as often happens with optical sensors (Bogdanovski et al., 2023).

In one study, Yu et al. (2023) employed SAR data, optical data and meteorological data for improved prediction of rice yield at field and county levels. In a different study, Singha

& Swain (2023) used crop biomass data, VV and VH polarization values for monitoring rice crop phenological characteristics and yield estimation with ML. In another work, Wang et al. (2019) estimated rice yield at field scale using S1 SAR backscatter data in coastal saline region of Jiangsu province in China. It was established that the proposed SAR simple difference index (SSD_{VH}) derived from the difference of VH polarization between end of tillering and end of grain filling stages exhibited a strong exponential relationship with rice yield, hence producing accurate yield estimation with a RMSE of 0.74 t ha^{-1} and relative error (RE) of 7.93%.

In another study, Li et al. (2023) acknowledges that despite some studies attempting the extraction of yield related data from satellite data, recent studies have concluded that using satellite data only is insufficient in yield estimation studies. Many studies have put forward the premise that methods involving combinations of several data sources such as satellite data, climatic data, soil data, field management datasets as well as historical yield data performs better than approaches relying on single types of data (Cedric et al., 2022; Yli-Heikkilä et al., 2022). This study therefore involved working with SAR data and auxiliary climatic data in the study area by enhancing the work done by Islam et al. (2023), Aworka et al. (2022)and Jhajharia et al. (2023) in their studies.

2.2 Mapping Crop Extent Using Sentinel-2 Data

Satellite systems such as S2 A and B with a high revisit frequency of 5 days and improved spatial resolution (10 m, 20 m, and 60 m) offer a solution for evaluating areas with smallholder agricultural landscapes (Elders et al., 2022). S2 and its Multispectral Instrument (MSI) is well suited for land surface and agriculture monitoring. S2 has a high revisit, which presents an opportunity to address limited temporal and spatial resolution gaps in conventional remote sensing platforms, which are required for continuous and dynamic monitoring of crops particularly during the growth stages.

In one study, Ali et al. (2020) utilized NDVI and LAI derived from S2 satellite data to monitor and estimate the rice spatial distribution and yield under intensive agriculture system in Sakha station Egypt. The findings established that multi-temporal resolution NDVI covering the entire rice growing period was effective in distinguishing plot boundaries which helped in crop mapping with an accuracy of 95% and a kappa coefficient of 0.93. Yield estimation was determined by employing an empirical yield-forecasting model using NDVI and LAI. The study concluded, that NDVI obtained during the rice heading stage was the most representative in forecasting yield. In another study, Elders et al. (2022) used multi-temporal moderate resolution S2 imagery in assorted smallholder landscapes in Burkina Faso to evaluate crop type and yield prediction skill of RF models. The study established that using RF models on multiday moderate resolution S2 image can predict crop type and rice yield in the study area with an accuracy of greater than 80% during the rainy season. These studies showed that optical remote sensing data could be effective in the identification and mapping of rice crop. However, these studies do not identify rice growth stages, which would have been ideal in estimation of yield potential as well as monitoring crop health and stress. Many studies using remote sensing have focused on only mapping rice field extents and fail to identify the rice growth stages which would be valuable especially in regions where the period or frequency of cropping is variable (Fatchurrachman et al., 2022). Therefore, the incorporation of S2 optical data could provide important spectral data for the mapping of rice fields and identifying rice growth stages effectively.

2.3 Mapping Crop Extent Using Sentinel-1 SAR Data

The S1 mission holds two polar orbiting satellites, (1A and 1B) which constitutes the Copernicus initiative of the European Radar Observatory. The satellites can operate day and night allowing imaging of the earth's surface regardless of the weather through the utilization of the C-band SAR operating at a centre frequency of 5.405 GHz (Bogdanovski et al., 2023). The interferometric wide swath (IW) is the observation mode used over land surfaces and is mostly used for agricultural studies. S1 has a 12-day revisit cycle with a nominal resolution of 10 m in the interferometric wide-swath (IW) (J. Wang et al., 2019). The IW mode supports a swath-width of 250 km and a geometric resolution of 5 m × 20 m in range and azimuth, respectively (Mansaray et al., 2020).

The S1 C-band SAR instruments supports operation in single polarization (HH / VV) and dual polarization (HH+HV / VV+VH), executed through one transmit chain (switchable to H or V) and two parallel receive chains for H and V polarization (Rosenqvist, 2018). The possible integrations are single band VV or HH, and dual band VV+VH and HH+HV are shown in Table 2.2.

Table 2.2: Single Band and Dual Band Polarizations

Polarization	Description
VV	single co-polarization, vertical emission/vertical reception
HH	single co-polarization, horizontal emission /horizontal reception
VV + VH	dual-band cross-polarization, vertical emission /horizontal reception
HH + HV	dual-band cross-polarization, horizontal emission /vertical reception

Source: (Rosenqvist, 2018)

Studies have shown that SAR data can provide vital information on crop characteristics such as canopy height, biomass and density, which are essential parameters in estimating crop yield (Amankulova et al., 2024; Wessels et al., 2023).

In one study, Luo et al. (2021) used monthly composites from S1 and S2 images for major crop mapping at regional scale with GEE in North-eastern China. The study utilized several reflective bands and indices covering the critical crop growth period derived from S2 imagery including red, blue, green, NIR, red edge, SWIR, NDVI, EVI and Land surface water index (LSWI). Additionally, VV and VH polarization were obtained from S1 dual polarized C-band SAR instrument sensors. The study integrated monthly composites of reflective bands, vegetation indices and polarization bands as input features before undertaking crop classification using the RF classifier. The findings indicated that integrating S1 and S2 monthly composites combined with the RF classifier could be used to develop a crop distribution map with good overall accuracy of 89.75%. It was substantiated that classification performance using time series images is significantly superior as opposed to utilizing single period images. The study recommended that future studies should explore the feasibility of the proposed methodology in other regions and settings. In a different study, Xu et al. (2023) proposed a novel SAR-based Paddy Rice Index (SPRI) technique to extract paddy fields in cloudy regions from SAR backscatter time series. The SPRI method was tested at five sites with variable climatic conditions, landscape complexity and cropping systems. The proposed method achieved an accurate classification map with an overall accuracy of greater 88% and an F1 score of over 0.86 pointing to the effectiveness of SAR in rice crop identification. These studies only managed to map rice area extents and did not identify the rice growth stages, which would provide useful data for planning, handling, redistribution, processing and export of rice crop.

In another work, Thorp & Drajat (2021) evaluated different deep learning approaches for the mapping and classification of rice growth stages using SAR and optical remote sensing data from S1 and S2 satellites in Indonesia. The study incorporated a recurrent neural network (RNN) model with long short-term memory (LSTM) nodes, which showed superior performance achieving classification accuracies of 79.6% for model training and 75.9% for model testing. These findings brought to the fore the ability of deep learning models in the classification of rice growth stages including tillering, heading and harvesting. The study recommended further research in the evaluation of ML approaches using ground-based data and satellite datasets covering longer periods for other crop types. The study points to the effectiveness of deep learning approaches in mapping rice extent and cropping calendars, however, the study used 21,696 observations for model training which has logistical, time and cost implications.

2.4 Crop yield Prediction using Conventional Methods

Crop yield estimation is an essential aspect of modern agriculture that integrates historical records with current environmental data to predict future crop production (Bauder &

Randall, 1982). In the recent past, majority of the yield estimation approaches including statistical-based models and crop simulation models (CSMs) still rely on traditional methods in establishing relationships between crop yields and explanatory variables (Cao et al., 2021). In particular, CSMs such as APSIM and DSSAT-CERES-Rice allows for the simulation of crop growth, development and yield (Gavasso-Rita et al., 2024). However, these models often require extensive information on climate, soil, crop varieties, and management practices, making them costly to implement where such data are limited or unavailable (Cao et al., 2021). Furthermore, the accuracy of CSMs relies heavily on input data quality, particularly weather, soils and management information which is often course and unavailable in some instances (Hashemi et al., 2024; Lokesh et al., 2026).To address these limitations, remote sensing has transformed agricultural monitoring by providing continuous, large-scale, and non-invasive observations of crop conditions throughout the growing season (Lokesh et al., 2026). As such, many studies often rely on satellite data with machine learning algorithms for yield prediction (Hashemi et al., 2024; Islam et al., 2023; Sánchez et al., 2025; Yu et al., 2023). Additionally, Betew et al. (2026) established in their review that machine learning algorithms consistently achieved higher prediction accuracy than traditional empirical or statistical models.

2.5 Crop Yield Prediction Using Machine Learning Models

Machine learning (ML) is a field of study in artificial intelligence, which comprises a set of tools and structures that have the capability to gather and assimilate information from extensive observations, enhancing and expanding itself by acquiring new knowledge through learning rather than direct programming (Woolf, 2009). ML models are being employed in agriculture landscapes for improved yield prediction due to their ability to determine patterns and correlations and discover knowledge from a given dataset. ML models can be descriptive or predictive where descriptive models are utilized in gaining knowledge from collected data and describe a given occurrence while predictive models are used in predicting the future (van Klompenburg et al., 2020).

Commonly applied ML algorithms for crop yield prediction include empirical regression models (e.g. Ridge), ensemble learning approaches (e.g. Random Forest, Gradient Boosting) Support Vector Regression and Artificial Neural Networks (Desloires et al., 2023).

Deep learning algorithms is a form of ML with the capability of predicting outcomes given a variety of raw data combinations. Deep learning utilizes artificial neural networks containing several layers to solve complex problems. Neural networks are composed of layers of nodes where nodes in a given layer are linked to nodes contained in another layer thus allowing signal travel between the nodes and assignment of weights (Jhajharia et al., 2023).

2.5.1 Crop Yield Prediction with Random Forest Model

The Random Forest (RF) algorithm operates based on ensemble learning, a method that combines multiple classifiers to tackle complex problems, boosting the accuracy and performance of the model. RF is a classifier which employs numerous decision trees on a random subset of training data and a random subset of the input features (Sharma et al., 2024). In simpler terms, instead of relying on just one tree, this algorithm considers the predictions of each tree and makes its final decision based on what most of the trees agree upon (Breiman, 2001; Jhajharia et al., 2023). This approach is particularly fitting for agronomic data, which typically exhibits high dimensionality and variability (Santos et al., 2024). Additionally, RF models are increasingly being applied in remote sensing studies due to their ease of implementation, speed and overall accuracy (Elders et al., 2022).

The RF model allows for strong prediction while reducing the loss error between crop targets and the reality (Aworka et al., 2022). For a given crop training dataset denoted by $D = \{(x_i + y_i)\}_{i=1}^{N}$ where N represents number of instances and B = Number of crop prediction trees. Each crop tree denoted by $\{T_b\}$ is developed from a random sample of the crop training data with replacement D* = (x_b, y_b) through the bootstrapping method.

Therefore, for each crop decision tree $\{T_b\}$ to fit on a particular sample dataset to forecast yield, the training data D is resampled with replacement to create individual subsets D*. The decisions obtained by each tree $\{T_b\}$ are combined by taking the mean outputs of all crop trees associated with the model to give the final crop yield. Hence, for any new crop input data *x*, the yield forecast is obtained by determining the mean prediction values from all the individual crop trees $\{T_b\}$ using Equation 2.1 obtained from the work by Aworka et al. (2022)

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x). \tag{2.1}$$

In a recent study, Santos et al. (2024) investigated the potential of RF as effectual ML model for predicting soybean yield for limited datasets in South of Alexandria, LA from 2016 to 2018. Gaps were identified in the original dataset containing 120 points and hence the Gaussian Copula and Simple Imputer methods were used to extend the dataset while maintaining the information, structure and variability of the original dataset. Applying the simple imputation technique for data enhancement, the findings revealed that an optimized RF achieves an average MSE of 0.1 while LR obtains an average MSE of 0.32. However, a reversal was obtained when Copula method was incorporated where RF achieved an MSE of 0.42 which was greater than for LR which attained an average MSE of 0.24. It was concluded that RF was more robust to changes in the sample size of the training sample dataset having a lower MSE value compared to LR independent of the sample size under consideration. Further, the study proposed exploration of the data enhancement and resampling techniques through replicating the analysis over multiple years and locations. This study does not consider the incorporation of SAR satellite data in its datasets, which would capture crop canopy characteristics and perhaps improve ML model performance. There is also need to test the performance of RF model in predicting other crops such as rice in other settings and locations.

In another work, Jhajharia et al. (2023) compared crop yield prediction utilizing ML and deep learning methods in Rajasthan state of India for rapeseed and mustard, wheat, barley, bajra and jowar. The study employed five different algorithms including ML models (RF, SVM and Lasso Regression) and Deep learning models (Gradient Descent and Long Short-Term Memory (LSTM). The findings of the analysis undertaken showed that RF outperformed the other models achieving 0.963 R^2, 0.035 RMSE and 0.0251 MAE. From the algorithms used ML models showed superior performance compared to deep learning models. However, the study recommends inclusion of remote sensing data combined with district level agricultural statistics to improve model performance in crop yield estimation. There is also need to test the performance of RF model in predicting other crops such as rice in other settings and different conditions.

In different study, Yu et al. (2023) improved the prediction of rice yield at field and county levels in China through the synergetic utilization of SAR backscatter, optical and meteorological data. The study utilized four backscattering S1 SAR metrics including VH and VV polarizations, ratio of polarization VV to VH, addition of polarization VV and VH and subtraction of polarizations VV from VH. The SAR data was combined with optical data from S2 and meteorological data as input variables for regression algorithms. The study proposed the utilization of meta-learning ensemble regression (MLER) which was developed by integrating three machine learning models algorithms (RF, XGBoost and SVR) for rice yield prediction. The findings established that S1 SAR could substitute the lack of optical data from S2 for rice yield prediction which was further enhanced with the incorporation of meteorological data. The MLER algorithm achieved superior performance in field level prediction accuracy (R^2=0.89, RMSE=0.54t/ha) compared with individual base algorithms (RF: R^2 = 0.80, RMSE = 0.72 t/ha; XGBoost: R^2 = 0.76, RMSE = 0.80 t/ha; SVR: R^2 = 0.74, RMSE = 0.83 t/ha) and the long short-term memory (LSTM) (R^2 = 0.76, RMSE = 0.80 t/ha) with a ten-fold cross validation. The study proposed further research into the incorporation of critical rice crop growth stages and use of phenological information in yield prediction. This study points to the fact that SAR and Optical data from S1 and S2 satellites respectively can be effective in predicting rice yield especially when coupled with meteorological data. However, there is need to test this methodology in other settings and localities to establish the performance of ML algorithm effectiveness in predicting rice yield.

2.5.2 Crop Yield Prediction with Extreme Gradient Boosting Model

Extreme Gradient Boosting (XGBoost) is a decision ensemble technique that utilizes the gradient boosting approach by constructing an additive model by minimizing a specified loss function. Since XGBoost exclusively utilizes decision trees as its base classifiers, it incorporates a modified loss function to manage the complexity of these trees (Bentéjac et al., 2019). In the recent past, XGBoost has become a popular ML algorithm being applied for crop yield prediction (Clarke et al., 2024). Further, Li et al. (2023) stipulates that when the classification and regression trees (CARTs) in XGBoost are developed sequentially, highly correlated features do not impact the model, thereby minimizing issues related to feature multicollinearity.

For the XGBoost model, weak crop leaners are built iteratively in such a way that each new crop learner is trained utilizing the errors or residuals of the prior crop learner until residuals of zero are attained as the model improves (Aworka et al., 2022). The XGBoost model is developed with parallelization in computational feature gain, algorithmic improvements, and depth-first tree pruning to increase training speed and ensure high accuracy (Li et al., 2023). Mathematically, XGBoost can be expressed using Equation 2.2 obtained from the work by Li et al. (2023)

$$.\hat{y_i} = \theta(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathrm{F} \tag{2.2}$$

Where $\hat{y_i}$ is the prediction term, x_i is the vector of the i^{th} samples, K is the number of trees, f_k is a function in functional space F, and $\mathrm{F} = \{f(x) = w_{q(x)}\}(q: \mathbb{R}^m \longrightarrow T, w \in \mathbb{R}^T)$ is the set of all possible classification and regression trees (CARTs). Each function f_k has an independent tree structure q and leaf weights w. The objective function of XGBoost can be optimized using Equation 2.3 obtained from the work by Li et al. (2023)

$$.\mathcal{L}(\theta) = \sum_{i}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.3}$$

Where l is a differentiable convex loss function that estimates the difference between the prediction $\hat{y_i}$ and the target yi and Ω is a regularization term penalizing the complexity of the model. Equation 2.3 is decomposed in an additive way into an optimization for each tree in the model because it cannot train it directly.

For the prediction $\hat{y}_i^{t}$ of the i^{th} samples at the t^{th} iteration, a new tree *ft* needs to be added to minimize objectives given in Equation 2.4 obtained from the work by between predicted and actual values (Altork, 2025). Table 3.5 obtained from obtained from the work by Li et al. (2023).

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (2.4)$$

Which means that the new tree *ft* is added to improve the XGBoost model.

Second order approximation with the Taylor expansion of the loss function is computed to speed up the optimization of each tree as shown in Equation 2.5 obtained from the work by Li et al. (2023).

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \quad (2.5)$$

Where g_i and h_i are the first and second order gradient statistics on the loss function respectively and may be expressed by Equations 2.6-2.7 obtained from the work by Li et al. (2023).

$$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) \quad (2.6)$$

$$h_i = \partial_{\hat{y}^{(t-1)}}^2 l\left(y_i, \hat{y}^{(t-1)}\right) \quad (2.7)$$

The complexity of the model Ω is defined in XGBoost as shown in Equation 2.8 obtained from the work by Li et al. (2023).

$$\Omega(f) = \Upsilon T + \frac{1}{2} \lambda \|w\|^2 \quad (2.8)$$

Where T is the number of leaves, w is the leaf weight and the parameters λ and Υ determine the strength of penalization. This complexity control approach is vital in preventing overfitting of the XGBoost model.

In one study, Singha & Swain (2023) monitored rice growth with S1 SAR backscatter data using ML models in GEE cloud in West Bengal, India. The study utilized VV, VH polarization values and crop biomass data to monitor crop phenological characteristics and crop yield. The study employed various ML models including RF, SVM, XGBoost, Decision Tree (DT), K-Nearest Neighbours (KNN) and Logistic Regression (LR) integrated with GIS in GEE cloud. The findings from the study revealed that rice yield prediction through SAR data with ML models RF, XGBoost and DT showed good correlation achieving an R^2 of greater than 0.8. The study recommended further research using the proposed yield prediction methodology by incorporating additional datasets such as meteorological data, soil information biophysical data and socio-economic data among others. This study, utilizes SAR data effectively in rice yield prediction using ML in India, however, there is need to try this methodology in other locations or settings and assess the performance of the XGBoost compared to other ML models.

In another work, Islam et al. (2023) estimated rice yield utilizing NDVI remote sensing data and meteorological data with ML in Nepal. The study established that factors including NDVI, rainfall, soil moisture, and evapotranspiration had a strong correlation with rice yield. Further, it was confirmed that tree-based regression ML models particularly the hybrid stack ensembles of multiple models (Random Forest, Light Gradient Boost, Gradient Boost and XGBoost) showed superior performance with limited data points achieving a RMSE of 288 kg/ha compared to the RMSE of 685 kg/ha attained by the LR model. The study recommended use of Sentinel-1 derived datasets for rice fields delineation and estimation of a vegetation index that could impact prediction accuracy. This study does not utilize additional datasets such as SAR backscatter, which could be combined with auxiliary climatic datasets to potentially improve the performance of ML models in yield prediction. Further, there is need to explore this methodology in other locations or settings and assess the performance of the XGBoost compared to other ML models.

In another study, Li et al. (2023) developed a framework for soybean yield prediction through the incorporation of XGBoost model together with multidimensional feature engineering at county level in the US. The findings of the train-validate-test assessment indicated, excellent accuracy values achieving an R^2 of 0.82 and RMSE of 0.246 t/ha, using the XGBoost algorithm. From the study, the XGBoost algorithm achieved superior performance compared to other county-level soy prediction models including random forest (RF), linear regression (LR), k-nearest neighbor (KNN), artificial neural network (ANN), support vector regression (SVR), long short-term memory (LSTM), and deep neural network (DNN). It was further established, that the incorporation of principal component analysis (PCA) to minimize dimensionality during development of yield prediction framework greatly reduced the accuracy of XGBoost model yield prediction when other factors are held constant. Hence, the study recommends further research to validate yield prediction using the XGBoost model. This study does not utilize SAR data in its dataset, which could potentially improve model performance by capturing crop canopy characteristics not captured by the optical satellite datasets. Further, there is need to explore this methodology using other crops such as rice in other locations and settings.

2.5.3 Crop Yield Prediction with Support Vector Machine Model

Support Vector Machine (SVM) utilizes various types of kernel functions such as sigmoid, radial basis function (RBF), linear and polynomial to reconstruct data into a higher dimensional space with the aim of reducing time complexity and improving the accuracy of the model (Sharma et al., 2024). The aim of SVM algorithm is to identify a hyperplane in N-dimensional space (where N is the number of features present in a given dataset) that separates the data points into unique classes. Multiple potential hyperplanes maybe considered to separate the two classes of data points. The ultimate goal is to locate a hyperplane with the optimal margin, meaning the greatest distance between data points belonging to different classes is identified (Jhajharia et al., 2023). The SVM model has the advantage of being less susceptible to overfitting problems and degradation of computational performance in high dimensions (Son et al., 2022).

The SVM model identifies hyperplanes in *m* dimensional space that distinctly isolates crop data points and the hyperplane is utilized to forecast crop yield. Generally, the aim is to develop a hyperplane such that the crop data points closest to the hyperplane are within the hyper tube (decision boundaries) a distance denoted by ε from the hyperplane (Aworka et al., 2022). Therefore, SVM model can be expressed linearly by Equation 2.9 obtained from the study by Aworka et al. (2022).

$$. -\varepsilon \leq y_i - \omega^T . x_i - b \leq \varepsilon \qquad (2.9)$$

Where ε is the user-defined error, *xi* is the vector of crop features of the *i*th yield training instance, *yi* is the crop yield target of the *i*th training instance and ω is a vector of learning parameters that will fit to the crop yield prediction.

Given that not all points lie within the hyper tube then the crop yield dataset is not linearly separable. Therefore, slack variables ξ_i, ξ_i^* for each crop data point *i* are introduced to allow for particular violations of certain limits. This means an optimization problem can be defined as a reduction of the crop data points outside the hyper tube while relaxing some constraints. Further, it is expected that the amount of data points inside the margin to be few as possible and have as little penetration of the margin as possible. As such the optimization problem can thus be defined as:

$$\text{Minimize} \qquad \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \qquad (2.10)$$

Subject to: $y_i - \omega . x_i - b \leq \varepsilon + \xi \, ; \, \omega . x_i + b - y_i \leq \varepsilon + \xi_i^* \, ; \, \xi_i , \xi_i^* \geq 0$

Where: ω is the weight vector, *b* is the bias term, ξ_i and $\xi_i{}^*$ are slack variables for points outside the hype tube, *C* is a regularization parameter.

In a recent study, Rajakumaran et al. (2024) employed a multi-attribute weighted tree-based support vector machine (MAWT-SVM) approach to predict corn yield in India. The study utilized datasets on agricultural productivity and meteorological statistics spanning 8 years (1999-2007). The obtained dataset was pre-processed using the z score normalization approach and the PCA was utilized for feature extraction with the Genetic Algorithm being used for feature selection. Performance indicators analysis of the MAWT-SVM study included accuracy (98.5%), precision (97%), recall (95%), Dice coefficient (98.5%) and Jaccard co-efficient (98.6%). The study concluded that the MAWT-SVM system outperformed other models such as CNN, DNN and MLR-ANN used in past similar studies. The study suggested future investigations in exploring crop production prediction across a variety of crops while employing dependent and independent parameters. This study does not incorporate SAR data in its dataset, which could potentially improve SVM model performance by capturing crop characteristics not captured by the agricultural and meteorological statistical data. Further, there is need to explore this methodology using other crops such as rice in other locations and settings.

In another study, Aworka et al. (2022) proposed a decision tool at country level for East African countries based on ML models for crop yield estimation. Crop Random Forest (CRF), Crop Gradient Boosting Machine (CGBM) and Crop Support Vector Machine (CSVM) models were trained using climatic data, crop data and pesticide data to propose a decision system to estimate crop yield. The model performance evaluation indicated that CRF had the highest coefficient of determination (R^2) (92.272%) outperforming CGBM (90.186%) and CSVM (86.377%). The aforementioned findings revealed that the proposed ML models predicted crop yield within acceptable levels of precision and accuracy. It is suggested that, perhaps the addition of more predictor variables such as water data, wind data, pollution data meteorological variability as well as economic data could enhance the model quality. This study does not incorporate SAR data in its dataset, which could potentially improve ML model performance by capturing crop characteristics and hence augment the climatic data, crop data and pesticide data used in yield estimation.

Further, there is need to explore this methodology using other crops such as rice in other locations and settings.

In a different study, Son et al. (2022) attempted to use monthly composites derived from S2 imagery to forecast rice yield one month prior to the harvesting period at field level using ML in Taiwan. The study applied RF, SVM and ANN in four consecutive seasons from 2019 to 2020 using field data. The study's findings on yield modelling and predictions indicated that the SVM model performed marginally better than RF and ANN. Validation results from the SVM models, using data from the transplanting to ripening stages, revealed RMSPE and MAPE values of 5.5% and 4.5% for the 2019 second crop, and 4.7% and 3.5% for the 2020 first crop, respectively. This study does not incorporate SAR data in its dataset, which could potentially improve ML model performance by capturing crop characteristics not captured by the optical S2 data. Further, there is need to explore this methodology in other locations and settings.

2.5.4 Crop Yield Prediction with Linear Regression Model

Linear Regression (LR) may be defined as a prediction technique that utilizes the relationship between a dependent and one or more independent variables (Altork, 2025). The method fits a linear equation to find the line of best fit that minimizes the difference between simulated and actual values (Sharma et al., 2024). Hence, it is a useful method for anticipating the measure of a response variable from a particular estimate of explanatory variable. The linear regression expression is given by Equation 2.11.

$$\mathcal{Y} = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n + \varepsilon \qquad (2.11)$$

Source: (Altork, 2025)

Where ω_0 is the intercept, $\omega_0, \omega_1, \ldots, \omega_n$ are the coefficients of the independent variables, $x_1, x_2, \ldots, x_n$ are the independent variables and ε is the error term.

In one study, Ansarifar et al. (2021) used multiple linear regression model to develop an interaction regression model for crop yield prediction. However, according to van Klompenburg et al. (2020), LR is widely used as a benchmarking algorithm to assess whether proposed ML algorithms perform better than LR or not.

2.6 Performance Evaluation of Machine Learning Models

Using models developed from prior year data to predict future yields is widely used in agricultural studies and has been utilized as an effective approach of estimating potential crop yields for the upcoming season aiding farmers to make informed crop management decisions (Guo et al., 2023). Error measurements are often utilized in model comparison as well as assessment of model accuracy. Statistical methods including Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) are widely used for error measurements (Ali et al., 2021). The RMSE is widely applied in accuracy analysis depicting the standard deviation of the prediction errors when dealing with differences between observed and simulated values. Generally, a lower value in RMSE is preferred to a higher value as this indicates less error in model prediction (Guo et al., 2023; Singh Boori et al., 2023). The RMSE is considered useful and valuable when large discrepancies between datasets are undesirable, whereas the MAPE is less affected by significant errors, providing a clearer measure of the average error. The MAPE may be utilized to illustrate an interpretable assessment of the average accuracy of predictions with respect to average observed yield (Desloires et al., 2023). Model accuracy is rated as very good if the MAPE values are under 10%, good if they fall between 10% and 20%, fair if they range from 20% to 30%, and poor if they exceed 30% (Son et al., 2022). The MAE may be defined as average differences between predicted and observed yield values (Singha & Swain, 2023). Further, these parameters (MAE, MAPE and RMSE) have also been successfully utilized in different studies for ML model evaluation (Ali et al., 2021; Hashemi et al., 2024; Jhajharia et al., 2023). As such these parameters were deployed in this study due to their effectiveness in accuracy analysis for evaluating the performance of ML models in rice yield prediction.

2.6.1 Performance Evaluation of ML Models with Out of Training Sample Data

Yield forecasts can be utilized to evaluate variations in performance between parcels in a given study area where reference or ground truth data is sparsely available. Nevertheless, a practical application for food security would involve the prediction of yield for an ongoing season prior to harvesting which would mean extrapolating yield to test samples from periods outside the training samples (Desloires et al., 2023).

In light of this, Desloires et al. (2023) used a novel approach leave one year out cross validation ($LOYO_{CV}$) approach which entailed considering four years for model training and one for testing out of a dataset consisting of five years. The four years were then divided into three years for model training while one year was used for model validation. Hence, after model training using the four years of data the yield simulations for each test year were compared to the crop yields for that season. For the evaluation matrix adopted for the extrapolation of corn yield for each season, utilizing the RF, SVR, XGBoost and MLP_{ENS} there was no clear best performing model. However, the models described much of the variation in corn yields. The study recommended further research in exploration of in season yield prediction using ground truth data at field level due to the complexity and lack of studies on extrapolation of crop yields. Even though this study attempts to evaluate the performance of ML models in extrapolating corn yield, more work is required in applying this methodology using other crops such as rice in other settings and localities.

In another study, Guo et al. (2023) estimated out of sample prediction performance of the best performing ML model (neural networks model) in a study involving smallholder maize yield estimation using satellite data with ML in Ethiopia. The ML model was trained utilizing datasets from the previous year (2017) and its predictive ability was tested on data from the following year (2018). The neural networks model achieved an R^2 value of 0.43 indicating moderate ability in forecasting future yields. The study presumed that the model accounted for about 43% of the variability in yield for a different year. Even though the study showed that a ML model developed and calibrated on previous year's data could give reasonable maize yield estimates in the subsequent year it was recommended that future work explores the robustness of the out of sample predictions in other settings. In spite, of this study attempting to evaluate the performance of ML models using out of sample corn yield forecasting, more work is needed in applying this methodology using other crops such as rice in other settings and localities.

2.7 Research Gap

In Africa, studies utilizing artificial intelligence for agricultural prediction are still limited. In East Africa, agricultural prediction methods largely rely on traditional approaches such as crop cuts and farmers estimates often rely on expert evaluation of sample areas making the prediction exercise time consuming, labour intensive and prone huge errors (Aworka et al., 2022). Additionally, few studies have been undertaken involving yield estimation of rice or other crops using satellite SAR backscatter data (Wang et al., 2019). In one study, Fatchurrachman et al. (2022) noted that even though remote sensing technology has been employed widely to map rice fields worldwide, most studies have only mapped the rice field extent and did not attempt to identify the growth stages. Timely and accurate data on rice growth stages is vital for planning and management of rice farming systems at both regional and national levels (Ramadhani et al., 2020). Therefore, this study, sought to characterize rice growth stages i.e. transplanting, vegetative, reproductive and maturity stages using spectral data derived from Sentinel satellites. Furthermore, this study utilized time series S1 SAR backscatter data (VV and VH polarization data) coupled with auxiliary climatic data (rainfall data, minimum and maximum temperature data) to predict rice yield using four machine learning models including RF, XGBoost, SVM and LR. Finally, evaluation of the performance of RF, XGBoost, SVM models in predicting rice yield was assessed using LR as a benchmark.

2.8 Conceptual Framework

The conceptual framework for this study was as shown in Figure 2.1 illustrating the interrelation between satellite and climatic auxiliary datasets used in the identification of rice growth stages and development of a rice yield-prediction model. The intervening variables include: satellite image quality, method used in image classification, ML model used in yield prediction and hyperparameter tuning of the ML model.

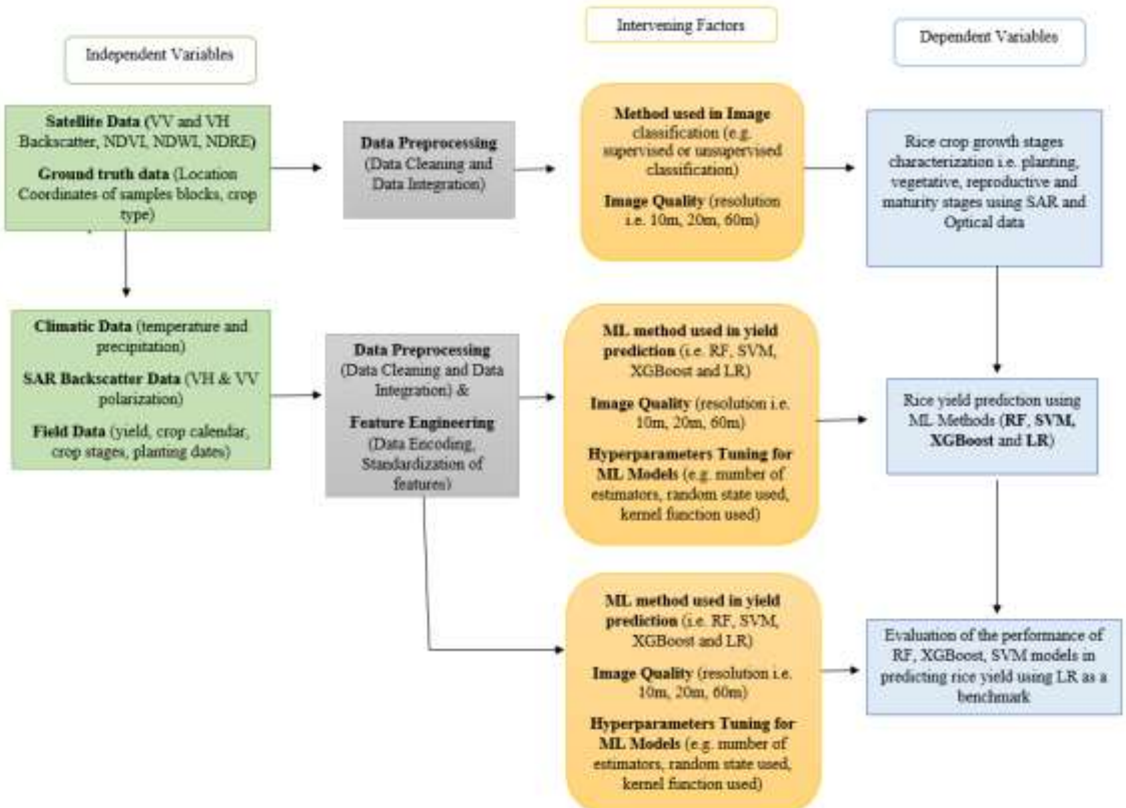

Figure 2.1: Conceptual Framework for the Study

CHAPTER THREE

MATERIALS AND METHODS

3.1 Description of the Study Area

Figure 3.1 illustrates the study area where the research was undertaken. The Mwea Irrigation Scheme (MIS) is located in Mwea West and Mwea East Sub counties in Kirinyaga County about 100 km north east of Nairobi, the capital city of Kenya and extends to the south of Mount Kenya. The scheme has a total area of about 12,140 ha (Akoko et al., 2020). Further, MIS has about 8500 ha of irrigated paddy area with improved Basmati rice being the main crop grown (Njagi & Mano, 2023).

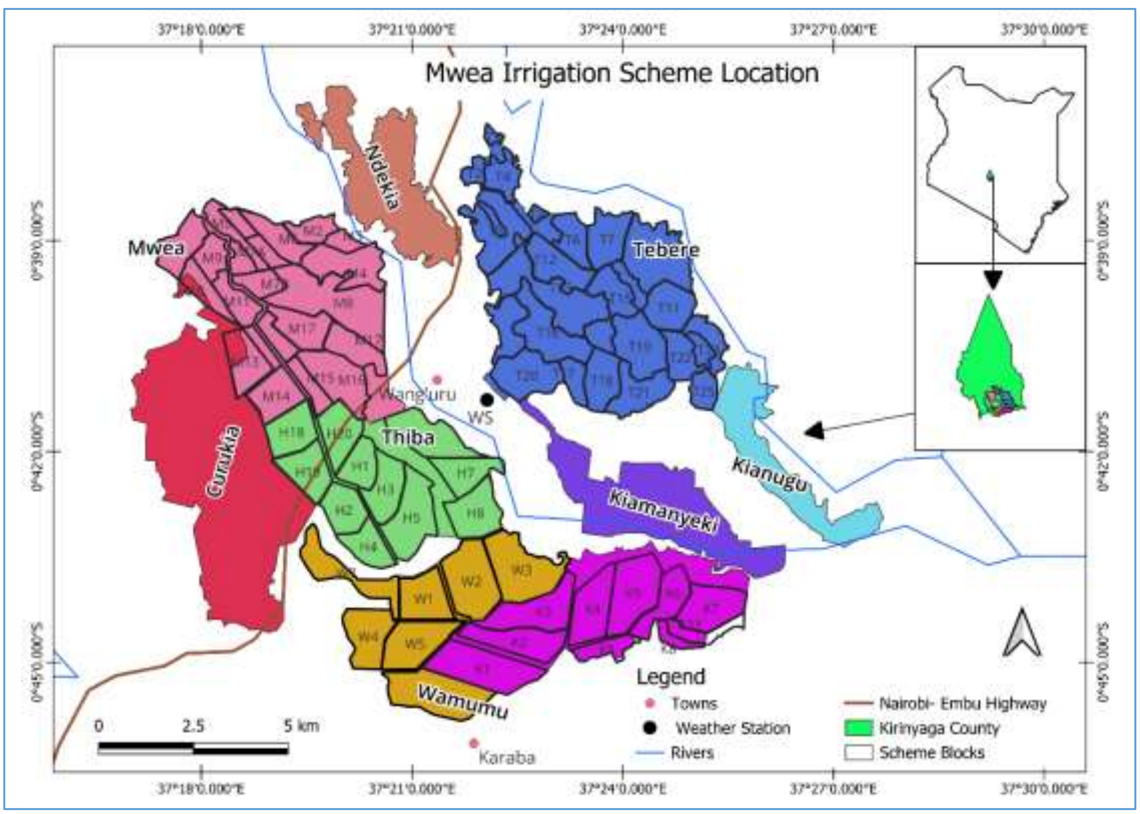

Figure 3.1: A Map Showing Area of Study in Mwea Irrigation Scheme, in Kirinyaga County, Central Kenya

The Mwea region has an altitude of approximately 1159 masl, with an average annual temperature of about 22 ^{0}C and an average annual rainfall of about 930 mm. The local climate is of tropical type and experiences two wet seasons: a long rains season from March to May and the short rains season from October to November. The predominant soil type is Vertisol, which covers most of the paddy fields (Watanabe et al., 2021). The scheme is divided into seven key sections as shown in Figure 3.1. The MIS sections include: Mwea, Tebere, Wamumu, Thiba, Karaba, Ndekia and Curukia and is served by River Nyamindi and River Thiba (Musila et al., 2024). The scheme is considered the largest region with irrigated rice production in Kenya (Uma et al., 2023). Hence, this makes it an ideal site for developing and testing a rice yield prediction ML model.

3.2 Data Collection

For this study, various datasets were required to accomplish the specific objectives of this study as outlined in Table 3.1.

Table 3.1: Summary of Data Collected for this Study for the Period 2021-2024

Data	Variable(s)	Period	Location	Source
Rice Production Data	Block location coordinates, block size, rice phenological data	2021, 2022, 2023 & 2024	Mwea Irrigation Scheme	National Irrigation Authority (NIA) Mwea Office
Rice yield data	Average block yields	2021, 2022, 2023 & 2024	Mwea Irrigation Scheme	National Irrigation Authority (NIA) Mwea Office
Climatic Data (Rainfall and Temperature Data)	Mean monthly time series	2021, 2022, 2023 & 2024	NIA Weather Station Mwea	NIA Weather Station-Mwea climateengine.org online database,
S1-SAR Backscatter (VH, VV polarization)	Median monthly time series	2021, 2022, 2023 & 2024	Mwea Irrigation Scheme	Sentinel-1 mission image collection on GEE (Copernicus Program)
S2 Optical Data (NDVI, NDWI, NDRE)	Maximum monthly time series	2024	Mwea Irrigation Scheme	Harmonized Sentinel-2 MSI image collection on GEE (Copernicus Program)

3.2.1 Sample Size

To determine the number of Scheme blocks to work with for this study, the sample size was calculated using the Cochran (1977) formula presented in Equation 3.1 and Equation 3.2. This approach is consistent with the methodology adopted by Abebe et al. (2022).

Infinite Population

$$n_0 = \frac{Z^2 pq}{e^2} \tag{3.1}$$

Where: n_0= required sample size for an infinite (or very large) population, Z = Z-score corresponding to the desired confidence level, 1.645 = 90% confidence; 1.96 = 95% confidence; 2.576 = 99% confidence, p = estimated proportion of the population possessing the characteristic of interest, q=1−p, e = desired level of precision (margin of error)

Finite Population Correction

$$n = \frac{n_0}{1+\frac{n_0+1}{N}} \tag{3.2}$$

Where n = adjusted sample size, n_0= initial sample size from Cochran's formula, N = total population size.

The calculated sample size was 55 Scheme blocks at a 95% confidence level from a finite population of 64 blocks distributed across the five main sections of the MIS, namely Karaba, Tebere, Wamumu, Thiba, and Mwea. However, this study included 62 irrigation blocks covering all five sections as shown in Figure 3.1 to maximize the amount of data available for the development and validation of the rice yield prediction models. The two blocks in Wamumu and Karaba sections were excluded due to serious data gaps covering the period 2021-2024. The allocation of blocks within the five scheme sections is as shown in Table 3.2.

Table 3.2: Blocks Distribution in the Scheme Sections

Mwea Scheme Sections	**Total Blocks**	**Total Blocks Used for this Study**
Karaba	10	9
Tebere	18	18
Wamumu	8	7
Thiba	11	11
Mwea	17	17
Total	**64**	**62**

3.2.2 Ground Truth Data

This study worked in conjunction with the National Irrigation Authority (NIA) Mwea office, which oversees the management of the scheme focusing on the operation and maintenance of the primary and secondary infrastructure while farmers are responsible for maintaining the tertiary infrastructure. Primary data on rice yield, cropping calendar, rainfall and temperature datasets were collected from the NIA Mwea office covering the period July to December for the years 2021 to 2024. During field visits to the scheme, GPS coordinates of sample blocks were obtained and recorded using a hand-held GPS receiver along with field photos of the sample parcels. Additionally, auxiliary datasets including maps of vegetation classes, soil, and land use of the study area were accessed from Google earth and earthmap.org online databases. The collected data aided in identification of the spatial distribution of crop cover in the study area according to their natural characteristics. These datasets proved critical in validating the rice growth stages identified as well as evaluating the accuracy of ML models in rice yield prediction.

3.2.3 Rice Yield Data

Rice yield data was obtained from NIA Mwea office covering the main season period July to December covering the years 2021, 2022, 2023 and 2024. The yield data covered the five main sections listed in Table 3.2 for the 62 blocks under consideration. This was informed by the review by De Clercq & Mahdi (2025) that analysed 153 rice yield prediction studies with ML and established that majority of the studies relied on existing historical yield datasets for yield prediction. As such, several studies have relied on existing datasets from state authorities for yield prediction utilizing machine learning with the aim of maximizing data available for ML model development (De Clercq & Mahdi, 2024; Jhajharia et al., 2023; Son et al., 2022).

3.2.4 Satellite Data

This study utilized satellite imagery i.e. S1-SAR backscatter data (VV and VH polarization) and S2 Optical data (NDVI, NDRE and NDWI) which were obtained from the European Space Agency's (ESA) repository on the cloud-based computing platform GEE. The spatial extents of image retrieval were determined using the minimum and maximum geographic coordinates of the study area. The temporal extents were defined with the main rice growing season at MIS, which spans from July to December (Watanabe et al., 2021). A GEE script was prepared and executed in the GEE code editor to obtain dual polarized, C-band, ground range detected (GRD) SAR product from S1 in the IW swath mode. A different GEE script was prepared and executed in the GEE code editor to obtain access to S2 Level-1C (Top-of-Atmosphere Reflectance, TOA). The characteristics of the satellite datasets utilized in this study are given in appendix I under the appendices section.

3.2.5 Climatic Data

Studies have evaluated the effects of numerous variables on agricultural yield. Temperature and precipitation variables have been established as having a higher impact on crop yields (Kang et al., 2020; Rajakumaran et al., 2024). Therefore, climatic datasets listed in Table 3.1 were obtained from NIA weather station shown in Figure 3.1 located in Mwea covering the period July to December for the years 2021-2024. To deal with missing data covering the period July to December for the years 2021-2024, daily rainfall data were obtained from the Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS) dataset, while daily minimum and maximum air temperature data were obtained from the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). Both datasets were accessed and extracted through the Climate Engine web platform (https://app.climateengine.org), which provides on-demand processing and retrieval of geospatial climate datasets using the Google Earth Engine cloud-computing environment.

3.3 Characterization of Rice Growth Stages Using SAR and Optical Data

The characterization of rice growth stages was accomplished through a phenology-based approach to map the rice field extents and estimate the cropping calendar for the scheme. This was attained through the utilization of S1- SAR backscatter and S2 optical time series datasets using unsupervised classification on the GEE platform. For characterization of rice growth stages, the satellite data time series covering the period between 1st July and 31st December 2024 was used. According to Fatchurrachman et al. (2022), rice growth stages can generally be divided into four distinct stages as shown in Table 3.3.

3.3.1 Generating Regions of Interest

The study area boundary was generated by digitizing the scheme boundary using the QGIS software utilizing data and maps obtained from the NIA Mwea office. The boundary was also checked using very high-resolution (VHR) images from Google Earth as well as ground truth data collected during field visits to the scheme. The VHR satellite imagery from Google Earth have been reported to have a spatial resolution finer than 5 m (Lesiv et al., 2018). The generated MIS boundary was saved in the shapefile format and uploaded to the GEE platform.

Table 3.3: Rice Growth Stages and their Descriptions

Growth stage	Days after Transplanting	Description
Transplanting Stage	20-30	This initial stage involves the flooding of fields and seed germination. Rice plants are just emerging.
Vegetative Stage	45-60	Characterized by rapid leaf development and canopy growth. The plants are establishing their structure, but no panicles are visible.
Reproductive Stage	20-30	The plants enter a reproductive phase, where panicles form, and flowering occurs.
Ripening Stage	35-65	The grain filling and maturation phase, where the canopy starts to senescence, and the leaves turn yellow before harvest.

3.3.2 Satellite Data Pre-processing

The GEE platform was utilized in accessing and pre-processing S1 SAR backscatter and S2 optical datasets for the study area covering the period of interest July-December for the year 2024. The image product details for the images accessed are given in appendix II under the appendices section. To obtain optical data, the S2 Level-1C (Top-of-Atmosphere Reflectance, TOA) image collection was utilized in this study as it has been widely applied successfully in several crop mapping research studies (Fatchurrachman et al., 2022; Inoue et al., 2020; Luo et al., 2021; Ramadhani et al., 2020). The S2 satellites are equipped with the Multispectral Instrument (MSI), a passive optical sensor which captures data across 12 spectral bands (Inoue et al., 2020). These sensors capture images having a swath distance of 290 km wide and offers resolutions ranging from 10 m-60 m (Inoue et al., 2020). Images from both Sentinel -2A and Sentinel 2B satellites were accessed from the GEE platform acquired during the main rice growing season from 1st July 2024 to 31st December 2024. Band 3 (Green), Band 4 (Red), Band 5 (Red-Edge) and Band 8 (Near-Infrared) from the S2 MSI data were used in the computation of Normalized Difference Water Index (NDWI), Normalized Difference Vegetation Index (NDVI) and Normalized Difference Red Edge (NDRE). A cloud mask algorithm script was prepared and executed on the GEE code editor to filter spectral data based on cloud cover. The image collections employed in this study had a suitable spatial resolution of 10 m.

This study also utilized S1 images spanning 1st July and 31st December 2024 obtained from the Sentinel-1 ground-range-detected (S1_GRD) image collection. In particular, the level 1 Ground Range Detected (GRD) products in Interferometric Wide (IW) swath mode were selected. The S1 satellite operates using the C-band SAR active sensor which allows for data acquisition during cloud cover conditions making it well-suited for rice paddy monitoring assignments (Inoue et al., 2020). The IW mode image collection provided dual polarization with vertical transmit, vertical receive (VV) and vertical transmit, horizontal receive (VH). However, only the VH polarized backscatter data was utilized in this characterization process because of its enhanced sensitivity to rice phenology and growth patterns compared to VV backscatter polarization (Fatchurrachman et al., 2022; Inoue et al., 2020; Minh et al., 2019; Rudiyanto et al., 2019; Zhang et al., 2024) . Given that only Sentinel-1A data was available for the study area, it was therefore employed with a spatial resolution of 10 m with the S1 toolbox being utilized for pre-processing. This included thermal noise removal, radiometric calibration, terrain correction and ortho-correction. After preprocessing the images were deemed ready for generation of monthly composites for both SAR and Optical data.

3.3.3. Ground Truth Data Cleaning

This entailed examining the raw datasets of yield , rainfall and temperature collected during field visits to MIS and removal of any data that was deemed irrelevant, repeated or redundant as this could skew the results of the subsequent classification task. The raw data covered the period of interest (July -December 2024) and included the datasets as listed in Table 3.1. This was implemented through visual inspection of the data. Further, any missing data was identified and handled statistically wherever possible.

3.3.4 Monthly Imagery Compositing

This study considered SAR backscatter -VH polarization from S1 as well as NDVI, NDRE and NDWI from S2 covering the main rice growing season of July-December 2024 period. According to Thorp & Drajat (2021), many studies have relied on spectral indices in mapping paddy rice production. As such, monthly composite data from S1 (6 composites) and S2 (18 composites) were stacked together. Since the two satellites acquired images at different times, the use of monthly composites addressed the time alignment issue between the datasets from the two satellites. According to Rudiyanto et al. (2019), S1 SAR datasets are prone to high noise to signal ratio particularly in the border sections with overlapping scenes and hence the Median Value was utilized to generate the monthly composite time series for the VH polarization data. The NDVI, NDRE and NDWI indices were calculated using Band B3 (Green), Band B4 (Red), Band B5 (Red Edge) and Band B8 (Near-Infrared- NIR) from S2 data utilizing Equations 3.3-3.5 obtained from the work by Carneiro et al. (2023) as follows:

$$NDVI = \frac{(B8-B4)}{(B8+B4)} \tag{3.3}$$

$$NDWI = \frac{(B3-B8)}{(B3+B8)} \tag{3.4}$$

$$NDRE = \frac{(B8-B5)}{(B8+B5)} \tag{3.5}$$

Optical data obtained from S2 can be affected with several factors such as existence of cloud cover as well as changes of atmospheric conditions. Hence, utilization of Maximum Value Composites (MaxVCs) approach is widely accepted as an efficient way to minimize these unwanted effects (Fatchurrachman et al., 2022). This therefore necessitated the use of Maximum Values Composites in developing the monthly composite time series for NDVI, NDRE and NDWI indices.

3.3.5 Unsupervised Classification of Time Series Spectral Data

The K-means algorithm is one of the most widely used unsupervised ML models which aims to arrange data into *k* clusters where the number of clusters (*k)* is defined by the user (Sitokonstantinou et al., 2021). In this study, the time series imagery data from S1 and S2 were classified utilizing the K-Means clustering method on the GEE platform. The classification was accomplished through incorporating the *ee.Clusterer.wekaKMeans()* function on the GEE platform which standardizes numerical input features and applies the Euclidean distance method to evaluate distances between cluster points with the objective of reducing variation within each cluster (Rudiyanto et al., 2019). The GEE platform employs a random-point sampling method to identify centroids and form clusters. As these points are distributed randomly throughout the target region of interest (ROI), it is reasonably presumed that this technique effectively captures the spectral data variations across the target area ultimately differentiating various land cover types present in the study area (Fatchurrachman et al., 2022).

A total of 24 monthly composites comprising of VH polarization backscatter from S1 as well as NDVI, NDRE and NDWI from S2 were stacked together and utilized as input into the K-means clustering model on the GEE platform. To train the model, a total of 20,000 random sampling points were generated from the ROI. This number of sampling points were selected due to the fact that it has been applied successfully in other mapping studies (Li et al., 2025; Zhu et al., 2016). The number of clusters *(k)* was set to 20 to reflect the different land cover types present in the study area including rice parcels, built up regions, water bodies among others.

3.3.6 Extracting Spectral Time Series Data

A total of 20,000 samples were generated randomly for each cluster on the GEE platform to extract spectral data time series for both SAR backscatter (VH polarization) and optical (NDVI, NDWI and NDRE) variables. This total number of samples entailed random generation of 500 samples for VH polarization, NDVI, NDWI and NDRE. The spatial median values were computed for the monthly time series for each cluster which were then utilized as representative time-series profiles for VH polarization, NDVI, NDRE, and NDWI.

3.3.7 Identification of Rice and Non-Rice Parcels

Rice parcels and non-rice parcels were identified visually based on the representative time series profiles for VH polarization, NDVI, NDRE and NDWI. Rice parcels display a unique time series cluster profile in which the VH polarization, NDVI, NDRE and NDWI vary during the growing season. In addition, spectral temporal patterns of SAR and Optical datasets illustrate typical dynamics related to the distinct rice growth stages (Fiorillo et al., 2020). Non-Rice parcels included built up areas, areas with tree cover as well as water bodies. After image classification, a rice extent map for the 2024 main growing season was generated showing the spatial distribution of the rice.

3.3.8 Identification of Rice Growth Stages

Representative time series profiles for the VH polarization, NDRE, NDWI, and NDVI were analysed to establish the different stages of rice growth. The growth stages included transplanting, vegetative, reproductive and maturity phases. The transplanting phase was determined by employing the SAR backscatter – VH polarization data and preceded with the subsequent rice growth stages. The rice growth stages were identified through the changes in the VH polarization, NDVI, NDRE and NDWI time series data.

3.3.9 Accuracy Assessment of Rice Extent Map

The accuracy of the 2024 rice extent area map produced in this study was evaluated using two types of assessments; validation using Google Earth VHR imagery as well as existing maps from the NIA Mwea office and total area accuracy (TA) assessed by comparing the extracted area with the actual cultivated area for the current season. Finally, the rice growth stages were compared to the scheme's documented planting and harvesting schedules. This study utilized 100 validation points, due to its successful application in other similar studies (Du et al., 2019). The points were randomly generated in the five main sections of the scheme i.e. Wamumu, Thiba, Tebere, Mwea and Karaba Sections. The accuracy of the rice extent map generated for the 2024 main season was then assessed using the confusion matrix, kappa coefficient value and total area accuracy. Finally, the identified rice growth stages were compared to the reference data obtained from the NIA Mwea office. The Kappa coefficient was obtained through the formula from the work by Congalton (1991) presented in Equation 3.6.

$$K = \frac{N\sum_{i=1}^{r} x_{ij} - \sum_{i=1}^{r}(x_{i+} * x_{j+})}{N^2 - \sum_{i=1}^{r}(x_{i+} * x_{j+})} \quad (3.6)$$

Where *K*=KHAT Statistic (an estimate of KAPPA), *r* is the number of rows in matrix, x_{ij} is the number of observations in row *i* and column *j*, x_{i+} and x_{j+} are the marginal totals of row *i* and column *j* respectively and *N* is the total number of observations.

The Kappa coefficient was assessed in accordance with the classification range used in the work by Ismail & Jusoff (2008) which is presented in Table 3.4.

Table 3.4: Kappa Coefficient Agreement Criteria

Agreement Criteria	Kappa Coefficient
Poor	< 0.40
Good	0.40 to 0.70
Excellent	> 0.75

3.4 Rice Yield Prediction with Machine Learning Models Using Sentinel-1 SAR Data and Auxiliary Data.

3.4.1 Dataset Compilation

Overall, in terms of ML model inputs the independent variables (features) included rice field number, field size, rainfall mean monthly data, minimum and maximum temperature mean monthly values, S1-SAR backscatter i.e. VH and VV polarization median monthly values, covering the critical rice growing phases for the main season. The dependent variable (target) was rice yield. The datasets for training and testing the ML models included datasets for the period 2021-2024. Part of the datasets was isolated for evaluating the performance of ML models in rice yield prediction using LR model as a benchmark. To undertake rice yield prediction various datasets were retrieved as summarized in Table 3.5.

Table 3.5: Summary of Datasets for the Rice Yield Prediction ML Models Covering the 2021-2024 Period

Data	Variable(s)
Rice production data	Block number, block size, yield
Climatic Data (Rainfall, Minimum and Maximum Temperature data)	Mean Monthly Rainfall & Temperature
SAR Backscatter (VV and VH polarization)	Monthly Median VV and VH Polarization Values

3.4.2 Data Pre-processing

Missing values in the obtained datasets were imputed using median values and abnormal values treated using z-core analysis. The datasets were also separated into feature variables (SAR backscatter datasets and climatic datasets) and target variables (yield datasets). Crop production related data, SAR backscatter data and climatic data were integrated into a unified dataset. In addition, all the datasets were put into a standardized CSV format for easier working with the ML models.

3.4.3 Feature Engineering of Datasets

This is a critical step as it involves getting data ready for training the ML models and involves data encoding and standardization of features in the datasets. The resulting final Dataframe had both continuous and categorical variables such as Farm No. and Block Name. Categorical variables take one unique value from a limited set of variables and is not quantifiable. The categorical data types present in the resulting Dataframe were converted into numerical format using one-hot encoding and label encoding techniques. This was necessitated by the fact that ML algorithms cannot work with labelled data. Standardization of features aids in bringing all the features in the dataset into a single scale for increased accuracy in the model output (Jhajharia et al., 2023). This aids in making the dataset features suitable for modelling and ensuring that no single feature disproportionately influences the ML algorithm due to differences in scale. The StandardScaler algorithm from the Scikit Learn Library was applied to the dataset to standardize features by removing mean and scaling to unit variance. The StandardScaler algorithm has also been applied successfully in other crop yield prediction studies using ML models (Pukrongta et al., 2024; Rufaioğlu et al., 2025).

3.4.4 Splitting Dataset into Training and Testing Sets

From the Scikit learn library, the train_test_split module was imported for splitting the dataset into X_test (testing set of features), X_train (training of features), Y_test (testing set of target variables) and Y_train (training set of target variables). For this study, 80% of the dataset was used for model training while 20% was reserved for model testing similar to what Maseko et al. (2024) used in a recent study. This training and testing ratio split of 80/20 has also been applied successfully in other studies (Maseko et al., 2024; Panigrahi et al., 2023; Ramadhan et al., 2024).

3.4.5 Hyperparameter Tuning

To reduce overfitting, hyperparameters of the ML models (RF, SVM and XGBoost), were tuned using the GridSearchCV module from Scikit learn library similar to what Desloires et al. (2023) used in their work. The optimal parameter setting was determined using a training validation strategy via the Grid Search approach. The LR model lacks hyperparameters and hence no tuning was done during yield predictions. The LR model finds the best-fitting line by minimizing the residual sum of squares (RSS) between predicted and actual values (Altork, 2025). Table 3.6 obtained from the work by Desloires et al. (2023) gives a summary of the key hyperparameters utilized in optimizing the rice yield predictions of the non-linear ML models.

Table 3.6: Summary of Hyperparameters Ranges for the ML Models

Model	Hyperparameters	Range
RF	Number of trees	{50, 100, 150}
	Maximum Depth	{10, 20, 30}
	Minimum Samples Split	{3,6,9}
	Minimum Samples Leaf	{1,2,3}
	Maximum Features	{'sqrt', 'log2'}
	Bootstrap	{True, False}
SVM	C- regularization parameter	{0.1, 1, 10}
	Kernel	{'linear', 'poly', 'rbf', 'sigmoid'}
	gamma	{'scale', 'auto', 0.001, 0.01}
	epsilon	{0.001, 0.01, 0.1, 0.2}
XGBoost	Number of trees	{100, 200, 300}
	Maximum Depth	{3, 6, 10}
	Learning Rate	{0.01, 0.05, 0.1}
	Minimum child Weight	{1, 3, 5}
	Subsample ratio samples	{0.6, 0.7, 0.8}
	Subsample ratio columns	{0.6, 0.7, 0.8}
	alpha	{0, 0.01, 0.1}
	gamma	{0, 0.1, 0.2}
	lambda	{0, 0.01, 0.1}

3.4.6 Rice Yield Prediction with Machine Leaning Models

The analysis was conducted on a computer with adequate computational resources to handle large datasets and complex computations i.e. a Hewlett-Packard (HP) system of multi-core processor with at least 4 cores, at least 16 GB RAM, a dedicated Graphics Processing Unit (GPU) with a minimum of 4 GB of VRAM and a Solid State Drive (SSD) with at least 1 TB of storage. The rice yield prediction was implemented on the web-based Jupyter Notebook Integrated Development Environment (IDE) using Python programming language with the Pandas, NumPy, Scikit learn, XGBoost and Matplotlib libraries for data processing, model development, and visualization. The requisite libraries including numpy, pandas, Matplotlib.pyplot were imported on Jupyter Notebook. The Pandas library is essential for data manipulation and analysis and is particularly useful when working with structured data such as tables and DataFrames (Kabir et al., 2024). The numpy library provides support for large, multi-dimensional arrays and matrices along with a collection of mathematical functions to operate on these arrays (Kabir et al., 2024). The matplotlib library is used for importing the 'pyplot' module that can be used in plotting graphs, charts and visual representations of data (Kabir et al., 2024). Four machine learning models— RF, SVM, XGBoost and LR —were utilized to predict rice yield. The models utilized 80% of the data as training dataset comprising of VV and VH polarization, minimum and maximum temperature, rainfall and crop yield spanning four seasons (2021-2024). Each ML model was trained and optimized using a training dataset, with hyperparameters tuning.

A python code was written and executed on the Jupyter Notebook online IDE using the pd.read_csv() function from the pandas library to read the CSV file containing all the variables. Once the code was executed, this function returned the contents of the CSV file as a DataFrame. Another python code line was written for execution to extract the independent variables (features) from the DataFrame and store them in a variable say X. An additional python code line was written and executed to not only extract the dependent variable (target) from the DataFrame but also store it as variable say Y.

Prediction using the RF algorithm was set up in the Python environment in Jupyter Notebook online IDE. The RandomForestRegressor class was imported from the ensemble module in the Scikit learn library. The Scikit learn is a library for ML, having tools that allow diverse operations and provide numerous algorithms (Sadenova et al., 2023). The RandomForestRegressor() class was assigned to an object and an appropriate number of trees was assigned to the ensemble to ensure accuracy of the prediction model. Afterwards the object was fitted using the fit(X, y) method that trains the RF model by learning the relationship between the independent variables (X) and dependent variable (Y). Once the fitting was completed, the model had learnt from the training dataset and was ready to make predictions. The predict(X) method was used on the trained RF model to predict the target variable (rice yield) based on pre-processed isolated testing data. Tuning of the RF model was undertaken using the hyperparameters and ranges listed in Table 3.5. Finally, the key hyperparameters for the RF model were set in the model and rice yield prediction was done with the tuned model.

Prediction using the XGBoost algorithm was set up in the Python environment in Jupyter Notebook online IDE. The XGBRegressor class was imported from the XGBoost library, which is a powerful and efficient open-source library that implements gradient boosting algorithms. The XGBRegressor() class was assigned to an object and an appropriate number of trees, learning rate and maximum depth was selected to train the XGBoost model and ensure accuracy of the prediction model. The object assigned to the XGBRegressor() class was fitted using the fit(X,y) method that trains the XGBoost model by learning the relationship between the independent variables (X) and dependent variable (Y). Once the fitting was completed, the XGBoost model had learnt from the training dataset and was ready to make predictions. The predict (X) method was used on the trained XGBoost model to predict the target variable (rice yield) based on pre-processed isolated testing data. Tuning of the XGBoost model was undertaken using the hyperparameters and ranges listed in Table 3.5. Finally, the key hyperparameters for the XGBoost model were set in the model and rice yield prediction was done with the tuned model.

Prediction using the SVM algorithm was set up in the Python environment in Jupyter Notebook online IDE. The SVR (Support Vector Regression) class was imported from the svm (Support Vector Machine) module in the Scikit learn library. The SVR() class was assigned to an object with the appropriate kernel option being selected along with the parameters for C, gamma and epsilon to ensure model accuracy in prediction. Afterwards an object was assigned to the SVR() class and fitted using the fit(X,y) method. This ensured that the SVM model was trained by finding the best fitting hyperplane (in higher-dimensional space) that approximates the relationship between the independent variables (X) and dependent variable (Y). Once the fitting was completed, the SVM model had learnt from the training data and was ready to make predictions. The predict(X) method was used on the trained SVM model to predict the target variable (rice yield) based on pre-processed isolated test data. Tuning of the SVM model was undertaken using the hyperparameters and ranges listed in Table 3.5. Finally, the key hyperparameters for the SVM model were set in the model and rice yield prediction was done using the tuned model.

Finally, prediction using the LR algorithm was set up in the Python environment in Jupyter Notebook online IDE. The LinearRegression class was imported from the linear_model module in the Scikit learn library. The LinearRegression() class was assigned to an object after which the object was fitted using the fit(X,y) method that trains the LR model by finding the best fitting linear equation that models the relationship between the independent variables (X) and dependent variable (Y). Once the fitting was completed, the LR model had learnt from the training data and was ready to make predictions. The predict(X) approach was then be used on the trained LR model to predict the target variable (rice yield) based on isolated test data.

3.4.7 Model Evaluation Using Testing Data

The 20% of the dataset isolated from model training was utilized to validate model performance as it provides a good trade-off between efficiency in evaluating model accuracy and examining overfitting problems. Therefore, the performance of the developed models was evaluated based on their accuracy in predicting rice yield using the 20% isolated testing dataset. Three metrics including: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) scores were used to assess model performance similar to the evaluation performed by Hashemi et al. (2024), Islam et al. (2023) and Son et al. (2022). These metrics are calculated using Equations 3.7-3.9. To undertake model evaluation of ML models using the isolated testing data the metrics module was used to import functions including mean_absolute_percentage_error and mean_absolute_error from the Scikit learn library on Jupyter Notebook online IDE. These functions were used to calculate the MAPE and MAE for RF, XGBoost, SVM and LR models in predicting rice yield using the isolated test data. However, the Scikit learn library does not provide direct functions for calculating the RMSE, hence, this parameter was calculated using basic operations in Python using Equation 3.7 as used by Hashemi et al. (2024).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \tag{3.7}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i| \tag{3.8}$$

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{P_i - O_i}{O_i}\right| \tag{3.9}$$

where P_i and O_i represent the predicted and observed rice yield for the i^{th} instance respectively.

3.5 Performance Evaluation of Machine Learning Models in Predicting Rice Yield Using Linear Regression as a Benchmark

The final stage of this study involved employing trained ML models (RF, XGBoost, SVM and LR) to forecast the 2024 rice yield according to the methods utilized by Guo et al. (2023) and Desloires et al. (2023). The trained models were validated using an independent test dataset not utilized in training and testing of the ML models to ensure their generalizability and robustness. The forecasted yield was compared against actual yield for the 2024 main season dataset previously isolated from the training and testing datasets (2021-2023). The essence was to evaluate the practical applicability of the models in real-world scenarios using LR as a benchmark.

3.5.1 Dataset Compilation

Relevant datasets including satellite data and climatic data, for the 2024 rice growing season was retrieved for purposes of assessing the ML models ability in extrapolating rice yield. These datasets were unseen for all the four ML models, which in essence means data not included in the training and testing datasets. As ML model inputs, the trained RF, XGBoost, SVM and LR models used independent variables (features) that included Rice Field Number, Field Size, Rainfall Mean Monthly Data, Minimum and Maximum Temperature Mean Monthly Values, S1-SAR backscatter i.e. VH and VV Polarization for the 2024 rice-growing main season. The trained ML models utilized these features to forecast the rice crop yield for the 2024 growing season. The datasets that the trained ML models used for rice yield prediction for the 2024 main season are summarized in Table 3.7.

Table 3.7: Datasets used for Rice Yield Prediction with ML Models for the 2024 Season

Data	Variable(s)
Rice production data	Block number, Block size, yield
Climatic Data (Rainfall and Temperature data)	Mean Monthly Values
SAR Backscatter (VV and VH polarization)	Monthly Median VV and VH Polarization Values

3.5.2 Data Pre-processing

Data pre-processing was undertaken using similar methods utilized previously in the preparation of datasets in the development of the ML models. This entailed data cleaning, image processing, SAR backscatter data feature extraction and data integration.

3.5.3 Rice Yield Prediction with Trained ML Models

The trained ML models were used in predicting rice yield outcomes of the isolated year 2024 on Jupyter Notebook online IDE using Python programming language. It was presumed that the isolated 2024 dataset would allow for the assessment of the model's performance in accurately predicting the actual rice yield for the upcoming growing season using datasets from previous seasons for model training (2021-2023). The comparison of forecasted yield with observed yield in 2024 assisted in the evaluation of model effectiveness in reflecting any potential variations and trends in rice yields.

On the Jupyter Notebook, a python code was written and executed using the pd.read_csv() function to read a CSV file containing all the variables from the isolated 2024 dataset. This function returns the contents of the CSV file as a DataFrame. Another python code line was written for execution to extract the independent variables (features) from the DataFrame as well as storing them in a variable say X_2024. An additional python code line was written and executed to not only extract the dependent variable (target) from the DataFrame but also store it as a variable say Y_2024.

The RandomForestRegressor() class previously trained with the 2021-2023 datasets was used in making rice yield predictions for the 2024 main season on Jupyter Notebook. The predict(X_2024) method was used on the trained RF model to predict the target variable i.e. rice yield (Y_2024) based on pre-processed independent variables (X_2024) data. Tuning of the RF model was undertaken using the hyperparameters and ranges listed in Table 3.5. Finally, the key hyperparameters for the RF model were set in the model and rice yield predictions done using the tuned model.

The XGBRegressor() class previously trained with the 2021-2023 datasets was used in predicting rice yield for the 2024 main season on Jupyter Notebook. The predict(X_2024) method was used on the trained XGBoost model to predict the target variable (rice yield) based on pre-processed independent variables (X_2024) data. Tuning of the XGBoost model was undertaken using the hyperparameters and ranges listed in Table 3.5. Finally, the key hyperparameters for the XGBoost model were set in the model and rice yield predictions done using the tuned model.

The SVR() class previously trained with the 2021-2023 datasets was used in predicting the rice yield for the 2024 main season on Jupyter Notebook. The predict(X_2024) method was used on the trained SVM model to predict the target variable (rice yield) based on pre-processed independent variables (X_2024) data. Tuning of the SVM model was undertaken using the hyperparameters and ranges listed in Table 3.5. Finally, the key hyperparameters for the SVM model were set in the model and rice yield predictions were undertaken using the tuned model.

The LinearRegression() class previously trained with the 2021-2023 datasets was used in predicting rice yield for the 2024 main season on Jupyter Notebook. The predict (X_2024) method was used on the previously trained LR model to predict the target variable (rice yield) based on pre-processed independent variables (X_2024) data.

3.5.4 Evaluation of Accuracy Performance of ML Models

To evaluate the performance accuracy of ML models in predicting the 2024 rice yield using the unseen 2024 dataset three accuracy statistical indicators were employed. These included Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) scores obtained using Equations 3.5-3.7. This entailed comparison between predicted and actual yields. Similar performance evaluation metrics were used as those used in model evaluation using test data (section 3.4.7) for consistency and ease of comparison of the ML models. The obtained accuracy values for LR model i.e. the benchmark, were compared to the accuracy values obtained using the RF, XGBoost and SVM models.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 General Agronomic Practices for Rice Growing in Mwea

Rice is predominantly grown as a main crop once per year, although some farmers cultivate a second (ratoon) crop immediately after the main harvest (Mwangangi et al., 2008). Additionally, rice cultivation in MIS is closely coordinated with irrigation scheduling to ensure equitable water distribution throughout the entire growth cycle (Musila et al., 2024).Therefore the rice production cycle begins with land preparation followed by nursery establishment in June–July. Seedlings are then transplanted in July-August, after which the crop progresses through the vegetative stage (August–October), flowering/reproductive stage (October–November), and maturation/ripening stage (November–December). Harvesting is typically carried out between mid-December and early January, while the ratoon crop is grown from January to May (Mwangangi et al., 2008). In general, during transplanting in MIS triple superphosphate (TSP) is commonly applied as a basal fertilizer. Approximately, 10 and 35 days after transplanting sulfate of ammonia is typically applied as a topdressing fertilizer to support crop growth and nitrogen uptake (Mwangangi et al., 2008).

According to NIA mwea office the rice varieties planted for the 2024 season included *Basmati*, *Komboka* and *Arize*. For the main sections Mwea, Tebere, Wamumu, Thiba, and Karaba transplanting occurred between 20th July 2024 and 2nd August 2024 with the first spraying and first fertilizer top dressing being undertaken between 11th August 2024 and 20th August 2024. The second fertilizer top dressing occurred between 12th September 2024 and 2nd October 2024. Finally harvesting occurred between 18th November 2024 and 9th December 2024 as shown in the 2024 cropping program from NIA attached in Appendix -III.

4.2 Climatic Datasets Collected from Mwea Irrigation Scheme Weather Station

Mwea Irrigation Scheme experiences two wet seasons March -May and October - December with the peak being in April and November. The temperatures are highest in March and October with the coldest month being experienced in July as shown in Figure 4.1. In particular, for the year 2021 the mean annual minimum and maximum temperatures was 21.37 ^{0}C and 22.72 ^{0}C respectively while the total annual precipitation was 951 mm as shown in Figure 4.1 (d). For the year 2022 the mean annual minimum and maximum temperatures was 21.37 ^{0}C and 22.73 ^{0}C respectively while the total annual precipitation was 746 mm as shown in Figure 4.1 (c). For the year 2023 the mean annual minimum and maximum temperatures was 21.821 ^{0}C and 23.185 ^{0}C respectively while the total annual precipitation was 1184 mm as shown in Figure 4.1 (b). For the year 2024 the mean annual minimum and maximum temperatures was 20.307 ^{0}C and 21.590 ^{0}C respectively while the total annual precipitation was 1411 mm as shown in Figure 4.1 (a). Given that this study was working with temperature and rainfall covering only the main season period (July -December) spanning 4 years very few data gaps were found which were filled using CHIRPS and MERRA-2 datasets for rainfall and temperature respectively.

4.3 Characterization of Rice Growth Stages Using SAR and Optical Data

The following sections present detailed characterization of the rice growth stages for the 2024 main season namely transplanting, vegetative, reproductive, and maturity phases within the Mwea Irrigation Scheme (MIS). Similar rice growth stages have also been identified in other studies (Fatchurrachman et al., 2022; Kobayashi & Ide, 2022; Rudiyanto et al., 2019). The aim was to establish how SAR backscatter and optical datasets from Sentinel satellites can be effectively utilized to distinguish between the different rice growth stages.

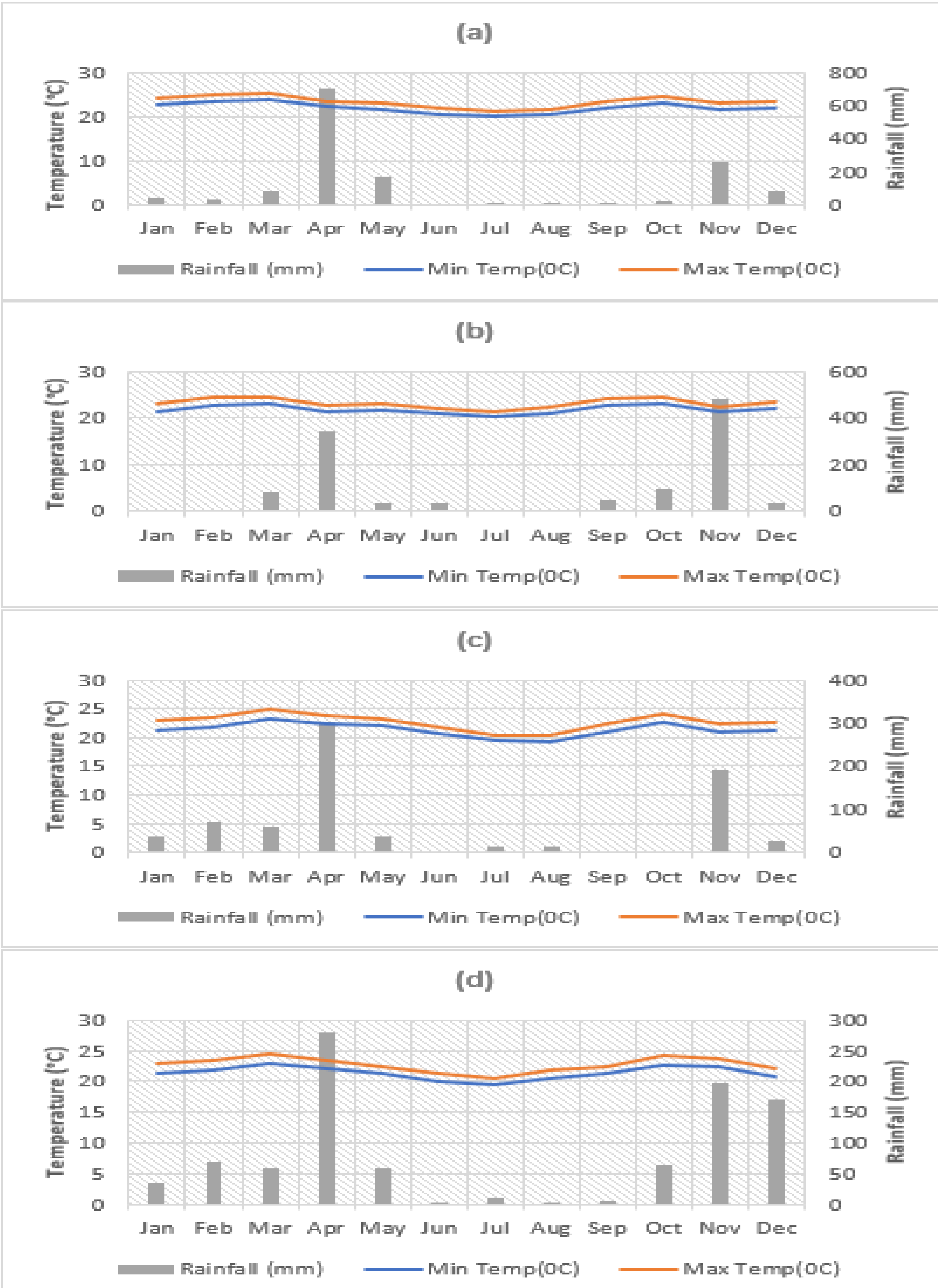

Figure 4.1: Mwea Irrigation Scheme Mean Monthly Rainfall and Temperature Data for (a) 2024, (b) 2023, (c) 2022, and (d) 2021

4.3.1 Time Series Spectral Cluster Profiles

Based on the K-means clustering approach spectra profiles for the VH polarization, NDVI, NDRE and NDWI two rice types were identified during the 2024 main rice growing season i.e. rice transplanted planted in July herein referred to as Rice-1 and rice transplanted in August herein referred to as Rice-2. This naming has been assigned for purposes of this study to illustrate when their transplanting takes place.

Figure 4.2 illustrates the time series cluster profiles for the VH polarization covering the 2024 main rice growing season in MIS. The VH polarization for chart (a) and (b) representing Rice-1 and Rice-2 respectively, generally increase from a local minimum obtained in July and August and increases to attain a local maximum in November and December. Conversely, chart (c) generally lacks the distinct VH polarization troughs observed during the July–August transplanting window, indicating that the area is likely non-rice vegetation or other non-rice land cover. The low backscatter values in July and August for chart (a) and (b) are most likely due to the smooth water surface after flooding the fields during transplanting. There is then a gradual backscatter increase in VH polarization as rice canopy grows and becomes denser. These results align with the findings by Kobayashi & Ide (2022) illustrating that structural changes in rice crop leads to gradual backscatter increase in VH polarization. This is possibly due to volume scattering within the rice canopy as the large number of stems and leaves interact with the VH polarization backscatter signal (He et al., 2018).

Figure 4.3 illustrates the time series cluster profiles for the NDWI water index covering the 2024 main rice growing season in MIS. The NDWI charts generally decrease from a local maximum obtained in July to a minimum attained in November and December. Initially the NDWI values are high particularly in the identified rice clusters charts (1) and (2) shown in Figure 4.3 likely due to flooding of the paddy fields. In one study, Acharya et al. (2016) noted that the NDWI and Modified Normalized Difference Water Index (MNDWI) water indices derived using spectral index approaches are effective in the identification of water. As such, Sun et al. (2024) noted that the (MNDWI) values were higher for paddy rice during the transplanting period due to flooding and could be used as an identifying feature. As the rice canopy becomes denser and moisture content decrease there is reduced water reflectance possibly leading to the gradual reduction of NDWI values. It can also be noted that cluster 9 on chart (a) and cluster 12 on chart (b) have relatively lower NDWI values compared with the other clusters in the month of September. This was likely due to reduced irrigation within the clusters, which led to lower moisture content. Approaching the month of October, the NDWI values increases in these two clusters possibly due to increased irrigation or precipitation resulting in moisture content increment into the two clusters. As the season progresses, the NDWI values reduce even further as the crop reaches maturity presumably in readiness for harvesting.

Figure 4.4 illustrates the time series cluster profiles for the NDVI vegetation index covering the 2024 main rice growing season in MIS. The NDVI for chart (a) and (b) representing Rice-1 and Rice-2 respectively, generally increase from a local minimum obtained in July and August and increases to attain a local maximum in October and November. On the contrary the chart (c) generally lacks the distinct NDVI troughs observed during the July–August transplanting window, indicating that the area is likely non-rice vegetation or other non-rice land cover. Initially NDVI values are low likely due to sparse/ absent vegetation particularly in the months of July and August. Afterwards there is a significant increase in NDVI values presumably due to increase in rice canopy cover which gives an indication of healthy vegetation growth. In the month of October and November the NDVI values begin to reduce possibly due to decrease in chlorophyl and the onset of senescence. These observations align with the findings of Zhang et al. (2019) showing that NDVI values increases gradually before stem elongation followed by rapid increase till booting is complete and finally a slow / static period ensues.

Figure 4.5 illustrates the time series cluster profiles for the NDRE index covering the 2024 main rice growing season in MIS. The NDRE for chart (a) and (b) representing Rice-1 and Rice-2 respectively, generally increase from a local minimum obtained in July and August and increases to attain a local maximum in October and November. On the contrary the chart (c) generally lacks the distinct NDRE troughs observed during the July–August transplanting window, indicating that the area is likely non-rice vegetation or other non-rice land cover. As the rice crop develops there is presumably increased chlorophyl content resulting in increased NDRE values. These observations align with the findings of Jiang et al. (2021) indicating that NDRE increases rapidly as heading takes place and the rice grows vigorously causing the reflectance of the red edge band and the near-infrared band to increase. Beginning in October the NDRE values begin to decrease likely indicating reduced chlorophyll content as the rice crop ripens.

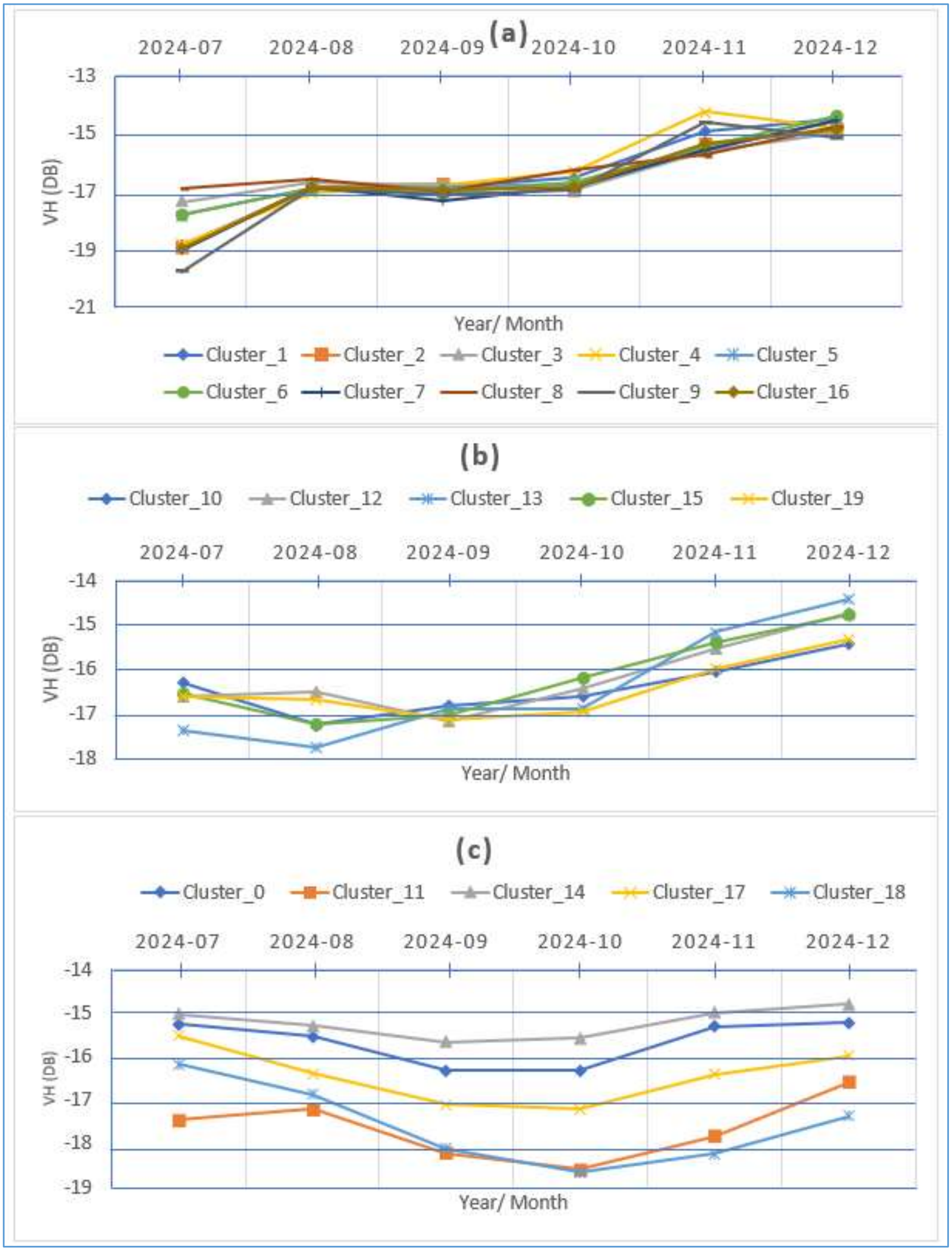

Figure 4.2: VH Polarization Time Series Profiles for (a) Rice-1, (b) Rice-2 and (c) Non-Rice Areas

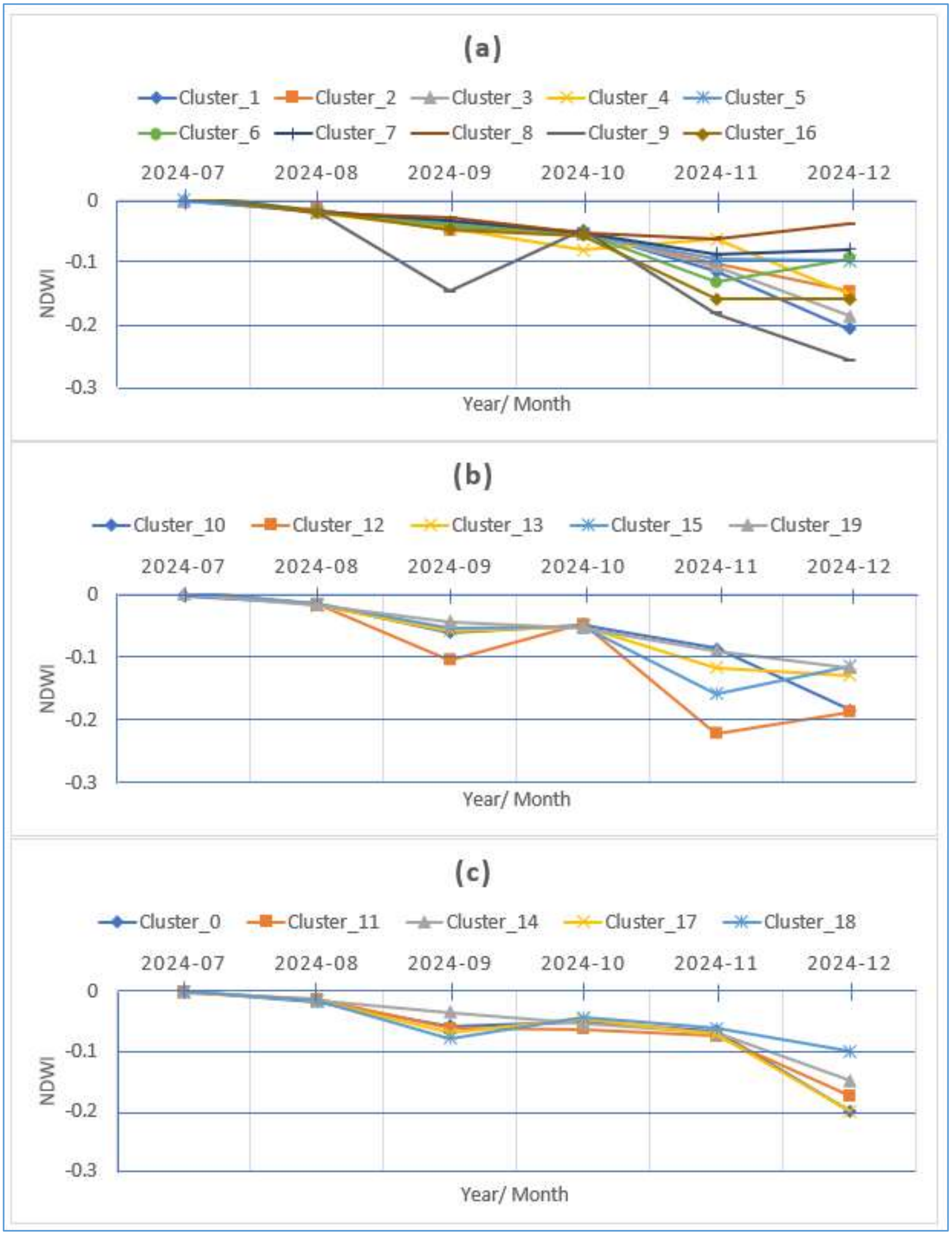

Figure 4.3: NDWI Time Series Profiles for (a) Rice-1, (b) Rice-2 and (c) Non-Rice Areas

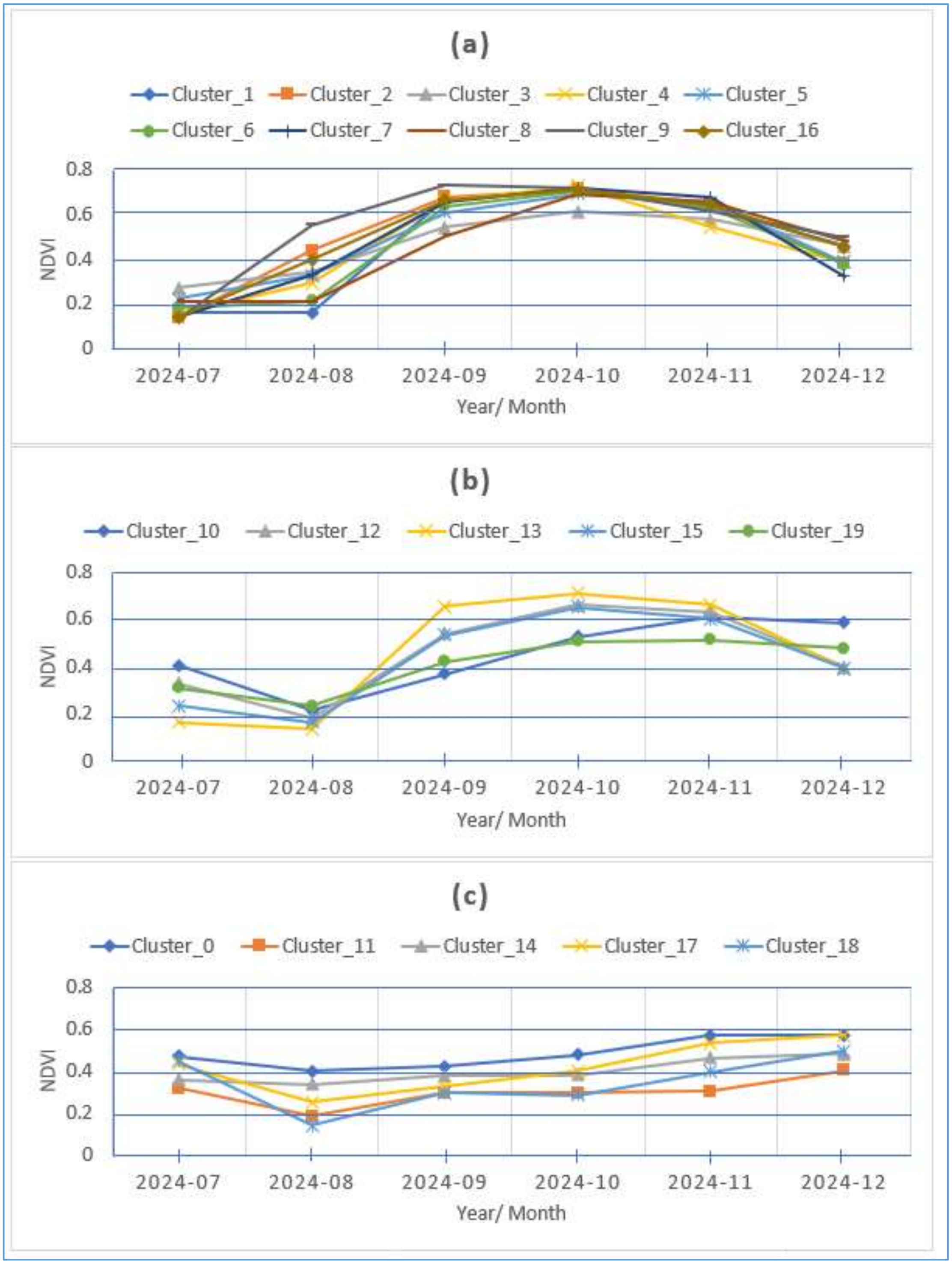

Figure 4.4: NDVI Time Series Profiles for (a) Rice-1, (b) Rice-2 and (c) Non-Rice Areas

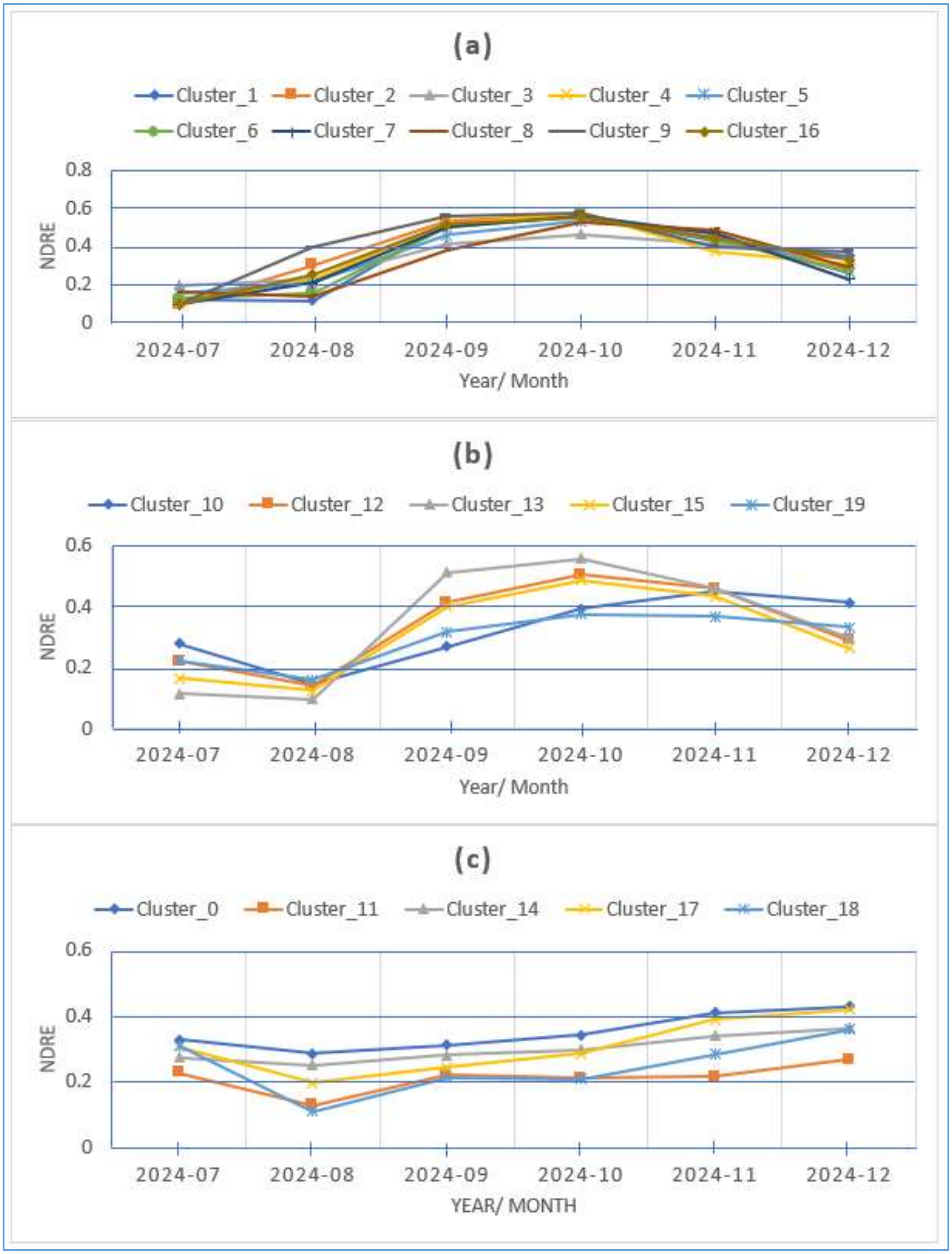

Figure 4.5: NDRE Time Series Profiles for (a) Rice-1, (b) Rice-2 and (c) Non-Rice Areas

Table 4.1: Ranges for the Median Values for Spectral data Covering the Main Rice Growing Season for the year 2024

Index/ Polarization	Range
Vertical transmit/ Horizontal receive (VH)	-19.74 to -14
Normalized Difference Vegetation Index (NDVI)	0.13 to 0.73
Normalized Difference Red Edge (NDRE)	0.09 to 0.58
Normalized Difference Water Index (NDWI)	-0.26 to 0.02

4.3.2 Identified Key Phenological Phases

On the basis of rice phenology and spectral time series data applied in this study, three key phases were identified. They included the transplanting (T) phase, vegetative (V) and reproductive (R) as well as maturity (M) phases. Details of changes in the time series spectral data during the season are presented in the sections that follow.

Transplanting Phase which is characterized with inundation of rice parcels. According to the VH polarization cluster profiles data in Figure 4.2 the rice parcels clusters identified had a median value range of approximately -17 to -20 dB. These values are similar to the results obtained in the studies by Fatchurrachman et al. (2022) and Rudiyanto et al. (2019). However, it should be noted that these values are dependent on moisture content during the transplanting stage which in essence means that more negative VH polarization values insinuate high moisture levels (Rudiyanto et al., 2019). According to Figure 4.3 NDWI values attain a high local maximum value of 0.0154 indicating the presence of moisture in the identified rice clusters during this phase. This is substantiated by Zhang et al. (2024) who established that the rice crop tends to have a high-water index during the transplanting phase. In contrast, the NDVI of identified rice cluster profiles shown in Figure 4.4 indicate low values during this phase which give an indication that vegetation may either sparse or absent. Similarly, Figure 4.5 indicate minimum values in the NDRE time series cluster profiles which would signify low vegetation chlorophyl content in the identified rice clusters.

Vegetative and Reproductive Phases are usually characterized by rapid growth, canopy formation and closure of the rice crop canopy. The NDVI values for the identified rice clusters get to a local maximum value of 0.729 in cluster 9 chart (a) as shown in Figure 4.4 which agrees with the findings obtained by Fatchurrachman et al. (2022). Furthermore, this phase is characterized by rapid increase in chlorophyll signals (Fatchurrachman et al., 2022). This is captured in the NDRE time series cluster profiles shown in Figure 4.5. Both NDVI and NDRE values show significant increase, peaking when the rice crop enters maturity/heading phase. Similarly, VH polarization in the identified rice clusters also increases in value during this phase as illustrated in Figure 4.2. In contrast, NDWI values during this phase tend to decline which can be attributed to canopy development and a reduction in moisture content during this phase as depicted in Figure 4.3. This aligns with observations by Zhang et al. (2024) which indicate that water related indices become smaller as the vegetation indices become larger up to the end of rice heading stage.

Maturity Phase which is characterized by a remarkable decline in chlorophyll content while the carotenoid content increases as the rice crop matures (Fatchurrachman et al., 2022). During this phase there is a noticeable reduction in NDVI (Figure 4.4) and NDRE (Figure 4.5) indicating the onset of rice crop maturity. The NDWI values (Figure 4.3) continue to drop further, suggesting that the rice fields are drying out as the crops reach maturity ready to be harvested. This phase typically occurs in the final month of the rice growing cycle.

The cluster profiles (Figure 4.2-Figure 4.5) under the identified key phenological phases i.e. transplanting, vegetative, reproductive and maturity show the change in the spectral data during the season. The ranges for median values time series cluster profiles for VH polarization, NDVI, NDRE and NDWI are presented in Table 4.1. The S1 SAR backscatter data (VH polarization) exhibited higher sensitivity in the identification of the transplanting phase while S2 Optical data (NDVI, NDRE and NDWI) provided insights into the rice growth stages and maturity phases. From the identified rice clusters, it can be shown that transplanting takes place in July (Rice - 1) and August (Rice - 2) while the maturity is attained in the October and November period, giving an indication that the main cropping season is approximately 120 days/ 4 months. This is similar to what Musila et al. (2024) illustrates in their handbook on paddy rice cultivation in Mwea, Kenya. The approach proposed in this study would be ideal in identifying rice growth stages for the *Komboka* and NIBAM11-*Pishori370*[1] rice varieties that attain maturity in 110-120 days and 115-120 days (Musila et al., 2024). The two varieties are known for being aromatic, having good yield potential, and possessing excellent ratoon attributes (Musila et al., 2024). Ratooning is the agronomic practice in which new rice shoots emerge from the stubble that remains after the main crop has been harvested, allowing the production of a subsequent crop without replanting (Wang et al., 2020). Figure 4.6 shows the transplanting, vegetative, reproductive and maturity rice growth stages identified at MIS during the 2024 main rice growing season.

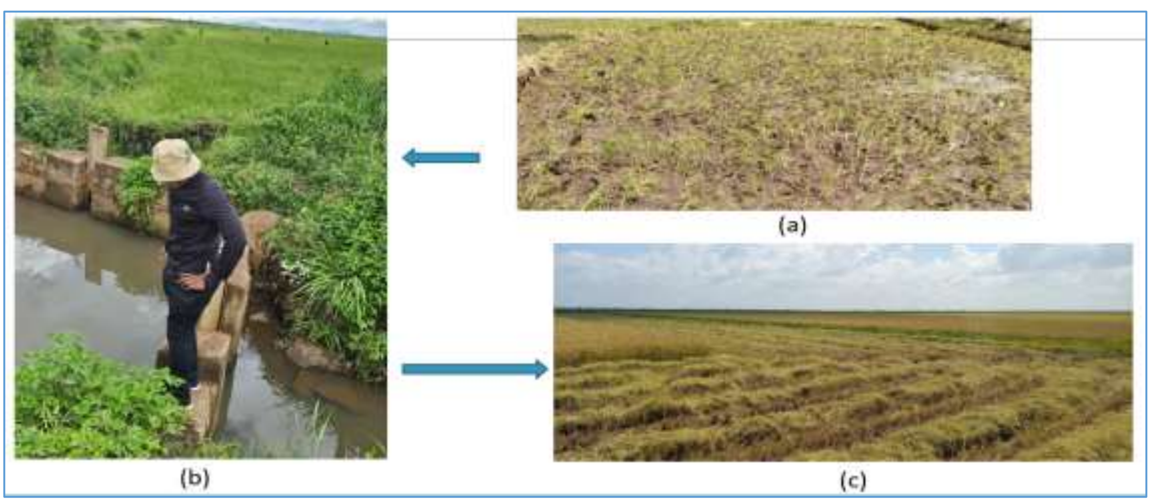

Figure 4.6: Transplanting (a), Vegetative & Reproductive (b) and Maturity Rice Growth Stages Identified at Mwea Irrigation Scheme

[1] Local name of NIBAM11 at Mwea Irrigation Scheme

4.3.3 Accuracy of Rice Maps

The rice parcels extent map for MIS covering the main growing season for 2024 was generated on GEE and is presented in Figure 4.7. Rice parcels within the MIS were identified based on the unsupervised classification approach. The map generated in this study demonstrates success in distinguishing rice fields from other land cover types. The generated map was validated using random sampling in the ROI for rice fields to evaluate the classification accuracy. The resulting map had an overall accuracy of 83% and a Kappa Coefficient of 0.53 as shown in Table 4.2. The generated rice map had an area of 28,354 acres against an area of 26,485 acres that had been targeted for consideration by NIA for the main rice growing season. This demonstrated a disparity of 1869 acres with a corresponding TA of 93%. This shows more areas could have been misclassified as rice areas resulting in higher commission errors than omission errors. Commission errors tend to occur when non-rice parcels are mistakenly categorized as rice fields whereas omission errors arise when actual rice parcels are incorrectly labelled as non-rice parcels (Rudiyanto et al., 2019). For this study commission and omission errors for rice fields were 13.92% and 8.11% respectively.

Table 4.2: Confusion Matrix for Rice Classification in MIS for the Year 2024

		Reference			User Accuracy	
		R[2]	**NR**[3]	**Total**	**Percentage Correct**	**Commission Error**
Predicted	**R**	68	11	**79**	86.08	13.92
	NR	6	15	**21**	71.43	28.57
	Total	**74**	**26**	**100**		
Producer Accuracy Percentage Correct		91.89	57.69			
Omission Error (%)		8.11	42.31			
Overall Accuracy				**83.00**		
Kappa				**0.53**		

[2] Rice Areas
[3] Non-Rice Areas

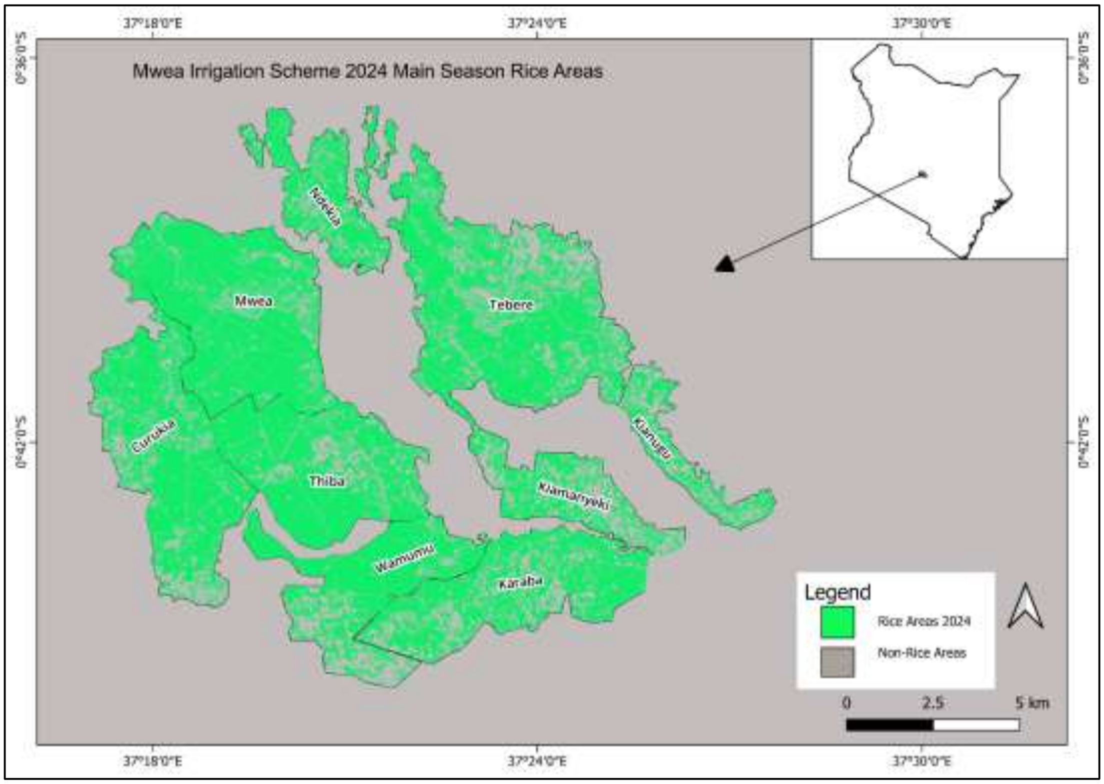

Figure 4.7: Rice Area Extent Map for the Year 2024 Main Season July-December

4.3.4 Rice Growth Stages in the Different K-Means Clusters

Using the spectra cluster profiles for the VH polarization, NDVI, NDRE and NDWI (Figure 4.2-Figure 4.5), the crop phenological parameters for each rice group for the 2024 main rice growing season were estimated as presented in Table 4.3. The data from this study agreed with the 2024 NIA cropping schedule for MIS. The Scheme implements a cropping calendar for purposes of attaining equitable distribution of water resources (Musila et al., 2024). The main sections of the scheme will have the transplanting occurring in late July and early August while maturity occurs in November with harvesting commencing in late November to December (Musila et al., 2024).

Table 4.3: 2024 Rice Growth Stages for Rice Clusters in MIS

Cluster	Classification	2024 Season Rice Growth Stages					
		Jul	Aug	Sep	Oct	Nov	Dec
0	NR[4]						
1	R[5]	T[6]	V[7]	R[8]	M[9]		
2	R	T	V	R	M		
3	R	T	V	R	M		
4	R	T	V	R	M		
5	R	T	V	R	M		
6	R	T	V	R	M		
7	R	T	V	R	M		
8	R	T	V	R	M		
9	R	T	V	R	M		
10	R		T	V	R	M	
11	NR						
12	R		T	V	R	M	
13	R		T	V	R	M	
14	NR						
15	R		T	V	R	M	
16	R	T	V	R	M		
17	NR						
18	NR						
19	R		T	V	R	M	
		Rice 1					
		Rice 2					

This study employed both S1-SAR backscatter and S2 optical time series data to capture the rice extent and identify rice growth stages for the 2024 main growing season. The use of vegetation indices is considered to be an effective approach in distinguishing between vegetation and other land covers (Zhang et al., 2024). This study utilized the NDVI,

[4] Non-Rice Area
[5] Rice Area
[6] Transplanting
[7] Vegetative
[8] Reproductive
[9] Maturity

NDRE and NDWI indices which have successfully been used in other studies involving the monitoring of paddy rice. For instance, Samsuddin Sah et al. (2023) used the NDVI and NDRE indices in their study while Zhang et al. (2024) applied the NDVI and NDWI vegetation indices in another study. The maximum value monthly composites were used similar to what Fatchurrachman et al. (2022) utilized in their study to reduce the effect of cloud cover. Although their study utilized only NDVI from S2 this study employed maximum value composites techniques to NDWI and NDRE since this approach has been applied successfully in other crop mapping studies. For instance, Tian et al. (2019) utilized maximum NDWI to map spring canola and spring wheat in China while in another study Thompson & Puntel (2020) used maximum NDRE for precision nitrogen management in corn in USA.

The incorporation of S1 SAR backscatter data has been successfully employed in several rice monitoring studies (Clauss & Kuenzer, 2018; Minh et al., 2019; Samsuddin Sah et al., 2023). For this study the median monthly composites were utilized to develop the time series datasets similar to what Rudiyanto et al. (2019) used in their work. This technique is preferred as it aids in reducing noise in the data scenes (Rudiyanto et al., 2019). The VH polarization data was found to be effective in the detection of rice fields during the transplanting period while the optical data NDVI, NDRE and NDWI was effective in showing the changes during the other rice growth stages. It was however, noticed that in some of the clusters the local minimum VH polarization value occurred in the month of September while the corresponding NDVI and NDRE values for these clusters showed a minimum value in the month of August. This could likely be attributed to late irrigation or precipitation events in the said clusters as indicated by a slight increase in NDWI hence causing increased surface smoothness resulting in specular reflection phenomena. This could have then resulted in the low backscatter values occurring in September rather than in August.

This study is based on the K-means clustering technique coupled with the identification of rice parcels from expert interpretation of the resulting unique spectral characteristics of SAR backscatter and optical data from Sentinel satellites. A 10 m resolution rice extent map for the main growing season was generated having an overall accuracy of 83%. Compared to the findings from similar studies utilizing SAR data in mapping rice parcels, this study distinctly highlights the potent capability of SAR backscatter data in rice monitoring studies. For instance, Minh et al. (2019) utilized SAR data with Support Vector Machines (SVM) classification technique to produce a rice cropping pattern map with an overall classification accuracy of 80.7% in Vietnam. In another study, Clauss & Kuenzer (2018) employed superpixel segmentation technique on SAR time series data to map rice parcels in six different regions and achieved an average overall accuracy of 83%. While this study aligns with the predictive accuracies of the aforementioned studies it distinguishes itself through the use of both SAR backscatter and optical data with unsupervised classification technique for paddy rice mapping.

This study achieves a kappa coefficient of 0.53 which reflects the level of agreement between the classified and the reference observations. In one study, Fiorillo et al. (2020) attained a kappa coefficient of 0.68 using object-based RF classification to map paddy rice in South Senegal. This study's kappa coefficient is comparable to the one attained in the aforementioned study, however, this study distinguishes itself with the use of phenological based approach with unsupervised classification for rice extent mapping. Furthermore, according to the agreement scale defined in Table 3.4 our classification may be considered fairly good as the kappa coefficient of 0.53 satisfies the criteria $0.4<k<0.7$.

Further, machine learning techniques including random forest (RF), decision trees (DT) and SVM can also be utilized in rice crop extent mapping (Fikriyah et al., 2025). For instance, in one study, Ma et al. (2020) applied both supervised and unsupervised crop classification in China attaining accuracies of 79% (random forest), 74% (k-means) and 75% (Iterative Self-Organizing Data Analysis -ISODATA) for one year and achieved almost similar findings in the preceding year achieving accuracies of 82% (RF), 76% (k-means) and 78% (ISODATA). In comparison to similar studies, the accuracy achieved in this study (83%) appears to be comparable with studies utilizing other machine learning approaches. The K-means algorithm utilized in this study provides a practical approach to map rice extents without requiring large training samples which makes it valuable in data-scarce environments. Additionally, this could be a cost saving approach particularly if obtaining reference datasets has high-cost implications. The K-Means clustering method, effectively identifies unique patterns in spectral time series data and forms clusters related to rice phenology. Furthermore, the method proposed in this study is able to identify pixels due to rice growth stages including transplanting, vegetative, reproductive and maturity as shown in Table 4.3. Similar rice growth stages were identified in other studies (Fatchurrachman et al., 2022; Rudiyanto et al., 2019). Nonetheless, the interpretation of these reflectance patterns accurately, requires prior knowledge (Fatchurrachman et al., 2022). This study establishes that unsupervised classification of S1 SAR backscatter and S2 Optical time series data on GEE is a reasonably cost-effective approach for mapping rice cultivation areas and identifying rice growth stages.

Even though this study was effective in implementing the rice mapping approach, some limitations could have arisen particularly when utilizing spectral datasets. Misclassification errors could have occurred along the edges separating rice and other land cover types resulting in omission errors in rice fields. This discrepancy is largely attributed to mixed pixel effect at the edge of rice fields and has been reported in past studies (Fatchurrachman et al., 2022; Clauss & Kuenzer, 2018; Rudiyanto et al., 2019). On the other hand, commission errors could have occurred due to non-rice features such as roads or bunds being misclassified as rice parcels. Figure 4.8 shows some of the non-rice features in the study area that may have resulted in misclassification.

Figure 4.8: Non-Rice Features such as Bunds and Roads in the Study Area

Another possible source of misclassification is the possible occurrence of flooding events that are not associated with rice cultivation but could have been as a result of precipitation events. This type of error could also feature in the reference dataset used for validation as non-agronomic and agronomic flooding may not be distinguished from remote sensing reference data (Clauss & Kuenzer, 2018). In addition, rice parcels at certain growth stages may show similar spectral signatures (such as flooded rice versus natural wetlands) that might not be distinguishable by the clustering algorithm applied in this study resulting in misclassification errors.

When multiple paddy rice parcels are divided by narrow features such as roads or bunds then Sentinel images with 10 m resolution might fail to detect them. Consequently, this causes our clustering algorithm to fail in differentiating those parcels. However, this effect can be resolved through the use of VHR aerial or satellite images as well as the utilization of superior clustering algorithms. The method proposed in this study is dependent on detecting flooding during the transplanting rice growth stage using the SAR-VH polarization datasets. Flooding of parcels during transplanting is a common occurrence in the study area. Hence, this means that rice cultivated in non-flooded environments may not be detected using the method proposed in this study since SAR derived datasets will show no major distinctions between rice and other crops. Although the proposed phenology driven approach provided acceptable accuracy, more advanced machine learning techniques such as RF are expected to deliver higher performance. Therefore, the approach used in this study can be considered as a baseline approach offering a foundation that can further be enhanced in future rice mapping applications.

4.4 Prediction of Rice Yield Using Machine-Learning Models based on Sentinel-1 SAR and Auxiliary Data

The first combination set included rice yield predictions using ML models with only S1-SAR backscatter datasets (VH and VV polarization) covering the main rice growing season period of four years (2021-2024). The study's findings are presented in Figure 4.9 (a), Figure 4.10 (a) and Figure 4.11 (a), highlighting the comparative performance of four ML methods-LR, RF, XGBoost and SVM in estimating rice crop yield utilizing only S1 SAR backscatter data. From this combination the RF model emerged with the highest accuracy having the lowest RMSE of 255.47 kg/acre, MAE of 214.82 kg/acre and MAPE of 8.11% indicating strong predictive accuracy. The LR model emerged second having a RMSE of 260.24 kg/acre, MAE of 221.15 kg/acre and MAPE of 8.36%. The SVM model emerged third with a RMSE of 263.74 kg/acre, MAE of 222.37 kg/acre and MAPE of 8.44%. Finally, XGBoost performed poorest of the four models as indicated by its higher RMSE of 269.83 kg/acre, MAE of 224.94 kg/acre and MAPE of 8.47% illustrating suboptimal performance.

The second combination entailed performing rice yield predictions utilizing ML models with auxiliary data sets (rainfall, minimum and maximum temperatures) combined with the SAR backscatter data covering the main rice growing season period spanning four years (2021-2024). The performance of the four machine learning models using the SAR and auxiliary climatic data combinations are indicated in Figure 4.9 (b), Figure 4.10 (b) and Figure 4.11 (b). From this combination the RF model emerged with the highest accuracy achieving the lowest RMSE of 246.91 kg/acre, MAE of 202.90 kg/acre and MAPE of 7.67% indicating superior predictive accuracy. The XGBoost model emerged second having a RMSE of 257.29 kg/acre, MAE of 212.84 kg/acre and MAPE of 7.98%. The SVM model emerged third with a RMSE of 257.64 kg/acre, MAE of 214.27 kg/acre and MAPE of 8.08%. Finally, LR performed poorest of the four models as indicated by its higher RMSE of 384.09 kg/acre, MAE of 290.86 kg/acre and MAPE of 11.06% indicating inferior predictive performance.

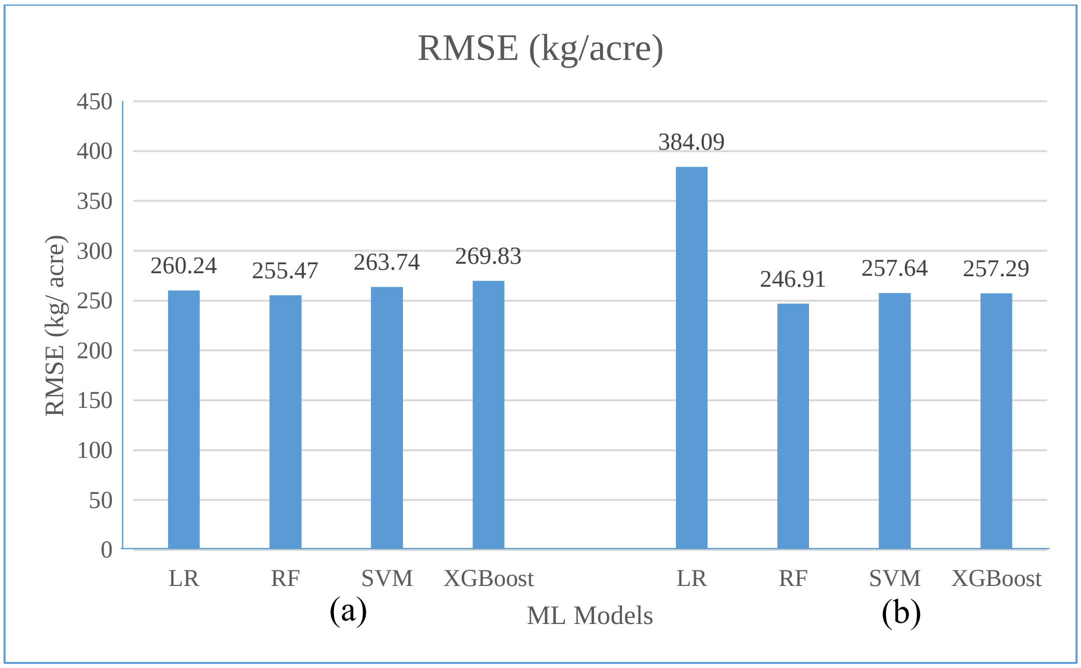

Figure 4.9: RMSE Values Obtained from ML Models Using SAR Only (a) and Combination of SAR and Auxiliary Climatic (b) Datasets

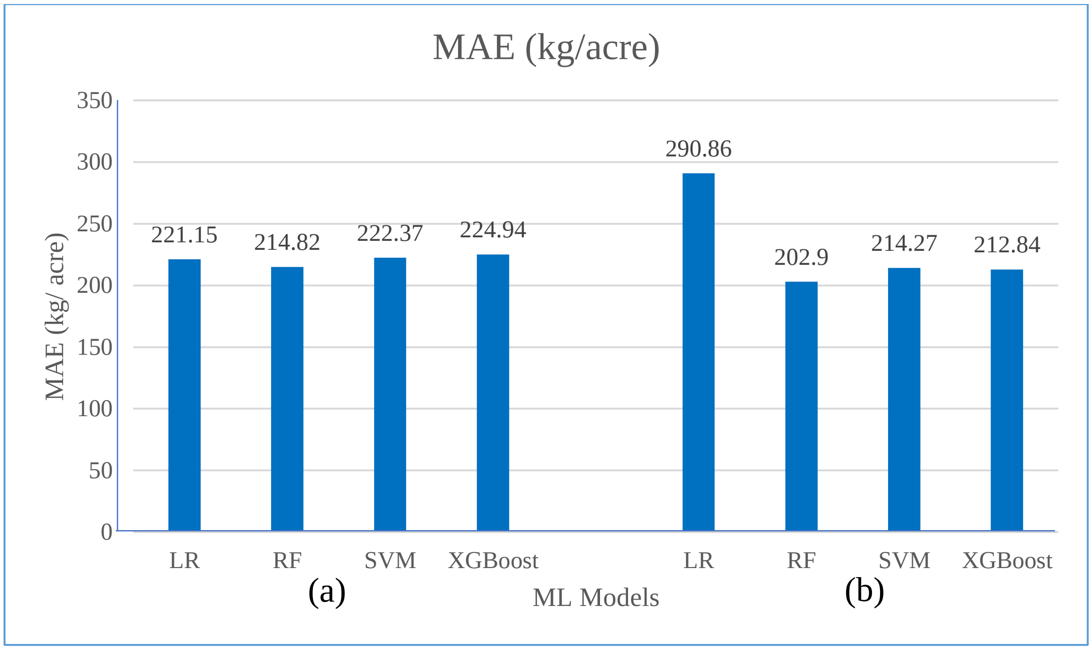

Figure 4.10: MAE Values Obtained from ML Models Using SAR Only (a) and Combination of SAR and Auxiliary Climatic (b) Datasets

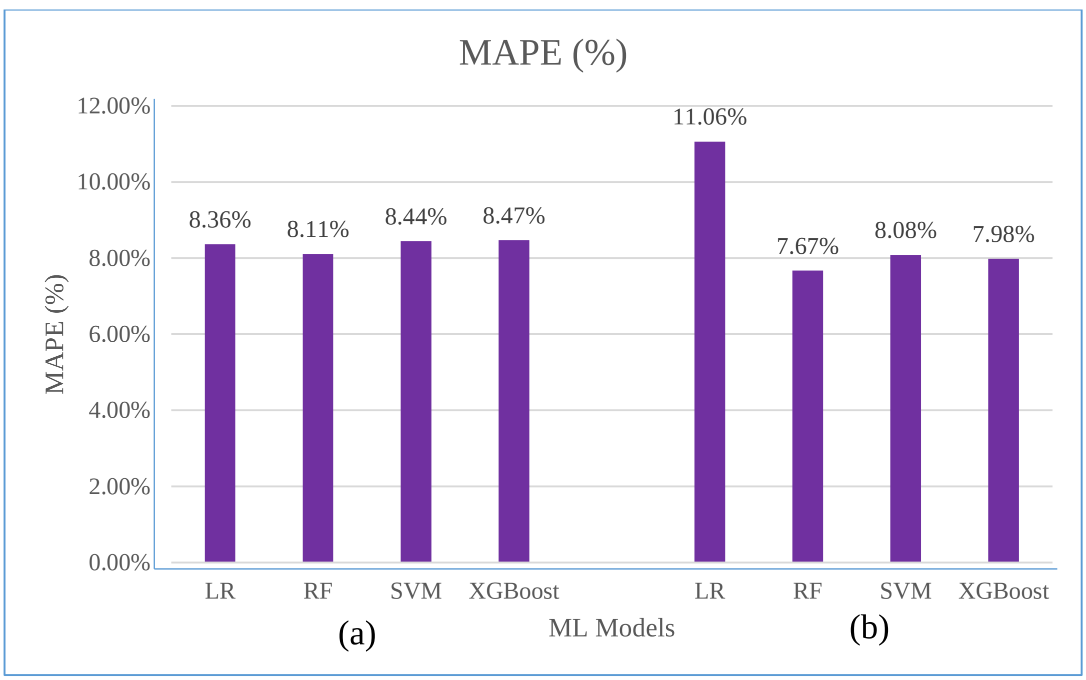

Figure 4.11: MAPE Values Obtained from ML Models Using SAR Only (a) and Combination of SAR and Auxiliary Climatic (b) Datasets

This study utilized two combinations of predictors which were assessed individually together with corresponding yield data covering the rice main growing season for the years 2021 to 2024 in MIS. These combinations included: S1-SAR backscatter time series datasets (VH and VV polarization) and auxiliary climatic datasets (rainfall data as well as minimum and maximum temperature). The aim was to establish how different ML algorithms compare in their ability to predict rice yield based on SAR and auxiliary climatic datasets. To ensure consistency, the same partitioning strategy was utilized for each combination in the development of training data and testing data. The four machine learning models were built using 80% of the datasets while the remaining 20% was isolated for assessing model performance. The VV and VH polarization datasets acquired from S1 were used due to the fact that studies have reported their reliability as vegetation indicators for monitoring agricultural crops (Khabbazan et al., 2019; Villarroya-Carpio et al., 2022). Additionally, rainfall, minimum and maximum temperature were utilized since they are among the features widely used in crop yield prediction with machine learning and deep learning algorithms (Jabed & Azmi Murad, 2024).

The satellite dataset combination of only S1-SAR backscatter had relatively higher error values in RMSE, MAE and MAPE compared to the combination involving SAR backscatter and auxiliary climatic datasets which suggests that spectral datasets alone had limited capacity in crop yield prediction. This aligns with the findings of Guo et al. (2023) and Islam et al. (2023) in their work that involved crop yield predictions with machine learning using spectral data and different auxiliary datasets including rainfall, soil moisture, soil properties as well as evapotranspiration. It can be noted that apart from LR whose structure is linear all the other non-linear machine learning models i.e. RF, SVM and XGBoost indicated lower model error values in estimating rice yield using the combination of S1-SAR backscatter and auxiliary climatic datasets. The higher model errors in the linear regression method may insinuate that this regression problem proves too complex to solve using linear determination. The linear regression model's poor performance in regression problems involving crop yield prediction has also been reported in other studies in different settings (Islam et al., 2023; Pejak et al., 2022).

The findings from this study indicate that the ensemble-based RF model is capable of integrating datasets from different sources (satellite-SAR & climatic) for rice yield prediction. Further, the RF model copes with complex non-linear relationships better than the other models. These findings align with observations by Islam et al. (2023) who established that ensemble model combinations involving stacking four different tree-based machine learning models (XGBoost, Light GBM, Gradient Boost and Random Forest) were more effective in accurate rice yield estimation attaining lower RMSE and MAE values despite dataset and methodology variance. Our study distinguishes itself through the utilization of S1-SAR backscatter datasets. Other studies in different settings and conditions have also established the RF model to be superior and robust in crop yield predictions such as the work by Singh Boori et al. (2023), Maseko et al. (2024) , Panigrahi et al. (2023) and Jhajharia et al. (2023).

The S1-SAR backscatter and auxiliary climatic dataset combinations indicated an improvement in the non-linear ML models performance resulting in lower RMSE, MAE and MAPE values suggesting that the inclusion of additional variables related to climate enhances machine learning model's predictive capabilities. This aligns with observations from other rice yield monitoring studies. For instance, in one study, Islam et al. (2023) established that the inclusion of auxiliary data types including rainfall, soil moisture, evapotranspiration and Land surface temperature (LST) with spectral data improved the rice yield prediction performance of machine learning models. Other studies have also illustrated that combining satellite imagery with weather information improves the accuracy of crop yield predictions with machine learning models (Huang et al., 2022; Mia et al., 2023).

Crop yield prediction robustness and accuracies can also be enhanced using more advanced techniques such as deep learning approaches which have frameworks with the ability to draw out complex spatial and temporal features (Lu et al., 2025; Mia et al., 2023). As such, deep learning approaches have been applied for rice yield prediction in other studies (Mia et al., 2023; Zhou et al., 2023). In particular, Lu et al. (2025) utilized Bayesian-optimized Convolutional Long Short-Term Memory with Attention (BCLA) model for rice yield prediction in Northern, China which attained higher prediction accuracies compared RF and XGBoost ML models. However, this study utilized a total of 108,195 survey samples which points to the data demands of deep learning approaches. Moreover, deep learning-based models like the CNN demand substantial computational power and memory, which can be a drawback in resource-limited settings (Mia et al., 2023). Additionally, the systematic review by Meghraoui et al. (2024) states that even though deep leaning approaches can achieve highly precise crop yield predictions, comprehension of how specific factors influence model predictions remains an enigma.

Overall, different studies have reported superior performance of deep learning approaches in crop yield predictions (Cao et al., 2021; Liu et al., 2022; Zhou et al., 2023). However, the findings of this study illustrate that ML models including RF, SVM, and XGBoost proposed in this study could be instrumental for rice yield prediction particularly in resource-constrained environments, because they require less data, are easier to train with limited computational resources, and offer more interpretability.

4.5 Performance Evaluation of Machine Learning Models in Predicting Rice Yield Using Linear Regression Model as a Benchmark.

Figures 4.12–4.15 present the actual and predicted spatial distribution of rice yield in Mwea Irrigation Scheme for the 2024 main season. Figure 4.12 shows the actual observed rice yield, while Figure 4.13, Figure 4.14, and Figure 4.15 present rice yield predicted using the RF, XGBoost, and SVM ML models, respectively. The predicted yield was compared with the actual observed yield to evaluate the prediction performance of the ML models. The actual average yield observed from the 62 blocks was 2622.44 kg/acre while the average predicted yield by the ML models was 2673.67kg/ acre from RF, followed by 2717.64 kg/acre from XGBoost and finally 2991.59 kg/acre from SVM. Furthermore, the performance metrics of the ML models in predicting rice yield for the 2024 main season was evaluated using the RMSE, MAE and MAPE and the findings are presented in Figure 4.16 (b), Figure 4.17 (b) and Figure 4.18 (b).

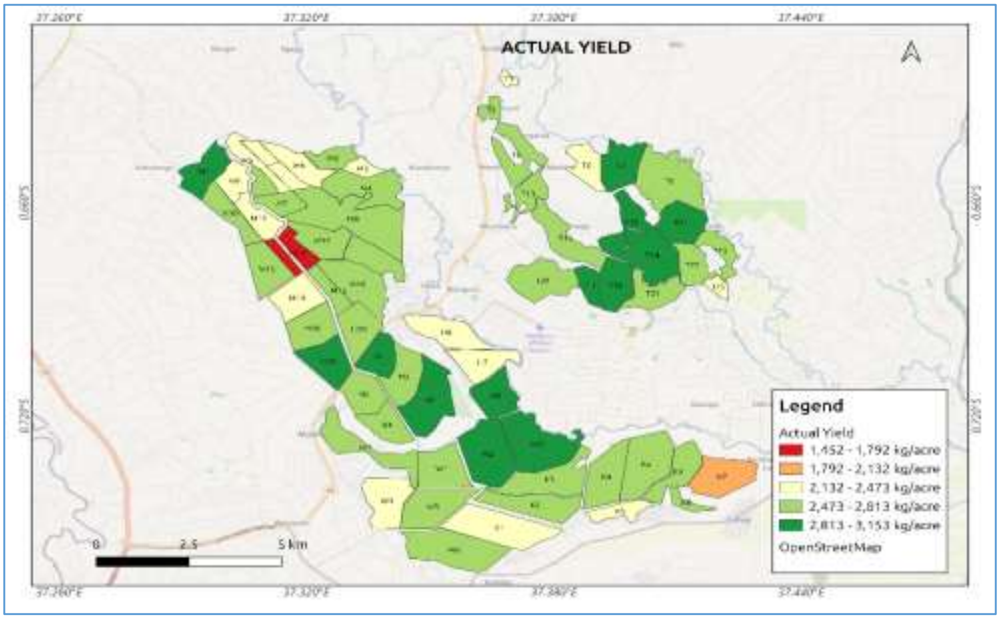

Figure 4.12: Map for the Spatial Distribution of Actual Rice Yield for the 2024 Main Season

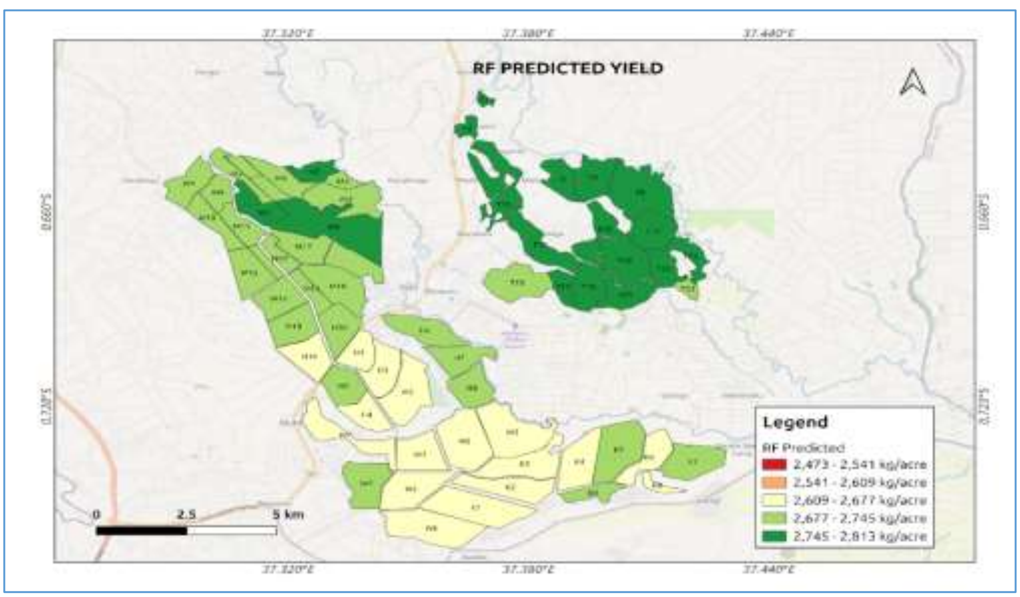

Figure 4.13: Map for the Spatial Distribution of Predicted Rice Yield for the 2024 Main Season Using the RF Model

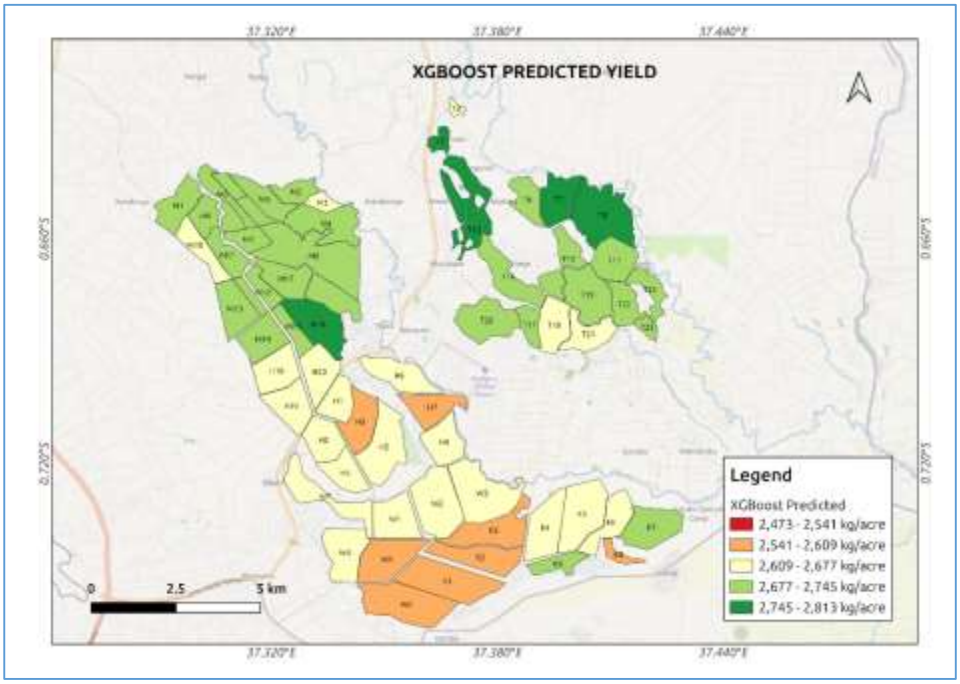

Figure 4.14: Map for the Spatial Distribution of Predicted Rice Yield for the 2024 Main Season Using the XGBoost Model

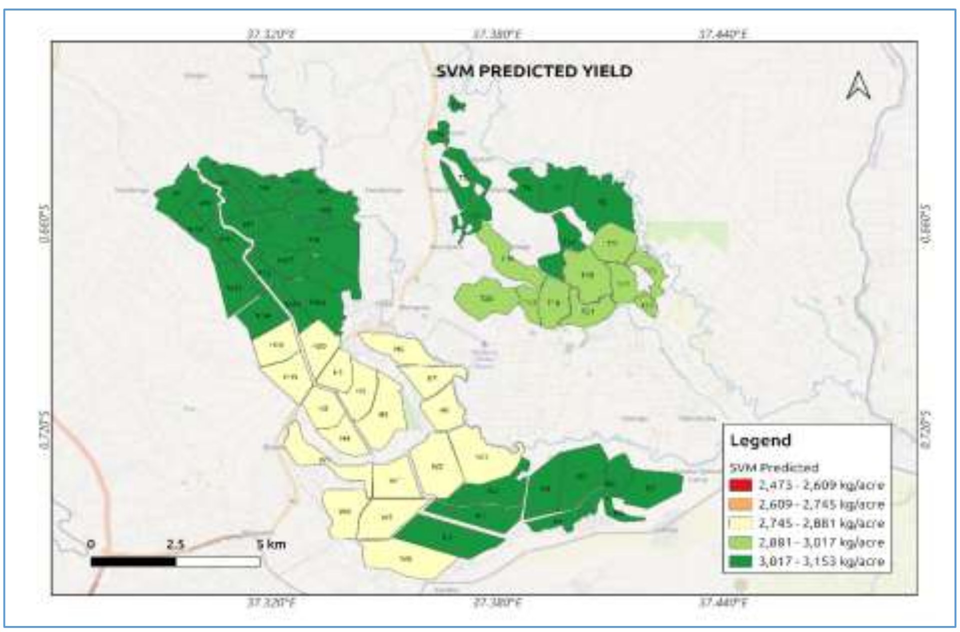

Figure 4.15: Map for the Spatial Distribution of Predicted Rice Yield for the 2024 Main Season Using the SVM Model

The performance evaluation of the trained ML models using the 20% isolated datasets (2021-2023) for testing is captured in Figure 4.16 (a), Figure 4.17 (a) and Figure 4.18 (a). The RF model emerged with the highest accuracy having the lowest error values with a RMSE of 209.66 kg/acre, MAE of 167.167 kg/acre and MAPE of 6.407% indicating strong predictive accuracy. The XGBoost model emerged second having a RMSE of 211.08 kg/acre, MAE of 169.41 kg/acre and MAPE of 6.40%. The SVM model emerged third with a RMSE of 212.37 kg/acre, MAE of 177.06 kg/acre and MAPE of 6.78%. Finally, LR model the benchmark performed poorest of the four models as illustrated by its higher RMSE of 227.18 kg/acre, MAE of 184.80 kg/acre and MAPE of 7.05%.

In predicting the 2024 main rice growing season, the performance metrics of the ML models are presented in Figure 4.16 (b), Figure 4.17 (b) and Figure 4.18 (b). The RF model emerged with the highest accuracy having the lowest RMSE of 288.52 kg/acre, MAE of 195.13 kg/acre and MAPE of 8.43%, followed closely by XGBoost model with a RMSE of 290.44 kg/acre, MAE of 201.45 kg/acre and MAPE of 8.56% and finally the SVM model attained a RMSE of 493.66 kg/acre, MAE of 395.21 kg/acre and MAPE of 16.73%. The, LR model gave negative yield results which was an indication of model failure in predicting the 2024 crop yields.

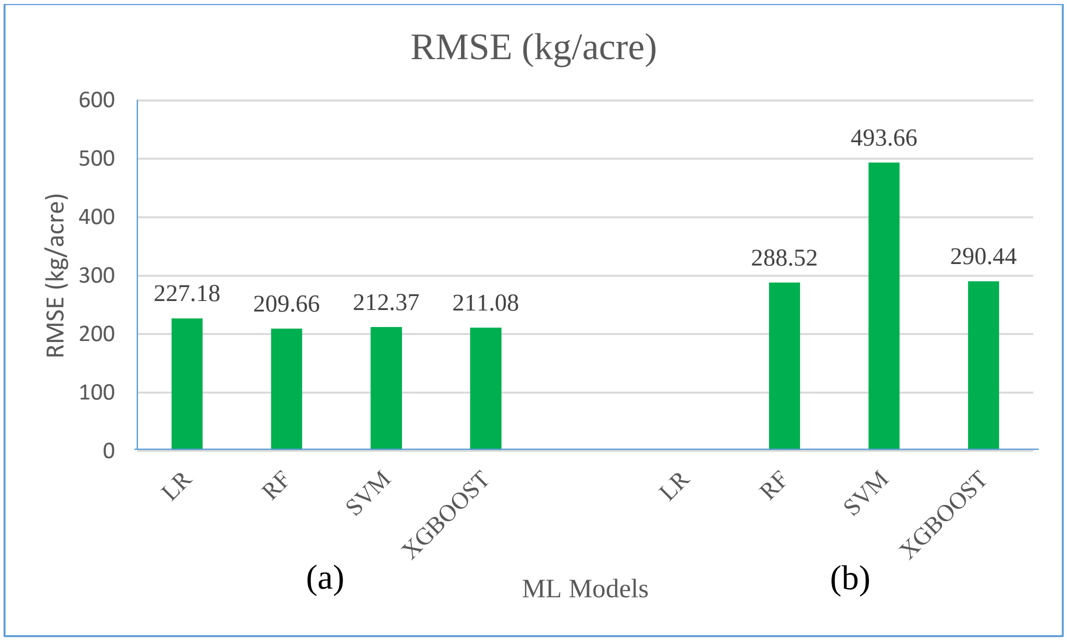

Figure 4.16: RMSE Values after Model Testing (b) and 2024 Yield Prediction Validation (b)

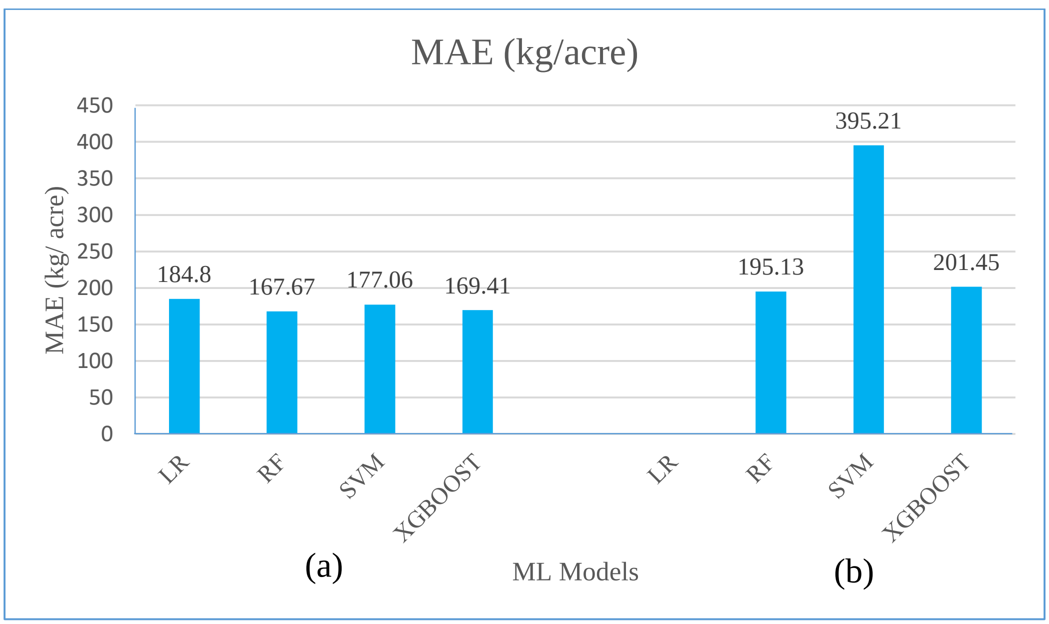

Figure 4.17: MAE Values after Model Testing (a) and 2024 Yield Prediction Validation (b)

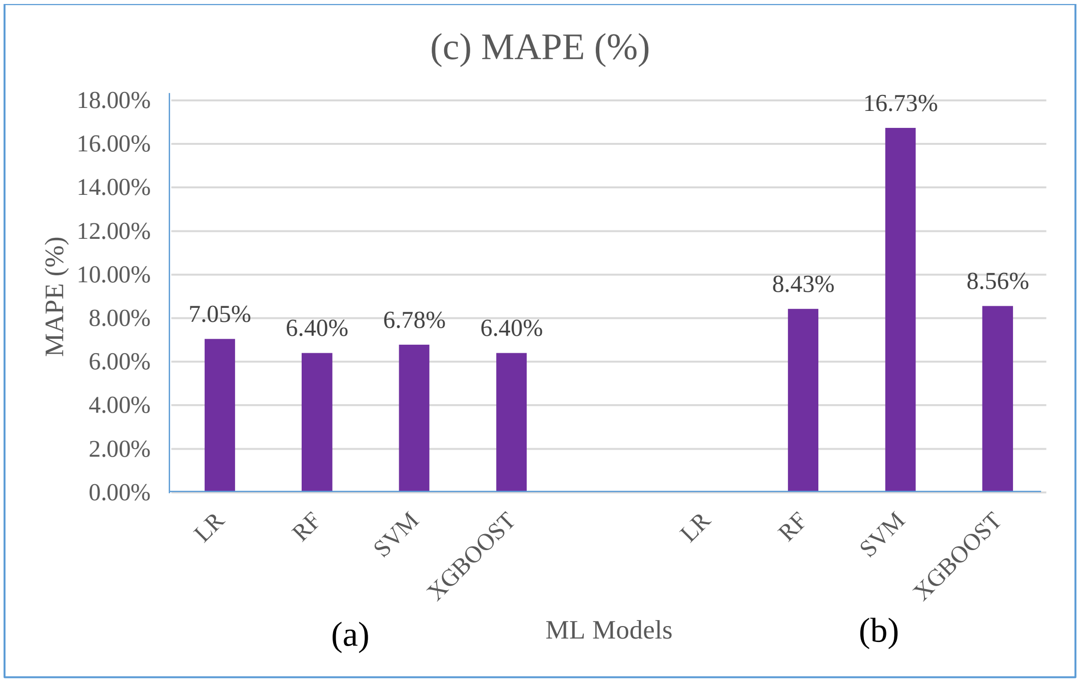

Figure 4.18: MAPE Values after Model Testing (a) and 2024 Yield Prediction Validation (b)

Four machine learning models were utilized in predicting future yield outcomes. To accomplish this, the four ML models were developed using data from the previous seasons covering three years (2021-2023) to predict the yields for the following year (2024). The models were developed using datasets covering the main growing season (July-December) for the years 2021-2023 which included a range of feature variables including S1-SAR backscatter datasets (VH and VV polarization) and auxiliary climatic datasets (rainfall, minimum and maximum temperature). The entire dataset was split such that 80% of the data was used for model development with the remaining 20% being isolated for testing performance of trained ML models before commencing crop yield prediction for the 2024 main rice growing season.

Once the models were developed, their predictive capabilities were validated on actual yield data for the 2024 main season. This isolated dataset was not utilized in model development, hence allowing for the assessment of ML model performance in accurately predicting the observed 2024 crop yields. This provided a glimpse into the

ML models' effectiveness in capturing variations and trends in rice yield for the 2024 main rice growing season. The utilization of feature variables improved the models' capacity to represent the complex relationships between environmental conditions and crop productivity, hence making ML an asset in agricultural forecasting.

The RF model had a higher accuracy in its ability to predict future yields for the upcoming season using datasets from previous seasons. These findings align with observations by Desloires et al. (2023) where non-linear ML models i.e. RF, SVM & XGBoost were found effective in predicting yields for years outside their training sample despite the methodology and dataset disparity. This study distinguishes itself through the use of S1-SAR backscatter and auxiliary climatic datasets for rice yield prediction outside the training sample dataset. Further, it has been demonstrated that nonlinear ML models including RF can effectively model the complex relationships between spectral data and crop yield due to its simplicity and strong performance (Deines et al., 2021). Additionally, other studies have attempted to undertake out of sample rice yield predictions using ML models (De Clercq & Mahdi, 2024; Son et al., 2022). In one study, De Clercq & Mahdi (2024) found that RF and XGBoost consistently outperformed SVR and ANN for out of sample rice yield predictions. This could likely be attributed to the RF and XGBoost models being less sensitive to the collinearity of input variables and hence resulting in better model performance (H. Huang et al., 2022). Further, the RF ensemble learning approach combines multiple decision trees, providing fast computation, strong generalization ability, and robust predictive accuracy (Jabed & Azmi Murad, 2024).

The results illustrate that LR model fails to capture the non-linear interaction relationships between satellite and climatic datasets with rice yield. This could likely have been caused by the LR model working approach of trying to fit a straight line between input variables and yield despite the fact that yield is influenced by nonlinear responses. Machine learning model failure has also been reported in other studies. For instance, Guo et al. (2023) reported failure of stepwise linear regression to produce maize crop yield results. Furthermore, other studies have reported poor performance of linear regression type models for rice yield predictions using satellite and climatic datasets (Cao et al., 2021; Liu et al., 2022). Specifically, the findings by Cao et al. (2021) indicate that machine learning and deep learning models significantly outperformed the linear regression approach for rice yield prediction utilizing multi-source datasets in China.

Time-series yield forecasting enables ongoing predictions throughout the growing season by incorporating up-to-date satellite and weather data (Potgieter et al., 2022). Earlier studies have shown that prediction accuracy tends to be lower early in the season but improves closer to harvest (Brinkhoff et al., 2024; Potgieter et al., 2022). Hence, this study focused on time series datasets (satellite and climatic) capturing the critical rice growth stages of transplanting, vegetative, reproductive and maturity. This put the lead time of the forecast about a month to the commencement of the rice harvest season which occurs in late November and December. The approach used in this study evaluates the capacity and performance of the ML models in real world scenario where the models are forced to generalize well on a different year/season utilizing only datasets from previous years/ seasons.

To assess the performance of ML models in rice yield prediction, the ML models were ranked according to the magnitude of the prediction errors they produced. A review by De Clercq and Mahdi (2025), which assessed 153 AI-based rice yield prediction studies, established that the majority of the studies did not report statistical significance tests when comparing machine learning models. Instead, model performance was primarily evaluated using descriptive performance metrics such as the RMSE, MAE, and MAPE. Further, the study identified the need for more robust and standardised ML model evaluation frameworks. As such, this study ranked the performance of the different ML models in predicting rice yield based on the RMSE, MAE and MAPE performance metrics.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The following conclusions can be made from this study:

i. The findings demonstrate that integration of Sentinel-1 SAR backscatter and Sentinel-2 optical time series data applying a phenology-driven framework on the Google Earth Engine platform provides an effective approach for rice mapping and growth stage identification. The method successfully generated a rice extent map accurately and captured the temporal progression of rice growth stages during the 2024 main season. The study therefore confirms that multi-sensor satellite data can provide timely and reliable information for monitoring rice cultivation at scheme level, thereby supporting irrigation planning and water resource allocation activities.

ii. The findings from this study suggest that integrating Sentinel-1 SAR backscatter with rainfall and temperature datasets improves rice yield prediction compared to using SAR data alone. Among the evaluated ML algorithms, RF consistently achieved the highest prediction accuracy having lower error values, indicating that ensemble learning based models are better suited for modelling the complex relationships between rice yield, SAR backscatter data and climatic variables of rainfall and temperature than conventional LR model.

iii. The study demonstrates that ML models calibrated using previous seasons' datasets can reasonably predict rice yield for subsequent growing seasons. The RF model provided the most accurate predictions one month prior to harvesting, followed closely by XGBoost, while SVM produced acceptable but lower predictive accuracy. The LR model failed to generalize to an unseen growing season, indicating that linear relationships are inadequate for capturing seasonal variability in rice production. These findings demonstrate the potential of ML models as operational tools for pre-harvest rice yield prediction using historical remote sensing integrated with rainfall and temperature datasets.

5.2 Recommendations

i. It is recommended that government agencies including the National Irrigation Authority (NIA), Ministry of Agriculture and County Governments should adopt Sentinel-1 and Sentinel-2 time-series datasets for routine mapping of rice extent and identification of growth stages. This would provide timely information for irrigation scheduling, monitoring cropped area and planning seasonal water allocation.

ii. The combination of Sentinel-1 SAR, rainfall and temperature variables with ML models should be integrated into operational yield prediction systems since they improve prediction accuracy compared to SAR data alone. Meteorological agencies and agricultural institutions should strengthen the collection and sharing of climate data to support national crop monitoring programmes. Future studies should consider investigating the contribution of additional explanatory variables such as soil characteristics, weed control, fertilizer application, pest and disease incidence, to improve yield prediction accuracy.

iii. Machine learning-based rice yield prediction, particularly using the RF model, should be incorporated into agricultural monitoring frameworks to provide early estimates of rice production before harvest. Reliable predictions would support government and state agencies in evidence-based decision-making on grain procurement, import and export planning, food reserve management and market stabilization. Future research could also compare conventional machine learning algorithms with deep learning approaches such as Long Short-Term Memory (LSTM), Temporal Convolutional Networks (TCN), Convolutional Neural Networks (CNN) and Transformer-based models to determine whether they provide superior prediction accuracy using multi-temporal satellite data with rainfall and temperature data.

REFERENCES

Abebe, G., Degefa, K., Biru, G., & Tsegaye, T. (2022). *Determinants of Adoption of Improved Rice Variety by Small Scale Farmers in Chewaka District of Buno Zone of Oromia Regional State*. *12*(23), 2022. https://doi.org/10.7176/JBAH/12-23-02

Acharya, T. D., Lee, D. H., Yang, I. T., & Lee, J. K. (2016). Identification of Water Bodies in a Landsat 8 OLI Image Using a J48 Decision Tree. *Sensors*, *16*(7). https://doi.org/10.3390/s16071075

Ajith, S., Vijayakumar, S., & Elakkiya, N. (2025). Yield prediction, pest and disease diagnosis, soil fertility mapping, precision irrigation scheduling, and food quality assessment using machine learning and deep learning algorithms. *Discover Food*, *5*(1), 67. https://doi.org/10.1007/s44187-025-00338-1

Akoko, G., Kato, T., & Tu, L. H. (2020). Evaluation of Irrigation Water Resources Availability and Climate Change Impacts—A Case Study of Mwea Irrigation Scheme, Kenya. *Water*, *12*(9). https://doi.org/10.3390/w12092330

Ali, A., Abouelghar, M., Belal, A. A., Saleh, N., Yones, M., Selim, A. I., Amin, M. E. S., Elwesemy, A., Kucher, D. E., Maginan, S., & Savin, I. (2022). Crop Yield Prediction Using Multi Sensors Remote Sensing (Review Article). *The Egyptian Journal of Remote Sensing and Space Sciences*, *25*(3), 711–716. https://doi.org/https://doi.org/10.1016/j.ejrs.2022.04.006

Ali, A. M., Savin, I., Poddubskiy, A., Abouelghar, M., Saleh, N., Abutaleb, K., El-Shirbeny, M., & Dokukin, P. (2020). Integrated method for rice cultivation monitoring using Sentinel-2 data and Leaf Area Index. *The*

Egyptian Journal of Remote Sensing and Space Science. https://doi.org/10.1016/j.ejrs.2020.06.007

Ali, A., Savin, I., Poddubskiy, A., Abouelghar, M., Saleh, N., Abutaleb, K., El-Shirbeny, M., & Dokukin, P. (2021). Integrated method for rice cultivation monitoring using Sentinel-2 data and Leaf Area Index. *The Egyptian Journal of Remote Sensing and Space Science*, *24*(3, Part 1), 431–441. https://doi.org/https://doi.org/10.1016/j.ejrs.2020.06.007

Altork, Y. (2025). Comparative analysis of machine learning models for wind speed forecasting: Support vector machines, fine tree, and linear regression approaches. *International Journal of Thermofluids*, *27*, 101217. https://doi.org/https://doi.org/10.1016/j.ijft.2025.101217

Amankulova, K., Farmonov, N., Omonov, K., Abdurakhimova, M., & Mucsi, L. (2024). Integrating the Sentinel-1, Sentinel-2 and topographic data into soybean yield modelling using machine learning. *Advances in Space Research*, *73*(8), 4052–4066. https://doi.org/https://doi.org/10.1016/j.asr.2024.01.040

Ansarifar, J., Wang, L., & Archontoulis, S. V. (2021). An interaction regression model for crop yield prediction. *Scientific Reports*, *11*(1), 17754. https://doi.org/10.1038/s41598-021-97221-7

Atera, E., Florence, N., & Majiwa, E. (2018). Production and marketing of rice in Kenya: Challenges and opportunities. *Journal of Development and Agricultural Economics*, *10*, 64–70. https://doi.org/10.5897/JDAE2017.0881

Aworka, R., Cedric, L. S., Adoni, W. Y. H., Zoueu, J. T., Mutombo, F. K., Kimpolo, C. L. M., Nahhal, T., & Krichen, M. (2022). Agricultural decision system based on advanced machine learning models for yield prediction: Case of East African countries. *Smart Agricultural Technology*, *2*, 100048. https://doi.org/10.1016/j.atech.2022.100048

Bauder, J. W., & Randall, G. W. (1982). Regression Models for Predicting Corn Yields from Climatic Data and Management Practices. *Soil Science Society of America Journal*, *46*(1), 158–161. https://doi.org/https://doi.org/10.2136/sssaj1982.03615995004600010039x

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). *A Comparative Analysis of XGBoost*.

Betew, A. G., Gebresenbet, G., Gelaw, G. K., Mengistu, D. A., & Yibre, A. M. (2026). Historical and contemporary crop yield prediction models: Key lessons and innovations. *Smart Agricultural Technology*, *13*, 101672. https://doi.org/https://doi.org/10.1016/j.atech.2025.101672

Bogdanovski, O. P., Svenningsson, C., Månsson, S., Oxenstierna, A., & Sopasakis, A. (2023). Yield Prediction for Winter Wheat with Machine Learning Models Using Sentinel-1, Topography, and Weather Data. *Agriculture*, *13*(4). https://doi.org/10.3390/agriculture13040813

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brinkhoff, J., Clarke, A., Dunn, B. W., & Groat, M. (2024). Analysis and forecasting of Australian rice yield using phenology-based aggregation of satellite and weather data. *Agricultural and Forest Meteorology*, *353*, 110055. https://doi.org/https://doi.org/10.1016/j.agrformet.2024.110055

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., & Xie, J. (2021). Integrating Multi-Source Data for Rice Yield Prediction across

China using Machine Learning and Deep Learning Approaches. *Agricultural and Forest Meteorology*, *297*, 108275. https://doi.org/https://doi.org/10.1016/j.agrformet.2020.108275

Carneiro, F. M., de Brito Filho, A. L., Ferreira, F. M., de Freitas Seben Junior, G., Brandão, Z. N., da Silva, R. P., & Shiratsuchi, L. S. (2023). Soil and satellite remote sensing variables importance using machine learning to predict cotton yield. *Smart Agricultural Technology*, *5*, 100292. https://doi.org/https://doi.org/10.1016/j.atech.2023.100292

Cedric, L. S., Adoni, W. Y. H., Aworka, R., Zoueu, J. T., Mutombo, F. K., Krichen, M., & Kimpolo, C. L. M. (2022). Crops yield prediction based on machine learning models: Case of West African countries. *Smart Agricultural Technology*, *2*, 100049. https://doi.org/https://doi.org/10.1016/j.atech.2022.100049

Chu, Z., & Yu, J. (2020). An end-to-end model for rice yield prediction using deep learning fusion. *Computers and Electronics in Agriculture*, *174*, 105471. https://doi.org/https://doi.org/10.1016/j.compag.2020.105471

Clarke, A., Yates, D., Blanchard, C., Islam, M. Z., Ford, R., Rehman, S., & Walsh, R. (2024). The effect of dataset construction and data pre-processing on the eXtreme Gradient Boosting algorithm applied to head rice yield prediction in Australia. *Computers and Electronics in Agriculture*, *219*, 108716. https://doi.org/https://doi.org/10.1016/j.compag.2024.108716

Cochran, W. G. . (1977). *Sampling techniques.* John Wiley & Sons.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, *37*(1), 35–46. https://doi.org/https://doi.org/10.1016/0034-4257(91)90048-B

Cravero, A., Sepúlveda, S., Gutiérrez, F., & Muñoz, L. (2026). From Precision Agriculture to Intelligent Agricultural Ecosystems: A Systematic Review of Machine Learning and Big Data Applications. *Agronomy*, *16*(5). https://doi.org/10.3390/agronomy16050516

De Clercq, D., & Mahdi, A. (2024). Feasibility of machine learning-based rice yield prediction in India at the district level using climate reanalysis and remote sensing data. *Agricultural Systems*, *220*, 104099. https://doi.org/https://doi.org/10.1016/j.agsy.2024.104099

De Clercq, D., & Mahdi, A. (2025). Modern computational approaches for rice yield prediction: A systematic review of statistical and machine learning-based methods. *Computers and Electronics in Agriculture*, *231*, 109852. https://doi.org/https://doi.org/10.1016/j.compag.2024.109852

Deines, J. M., Patel, R., Liang, S.-Z., Dado, W., & Lobell, D. B. (2021). A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. *Remote Sensing of Environment*, *253*, 112174. https://doi.org/https://doi.org/10.1016/j.rse.2020.112174

Desloires, J., Ienco, D., & Botrel, A. (2023). Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods. *Computers and Electronics in Agriculture*, *209*, 107807. https://doi.org/https://doi.org/10.1016/j.compag.2023.107807

Du, Z., Yang, J., Ou, C., & Zhang, T. (2019). Smallholder Crop Area Mapped with a Semantic Segmentation Deep Learning Method. *Remote Sensing*, *11*(7). https://doi.org/10.3390/rs11070888

Elders, A., Carroll, M. L., Neigh, C. S. R., D'Agostino, A. L., Ksoll, C., Wooten, M. R., & Brown, M. E. (2022). Estimating crop type and yield of small holder fields in Burkina Faso using multi-day Sentinel-2. *Remote Sensing Applications: Society and Environment*, *27*, 100820. https://doi.org/https://doi.org/10.1016/j.rsase.2022.100820

Fatchurrachman, Rudiyanto, Soh, N. C., Shah, R. M., Giap, S. G. E., Setiawan, B. I., & Minasny, B. (2022). High-Resolution Mapping of Paddy Rice Extent and Growth Stages across Peninsular Malaysia Using a Fusion of Sentinel-1 and 2 Time Series Data in Google Earth Engine. *Remote Sensing*, *14*(8). https://doi.org/10.3390/rs14081875

Fikriyah, V. N., Darvishzadeh, R., Laborte, A., & Nelson, A. (2025). Ratoon rice mapping based on Sentinel-1 and Sentinel-2 imagery. *Remote Sensing Applications: Society and Environment*, *38*, 101592. https://doi.org/https://doi.org/10.1016/j.rsase.2025.101592

Fiorillo, E., Di Giuseppe, E., Fontanelli, G., & Maselli, F. (2020). Lowland Rice Mapping in Sédhiou Region (Senegal) Using Sentinel 1 and Sentinel 2 Data and Random Forest. *Remote Sensing*, *12*(20). https://doi.org/10.3390/rs12203403

Gao, Y., Wang, S., Guan, K., Wolanin, A., You, L., Ju, W., & Zhang, Y. (2020). The Ability of Sun-Induced Chlorophyll Fluorescence From OCO-2 and MODIS-EVI to Monitor Spatial Variations of Soybean and Maize Yields in the Midwestern USA. *Remote Sensing*, *12*(7). https://doi.org/10.3390/rs12071111

Gavasso-Rita, Y. L., Papalexiou, S. M., Li, Y., Elshorbagy, A., Li, Z., & Schuster-Wallace, C. (2024). Crop models and their use in assessing crop production and food security: A review. *Food and Energy Security*, *13*(1), e503. https://doi.org/https://doi.org/10.1002/fes3.503

Guo, Z., Chamberlin, J., & You, L. (2023). Smallholder maize yield estimation using satellite data and machine learning in Ethiopia. *Crop and Environment*, *2*(4), 165–174. https://doi.org/https://doi.org/10.1016/j.crope.2023.07.002

Hashemi, M. G. Z., Tan, P.-N., Jalilvand, E., Wilke, B., Alemohammad, H., & Das, N. N. (2024). Yield estimation from SAR data using patch-based deep learning and machine learning techniques. *Computers and Electronics in Agriculture*, *226*, 109340. https://doi.org/https://doi.org/10.1016/j.compag.2024.109340

He, Z., Li, S., Wang, Y., Dai, L., & Lin, S. (2018). Monitoring Rice Phenology Based on Backscattering Characteristics of Multi-Temporal RADARSAT-2 Datasets. *Remote Sensing*, *10*(2). https://doi.org/10.3390/rs10020340

Huang, H., Huang, J., Feng, Q., Liu, J., Li, X., Wang, X., & Niu, Q. (2022). Developing a Dual-Stream Deep-Learning Neural Network Model for Improving County-Level Winter Wheat Yield Estimates in China. *Remote Sensing*, *14*(20). https://doi.org/10.3390/rs14205280

Huang, Y., Chen, Z., Yu, T., Huang, X., & Gu, X. (2018). Agricultural remote sensing big data: Management and applications. *Journal of Integrative Agriculture*, *17*, 1915–1931. https://doi.org/10.1016/S2095-3119(17)61859-8

Inoue, S., Ito, A., & Yonezawa, C. (2020). Mapping Paddy Fields in Japan by Using a Sentinel-1 SAR Time Series Supplemented by Sentinel-2 Images on Google Earth Engine. *Remote Sensing*, *12*(10). https://doi.org/10.3390/rs12101622

International Rice Research Institute(IRRI). (2020). *IRRI Annual Report 2022-Transforming Food systems in Africa and Asia*. https://www.irri.org/about-us

Islam, M. D., Di, L., Qamer, F. M., Shrestha, S., Guo, L., Lin, L., Mayer, T. J., & Phalke, A. R. (2023). Rapid Rice Yield Estimation Using Integrated Remote Sensing and Meteorological Data and Machine Learning. *Remote Sensing*, *15*(9). https://doi.org/10.3390/rs15092374

Ismail, M. H., & Jusoff, K. (2008). Satellite data classification accuracy assessment based from reference dataset. *International Journal of Computer and Information Science and Engineering*, *2*, 96–102.

Jabed, Md. A., & Azmi Murad, M. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*, *10*(24), e40836. https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e40836

Jeong, S., Ko, J., & Yeom, J.-M. (2022). Predicting rice yield at pixel scale through synthetic use of crop and deep learning models with satellite data in South and North Korea. *Science of The Total Environment*, *802*, 149726. https://doi.org/https://doi.org/10.1016/j.scitotenv.2021.149726

Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop Yield Prediction using Machine Learning and Deep Learning Techniques. *Procedia Computer Science*, *218*, 406–417. https://doi.org/https://doi.org/10.1016/j.procs.2023.01.023

Jiang, X., Fang, S., Huang, X., Liu, Y., & Guo, L. (2021). Rice Mapping and Growth Monitoring Based on Time Series GF-6 Images and Red-Edge Bands. *Remote Sensing*, *13*(4). https://doi.org/10.3390/rs13040579

Clauss K, M. O., & Kuenzer, C. (2018). Mapping rice areas with Sentinel-1 time series and superpixel segmentation. *International Journal of Remote Sensing*, *39*(5), 1399–1420. https://doi.org/10.1080/01431161.2017.1404162

Kabir, M. A., Ahmed, F., Islam, M. M., & Ahmed, Md. R. (2024). Python For Data Analytics: A Systematic Literature Review Of Tools, Techniques, And Applications. *Academic Journal on Science, Technology, Engineering & Mathematics Education*, *4*, 134–154. https://doi.org/10.69593/ajsteme.v4i04.146

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., & Anderson, M. (2020). Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environmental Research Letters*, *15*(6), 64005. https://doi.org/10.1088/1748-9326/ab7df9

Khabbazan, S., Vermunt, P., Steele-Dunne, S., Ratering Arntz, L., Marinetti, C., van der Valk, D., Iannini, L., Molijn, R., Westerdijk, K., & van der Sande, C. (2019). Crop Monitoring Using Sentinel-1 Data: A Case Study from The Netherlands. *Remote Sensing*, *11*(16). https://doi.org/10.3390/rs11161887

Kobayashi, S., & Ide, H. (2022). Rice Crop Monitoring Using Sentinel-1 SAR Data: A Case Study in Saku, Japan. *Remote Sensing*, *14*(14). https://doi.org/10.3390/rs14143254

Lesiv, M., See, L., Laso Bayas, J. C., Sturn, T., Schepaschenko, D., Karner, M., Moorthy, I., McCallum, I., & Fritz, S. (2018). Characterizing the Spatial and Temporal Availability of Very High Resolution Satellite Imagery in Google Earth and Microsoft Bing Maps as a Source of Reference Data. *Land*, *7*(4). https://doi.org/10.3390/land7040118

Li, B., Liu, K., Wang, M., Wang, Y., Zhuang, L., Zhu, W., Li, C., Zhang, L., & Chen, Y. (2025). Evaluating the performance of pixel-based and object-based multidimensional clustering algorithms for automated surface water mapping. *Geo-Spatial Information Science*. https://doi.org/10.1080/10095020.2025.2523993

Li, Y., Zeng, H., Zhang, M., Wu, B., Zhao, Y., Yao, X., Cheng, T., Qin, X., & Wu, F. (2023). A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering. *International Journal of Applied Earth Observation and Geoinformation*, *118*, 103269. https://doi.org/https://doi.org/10.1016/j.jag.2023.103269

Liu, Y., Wang, S., Chen, J., Chen, B., Wang, X., Hao, D., & Sun, L. (2022). Rice Yield Prediction and Model Interpretation Based on Satellite and Climatic Indicators Using a Transformer Method. *Remote Sensing*, *14*(19). https://doi.org/10.3390/rs14195045

Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, *164*, 324–333. https://doi.org/10.1016/j.rse.2015.04.021

Lokesh, C., Gumma, M. K., Susheela, R., Ronanki, S., Shankaraiah, M., & Panjala, P. (2026). Integrating Remote Sensing and Crop Simulation Models for Rice Yield Estimation: A Comprehensive Review. *AgriEngineering*, *8*(3). https://doi.org/10.3390/agriengineering8030088

Lu, J., Li, J., Fu, H., Zou, W., Kang, J., Yu, H., & Lin, X. (2025). Estimation of rice yield using multi-source remote sensing data combined with crop growth model and deep learning algorithm. *Agricultural and Forest Meteorology*, *370*, 110600. https://doi.org/https://doi.org/10.1016/j.agrformet.2025.110600

Luo, C., Liu, H., Lu, L., Liu, Z., Kong, F., & Zhang, X. (2021). Monthly composites from Sentinel-1 and Sentinel-2 images for regional major crop mapping with Google Earth Engine. *Journal of Integrative Agriculture*, *20*(7), 1944–1957. https://doi.org/https://doi.org/10.1016/S2095-3119(20)63329-9

Ma, Z., Liu, Z., Zhao, Y., Zhang, L., Liu, D., Ren, T., Zhang, X., & Li, S. (2020). An Unsupervised Crop Classification Method Based on Principal Components Isometric Binning. *ISPRS International Journal of Geo-Information*, *9*(11). https://doi.org/10.3390/ijgi9110648

Mansaray, L. R., Zhang, K., & Kanu, A. S. (2020). Dry biomass estimation of paddy rice with Sentinel-1A satellite data using machine learning regression algorithms. *Computers and Electronics in Agriculture*, *176*, 105674. https://doi.org/https://doi.org/10.1016/j.compag.2020.105674

Maseko, S., van der Laan, M., Tesfamariam, E. H., Delport, M., & Otterman, H. (2024). Evaluating machine learning models and identifying key factors influencing spatial maize yield predictions in data intensive farm management. *European Journal of Agronomy*, *157*, 127193. https://doi.org/https://doi.org/10.1016/j.eja.2024.127193

Meena, R. S., Mitran, T., Kumar, S., Yadav, G. S., Bohra, J. S., & Datta, R. (2018). Application of remote sensing for sustainable agriculture and forest management. *Information Processing in Agriculture*, *5*(3), 295–297. https://doi.org/https://doi.org/10.1016/j.inpa.2018.03.003

Meghraoui, K., Sebari, I., Pilz, J., Ait El Kadi, K., & Bensiali, S. (2024a). Applied Deep Learning-Based Crop Yield Prediction: A Systematic Analysis of Current Developments and Potential Challenges. *Technologies*, *12*(4). https://doi.org/10.3390/technologies12040043

Meghraoui, K., Sebari, I., Pilz, J., Ait El Kadi, K., & Bensiali, S. (2024b). Applied Deep Learning-Based Crop Yield Prediction: A Systematic Analysis of Current Developments and Potential Challenges. *Technologies*, *12*(4). https://doi.org/10.3390/technologies12040043

Mia, Md. S., Tanabe, R., Habibi, L. N., Hashimoto, N., Homma, K., Maki, M., Matsui, T., & Tanaka, T. S. T. (2023). Multimodal Deep Learning for Rice Yield Prediction Using UAV-Based Multispectral Imagery and Weather Data. *Remote Sensing*, *15*(10). https://doi.org/10.3390/rs15102511

Minh, H. V. T., Avtar, R., Mohan, G., Misra, P., & Kurasaki, M. (2019). Monitoring and Mapping of Rice Cropping Pattern in Flooding Area in the Vietnamese Mekong Delta Using Sentinel-1A Data: A Case of An Giang Province. *ISPRS International Journal of Geo-Information*, *8*(5). https://doi.org/10.3390/ijgi8050211

Mokesh Rayalu, & Saruk BS. (2025). Optimizing rice yield prediction: A policy-oriented approach to agricultural sustainability. *Results in Engineering*, *28*, 107844. https://doi.org/https://doi.org/10.1016/j.rineng.2025.107844

Musila, R., Gichuhi, E., & Menge, D. (2024). *Handbook on paddy rice cultivation in Mwea Kenya*.

Mwangangi, J., Muturi, E., Shililu, J., Muriu, S., Jacob, B., Kabiru, E., Mbogo, C., Githure, J., & Novak, R. (2008). Environmental Covariates 0f Anopheles Arabiensis In A Rice Agroecosystem In Mwea, Central Kenya. *Journal of the American Mosquito Control Association*, *23*, 371–377. https://doi.org/10.2987/5605.1

Nakalembe, C., Becker-Reshef, I., Bonifacio, R., Hu, G., Humber, M. L., Justice, C. J., Keniston, J., Mwangi, K., Rembold, F., Shukla, S., Urbano, F., Whitcraft, A. K., Li, Y., Zappacosta, M., Jarvis, I., & Sanchez, A. (2021). A review of satellite-based global agricultural monitoring systems available for Africa. *Global Food Security*, *29*, 100543. https://doi.org/10.1016/j.gfs.2021.100543

Ng'endo, M., Kinyua, M., Chebet, L., Mutiga, S., Ndung'u, J., Nyongesa, O., Njau, S., Panchbhai, A., Musila, R., & Murori, R. (2022). The importance of market signals in crop varietal development: lessons from Komboka rice variety. *CABI Agriculture and Bioscience*, *3*(1), 57. https://doi.org/10.1186/s43170-022-00122-6

Njagi, T., & Mano, Y. (2023). Toward a Green Revolution in Sub-Saharan Africa: Farm Mechanization in the Mwea Irrigation Scheme. In J. P. Estudillo, Y. Kijima, & T. Sonobe (Eds.), *Agricultural Development in Asia and Africa: Essays in Honor of Keijiro Otsuka* (pp. 63–73). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-5542-6_5

Nyangau, W., Mati, B., Kalamwa, K., Wanjogu, R., & Kiplagat, L. (2014). Estimating Rice Yield under Changing Weather Conditions in Kenya Using CERES Rice Model. *International Journal of Agronomy*, *2014*, 1–12. https://doi.org/10.1155/2014/849496

Panigrahi, B., Kathala, K. C. R., & Sujatha, M. (2023). A Machine Learning-Based Comparative Approach to Predict the Crop Yield Using Supervised Learning With Regression Models. *Procedia Computer Science*, *218*, 2684–2693. https://doi.org/https://doi.org/10.1016/j.procs.2023.01.241

Pejak, B., Lugonja, P., Antić, A., Panić, M., Pandžić, M., Alexakis, E., Mavrepis, P., Zhou, N., Marko, O., & Crnojević, V. (2022). Soya Yield Prediction on a Within-Field Scale Using Machine Learning Models

Trained on Sentinel-2 and Soil Data. *Remote Sensing*, *14*(9). https://doi.org/10.3390/rs14092256

Potgieter, A. B., Schepen, A., Brider, J., & Hammer, G. L. (2022). Lead time and skill of Australian wheat yield forecasts based on ENSO-analogue or GCM-derived seasonal climate forecasts – A comparative analysis. *Agricultural and Forest Meteorology*, *324*, 109116. https://doi.org/https://doi.org/10.1016/j.agrformet.2022.109116

Pukrongta, N., Taparugssanagorn, A., & Sangpradit, K. (2024). Enhancing Crop Yield Predictions with PEnsemble 4: IoT and ML-Driven for Precision Agriculture. *Applied Sciences*, *14*(8). https://doi.org/10.3390/app14083313

Radočaj, D., Jurišić, M., Plaščak, I., & Galić, L. (2026). Satellite Remote Sensing for Crop Yield Prediction: A Review. *Agriculture*, *16*(4). https://doi.org/10.3390/agriculture16040417

Rajakumaran, M., Arulselvan, G., Subashree, S., & Sindhuja, R. (2024). Crop yield prediction using multi-attribute weighted tree-based support vector machine. *Measurement: Sensors*, *31*, 101002. https://doi.org/https://doi.org/10.1016/j.measen.2023.101002

Ramadhan, A., Priya, K., Pavithra, V., Mishra, P., Dash, A., Abotaleb, M., Alkattan, H., & Albadran, Z. (2024). Machine Learning Techniques for Sugarcane Yield Prediction Using Weather Variables. *BIO Web of Conferences*, *97*, 157. https://doi.org/10.1051/bioconf/20249700157

Ramadhani, F., Pullanagari, R., Kereszturi, G., & Procter, J. (2020). Automatic Mapping of Rice Growth Stages Using the Integration of SENTINEL-2, MOD13Q1, and SENTINEL-1. *Remote Sensing*, *12*(21). https://doi.org/10.3390/rs12213613

Rejeb, A., Rejeb, K., & Zailani, S. (2021). Big data for sustainable agri-food supply chains: a review and future research perspectives. *Journal of Data, Information and Management*, *3*(3), 167–182. https://doi.org/10.1007/s42488-021-00045-3

Republic of Kenya. (2020). *National Rice Development Strategy-2 (2019–2030). Ministry of Agriculture, Livestock, Fisheries and Cooperatives (MoALF&C), State Department for Crop Development and Agricultural Research*.

Rosenqvist, A. (2018). *A Layman's Interpretation Guide to L-band and C-band Synthetic Aperture Radar data*.

Rudiyanto, Minasny, B., Shah, R. M., Che Soh, N., Arif, C., & Indra Setiawan, B. (2019). Automated Near-Real-Time Mapping and Monitoring of Rice Extent, Cropping Patterns, and Growth Stages in Southeast Asia Using Sentinel-1 Time Series on a Google Earth Engine Platform. *Remote Sensing*, *11*(14). https://doi.org/10.3390/rs11141666

Rufaioğlu, S. B., Bilgili, A. V., Savaşlı, E., Özberk, İ., Aydemir, S., Ismael, A. M., Kaya, Y., & Matos-Carvalho, J. P. (2025). Sensor-Based Yield Prediction in Durum Wheat Under Semi-Arid Conditions Using Machine Learning Across Zadoks Growth Stages. *Remote Sensing*, *17*(14). https://doi.org/10.3390/rs17142416

Sadenova, M., Beisekenov, N., Varbanov, P. S., & Pan, T. (2023). Application of Machine Learning and Neural Networks to Predict the Yield of Cereals, Legumes, Oilseeds and Forage Crops in Kazakhstan. *Agriculture, 13*(6). https://doi.org/10.3390/agriculture13061195

Samsuddin Sah, S., Abdul Maulud, K. N., Sharil, S., A. Karim, O., & Pradhan, B. (2023). Monitoring of three stages of paddy growth using multispectral vegetation index derived from UAV images. *The Egyptian Journal of Remote Sensing and Space Sciences*, *26*(4), 989–998. https://doi.org/https://doi.org/10.1016/j.ejrs.2023.11.005

Sánchez, J. C. M., Mesa, H. G. A., Espinosa, A. T., Castilla, S. R., & Lamont, F. G. (2025). Improving wheat yield prediction through variable selection using Support Vector Regression, Random Forest, and Extreme Gradient Boosting. *Smart Agricultural Technology*, *10*, 100791. https://doi.org/https://doi.org/10.1016/j.atech.2025.100791

Santos, L. B., Gentry, D., Tryforos, A., Fultz, L., Beasley, J., & Gentimis, T. (2024). Soybean yield prediction using machine learning algorithms under a cover crop management system. *Smart Agricultural Technology*, *8*, 100442. https://doi.org/https://doi.org/10.1016/j.atech.2024.100442

Sharma, V., Honkavaara, E., Hayden, M., & Kant, S. (2024). UAV remote sensing phenotyping of wheat collection for response to water stress and yield prediction using machine learning. *Plant Stress*, *12*, 100464. https://doi.org/https://doi.org/10.1016/j.stress.2024.100464

Singh Boori, M., Choudhary, K., Paringer, R., & Kupriyanov, A. (2023). Machine learning for yield prediction in Fergana valley, Central Asia. *Journal of the Saudi Society of Agricultural Sciences*, *22*(2), 107–120. https://doi.org/https://doi.org/10.1016/j.jssas.2022.07.006

Singha, C., & Swain, K. C. (2023). Rice crop growth monitoring with sentinel 1 SAR data using machine learning models in google earth engine cloud. *Remote Sensing Applications: Society and Environment*, *32*, 101029. https://doi.org/https://doi.org/10.1016/j.rsase.2023.101029

Sitokonstantinou, V., Koukos, A., Drivas, T., Kontoes, C., Papoutsis, I., & Karathanassi, V. (2021). A Scalable Machine Learning Pipeline for Paddy

Rice Classification Using Multi-Temporal Sentinel Data. *Remote Sensing*, *13*(9). https://doi.org/10.3390/rs13091769

Son, N.-T., Chen, C.-F., Cheng, Y.-S., Toscano, P., Chen, C.-R., Chen, S.-L., Tseng, K.-H., Syu, C.-H., Guo, H.-Y., & Zhang, Y.-T. (2022). Field-scale rice yield prediction from Sentinel-2 monthly image composites using machine learning algorithms. *Ecological Informatics*, *69*, 101618. https://doi.org/https://doi.org/10.1016/j.ecoinf.2022.101618

Sun, L., Yang, T., Lou, Y., Shi, Q., & Zhang, L. (2024). Paddy Rice Mapping Based on Phenology Matching and Cultivation Pattern Analysis Combining Multi-Source Data in Guangdong, China. *Journal of Remote Sensing*, *4*, 152. https://doi.org/10.34133/remotesensing.0152

Thompson, L. J., & Puntel, L. A. (2020). Transforming Unmanned Aerial Vehicle (UAV) and Multispectral Sensor into a Practical Decision Support System for Precision Nitrogen Management in Corn. *Remote Sensing*, *12*(10). https://doi.org/10.3390/rs12101597

Thorp, K. R., & Drajat, D. (2021). Deep machine learning with Sentinel satellite data to map paddy rice production stages across West Java, Indonesia. *Remote Sensing of Environment*, *265*. https://doi.org/10.1016/j.rse.2021.112679

Tian, H., Meng, M., Wu, M., & Niu, Z. (2019). Mapping Spring Canola and Spring Wheat using Radarsat-2 and Landsat-8 Images with Google Earth Engine. *Current Science*, *116*, 291–298. https://doi.org/10.18520/cs/v116/i2/291-298

Uma, A., Okello, D., & Opondo, F. (2023). Effect of paddy sourcing methods on the volume of rice milled by rice millers in Mwea, Kirinyaga county, Kenya. *Heliyon*, *9*(12), e22901. https://doi.org/https://doi.org/10.1016/j.heliyon.2023.e22901

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/https://doi.org/10.1016/j.compag.2020.105709

Villarroya-Carpio, A., Lopez-Sanchez, J. M., & Engdahl, M. E. (2022). Sentinel-1 interferometric coherence as a vegetation index for agriculture. *Remote Sensing of Environment*, *280*, 113208. https://doi.org/https://doi.org/10.1016/j.rse.2022.113208

Wang, B., Liu, Y., Sheng, Q., Li, J., Tao, J., & Yan, Z. (2022). Rice Phenology Retrieval Based on Growth Curve Simulation and Multi-Temporal Sentinel-1 Data. *Sustainability*, *14*(13). https://doi.org/10.3390/su14138009

Wang, J., Dai, Q., Shang, J., Jin, X., Sun, Q., Zhou, G., & Dai, Q. (2019). Field-Scale Rice Yield Estimation Using Sentinel-1A Synthetic Aperture Radar (SAR) Data in Coastal Saline Region of Jiangsu Province, China. *Remote Sensing*, *11*(19). https://doi.org/10.3390/rs11192274

Wang, W., He, A., Jiang, G., Sun, H., Jiang, M., Man, J., Ling, X., Cui, K., Huang, J., Peng, S., & Nie, L. (2020). *Chapter Four - Ratoon rice technology: A green and resource-efficient way for rice production* (D. L. Sparks, Ed.; Vol. 159, pp. 135–167). Academic Press. https://doi.org/https://doi.org/10.1016/bs.agron.2019.07.006

Watanabe, M., Sumita, Y., Azechi, I., Ito, K., & Noda, K. (2021). The Value Chain of Locally Grown Japonica Rice in Mwea, Kenya. *Agriculture*, *11*(10). https://doi.org/10.3390/agriculture11100974

Wessels, K., Li, X., Bouvet, A., Mathieu, R., Main, R., Naidoo, L., Erasmus, B., & Asner, G. P. (2023). Quantifying the sensitivity of L-Band SAR to a decade of vegetation structure changes in savannas. *Remote Sensing of Environment*, *284*, 113369. https://doi.org/https://doi.org/10.1016/j.rse.2022.113369

Woolf, B. P. (2009). Chapter 7 - Machine Learning. In B. P. Woolf (Ed.), *Building Intelligent Interactive Tutors* (pp. 221–297). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-373594-2.00007-1

Xu, S., Zhu, X., Chen, J., Zhu, X., Duan, M., Qiu, B., Wan, L., Tan, X., Xu, Y. N., & Cao, R. (2023). A robust index to extract paddy fields in cloudy regions from SAR time series. *Remote Sensing of Environment*, *285*, 113374. https://doi.org/https://doi.org/10.1016/j.rse.2022.113374

Yli-Heikkilä, M., Wittke, S., Luotamo, M., Puttonen, E., Sulkava, M., Pellikka, P., Heiskanen, J., & Klami, A. (2022). Scalable Crop Yield Prediction with Sentinel-2 Time Series and Temporal Convolutional Network. *Remote Sensing*, *14*(17). https://doi.org/10.3390/rs14174193

Yu, W., Yang, G., Li, D., Zheng, H., Yao, X., Zhu, Y., Cao, W., Qiu, L., & Cheng, T. (2023). Improved prediction of rice yield at field and county levels by synergistic use of SAR, optical and meteorological data. *Agricultural and Forest Meteorology*, *342*, 109729. https://doi.org/https://doi.org/10.1016/j.agrformet.2023.109729

Zhang, J., Gao, L., Liu, M., Dong, Y., Liu, C., Casa, R., Pignatti, S., Huang, W., Li, Z., Tian, T., & Hu, R. (2024). Accurate Paddy Rice Mapping Based on Phenology-Based Features and Object-Based Classification. *Remote Sensing*, *16*(23). https://doi.org/10.3390/rs16234406

Zhou, S., Xu, L., & Chen, N. (2023). Rice Yield Prediction in Hubei Province Based on Deep Learning and the Effect of Spatial Heterogeneity. *Remote Sensing*, *15*(5). https://doi.org/10.3390/rs15051361

Zhu, Z., Gallant, A., Woodcock, C., Pengra, B., Olofsson, P., Loveland, T., Jin, S., Dahal, D., Yang, L., & Auch, R. (2016). Optimizing selection of training and auxiliary data for operational land cover classification for the LCMAP initiative. *ISPRS Journal of Photogrammetry and Remote Sensing*, *122*. https://doi.org/10.1016/j.isprsjprs.2016.11.004

APPENDICES

Appendix I: Characteristics and Sources of Satellite Data used in this Study

Satellite	Sensor	Temporal Resolution (days)	Spatial Resolution (m)	Polarization/ Band	Wavelength(nm)	Use
Sentinel-2	MSI	5 days	10	B3	560	Green
			10	B4	665	Red
			20	B5	705	Vegetation Red Edge
			10	B8	842	Near Infrared
Sentinel-1	C-band SAR	12 days	10	C (VH)		Interferometric Wide Mode

Appendix II: Satellite Image Product ID Data used in Rice Growth Stages Identification for this Study

Satellite	Product
Sentinel 2	**COPERNICUS/S2_HARMONIZED/20240701T073621_20240701T075911_T37MBV**
	COPERNICUS/S2_HARMONIZED/20240830T073611_20240830T075015_T37MBV
	COPERNICUS/S2_HARMONIZED/20240909T073611_20240909T075950_T37MBV
	COPERNICUS/S2_HARMONIZED/20240914T073609_20240914T075345_T37MBV
	COPERNICUS/S2_HARMONIZED/20240914T073609_20240914T075345_T37MCV
	COPERNICUS/S2_HARMONIZED/20241004T073659_20241004T075531_T37MBV
	COPERNICUS/S2_HARMONIZED/20241009T073821_20241009T075433_T37MCV
	COPERNICUS/S2_HARMONIZED/20241029T074031_20241029T075434_T37MBV
	COPERNICUS/S2_HARMONIZED/20241029T074031_20241029T075434_T37MCV
	COPERNICUS/S2_HARMONIZED/20241118T074211_20241118T075516_T37MBV
	COPERNICUS/S2_HARMONIZED/20241208T074321_20241208T075544_T37MCV

Satellite	**Product**		
Interpretation of the first S1 Image in bold	**Component**	**Meaning**	**Interpretation**
	COPERNICUS/S2_HARMONIZED	GEE Image Collection	Sentinel-2 Harmonized Level-1C (Top-of-Atmosphere Reflectance) collection
	20240701T073621	Datatake sensing start time	**1 July 2024, 07:36:21 UTC**
	20240701T075911	Product discriminator (processing) time	**1 July 2024, 07:59:11 UTC** (used to distinguish products from the same datatake)
	T37MBV	MGRS tile	**Tile 37MBV**, covering part of central Kenya including the Mwea Irrigation Scheme

Sentinel 1	**COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20240712T154841_20240712T154906_054729_06A9E1_D9A6** COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20240724T154840_20240724T154905_054904_06AFF9_B498 COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20240805T154840_20240805T154905_055079_06B5FF_57AF COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20240817T154840_20240817T154905_055254_06BC63_3120 COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20240829T154841_20240829T154906_055429_06C2DB_84D2 COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20240910T154841_20240910T154906_055604_06C9B0_7D3C COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20240922T154841_20240922T154906_055779_06D08F_E4B6 COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20241004T154842_20241004T154907_055954_06D788_C324 COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20241016T154831_20241016T154856_056129_06DE66_B871 COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20241028T154831_20241028T154856_056304_06E558_E23A COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20241109T154830_20241109T154855_056479_06EC44_284A COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20241121T154830_20241121T154855_056654_06F34E_DFEA COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20241203T154829_20241203T154854_056829_06FA3F_7974 COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20241215T154828_20241215T154853_057004_070124_1720 COPERNICUS/S1_GRD/S1A_IW_GRDH_1SDV_20241227T154838_20241227T154903_057179_07081D_8349

<table>
<tr><td rowspan="12">Interpretation of the first S-1 Image in bold</td><th>Component</th><th>Meaning</th><th>Interpretation</th></tr>
<tr><td>COPERNICUS/S1_GRD</td><td>GEE Image Collection</td><td>Sentinel-1 Ground Range Detected (GRD) collection in Google Earth Engine</td></tr>
<tr><td>S1A</td><td>Satellite</td><td>Sentinel-1A</td></tr>
<tr><td>IW</td><td>Instrument mode</td><td>Interferometric Wide Swath (IW) mode, the standard land imaging mode with a 250 km swath</td></tr>
<tr><td>GRDH</td><td>Product type</td><td>Ground Range Detected (GRD), High Resolution (~10 m pixel spacing)</td></tr>
<tr><td>1</td><td>Processing level</td><td>Level-1 product</td></tr>
<tr><td>SDV</td><td>Product class and polarization</td><td>S = Standard product, DV = Dual polarization (VV and VH)</td></tr>
<tr><td>20240712T154841</td><td>Acquisition start time</td><td>12 July 2024, 15:48:41 UTC</td></tr>
<tr><td>20240712T154906</td><td>Acquisition end time</td><td>12 July 2024, 15:49:06 UTC</td></tr>
<tr><td>054729</td><td>Absolute orbit number</td><td>Orbit 54,729</td></tr>
<tr><td>06A9E1</td><td>Mission data-take ID</td><td>Unique identifier for the satellite acquisition</td></tr>
<tr><td>D9A6</td><td>Product unique identifier</td><td>Unique product checksum/identifier used to distinguish this processed product</td></tr>
</table>

Appendix III: MIS Cropping Program for the 2024 Rice Growing Season

MWEA IRRIGATION SCHEME

PROPOSED CROPPING PROGRAMME – 2024/2025 PRODUCTION SEASON

GROUP	BRANCH CANAL	UNITS	FLOODING	ROTAVATION	SEED ISSUE	NURSERY SOWING	TRANS-PLANTING	1^{ST} IRR	1^{ST} SPRAY	1^{ST} WEEDING	1^{ST} TOP DRESSING	2^{ND} TOP DRESSING	HARV'G
1	NMC		27/5/2024– 19/6/2024	27/5/2024– 19/6/2024	26/6/2024– 27/6/2024	29/6/2024– 30/6/2024	20/7/2024– 30/7/2024	30/7/2024– 10/8/2024	11/8/2024– 13/8/2024	27/7/2024– 1/8/2024	11/8/2024– 13/8/2024	12/9/2024- 15/9/2024	18/11/2024– 2/12/2024
	NBC 1	T6,T7,T10J/kali,T22,T23,T25 and Kianugu											
	NBC 2&3	T15,T19,T13,T18,T9J/kali,T15J/kali, T21,T16,T17,T1J/kali											
	TRW	T20											
	TMC	M14,M9,M10,M11,M12,M13,H18,H19,H20											
	TBC 1A&B	M5,M7,M8,M1,M2,M3,M4,M6											
	TBC 2	H1,H6,H7,H8											
	TBC 3	H3,H4,Mbui njeru,H5,H5Ext.,W2,W3											
	TBC 4	W7,W6											
	CURUKIA	Kiandegwa, MIAD, Kandongu and Mugaa											
2	NMC	T2,T3,T5	10/6/2024– 26/6/2024	10/6/2024– 26/6/2024	4/7/2024– 6/7/2024	7/7/2024 – 10/7/2024	28/7/2024 – 2/8/2024	7/8/2024- 17/8/2024	18/8/2024– 20/8/2024	4/8/2024 – 9/8/2024	18/8/2024– 20/9/2024	19/9/2024- 2/10/2024	25/11/2024- 9/12/2024
	NBC 1	T11,T8											
	NBC 2&3	T13E,T12,T14J/kali											
	TRW												
	TMC	M15,M16,Njoga,M17											
	TBC 1A&B												
	CURUKIA	Cumbiri 1 ,II, III, Riga , Ngothi											
	TBC 2												
	TBC 3	H2,W1,W8											
	TBC 4	W4,W5,K1,K2,K3,K4,K5,K6,K7,K8,K9,K10a &b,K11,Marurumo											
3	NDEKIA	Ndekia 1,2a,2b,3,4	4-16/9/2024	4-16/9/2024	1/9/2024	4/9/2024	25/9/2024	5/10/2024	16/10/2024	1/10/2024	15/10/2024	9/11/2024	23/1/2025
4	KIAMANYEKI	A,B,C,D	4/10/2024 16/10/2024	4/10/2024 17/10/2024	1/10/2024 3/10/2024	4/10/2024 5/10/2024	25/10/2024- 4/11/2024	5/11/2024 15/11/2024	28/11/2024	7/11/2024	17/11/2024	18/12/2024	22/2/2025 29/2/2025

Harvesting Date for different rice varieties;

1st group: – Basmati – 18/11/24, Komboka – 12/11/2024, Arize – 27/11/24
2nd group: – Basmati – 25/11/24, Komboka – 20/11/2024, Arize – 6/12/24
3rd group: – Basmati – 23/1/2025, Komboka – 18/1/2025, Arize – 2/2/2025
4th group: – Basmati – 23/2/2025, Komboka – 23/1/2025, Arize – 7/2/2025