

Wheat Yield Prediction Using Remote Sensing and Machine Learning



By

Fatima Khattak


(2019-NUST-MS -GIS-320534)

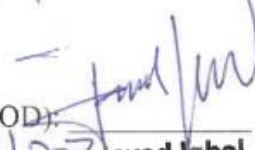
**A thesis submitted in partial fulfillment of the requirements for the degree
of Master of Science in Remote Sensing and GIS**

**Institution of Geographical Information Systems
School of Civil and Environmental Engineering
National University of Science & Technology
Islamabad, Pakistan
June 2023**

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by «Fatima Khattak » (Registration No. MSRSGIS-320534), of Session 2019 (Institute of Geographical Information systems) has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulation, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: 
Name of Supervisor: Dr. Javed Iqbal
Date: 18/8/23
Dr. Javed Iqbal
Professor & HOD IGIS. SCEE (NUST)
H-12, Islamabad

Signature (HOD): 
Date: 18/8/23
Dr. Javed Iqbal
Professor & HOD IGIS. SCEE (NUST)
H-12, Islamabad

Signature (Associate Dean): 
Date: 18.8.2023
Dr. Ejaz Hussain
Associate Dean IGIS. SCEE (NUST)
H-12, ISLAMABAD

Signature (Principal & Dean SCEE): 
Date: 26 AUG 2023
PROF DR MUHAMMAD IRFAN
Principal & Dean
SCEE, NUST

DEDICATION

This work is dedicated

To

My loving Parents, my caring and supporting siblings and my friends whose
tremendous cooperation led me to this wonderful accomplishment

ACADEMIC THESIS: DECLARATION OF AUTHORSHIP

I, Fatima Khattak declare that this thesis and the work presented in it are my own and have been generated by me as the result of my original research.

“Wheat Yield Prediction Using Remote Sensing and Machine Learning”

I confirm that:

1. This work was done wholly by me in candidature for an MS research degree at the National University of Sciences and Technology, Islamabad.
2. Wherever I have consulted the published work of others, it has been attributed.
3. Wherever I have quoted from the work of others, the source has been always cited.
4. I have acknowledged all main sources of help.
5. Where the work of the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
6. None of this work has been published before submission. This work is not plagiarized under the HEC plagiarism policy.

Signed:

Date:


Fatima
KIK
25/08/2023

ACKNOWLEDGEMENTS

I am thankful to Almighty Allah, who is the creator of all and is the Master of the day of judgment. No words of thanks can be appropriate for his immense blessings.

I would like to acknowledge the contribution of the following people without whose help and guidance this would not have reached completion. I would like to take this opportunity to sincerely express my highest gratitude to Dr, Javed Iqbal, my esteemed Supervisor for making this research possible. His support, guidance, and advice, throughout the research project, as well as his painstaking effort in proofreading the draft, are greatly appreciated.

I am deeply grateful to my thesis committee member for their valuable suggestions, support, and guidance during this research work. I am also feeling gratitude for all IGIS faculty members and staff for their assistance and encouragement. I would like to thank the Directorate of Crop Reporting Service for providing relevant data during the completion of this research. I am grateful to my family and friends for their moral support, inspiration and help during the whole MS.

Fatima Khattak

TABLE OF CONTENTS

CERTIFICATE	ii
DEDICATION	iii
ACADEMIC THESIS: DECLARATION OF AUTHORSHIP	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vi
LIST OF FIGURES.....	viii
LIST OF TABLES	ix
ABSTRACT	x
Chapter 1: INTRODUCTION.....	1
1.1 Background.....	1
1.2 Research Gap.....	3
1.3 Objectives	3
1.4 Significance of Research	4
1.5 Objectives	4
Chapter 2: LITERATURE REVIEW.....	5
Chapter 3: MATERIAL AND METHODS.....	11
3.1 Study Area	11
3.2 Types of Crops in Faisalabad Division	13
3.3 The Environmental and Geographical Characteristics of the Faisalabad Division ...	13
3.4 Wheat Crop Production.....	14
3.5 Remote Sensing Data Acquisition.....	14
3.6 Google Earth Engine	15
3.7 Crop Yield Ground Truth Data.....	15
3.8 Meteorological Data	15
3.9 Software and Library Used	15
3.10 Data Processing.....	18
3.11 Vegetation Indices	18
3.12 Vegetation Indices Calculation	19
3.13 Feature Selection.....	19

3.14 Machine Learning Model’s Implementation.....	19
3.15 The Random Forest Regression (RFR) Model.....	21
3.16 The Decision Tree Regression (DTR) Model	21
3.17 Experimental Setup.....	21
3.18 Model Accuracy Assessment.....	22
Chapter 4 : RESULTS AND DISCUSSION.....	24
4.1 Relationship of Variables with Yield	24
4.2 Subset Variables Selected from the Correlation Analysis	30
4.3 Models Training Results.....	32
4.4 Models Testing Results	32
4.5 Analysis of Models	35
4.6 Identify the Best Model.....	35
Chapter 5 : CONCLUSION AND RECOMMENDATIONS.....	36
5.1 Conclusion.....	36
5.2 Recommendations	36
REFERENCES	37
APPENDICES	44
Appendix-1. Yield correlation with time series T2MAX (the maximum temperature at 2 meters) (2019-20).....	44
Appendix-2. Yield correlation with time series T2MAX (the maximum temperature at 2 meters) (2020-21).....	45
Appendix-3. Yield correlation with time series relative humidity (relative humidity at 2 meters) (2019-20).....	46
Appendix-4. Yield Correlation with time series relative humidity (relative humidity at 2 meters) (2020-21).....	47

LIST OF FIGURES

Figure 3.1. Study area map.	12
Figure 3.2. NDVI time-series & phenological stages	20
Figure 3.3. Complete flow chart methodology of the study area.....	23
Figure 4.1. Model training results of decision tree regression (DTR) and random tree regression (RFR).....	33
Figure 4.2. Model testing results of decision tree regression (DTR) and random tree regression (RFR).....	34

LIST OF TABLES

Table 3.1. Landsat 8 acquisition dates for the study area in 2019-20.....	16
Table 3.2. Landsat 8 acquisition dates for the study area in 2020-21.....	16
Table 3.3. List of datasets with specification and their sources	17
Table 3.4. Tools and software used for the analysis and preprocessing of the data.	17
Table 4.1. NDVI time series correlation with the yield (kg) 2019-20.....	26
Table 4.2. NDVI time series correlation with the yield (kg) 2020-2021	27
Table 4.3. EVI time series correlation with the yield (kg) 2019-2020.....	28
Table 4.4. EVI time series correlation with the yield (kg) 2020-21	29
Table 4.5. Yield correlation with a subset of variables selected from the correlation analysis (2019-20)	31
Table 4.6. Yield correlation with a subset of variables selected from correlation analysis (2020-21).....	31

ABSTRACT

Crop management extensively utilizes remote sensing data for predicting crop yield. Freely available data products (Landsat, sentinel) have been used extensively. This study explores the potential of remote sensing and machine learning for wheat yield estimation of the Faisalabad division of Punjab, by utilizing Landsat-8 surface reflectance data. Time series of vegetation indices as Normalized Difference Vegetation indices (NDVI) and Enhanced Vegetation Indices (EVI) for the years 2019-2020 and 2020- 2021 were extracted. Several machine learning models were tested and two models were selected for the final yield prediction after feature selection using correlation analysis. Random Forest Regression (RFR) and Decision Tree Regression (DTR) are the two models that were used for wheat yield prediction. Feature selection was critical in reducing input data to avoid uncertainty, and only important data was used as input to the models. The 8th time step was found to have a high correlation with yield, and data from this step was used for model input. For the two years separately, separate feature selections were made and meteorological variables of other time steps were found to be correlated with yield. Training and testing results and model accuracy were based on the Root mean square Error (RMSE) and Root Square. The results of the Decision Tree Regression (DTR) in training (RMSE = 0.062, R2 = 0.952 t/ha, RMSE = 0.062, R2 = 0.952 t/ha) and testing (RMSE = 0.150, R2 = 0.700 t/ha, RMSE = 0.120, R2= 0.799 t/ha) for both of the years shows that the model overfitted in the training phase. The results of Random Forest Regression (RFR) in training (RMSE = 0.076, R2 = 0.929 t/ha, RMSE = 0.075, R2= 0.930 t/ha) and testing (RMSE = 0.144, R2=0.725 t/ha, RMSE = 0.106, R2 = 0.842 t/ha) for both of the years. The finding suggests that the RFR models resist overfitting and have strong adaptability for the variables and wheat yield prediction. This study demonstrates the potential of remote sensing and machine learning in precision agriculture and provides on sights into the selection of relevant input data for accurate yield prediction.

INTRODUCTION

1.1 Background

Accurate and real-time crop yield prediction is now very important in the context of food security and precision agriculture and has significant value in the formulation of food policies, decision-making, adjustment of food prices and management of agriculture (Mueller et al., 2012). Future population projections put Pakistan at a very high increasing rate which shows that the population in Pakistan would be 271 million by 2050. However, the rate of food production does not meet the increasing rate (Schafer and Victor, 2000). Due to climate change, Pakistan faces severe drought and floods which affects food availability. Access of planners and policymakers to real-time yield estimation and monitoring of the crop condition is very important to sustain and manage the availability of food for the ever-growing population of Pakistan (Briscoe and Qamar, 2006). Agriculture constitutes the largest sector of Pakistan's economy, and the main portion of our population directly or indirectly depends on the agriculture sector. Wheat is the major crop in Pakistan which is grown over a large area and a wide range of soil and climatic conditions. Up to 2.0 per cent of GDP and 9.9 per cent of the total agriculture productivity comprises wheat production (Faruque et al., 1996).

Wheat yield estimation before harvesting is important for guidance and timely decision. Traditional approaches are based on field surveys and high-quality measurements for crop yield estimation which are costly and time-consuming. Due to the inefficiency of traditional methods, it is sometimes very hard to predict yield on a larger scale. Modern approaches are based on remote sensing data, vegetation indices based on satellite data, and various crop models used for the early estimation of crop yield (Dorosh and Salam, 2008). Remote sensing data which is acquired with high spatial and temporal resolution can provide data in high volumes to study several terrestrial phenomena such as crop health and stress monitoring. Remote sensing data is obtained in several electromagnetic regions however, instead of using raw reflectance values in vegetation indices works well to study vegetation and agriculture since they highlight the properties of vegetation and suppress the secondary information. Vegetation indices like normalized difference vegetation index (NDVI), Enhanced vegetation index (EVI), and Land surface temperature (LST) are used to monitor crop yield, vegetation stress, and biomass (Friedl et al, 2002). For crop yield prediction, mainly two kinds of techniques are used: (1) physical crop simulation models, and (2) statistical models. Physical models due to their detailed nature are widely used for crop yield estimations. However, they require expert knowledge and very detailed data on several parameters which is sometimes

challenging to collect. Statistical regression methods due to their simplicity and ease of usage are widely used, but the limitation of statistical models is that it is not extendable to other areas, as typically localized (Pede et al., 2019). In recent years with some explicit cause-and-effect relationship, crop models are progressively replaced by statistical regression models due to their spatial generalization and explanatory power (Chen et al., 2017). In addition, soil data and climate data such as temperature and precipitation are the primary inputs for the model as they can capture important environmental information used for crop yield prediction. In most of the statistical models, regression equations are developed between measured yields at different spatial and temporal scales and climate variables like precipitation, temperature and solar radiation etc. For the yield prediction (Zhang et al., 2015). Many researchers are focused on improving crop prediction by using different methods and machine learning models (ML), which performed better than the traditional statistical models. Machine learning models, due to their efficiency in classification and predictions are well suited for such problems. Due to the availability of high spatial and temporal data of explanatory variables, the usage of machine learning is increasing for yield prediction. Compared to a statistical model, the machine learning model prefers weights rather than the probability or likelihood of any information prediction (Lee et al., 2018). Machine learning is a subfield of artificial intelligence which enables the machine (computer) to learn from the data and experiences without explicit instructions or programming.

Machine learning is self-learning based on algorithms, which means the system learns from its experience as the input data type grasps the pattern and the responding result is the model learning as output. It is a sub-class of artificial intelligence and automatically learns based on data representation without human help (Sharma et al., 2021). In recent decades, machine learning techniques have been used for data mining and have demonstrated their powerful performance in agriculture analysis, including yield prediction and crop type classification (Cai et al., 2018). Although in previous studies, the accuracy in crop yield prediction has improved from spatial and temporal domains, they have only focused on the smaller scale region due to the complicated data process (Aghighi et al., 2018). The larger-scale crop yield prediction generally requires large and complex data that need more time and cost to process huge data sets (Jin et al., 2019). Fortunately, the Google earth engine (GEE) is a cloud-based computing platform that freely provides data along with processing capabilities (at petabyte-scale). GEE contains raw and processed data such as Vegetation indices, Land surface temperature and satellite data for the geospatial analysis and also visualization of the geospatial data set (Gorelick et al., 2017). The research was carried out on wheat yield

prediction using remote sensing data acquired from the GEE of the study area. Landsat-8 satellite temporal (2019- 2020 to 2020 to 2021 with a gap of one year) imagery was used to calculate the vegetation indices such as normalized difference vegetation Indices (NDVI) and enhanced vegetation index (EVI). Wheat yield (kg) correlation with time series NDVI and EVI was generated from Landsat-8 surface reflectance of both of the years. NDVI and EVI time series and phenological stages of sowing, anthesis and maturity stages were generated. The correlation was carried out between time series yield and meteorological data of the study area's maximum temperature, minimum temperature, and precipitation. Machine learning models such as Decision Tree Regression and Random Forest Regression were trained and tested on predicted yield and measured yield. Finally, both model's performances and accuracy comparison were carried out using R², RMSE.

1.2 Research Gap

There are some studies on the usage of machine learning and crop yield prediction. However, most of the studies focused on the usage of statistical and machine learning for image classification and vegetation indices calculation to study crops. Based on the classification results, the studies extract different crops and subsequent areas of those crops. This further leads to yield prediction models for crop yield prediction. In this study, we do not use any crop type classification as it can lead to uncertainties in the actual crop fields and area. We directly use the plot scale data and extracted the remote sensing data based on the location of the plots. This study focuses on the advantage of using surface reflectance data from the GEE and machine learning models for wheat yield prediction. This study also compares the performance of different machine learning models, and the two best models were then used to predict the yield of the study area.

1.3 Problem Statement

As crop yield prediction is very important for precision agriculture, policymakers to make early decisions for the food stock and export and import of the wheat crop. As wheat is one of the main crops in Pakistan, facing the challenge of increasing population along with other climate crises. The study area has a high potential to cultivate more and a high yield of wheat crops. Early prediction of the wheat yield helps in the prior strategies that can help in improving the wheat yield.

1.4 Significance of Research

Crop yield prediction is very important from several perspectives such as food security, precision agriculture, farm management and crop breeding. The early prediction of yield can support policy decision-making. In this context, yield prediction is very important and challenging in agriculture decision-making. The study was done to compare the best machine learning model used to predict crop yield. The study outcomes will help policymakers in early decision-making about the food stocks, export and import of the respective crop. From the early prediction of crop yield farmers can be drawn decisions on what to do during the growing of crops and what crops to grow to meet the required target of food production according to the rate of increasing population. This study can help the farmers will take on new precision agriculture techniques and collaborate with the government. Furthermore, conventional techniques of yield prediction are costly and not very accurate. Another major challenge is that it is not possible to estimate large-scale yields in a short period. Remote sensing data and machine learning methods provide a quick and less expensive solution to resolve this and provide an early estimation of vegetation indices.

1.5 Objectives

- To predict wheat yield using Landsat-8 surface reflectance and machine-learning models.
- To compare the performance of different machine learning algorithms on multiple datasets such as remote sensing and meteorological data for wheat yield prediction.

LITERATURE REVIEW

At present, satellite remote sensing-based technologies are widely used for the monitoring of standing crops, their health condition, water stress, and the early prediction of crop yield. All of the above information is helped to the policy maker to make decisions for the betterment of the crops and to incorporate with the farmers to take the steps like precise fertilization on other management, which is an effective way to ensure food security (Bongiovanni and Lowenberg-Deboer, 2004). In the satellite, remote sensing vegetation indices are used to study the plant's health and growth with the help of different sensors that carry by the satellite. Normalized difference vegetation index (NDVI) and Enhanced vegetation index (EVI), as both of the indices, is based on near-infrared (NIR) and red portion (670nm, 800nm) of the spectrum help regarding vegetation analysis (Fu et al., 2014). High-resolution remote sensing data of the unmanned vehicle (UAV) were used for the crop yield prediction of winter wheat at field scale in Xuzhou City, Jiangsu Province, China. For this regional scale study, six machine learning (ML) models and ten different vegetation (VI) were used for the five key growth stages of the wheat crop. Among the five ML models the Gaussian Process Regression (GPR) achieved the highest accuracy of $R^2= 0.87$ to predict the crop yield at the field scale (Bian et al., 2022). For the management of crops and global food security, early and reliable crop yield prediction is very important. At the regional and national scales investigations of crop yield are made from the use of remote sensing data and climate data. The study was made to attempt the national level yield prediction of wheat of the years 2002-2010 in 1582 counties of China, by using the nine different variables of the climate, and three different machine learning models (RF, SVM, and LASSO). The result of the study was diverse, as Water-related and Temperature variables outperformed in the yield prediction. Random forest regression performed better with an R^2 of 0.79. the study demonstrated the effectiveness of the integration of data as both climate and remote sensing at the county level will help the researchers and advisors (Zhou et al., 2022).

Remote sensing satellite Landsat 5 TM and crop yield data at plot scale have been used for the early crop prediction (Maize) of the Russel Ranch Sustainable Agriculture Facility (RRSAF) near the University of California, Davis campus from 1994 to 2007. Other multi-sources including, soil data, monthly climate data, vegetation indices (VI) and fertilizers data were also used in the machine learning (ML) models to check the accuracy of

different inputs in crop yield prediction. Incorporating all of the datasets the results showed that Random Forest (RF) model and Adaptive Boosting (AD) model achieved the best accuracy of (R^2 : 0.85, 0.98). Besides that, the combination of the climate data, VI and soil data can predict crop yield more efficiently than the other combinations (Meng et al., 2021). To meet the challenges of increasing population, climate change and increasing food demand, accurate, reliable, and timely crop yield estimation at a large scale is immediately needed. The study was done on wheat yield estimation in thirteen provinces, in China, in which comparisons were made between the traditional machine learning models and three deep learning models. Satellite data, climate data, spatial information data, and soil properties were acquired from the publicly available data within the Google Earth Engine (GEE). Random Forest Regression (RFR) model from Machine Learning and Deep Neural Networks (DNN), 1D Convolutional Neural Networks (1D-CNN), and Long Short-term Memory Networks (LSTM) were used to predict the wheat crop yield estimation. The result of the model comparison showed that the performance of RF and DNN at the field level was relatively good. The study findings demonstrated a simple inexpensive, and scalable framework at various scales of crop estimation which is important for agriculture disaster monitoring, yield forecasting of crops, food security warnings, and food trade policy (Cao et al., 2021). Yield maps play an important role in guiding precision agriculture as it provides the necessary information. Many yield predictions studied were done by the researcher for different crops like corn, maize, rice, and wheat but not for sugarcane. The study was conducted to develop the yield estimation model, by integrating time-series images and machine learning models for sugarcane in the small areas comprises of four fields in Sao Paulo, Brazil. Sentinel-2 images were downloaded from the two consecutive cropping periods and used in the Random Forest (RF) and Multiple Linear Regression (MLR) models to generate the yield maps. Filtered original data was interpolated with the orbital images with the same spatial resolution. Before the execution of the machine learning models, the entire dataset was divided into two datasets as testing and training datasets. At the thriving stage of the crop, the near-infrared spectral band show a greater contribution in the prediction of sugarcane yield as compared to the derived spectral vegetation indices. The study results showed that the RF regression based on multiple spectral bands models shows performance was better than the MLR, with R^2 of 0.70 and Root Mean Square Error (RMSE) of 4.63 Mg/ha for the testing datasets (Canata et al., 2021).

Wheat is one of the main crops and its early yield prediction is very important for national food security and regional trade. For its increasing concerns that how to integrate

machine learning techniques and multi-sources data to establish an accurate, simple and timely model for crop yield estimation at an administrative unit. Much main focus was to use the whole growing period of the crop through remote sensing, climate data and expensive manual surveys. The study was done using only the effective different time windows of yield, which separates the whole growth period into four windows of wheat crop in China. Modelling frameworks were developed to integrate the remote sensing data with soil data, and climate data based on the Google Earth Engine (GEE) platform to predict the winter wheat yield. The study results showed that the models can predict the accurate yield with an error of less than 10% and the $R^2 > 0.75$ before 1-2 months of the harvesting of the wheat at the county level. Different Machine Learning (ML) models are used for the study area, Gaussian Process Regression (GPR), Support Vector Machine (SPV), and Random Forest (RF) methods performed best for the prediction of yield. Research work aimed to highlight a potentially powerful tool that helps in the prediction of crop yield using multi-source data and machine learning in other regions (Han et al., 2020). New technologies enable to analysis and synthesis the big data, and that accurately predicts crop yield. As compared to the typical simulation of crop modelling, Machine Learning provides faster and more reasonable yield prediction. The study was carried out to forecast the corn yield in the US Corn Belt states (Illinois, Indiana, and Iowa) at the county level scale. For the designing of the machine learning framework data used for the forecast of the corn yield were weather data, soil data, yield, and management data. Spatial and temporal correlations were checked between the yield and other data. Machine Learning (ML) models such as Multiple linear regression (MLR) and Random Forest (RF) achieved the result of 9.5% and 9.2% respectively. The study findings suggest that the weather data used in the forecasting of corn yield has also a very important feature (Shahhosseini et al., 2020).

In the previous studies researchers either used satellite data or climate data or usually a combination of both used in the building of the empirical models for the various crop prediction. However, the empirical model performance in the yield prediction was improved by feeding the climate and satellite data but the contribution is still not clear from the data sources. Similarly, the comparison of the performances between the machine learning models and regression-based models in yield prediction is still unclear and needs in-depth investigations. In the study, wheat yield predictions were made by integrating various sources of data from the years 2000 to 2014, at the statistical division (SD) level in Australia. Wheat is the staple and most important growing crop in Australia, and wheat is the top exporting product of Australia globally. The most well-known regression method Least Absolute

Shrinkage and Selection (LASSO) was adopted and three main machine learning methods as Random forest (RF), Support Vector Machine (SVM), and Neural Network (NN) were used for yield prediction and to build various empirical models. Approximate crop productivity was drawn from satellite data as Enhanced Vegetation Indices from the MODIS and Solar-induced Chlorophyll Fluorescence (SIF) from GOME-2. The results of the study confirm that the combination of both data as satellite and climate data gives the high performance of $R^2=0.75$ in yield prediction at the SD level. The machine learning-based models perform better in yield prediction than the regression methods. Crop growing conditions were tracked from the satellite data and it also gradually captures the variability of yield evolving during the growing season and usually at the peak of the growing season satellite data contributes to the yield prediction. The addition of climate variables to the empirical models shows that it exists over the whole season not only at certain stages, as Climate data provide unique and extra information for yield prediction. The study also finds that using the vegetation indices information as input achieves better performance in yield prediction. EVI gives good results in yield prediction than satellite-based SIF, due to course resolution both in time and space it consists of large noise. The study has the best results as it explored the potential for the optimal prediction of two-month-ahead wheat yield prediction in the study area (Cai et al., 2019). The study emphasizes the use of freely available satellite data of Sentinel-2 for the wheat yield estimation within a field of one year, in two different regions of 28 fields in the UK. Environmental data such as topographical and metrological data and soil data were combined with the Sentenil-2 data of different periods of the growing season. Using different combinations of input data helped in exploring the impact of data availability and resolution on yield estimation. The machine learning model Random Forest (RF) was used for data training and validation in yield estimation. Over 8000 points of data were collected from the 29 wheat fields with the help of a combine harvester yield monitor. The study results showed that it is possible to produce an accurate yield map with the field at a 10-meter resolution using the sentinel-2 data (Hunt et al., 2019).

Satellite remote sensing using optical sensors is used to study the growth of plants and to calculate different vegetation indices. Assessing crop yield using vegetation indices is the key significance of the satellite data. The study was conducted in Northern Italy in an area of 11.7 ha under the Mediterranean climate to explore the spatial relationship and variability between the grain yield and six remotely sensed vegetation indices (VI). Freely available data from Landsat 5, Landsat 7, and Landsat 8 images were used. All of the images were downloaded, during the crop growing period and six vegetation indices were extracted as the

Normalized difference vegetation index (NDVI), Enhanced Vegetation Index (EVI), Soil Adjusted Vegetation Index (SAVI), Green Normalized difference vegetation index (GNDVI), Green chlorophyll Index (GCI), and Sample Ratio (SR). Through the statistical analysis, different crops surveyed as wheat (2010), coriander (2013), sunflower (2011) and bread wheat (2012 and 2014), their geo-referenced grain yield and vegetation indices were used to generate the spatial trend maps across the experimental field. A correlation was performed between the grain yield and vegetation indices at the 30-meter spatial resolution. At the crop stages, the best-given correlation period was used for the grain yield prediction. The results of the study showed that the vegetation indices, SR, EVI, and NDVI give a high correlation with the respective crop as compared to SAVI, GNDVI, and GCI. Landsat imagery has proved a good potential for estimating the final grain yield with its spatial and temporal resolution over different crops in a rotation of a small field (Ali et al., 2019).

The study examined the maize yield prediction by using the time series of the NDVI vegetation index extracted from the Landsat-8 OLI imagery of Iran. Advance Machine Learning approaches such as Support Vector Regression (SVR), Boosted Vector Regression (BVR), Random Forest Regression (RFR), and Gaussian Process Regression (GPR) was used and their performances were compared. In the evaluation of their performances, RFR showed a higher R^2 of 0.87 and outperformed the other machine learning models. The demonstration of the study showed that RFR was the most stable model for the prediction of 2015 maize yield, by using it as it was trained and tested for the previous year's data (Aghighi et al., 2018). Timely acquisition of yield maps and crop yield estimation of high quality and low cost are required for the adaptation of precision agriculture. In the comparison of conventional approaches, the integration of machine learning models and remotely sensed data offers cost and time-effective approaches for crop yield prediction and soil properties. The study was conducted to evaluate the role of remote sensing data and comparison of the performances of the machine learning algorithms for the yield prediction of corn in Madison County, Ohio, USA. Soil properties, yield data, topographic data, and multispectral images were used to derive vegetation indices. Five machine learning models as Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF), Gradient Boosting Model (GBM), Cubist (CU), and Multiple Linear Regression (MLR) are trained and tested on the datasets for the yield prediction of corn of the study area. The accuracy and results were based on the Root Mean Square Error (RMSE) and Root Square (R^2). The output of the machine learning models showed that the results outperformed Multiple Linear Regression. Between the performances of the machine learning algorithms Random forest

(RF) performance best with the highest accuracy of $R^2 = 0.53$ and $RMSE = 0.97$. The study outcomes can be implemented in site-specific farming (Khanal et al., 2018).

Studies mentioned above used different techniques and sets of input features to predict yield on different scales. This study examined the feasibility of vegetation indices such as NDVI and EVI for crop yield estimations in the study area. Vegetation indices and surface reflectance data were obtained from GEE and pre-processed to match the study locations. Two machine learning approaches were used Random Forest Regression (RFR) and Decision Tree Regression (DTR) for the yield prediction and comparisons were made for better performance between the models. The best models were tested on data of growing seasons of both year and final training and testing results were derived and exported.

MATERIAL AND METHODS

3.1 Study Area

The Faisalabad division is an administrative division of Punjab and consists of four districts including Faisalabad, Jhang, Toba Tek Singh and Chiniot (figure 1). The geographical location of the district Faisalabad division lies between the longitude of 71° and 73° East, latitude 30° and 31.5° north. According to the 2017 census, the population of district Faisalabad is 7.87 million, Jhang is 2.74 million, Toba Tek Singh is 2.19 million and Chiniot is 1.37 million respectively. The study area receives water from river Ravi from the eastern and southern parts as it touches the district Toba Tek Singh, and the district Chiniot boundaries touch the bank of river Chenab. The northwest part of the district Jhelum receives the river Jhelum and flows along the southern-west part of the district. The Faisalabad district is located in the centre of the Rachna Doab, which lies between the two rivers as Ravi and Chenab. The overall study area topography consists of the local depressions, alluvial plains consisting of rocky sandstone and slate, high grounds, low land, valleys, and some parts receiving a semi- desert area of Thal. For irrigation, the study area utilizes a canal system originating from the nearby headworks as Trimmu headworks but some parts are also irrigated through tube wells. The Faisalabad division is famous for its good amount of wheat production because of its suitable soil, temperature, suitable pattern of rainfall and good irrigation system. The Faisalabad division contributes 45% of wheat production in the total national wheat production. The study area consists of 715.47 thousand hectares of cultivation land for wheat crops producing 2329.60 thousand tons of wheat yield per annum. The climate of the study area is semi-arid with very hot and humid summers starting from mid-April and last till late October and dry cool winters from November up to February. The temperature of the study area goes higher in summer and low in winter. The average maximum temperature of the study area is between 45.5 °C and 26.9 °C. June is the hottest month of the summer. The average minimum temperature is 19.4 °C and 4.1 °C. January is the coolest month of the winter month. The average rainfall received in monsoon in the month of the July and August is approximately 375 mm annually.

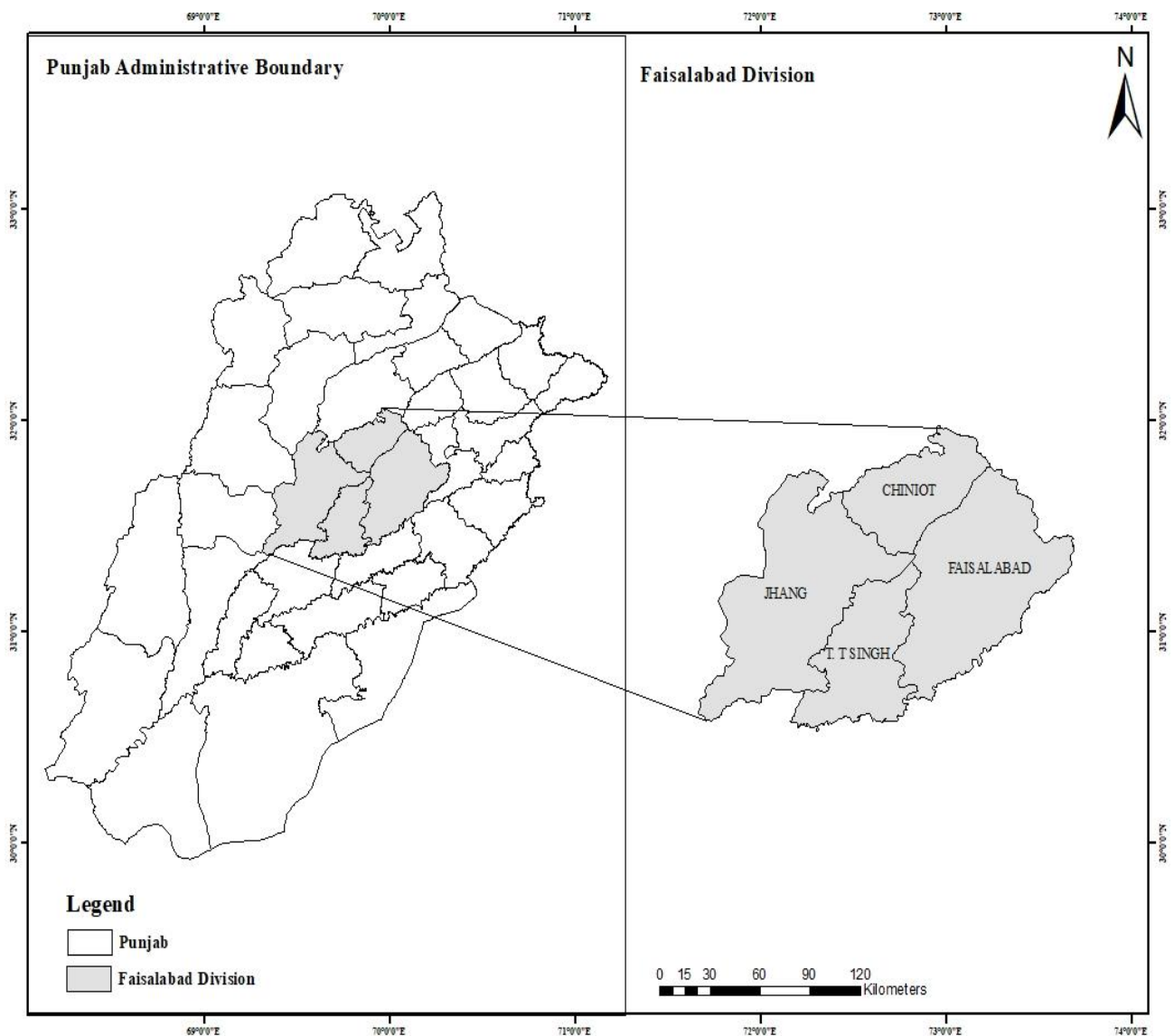


Figure 3.1. Study area map showing (a) Punjab administrative boundaries, (b) Faisalabad division in the Punjab province along with the selected study the Faisalabad division consists of four districts, Chiniot, Jhang, Toba Tek Singh, and Faisalabad.

3.2 Types of Crops in Faisalabad Division

In the Faisalabad division, mixed cropping pattern is followed. Major crops of the area are Cotton, Wheat, Maize, Grain, and other cash crops are grown. Rice and Sugarcane are mainly dominated in the riverine areas. In open fields and tunnels, a wide range of vegetables are grown in the districts. In orchards, Mangoes, Guava, and Citrus are major orchards. In the whole division, a wide range of fodder crops is grown to feed the cattle. The Faisalabad division produced a huge quantity of wheat and contributes about 45% of wheat to the national production. The area of cultivation for wheat in the whole division including four districts is 715.47 thousand hectares of which approximately 2329.60 thousand tons of wheat were produced annually (Ahmad et al., 2019).

3.3 The Environmental and Geographical Characteristics of the Faisalabad Division

In the Faisalabad division, the district Faisalabad is located at the centre of the lower Rechna Doab, the area is situated between the Ravi and Chenab rivers. The topography of the district of Faisalabad is marked by local depression, high grounds and valleys. The district Chiniot is located on the bank of the river Chenab. The topography consists of alluvial plains, spread with rocky sandstone and slate. The river Ravi runs along the southern part and southeastern borders of the district Toba Tek Singh. The topography consists of a large area of the lowlands. Usually, floods are generated by the river Ravi at the border. The fourth district of the Faisalabad division is Jhang, which received the river Chenab from the northeastern part and flows toward the southwest.

The river Jhelum enters in the northwest part and flows in the north-south direction. Finally, the joining point of both river Jhelum and river Chenab is Athara Hazari's north-trimmed headworks originating from the Trimmu headworks. The Chenab district has a variety of topography as it has a low-lying area along the Chenab and Jhelum rivers and a semi-desert area of Thal that is located at the west of the Chenab and Jhelum rivers (Ahmad et al., 2019).

The climate of the Faisalabad division is semi-arid with dry cool winters and very hot humid summers (Chaudhry and Rasul., 2004). The summer season starts in mid-April and continues until late October. The hottest months are June and May. At the end of June, July, August, and mid of September the climate becomes hot and humid, but when it rains the temperature comes down. In June dust storms are very common and as it is the hottest month the climate conditions are very hot. The winter season starts in November and early February. January is the coldest month with extremely dense fog in the early morning and

night hours (Shamshad., 1988). Springs start in the last February and March. The annual average rainfall is approximately 375 mm, which usually takes place during the monsoon season in July and August. The average maximum temperature in June is up to 45.5 °C and 26.9 °C and the temperature has downed a minimum of 19.4 °C and 4.1 °C in January (Pakistan meteorological department., 2013). The Faisalabad division received good fertile soil but some of the areas are affected by water logging and salinity.

3.4 Wheat Crop Production

In Pakistan “Atta” or Wheat flour is a very common food and supplies 72% of caloric energy in the average diet. Wheat is valued as the main nutritional food cereal crop in Pakistan than the other food and cereals. The wheat consumption rate per capita in Pakistan is estimated as 124 kg per year, which is the highest quantity in the world, reflecting the importance of wheat in Pakistan (Imran and Noureen., 2021). The total yield production of wheat varies every year due to various weather and climatic variations such as drought and flooding etc., which directly affects the social balance and economy of our country. As wheat is a Rabi crop which is grown in the winter season of October to December and March and May are the harvesting months. In the Faisalabad division total of 717.47 thousand hectares of area are used to cultivate the wheat crop, which produced 2329.60 thousand tons of wheat yield according to the 2019-2020 final report of the crop reporting services Punjab. The Faisalabad division consists of four districts each area has its contribution.

Wheat crop is grown on an area of 230.67 thousand hectares and the yield produced was 793.86 thousand tons in the Faisalabad district, 75.67 thousand hectare area in the Chiniot district and the yield produced was 238.51 tons, 267.09 thousand hectare area in the Jhang district and the yield produced was 799.18 thousand tons and 142.04 thousand hectare area in Toba Tek Singh and 498.05 thousand tons respectively. Up to 98 % of the area is irrigated through the irrigation system originating from the river in the respective districts. Besides the implementation of modern technology, the demand for wheat production is not met at the rate of increasing population.

3.5 Remote Sensing Data Acquisition

Remote sensing data i.e. surface reflectance of Landsat-8 satellite reflectance of years October 2019 to April 2020 and October 2020 to April 2021, was acquired from Google Earth Engine (GEE). Satellite Landsat-8 carries two types of sensors: Operational Land Imager (OLI), and Thermal Infrared Sensor (TIRS). The spectral resolution of the Landsat-8 is of 11 bands ranging from 433 nm to 13800 nm and thermal bands are of range between 10600 nm -12510 nm. The OLI sensor has nine spectral bands and TIRS have two spectral

bands. Landsat-8 imagery has a Spatial resolution of 30 meters in visible, near-infrared (NIR), shortwave infrared (SWIR), and 100 meters in Thermal. The resolution of the panchromatic band is 15 meters. Thermal bands are of range between 10600 nm – 12510 nm. The temporal resolution of the Landsat-8 is of 16 days. The images of the study area acquired of the wheat crop period are of growing till harvesting, with the specific dates of both of the years shown in tables 3.1 and 3.2.

3.6 Google Earth Engine

Google Earth Engine (GEE) is cloud base platform that provides freely preprocessed data of satellite, Vegetation indices, and Land Surface Temperature at a petabyte scale, which can be used for geospatial analysis. The GEE platform was announced in 2010 and the availability of the data was from 2015. The platform accepts the parameter as input in the format of CSV, Shapefiles, TF Record (simple format for storing a sequence of binary records), and Geo TIFF which should be provided as an asset. Java and Python language scripts can be used in the analysis. Data computation and data analysis both can be done at any scale. The GEE platform also provides some machine learning model implementation such as using tensor flow to resolve problems related to regression, classification and statistical analysis in TensorFlow (Gorelick et al., 2017).

3.7 Crop Yield Ground Truth Data

Wheat Crop yield ground data of the whole Faisalabad division was acquired from the Director of Agriculture Crop Reporting Service, Punjab. The data is in the point data taken from the standing wheat field during the survey. The plot cut technique was used which is of 30 * 30 feet area and a random sampling technique was used for the sampling of the wheat crops. All of the data with sources are mentioned below in table 3.3.

3.8 Meteorological Data

Maximum and minimum temperature and Precipitation data of monthly averages of the Faisalabad division is acquired from the Director of Agriculture Crop Reporting Service, Punjab. The metrological data was taken at the same time as a survey of the specific point in the field.

3.9 Software and Library Used

To analyze the data, mapping of the study area, preprocess the data and use machine learning models the tools and software used in the study are listed in table 3.4.

Table 3.1 Landsat 8 acquisition dates for the study area in 2019-20.

Dates of 2019 to 2020	
2019-10-01	2020-02-06
2019-10-17	2020-02-22
2019-11-02	2020-03-09
2019-11-18	2020-03-25
2019-12-04	2020-04-10
2020-01-21	2020-04-26

Table 3.2 Landsat 8 acquisition dates for the study area in 2020-21.

Dates of 2020 to 2021	
2020-10-12	2021-01-16
2020-10-28	2021-02-01
2020-11-13	2021-02-17
2020-11-29	2021-03-05
2020-12-14	2021-04-04
2020-12-31	2021-04-22

Table 3.3. List of datasets with specification and their sources.

SN	Data	Specification	Source
1	Satellite Imagery	Landsat-8 surface reflectance of years October 2019 to April 2020 & October 2020 to April 2021; Spatial resolution 30 m (visible, NIR, SWIR); 100 m (thermal); 15 m (panchromatic). Spectral resolution (11 bands) ranges from 433 nm -2300 nm; thermal bands are 10600 nm – 12510 nm. The temporal resolution of 16 days.	Google Earth Engine (GEE)
2	Wheat Yield (kg)	Crop yield ground truth data	Director of Agriculture Crop Reporting Service, Punjab
3	Meteorological data	Maximum and Minimum Temperature, Relative Humidity, Precipitation (monthly averages)	Director of Agriculture Crop Reporting Service, Punjab

Table 3.4. Tools and software used for the analysis and preprocessing of the data.

SN	Software	Version
1	ArcMap	10.8
2	Google Earth Engine (GEE)	Python version: 3.8.16
3	Jupyter Notebook	Python version 3.8.16
4	Machine learning libraries	Scikit-learn, Matplotlib

3.10 Data Processing

Remote sensing data of Landsat-8 surface reflectance was acquired from GEE for each yield point temporally. A buffer of 15 meters was drawn for each point as the yield collected from the wheat crop field was taken by plot cut method of 30 * 20 feet. The WRS (world reference system, which is a global national system for Landsat data) row path used was 149/38. The Shapefile of the wheat yield points was uploaded to GEE. Data for the whole season of both the years 2019-2020 and 2020 -2021 from October to April was selected for those points (from shapefile) as per the crop calendar of the wheat crop. All of the raw reflectance values from all bands along with surface temperature were downloaded. The surface temperature unit was Kelvin. For the elimination of the records with no data or duplication of the data, data was then filtered for each year for the specified dates. This is because we needed a consistent time series for every point in the shapefile. Since some of the points were covering two tiles, the only selected tile was 149/28 so the consistent time series for each point were acquired. Data for each point was combined temporally and cleaned. The final dataset contains 238 points for both the years 2019 to 2020 and 2020 to 2021, with 16 days of temporal data from October 2019 to April 2020 and October 2020 to April 2021.

3.11 Vegetation Indices

Remotely sensed Landsat-8 surface reflectance data were used throughout the Faisalabad division, of Punjab. Normalized difference vegetation index (NDVI) and Enhanced Vegetation Index (EVI) were used as explanatory variables in crop yield prediction based on correlation with yield. NDVI is used for remote estimation of plant health. It gives a special reflectance curve by the difference between two bands (visible red and near-infrared) ranging from -1 to 1 (Robinson et al., 2017). The mathematical equation to derive NDVI is illustrated in equation 1.

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad \text{Equation 1}$$

Where NIR is the reflectance in a near-infrared band while red is the reflectance in the red band of the electromagnetic spectrum. Enhanced vegetation index (EVI) is used to quantify vegetation greenness and is similar to the NDVI. EVI helps in the correction of canopy background noise, and atmospheric conditions and becomes more sensitive in an area with dense vegetation (Trujillo et al., 2020). The formula for the EVI used is in equation 2.

$$EVI = 2.5 \times \frac{(NIR - RED)}{(NIR + C1 \times RED - C2 \times BLUE + L)} \quad \text{Equation 2}$$

The “L” in the formula adjusts the canopy background, “C” values as coefficients for atmospheric resistance, and “B” from the blue band. As the ratio is the same as NDVI between the NIR and R along with the reducing effects of atmospheric noise, atmospheric noise and saturation in most cases.

3.12 Vegetation Indices Calculation

From the Landsat-8 surface reflectance vegetation indices NDVI and EVI were calculated. Time series graphs were formed with the phenological stages of the wheat crop of both the years 2019- 2020 and 2020- 2021 as shown in Figure 3.2 respectively.

3.13 Feature Selection

Based on correlation analysis, suitable features were selected. In machine learning, it is not always suitable to feed all the features to the model. This is because models trained with all features are prone to overfitting which makes the model generalizing abilities very poor and results in very poor performance on testing data. Furthermore, collecting data on a very large number of variables is not always feasible and is time-consuming and cost-expensive. In the case of remote sensing, very dense time-series data is sometimes not available. In such cases, feature selection becomes very important. In this study, a correlation-based feature selection technique was adopted. In this technique, the correlation of the dependent variable is calculated with several explanatory variables and based on a predefined threshold of correlation only those features are selected which fulfil the criteria of the correlation threshold.

3.14 Machine Learning Model’s Implementation

Machine learning models were used for the crop yield estimation by analyzing the data from various sources such as satellite imagery data, weather data and historical yield data of the Faisalabad division. The choice of machine learning model is very important for better results. There are several machine learning models which can be utilized for tasks such as classification and regression such as Random Forest Regression, Decision Trees, and Neural Network models. Since wheat yield prediction is a regression problem, several machine learning models were initially tested on the data and only the two best models were selected based on initial results. Those models were random forest regression (RFR) and decision tree regression (DTR). Both models and their details are explained in the subsequent section.

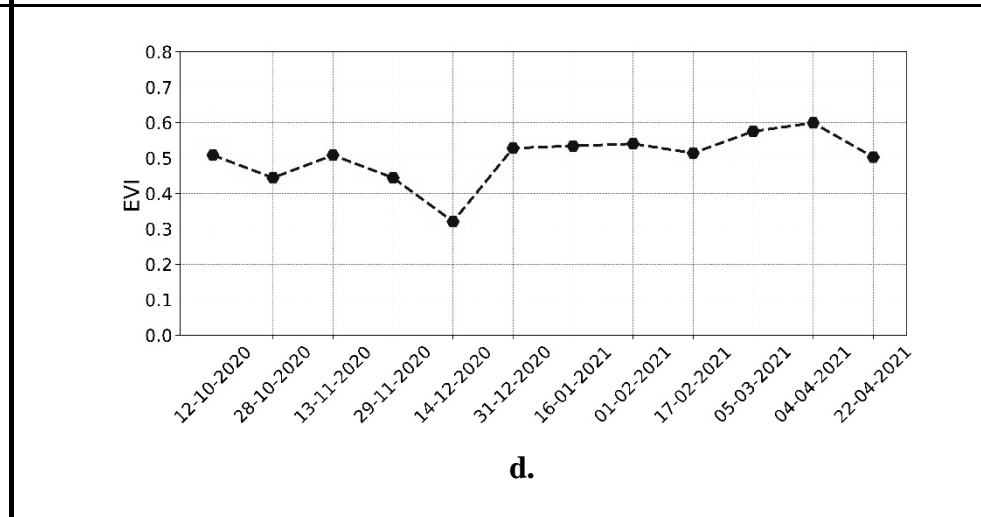
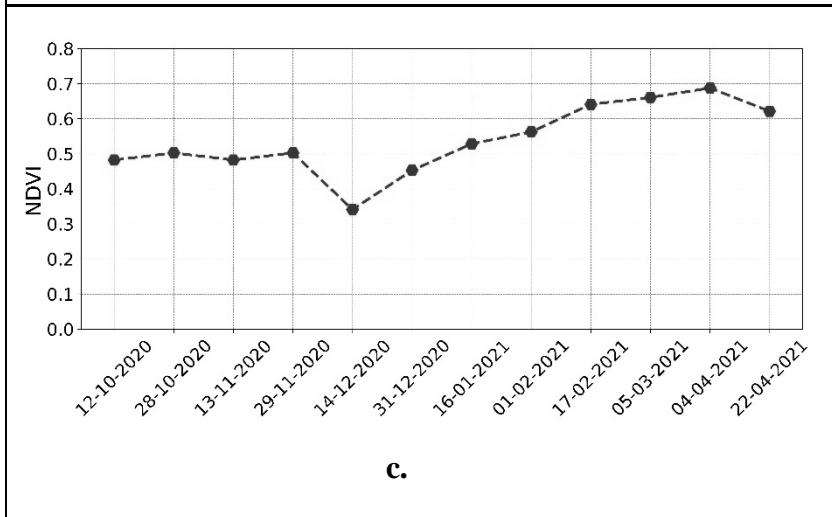
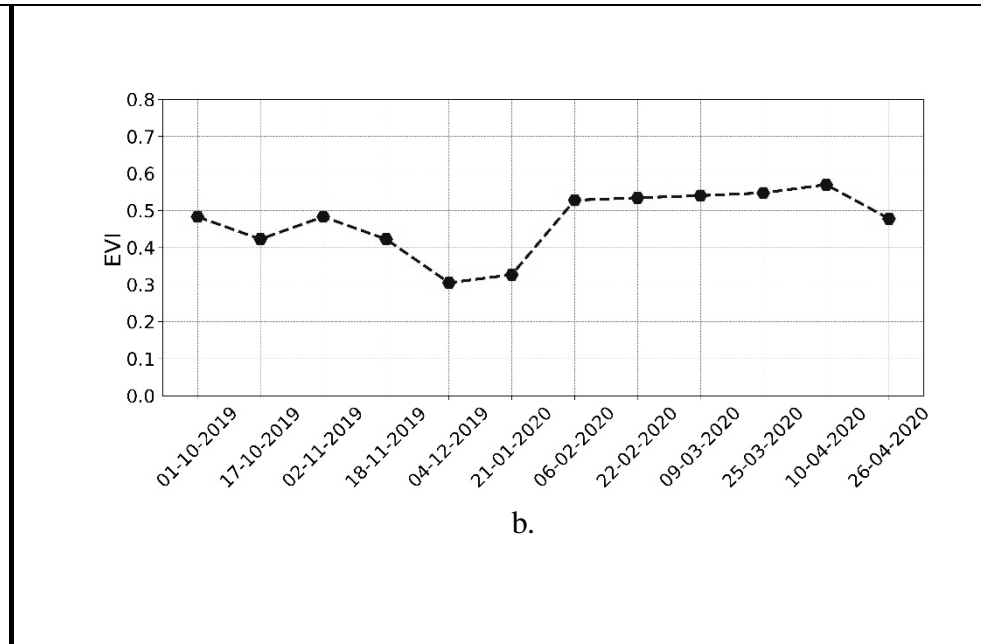
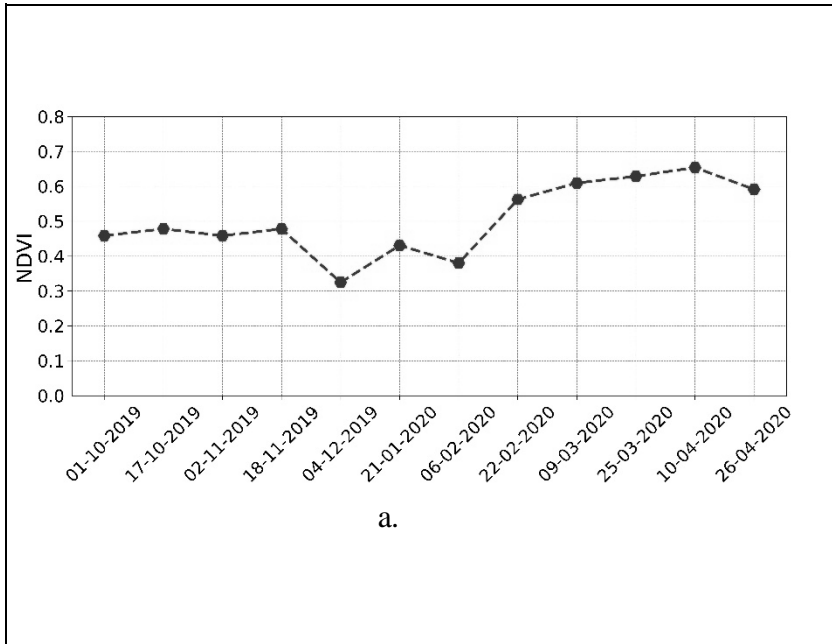


Figure 3.2. NDVI time-series & phenological stages (2019-20) (a), EVI time-series & phenological stages (2019-20) (b), NDVI time series and phenological stages (2020-21) (c), EVI time-series & phenological stages (2020-21) (d).

3.15 The Random Forest Regression (RFR) Model

The random forest regression model is a machine learning algorithm which was used for the crop yield estimation of the Faisalabad division. RFR is a supervised learning algorithm and model that uses an ensemble of decision trees to make decision predictions. In the case of crop yield estimation, firstly the model got trained on a dataset using a library such as sci-kit-learn in Python that consists of subsets variables of 8th-time steps (such as temperature, humidity, vegetation indices such as NDVI and EVI) and corresponding crop yield values. In this model, the algorithm builds multiple decision trees by selecting subsets of the features randomly and data points from the training set. In the RFR each decision tree predicts crop yield independently for the particular set of input features. The final prediction makes based on the average predictions of all the individual decision trees.

3.16 The Decision Tree Regression (DTR) Model

Decision tree regression is also a machine learning model used in this study for the prediction of wheat yield. The model used the relationship between the set of input features (temperature, humidity, vegetation indices, and vegetation indices NDVI and EVI) and a continuous target variable such as crop yield. It works by partitioning the input feature space into smaller and smaller subsets and assigning a prediction to each subset based on the average value of the target variable in that subset.

3.17 Experimental Setup

The step-by-step process of machine learning model building for crop yield prediction adapted in this study is explained as follows.

- The first step is the data collection process. It was done by collecting the data variables such as temperature and humidity data, and vegetation indices data (NDVI and EVI) from remote sensing data. All of the above data and input variables for the ML models.
- The second step is data preprocessing which is done by data cleaning, and time steps series of data were formed. The correlation of the data variables was performed for the most relevant feature selection between the yield and the meteorological data and vegetation indices, for predicting the crop yield from the available data sources.
- In the third step, the data is split into parts, training data and testing data. The training data was used to build the ML model, while the testing data was used for the evaluation of the ML models.
- In the model building the fourth step involves defining the model hyperparameter determination. The hyperparameters of Random Forest Regression are the number of trees in the forest, the number of features to consider at each split and the maximum

depth of each tree. In the case of Decision Tree Regression, the hyperparameters are the maximum tree depth, minimum samplers per leaf and the splitting criterion.

- The fifth step is the training of the model. Training of the models allows the models to learn how to predict crop yields based on the input data. In the case of Random Forest Regression, the algorithm would create a large number of decision trees, each of which would be trained on a different subset of the training data. Each decision tree used a random subset of the variable feature to make predictions about crop yield. This approach helps to reduce the risk of model overfitting. Decision Tree Regression involves partitioning the data space based on the selected features and assigning a prediction to each partition based on the average value of the crop yield.
- Evaluation of the model means the accuracy of the model's performance of the trained model on the testing data. For the present study, we used the Root Mean Squared Error (RMSE), and Coefficient of Determination (R-squared).
- The last steps consist of the testing of the model. Once the model is trained and optimized, the model got ready to make new predictions on new data remaining from the training data. The step involves comparing the predicted wheat crop yield values with the actual crop yield values for the testing data.

3.18 Model Accuracy Assessment

The model's accuracy assessment was done using the R^2 and RMSE. Training and testing R^2 of decision tree regression were 0.952 and 0.700, respectively. Training and testing R^2 of random forest regression were 0.929 and 0.725 for the estimated crop yield for the years 2019-20. Training and testing R^2 of decision tree regression were 0.952 and 0.799 respectively. Training and testing R^2 of random forest regression were 0.930 and 0.842 for the estimated crop yield for the year 2020-21.

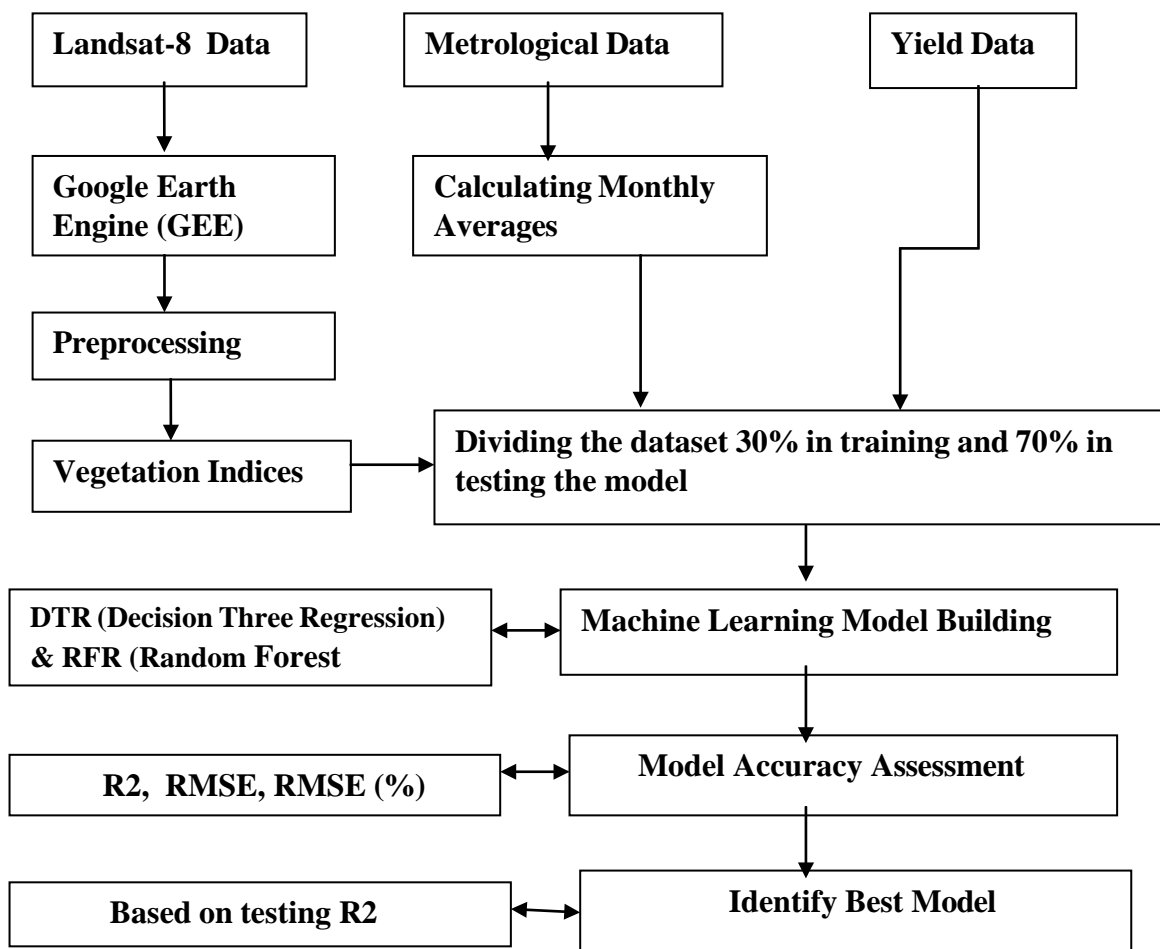


Figure 3.3. Complete flow chart methodology of the study area.

RESULTS AND DISCUSSION

4.1 Relationship of Metrological Variables with Yield

Performing correlation analysis involving NDVI, EVI, temperature, humidity, and wheat yield can help identify the extent to which these variables are related and provide insights into the potential impact they may have on wheat yield. NDVI is commonly used as an indicator of vegetation health and biomass production. Higher NDVI values generally indicate healthier and more productive vegetation, including crops like wheat. A positive correlation between NDVI and wheat yield suggests that as NDVI values increase, wheat yield is likely to increase as well. The correlation analysis of yield with time-series NDVI in the growing season of 2019-20 (table 4.1) and the season 2020-21 (table 4.2). The significance of each correlation coefficient was presented through several asterisks. The significance of the correlation coefficient was taken as (P-value \leq 0.01) high significance, (P-value \leq 0.05) moderate significance, and (P-value \leq 0.1) shows low significance. The 8th time series of NDVI was showing a high correlation with yield i.e. 0.602 (table 4.1) and 0.6 (table 4.2). The lowest correlation of yield was noted with the 7th and 11th timestep which is -0.07 and 0.07 respectively, for the year 2019-20 (table 4.1). For the year 2020-21, the lowest correlation of yield is with the 6th and 7th timestep which is -0.06 and 0.06 (table 4.2) respectively. EVI is another vegetation index that considers atmospheric influences and enhances the sensitivity to vegetation changes. Similar to NDVI, a positive correlation between EVI and wheat yield suggests that higher EVI values correspond to higher wheat yields. The correlation analysis of yield with time-series EVI in the growing season of 2019- 20 with a high correlation with yield in the 8th timestep was showing as 0.6 (table 4.3) and the lowest correlation with yield in the 5th and 12th timestep is -0.03 and 0.03 (table 4.3). For the season 2020-21, a high correlation with yield in the 8th timestep is 0.602 (table 4.4) and the lowest correlation of the 6th timestep is -0.01 (table 4.4) respectively.

Temperature plays a crucial role in determining crop growth and development. However, the relationship between temperature and wheat yield is not necessarily straightforward. Wheat has specific temperature requirements at different growth stages, and extreme temperatures (both high and low) can negatively impact yield. In general, an optimal temperature range during the growing season is associated with higher wheat yields. The correlation between temperature and wheat yield can vary depending on the specific temperature conditions experienced by the crop. Yield Correlation with time series T2MAX (Maximum temperature at 2 meters) of the years 2019-20 (table-1) and of the years 2020-21

(table -2). Overall the T2MAX shows weak relation however, we selected the 8th time series as it is comparatively better and also aligned temporally with vegetation indices as 0.22 (table-1) and the 10th time series as it is comparatively better with 0.20 (table-2). In the feature selection, the T2MIN was not considered overall as it has shown a weak correlation with yield for both of the years. Humidity, or relative humidity, refers to the amount of moisture present in the air. Wheat requires adequate moisture for optimal growth and yield. High humidity can create favourable conditions for diseases, such as fungal infections, which may negatively impact wheat yield. On the other hand, low humidity can lead to increased evaporation and water stress, also affecting yield. The relationship between humidity and wheat yield can be complex and dependent on other factors like precipitation patterns. Yield Correlation with time series RH2M (Relative humidity at 2 meters) for the year 2019-20 (table-3) and for the year 2002-21 (table -4). RH2M shows a weak relation, however, selected the 8th time series as it is comparatively better and aligned temporally with vegetation indices as 0.16 (table-3) and the 9th time series it still has some correlation is -0.25 (table-4).

The obtained results agree with the results obtained by previous studies (Li et al., 2019). NDVI and EVI data have been used for yield predictions very widely and different machine-learning models have been compared for yield predictions (Shammi & Meng, 2021; Sharifi, 2021). One of the main advantages of using NDVI and EVI for yield predictions is its ability to highlight information related to crop health and its response to growing conditions. This makes the vegetation indices such as NDVI and EVI very useful for yield predictions (Bannari et al., 1995). Although the vegetation indices perform well for yield prediction due to their ability to model the crop conditions; they are not a direct indicator of crop health status. Since vegetation indices are solely based on reflectance, they can sometimes be misleading and need to be used carefully. Furthermore, the usage of meteorological information as we have seen in this case can complement the other datasets and provides better performance for yield prediction (Schwalbert et al., 2020).

4.2 Subset Variables Selected from the Correlation Analysis

After the correlation analysis of yield data with all possible variables, the variables were selected based on the correlation coefficient value. Variables having a high correlation with yield mostly existed in the later stages of the crop which is mostly near to three maturities such as the 8th times step according to the correlation analysis. A subset of variables of both years (2019-20) and (2020-21) according to high correlation coefficient was chosen (table 4.5 & 4.6) with their significance levels. The variable minimum temperature at 2 meters (T2MIN) was excluded from the analysis based on correlation coefficient value and the 8th-time step of all other variables was selected for the year 2019-20. The p-value is $P <$

0.01, $P = 0.05$, and $P = 0.1$, where 8 at the end of every variable shows that the variable selected is from the 8th time step. ST_B10 is the surface temperature obtained from remote sensing data, NDVI is the normalized difference vegetation index, EVI is the enhanced vegetation index, T2MAX is the maximum temperature at 2 meters and RH2M is relative humidity at 2 meters.

4.3 Models Training Results

Machine learning model: (1) decision tree regression (DTR), and (2) random forest regression (RFR) were trained on the selected variables. The training results of both machine learning models (figure 4.1) for the crop year 2019-20 and 2020-21 respectively. For the crop year 2019-20, the training R2 for DTR and RFR was 0.952 and 0.929 respectively. The RMSE of both models was 0.062 and 0.076 respectively. From the training results, it is very clear that the DTR model is overfitting the training data. This is very common in machine learning models is overfitting the training data. This has a bad impact on testing results since machine learning models learn the training data too well and are not able to generalize based on that. The training results for the year 2020-21 (figure 4.1). The training results for the year 2019-20 achieved R2 of 0.952 and 0.930. the DTR model again overfits the training data in the year 2020-21. In machine learning, overfitting occurs when a model learns the training data too well. Learning training data too well impacts the testing performance of machine learning models when the model performs poorly on testing and cross-validation.

4.4 Models Testing Results

Decision tree regression (DTR), and random forest regression (RFR) were tested on the variables that remained from the training. The testing results of both machine learning models (figure 4.2) for crop years 2019-20 and 2020-21 respectively. For the year 2019-20, the testing R2 for DTR and RFR was 0.700 and 0.725 respectively. The RMSE of both of the models was 0.150 and 0.144. The testing result for the year 2020- 21, the testing result of DTR and RFR model R2 is 0.799 and 0.842 and RMSE is 0.120 and 0.106 respectively. From the testing result it's clear that the RFR model R2 is better than DTR with less RMSE error.

4.5 Analysis of Models

The training and testing results of machine learning models presented above indicate the performance measures of models for crop yield prediction. The RFR model outperformed the DTR model for wheat yield prediction. The DTR model overfitted the training data and the generalizing power was very poor for yield prediction which is why it achieved a perfect training R2 but lower testing R2 for both year testing results, the RFR model achieved better testing performance. This is because the RFR model is based on several decision tree models

and combines the output from several decision tree models. In classification problems, RFR takes a majority vote while in regression problems it averages the output of several decision tree models which is why it performs better than single DTR models. This is why RFR is also sometimes called assembled learning model. The overfitting issue of DTR is consistent with other research specifically when the number of samples is not very large (Kotsiantis, 2013). Other than DTR, the issue of overfitting with small samples is an overall problem in the machine learning community as machine learning models require increasing amounts of data for training purposes. The performance of the DTR model was also consistent with other studies in the context of crop yield prediction where the RFR model outperformed the DTR model (Khan et al., 2022). The results of the RFR model were also in agreement with other studies where it outperformed other machine learning models. This can be attributed to the working of RFR where instead of one decision tree it trains hundreds of decision trees and averages the outcome of all trees thus avoiding overfitting and improving the model performance. Although the overall performance of RFR was comparatively better, the models need to be tested in more diverse environments and different conditions to check their robustness.

4.6 Identify the Best Model

By comparing the results and based on the accuracy Random forest Regression performance is better than the Decision tree regression model because it reduces overfitting and captures more complex interaction between the features. However, a random forest model may be slower to train and harder to perform better than a single decision tree model. The random forest regression model has high R2 values and low RMSE in training (RMSE = 0.076, R2 = 0.929, RMSE = 0.075, R2 = 0.930) and testing (RMSE = 0.144, R2=0.725, RMSE = 0.106, R2 = 0.842) for both of the years. As discussed earlier, as compared to DTR, the RFR model performed relatively well due to its ability to use several decision trees instead of single tree fitting. RFR trains hundreds of decision trees instead of a few and then averages the predictions of all the trees to make the final predictions (Smith et al., 2013). This makes RFR more valuable in terms of avoiding overfitting and increasing performance.

Table 4.1. NDVI time series correlation with the yield (kg) 2019-20.

	yield	ndvi_1	ndvi_2	ndvi_3	ndvi_4	ndvi_5	ndvi_6	ndvi_7	ndvi_8	ndvi_9	ndvi_10	ndvi_11	ndvi_12
yield	1	0.198***	-0.102	0.198***	-0.102	-0.107*	-0.059	-0.07	0.602***	0.163**	0.103	0.072	0.108*
ndvi_1	0.198***	1	0.42***	1.0***	0.42***	0.08	0.068	-0.07	0.227***	0.112*	0.098	0.09	0.071
ndvi_2	-0.102	0.42***	1	0.42***	1.0***	0.461***	0.291***	0.045	0.028	0	0.045	0.028	0.02
ndvi_3	0.198***	1.0***	0.42***	1	0.42***	0.08	0.068	-0.07	0.227***	0.112*	0.098	0.09	0.071
ndvi_4	-0.102	0.42***	1.0***	0.42***	1	0.461***	0.291***	0.045	0.028	0	0.045	0.028	0.02
ndvi_5	-0.107*	0.08	0.461***	0.08	0.461***	1	0.306***	0.145**	0.221***	-0.123*	-0.109*	-0.126*	-0.139**
ndvi_6	-0.059	0.068	0.291***	0.068	0.291***	0.306***	1	0.807***	-0.045	-0.154**	-0.19***	-	-0.065
ndvi_7	-0.07	-0.07	0.045	-0.07	0.045	0.145**	0.807***	1	-0.065	-0.114*	-	-	-0.147**
ndvi_8	0.602***	0.227***	0.028	0.227***	0.028	0.221***	-0.045	-0.065	1	0.202***	0.145**	0.087	0.077
ndvi_9	0.163**	0.112*	0	0.112*	0	-0.123*	-0.154**	-0.114*	0.202***	1	0.897***	0.798***	0.618***
ndvi_10	0.103	0.098	0.045	0.098	0.045	-0.109*	-0.19***	-	0.145**	0.897***	1	0.932***	0.741***
ndvi_11	0.072	0.09	0.028	0.09	0.028	-0.126*	-	-	0.087	0.798***	0.932***	1	0.809***
ndvi_12	0.108*	0.071	0.02	0.071	0.02	-0.139**	-0.065	-0.147**	0.077	0.618***	0.741***	0.809***	1

Significance level *: p-value <= 0.1, **: p-value <= 0.05, ***: p-value <= 0.01

Table 4.2. NDVI time series correlation with the yield (kg) 2020-2021.

	yield	ndvi_1	ndvi_2	ndvi_3	ndvi_4	ndvi_5	ndvi_6	ndvi_7	ndvi_8	ndvi_9	ndvi_10	ndvi_11	ndvi_12
yield	1	0.19***	-0.11*	0.19***	-0.11*	-0.11*	-0.06	-0.07	0.6***	0.16**	0.09	0.06	0.1
ndvi_1	0.19***	1	0.42***	1.0***	0.42***	0.08	0.07	-0.06	0.22***	0.12*	0.1	0.09	0.07
ndvi_2	-0.11*	0.42***	1	0.42***	1.0***	0.47***	0.29***	0.05	0.02	0.01	0.05	0.03	0.02
ndvi_3	0.19***	1.0***	0.42***	1	0.42***	0.08	0.07	-0.06	0.22***	0.12*	0.1	0.09	0.07
ndvi_4	-0.11*	0.42***	1.0***	0.42***	1	0.47***	0.29***	0.05	0.02	0.01	0.05	0.03	0.02
ndvi_5	-0.11*	0.08	0.47***	0.08	0.47***	1	0.31***	0.15**	0.22***	-0.12*	-0.11*	-0.12*	-0.14**
ndvi_6	-0.06	0.07	0.29***	0.07	0.29***	0.31***	1	0.81***	-0.04	-0.14**	-0.18***	-0.23***	-0.07
ndvi_7	-0.07	-0.06	0.05	-0.06	0.05	0.15**	0.81***	1	-0.06	-0.1	-0.24***	-0.31***	-0.14**
ndvi_8	0.6***	0.22***	0.02	0.22***	0.02	0.22***	-0.04	-0.06	1	0.2***	0.14**	0.08	0.07
ndvi_9	0.16**	0.12*	0.01	0.12*	0.01	-0.12*	-0.14**	-0.1	0.2***	1	0.9***	0.8***	0.62***
ndvi_10	0.09	0.1	0.05	0.1	0.05	-0.11*	-0.18***	-0.24***	0.14**	0.9***	1	0.93***	0.74***
ndvi_11	0.06	0.09	0.03	0.09	0.03	-0.12*	-0.23***	-0.31***	0.08	0.8***	0.93***	1	0.81***
ndvi_12	0.1	0.07	0.02	0.07	0.02	-0.14**	-0.07	-0.14**	0.07	0.62***	0.74***	0.81***	1

Significance level *, p-value ≤ 0.1 , **, p-value ≤ 0.05 , ***, p-value ≤ 0.01

Table 4.3. EVI time-series correlation with the yield (kg) 2019-2020.

	yield	evi_1	evi_2	evi_3	evi_4	evi_5	evi_6	evi_7	evi_8	evi_9	evi_10	evi_11	evi_12
yield	1	0.15**	-0.05	0.15**	-0.05	-0.03	0	0.03	0.6***	0.21***	0.15**	0.08	0.03
evi_1	0.15**	1	0.65***	1.0***	0.65***	0.33***	0.07	-0.01	0.2***	0.05	0.07	0.07	0
evi_2	-0.05	0.65***	1	0.65***	1.0***	0.52***	0.15**	-0.04	0	-0.07	0.02	0.04	-0.01
evi_3	0.15**	1.0***	0.65***	1	0.65***	0.33***	0.07	-0.01	0.2***	0.05	0.07	0.07	0
evi_4	-0.05	0.65***	1.0***	0.65***	1	0.52***	0.15**	-0.04	0	-0.07	0.02	0.04	-0.01
evi_5	-0.03	0.33***	0.52***	0.33***	0.52***	1	0.3***	0.15**	0.14**	-0.09	-0.07	-0.06	-0.05
evi_6	0	0.07	0.15**	0.07	0.15**	0.3***	1	0.82***	-0.03	-0.03	-0.12*	-0.14**	0.08
evi_7	0.03	-0.01	-0.04	-0.01	-0.04	0.15**	0.82***	1	0.01	0.06	-0.13**	-0.18***	0.1
evi_8	0.6***	0.2***	0	0.2***	0	0.14**	-0.03	0.01	1	0.27***	0.21***	0.13**	0.09
evi_9	0.21***	0.05	-0.07	0.05	-0.07	-0.09	-0.03	0.06	0.27***	1	0.88***	0.77***	0.58***
evi_10	0.15**	0.07	0.02	0.07	0.02	-0.07	-0.12*	-0.13**	0.21***	0.88***	1	0.92***	0.69***
evi_11	0.08	0.07	0.04	0.07	0.04	-0.06	-0.14**	-0.18***	0.13**	0.77***	0.92***	1	0.79***
evi_12	0.03	0	-0.01	0	-0.01	-0.05	0.08	0.1	0.09	0.58***	0.69***	0.79***	1

Significance level *: p-value <= 0.1, **: p-value <= 0.05, ***; p-value <= 0.01

Table 4.4. EVI time-series correlation with the yield (kg) 2020-21.

	yield	evi_1	evi_2	evi_3	evi_4	evi_5	evi_6	evi_7	evi_8	evi_9	evi_10	evi_11	evi_12
yield	1	0.14**	-0.06	0.14**	-0.06	-0.04	-0.01	0.02	0.6***	0.2***	0.14**	0.08	0.02
evi_1	0.14**	1	0.65***	1.0***	0.65***	0.34***	0.07	0	0.2***	0.06	0.08	0.07	0
evi_2	-0.06	0.65***	1	0.65***	1.0***	0.53***	0.15**	-0.03	-0.01	-0.06	0.02	0.04	-0.01
evi_3	0.14**	1.0***	0.65***	1	0.65***	0.34***	0.07	0	0.2***	0.06	0.08	0.07	0
evi_4	-0.06	0.65***	1.0***	0.65***	1	0.53***	0.15**	-0.03	-0.01	-0.06	0.02	0.04	-0.01
evi_5	-0.04	0.34***	0.53***	0.34***	0.53***	1	0.3***	0.16***	0.14**	-0.09	-0.07	-0.06	-0.05
evi_6	-0.01	0.07	0.15**	0.07	0.15**	0.3***	1	0.82***	-0.03	-0.02	-0.12*	-0.14**	0.08
evi_7	0.02	0	-0.03	0	-0.03	0.16***	0.82***	1	0.02	0.07	-0.12*	-0.17***	0.11*
evi_8	0.6***	0.2***	-0.01	0.2***	-0.01	0.14**	-0.03	0.02	1	0.27***	0.21***	0.13**	0.09
evi_9	0.2***	0.06	-0.06	0.06	-0.06	-0.09	-0.02	0.07	0.27***	1	0.89***	0.78***	0.59***
evi_10	0.14**	0.08	0.02	0.08	0.02	-0.07	-0.12*	-0.12*	0.21***	0.89***	1	0.92***	0.69***
evi_11	0.08	0.07	0.04	0.07	0.04	-0.06	-0.14**	-0.17***	0.13**	0.78***	0.92***	1	0.79***
evi_12	0.02	0	-0.01	0	-0.01	-0.05	0.08	0.11*	0.09	0.59***	0.69***	0.79***	1

Significance level *: p-value ≤ 0.1 , **: p-value ≤ 0.05 , ***: p-value ≤ 0.01

Table 4.5 Yield correlation with a subset of variables selected from the correlation analysis (2019-20).

	yield	ST_B10_8	ndvi_8	evi_8	T2MAX_8	RH2M_8
Yield	1	0.62***	0.6***	0.6***	0.22***	-0.16***
ST_B10_8	0.62***	1	0.91***	0.81***	0.07	-0.09
ndvi_8	0.6***	0.91***	1	0.84***	0.08	-0.04
evi_8	0.6***	0.81***	0.84***	1	0.05	-0.06
T2MAX_8	0.22***	0.07	0.08	0.05	1	-0.05
RH2M_8	-0.16***	-0.09	-0.04	-0.06	-0.05	1

Table 4.6 Yield correlation with a subset of variables selected from correlation analysis (2020- 21).

	yield	ST_B10_8	ndvi_8	evi_8	T2MAX_10	RH2M_9
yield	1	0.52***	0.6***	0.6***	0.2***	-0.25***
ST_B10_8	0.52***	1	0.82***	0.71***	0.24***	-0.02
ndvi_8	0.6***	0.82***	1	0.84***	0.1	-0.08
evi_8	0.6***	0.71***	0.84***	1	0.05	-0.1
T2MAX_10	0.2***	0.24***	0.1	0.05	1	-0.04
RH2M_9	-0.25***	-0.02	-0.08	-0.1	-0.04	1

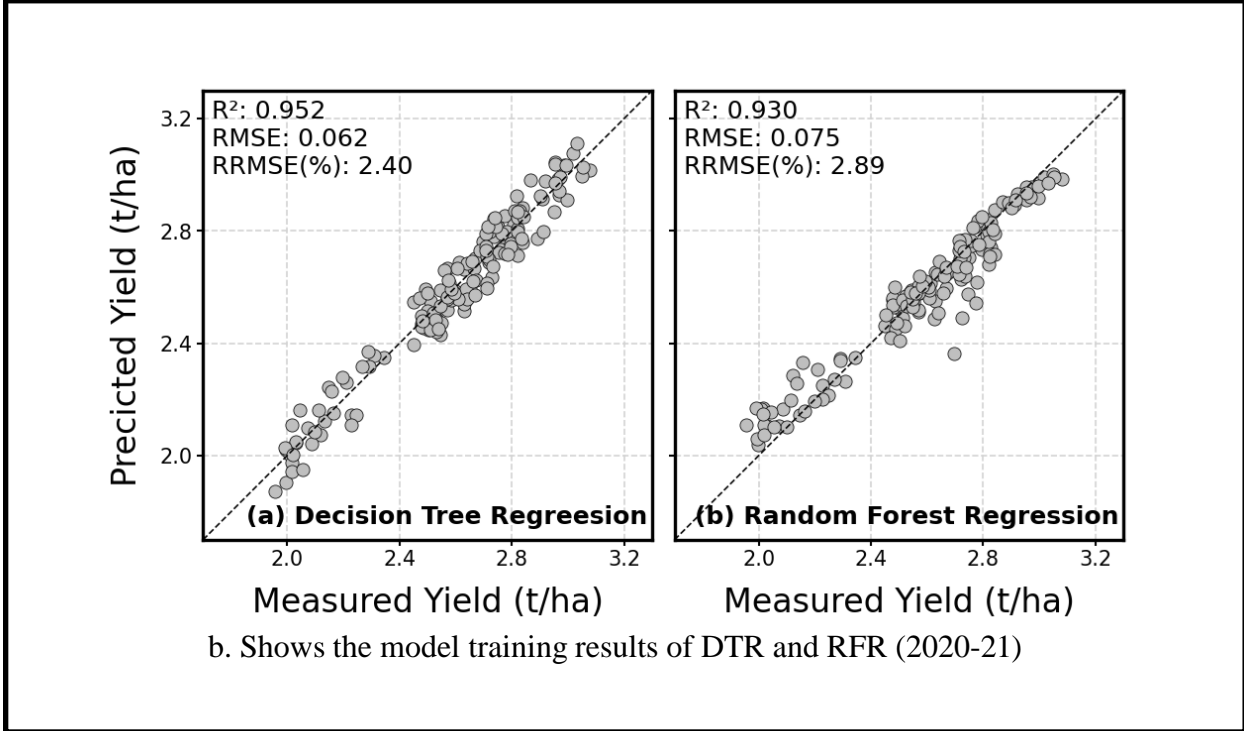
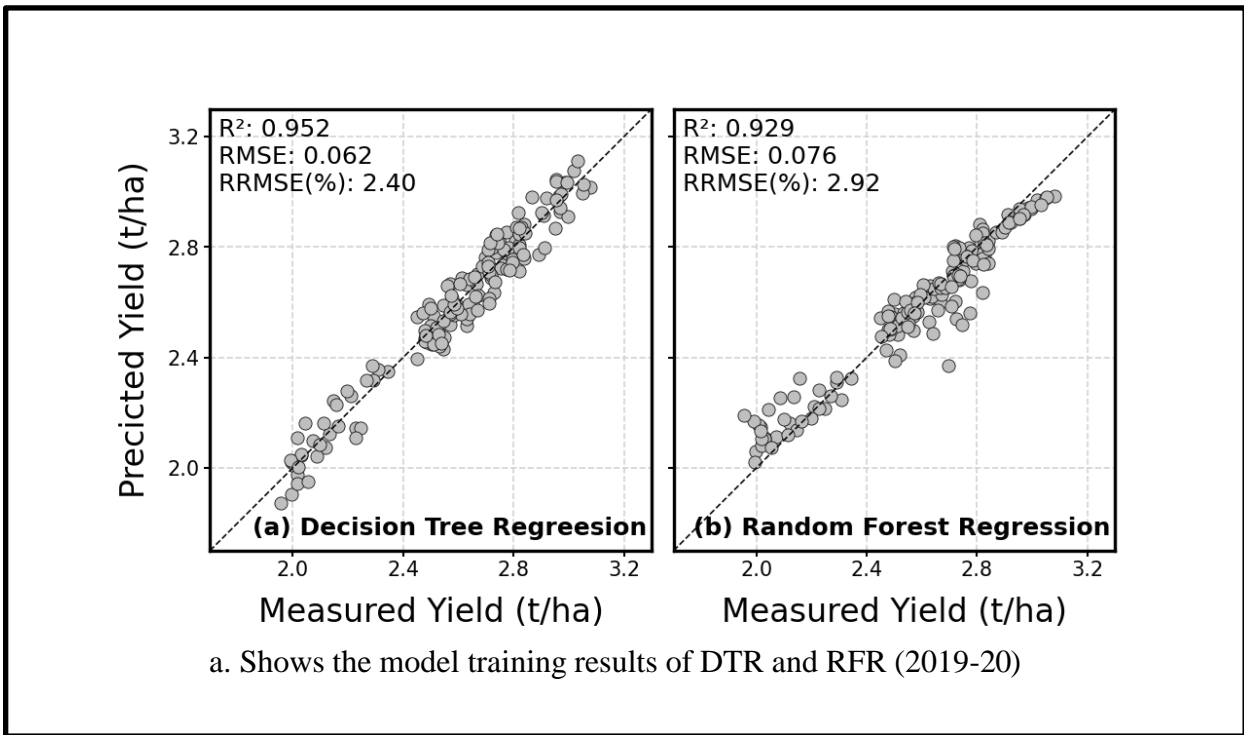
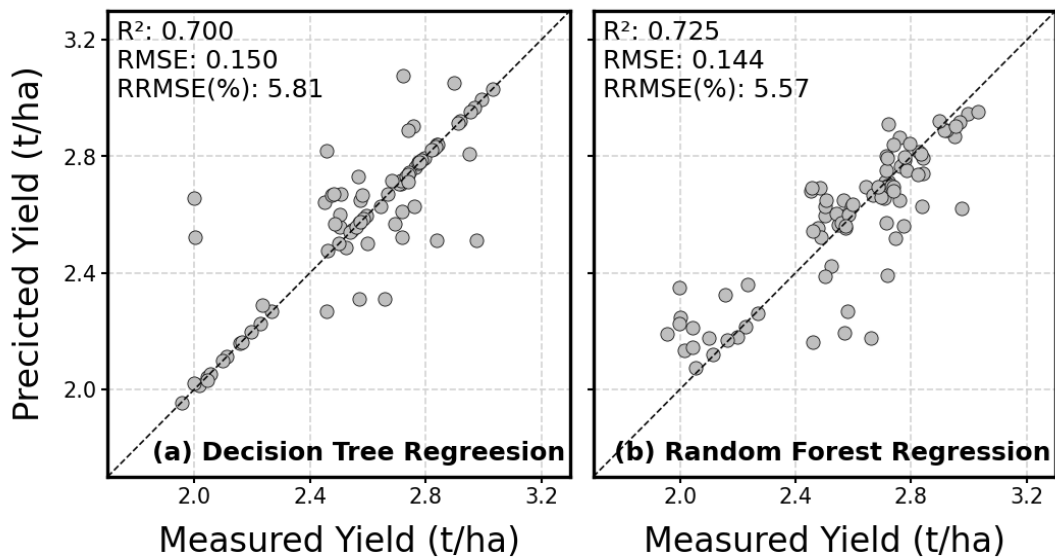
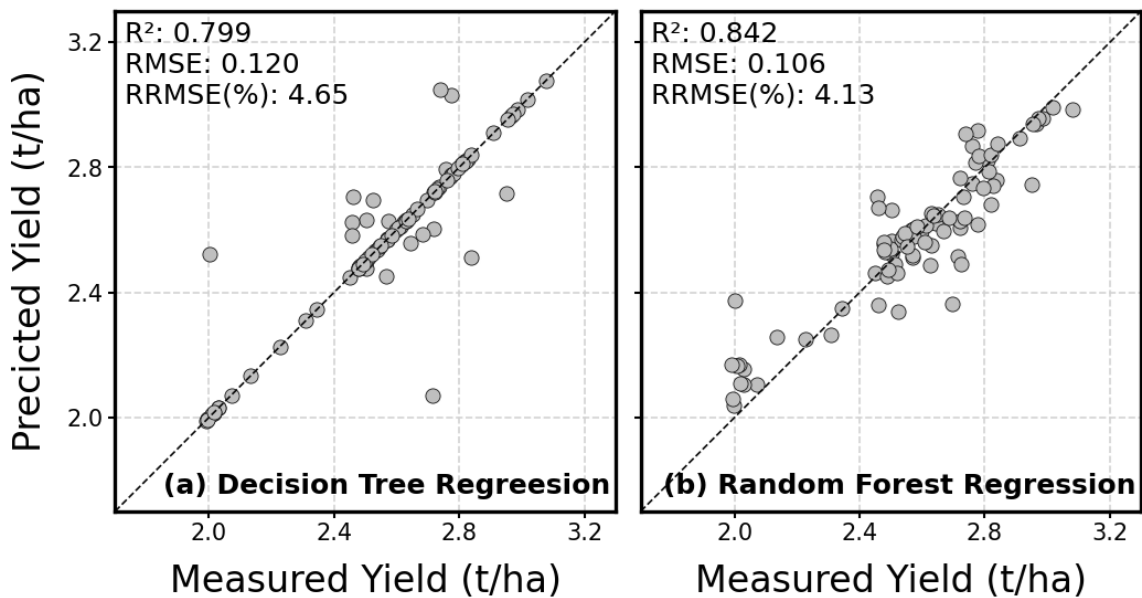


Figure 4.1. Model training results of DTR and RFR (2019-20) (a) & (2020-21) (b).



a. Shows the model testing results of DTR and RFR (2019-20)



b. Shows the model testing results of DTR and RFR (2020-21)

Figure 4.2. Model testing results of DTR and RFR (2019-20) (a) & (2020-21).

CONCLUSION AND RECOMMENDATIONS**5.1 Conclusion**

The study demonstrates the successful use of remote sensing and machine learning in wheat yield prediction. The use of Landsat-8 surface reflectance data acquired from GEE and the time series of vegetation indices enabled accurate prediction of wheat yield. Feature selection using correlation analysis was critical in reducing input data to avoid uncertainty, and only relevant data was used as input to the model. The 8th time step was found to have a high correlation with yield, and data from this step was used for model input. The study trained models for both of the years separately and made separate feature selections, providing insights into relevant input data for accurate yield predictions. Decision Tree Regression (DTR) and Random Forest Regression (RFR) results show that the RFR gives the best performance for wheat yield prediction. One of the advantages of random forest regression is that it is resistant to overfitting. This can occur when a model becomes too complex and starts to fit noise in the data rather than the underlying patterns. This is because the random forest algorithm uses multiple decision trees with different subsets of features and data points, which helps to reduce the variance in the predictions. Overall, machine learning models can help farmers and agronomists make informed decisions about crop management practices and improve crop yields by providing accurate yield estimation and predictions.

The study was able to successfully test machine learning models for yield predictions. However, there are some uncertainties related to this study. Firstly, the yield range was very narrow. Estimation of variables which has a narrow range prohibits machine learning models from better generalising the data, thus making some of the estimation uncertain or incorrect. Secondly, the sample size of the yield was small. Machine learning models required large amounts of data to better generalize. These issues need to be addressed in the future.

5.2 Recommendations

Based on the study, the following recommendations are made

- Further research should explore the potential of combining remote sensing and machine learning with other data sources, such as meteorological data, to improve accuracy in crop yield prediction.
- As more remote sensing data become available, researchers should investigate the use of deep learning techniques to improve yield prediction accuracy.

- The study highlights the importance of feature selection in reducing input data and avoiding uncertainty. Future studies should explore different feature selection techniques and their impact on yield prediction accuracy.
- Finally, the stakeholders in the agriculture industry should leverage the potential of remote sensing and machine learning to improve crop management practices, decision-making, the potential for precision agriculture and its ability to increase yields and improve overall efficiency in agriculture practices.

REFERENCES

1. Aghighi, H., Azadbakht, M., Ashourloo, D., Shahrabi, H. S., & Radiom, S. (2018). Machine learning regression techniques for the silage maize yield prediction using time-series images of Landsat 8 OLI. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12), 4563-4577.
2. A. Ahmad, M.R. Khan, S.H.H Shah, M.A. Kamran, S.A. Wajid, M. Amin, A. Khan, M.N Arshad, M.J.M Cheema, Z.A. Saqib, R. Ullah, K. Ziaf, A. ul Huq, S. Ahmad, I. Ahmad, M. Fahad, M.M. Waqas, A. Abbas, A. Iqbal, A. Pervaiz & I.A. Khan. (2019). *Agro-ecological zones of Punjab, Pakistan-2019*. Rome, FAO.
3. Ali, A., Martelli, R., Lupia, F., & Barbanti, L. (2019). Assessing multiple years' spatial variability of crop yields using satellite vegetation indices. *Remote sensing*, 11(20), 2384.
4. Aghighi, H., Azadbakht, M., Ashourloo, D., Shahrabi, H. S., & Radiom, S. (2018). Machine learning regression techniques for the silage maize yield prediction using time-series images of Landsat 8 OLI. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12), 4563-4577.
5. Briscoe, J., & Qamar, U. (2006). *Pakistan's water economy: running dry*. The World Bank.
6. Bongiovanni, R., & Lowenberg-DeBoer, J. (2004). Precision agriculture and sustainability. *Precision agriculture*, 5, 359-387.
7. Bian, C., Shi, H., Wu, S., Zhang, K., Wei, M., Zhao, Y., ... & Chen, S. (2022). Prediction of field-scale wheat yield using machine learning method and multi-spectral UAV data. *Remote Sensing*, 14(6), 1474.
8. Bannari, A., Morin, D., Bonn, F., & Huete, A. (1995). A review of vegetation indices. *Remote sensing reviews*, 13(1-2), 95-120.
9. Chen, Y., Zhang, Z., Tao, F., Wang, P., & Wei, X. (2017). Spatio-temporal patterns of winter wheat yield potential and yield gap during the past three decades in North China. *Field Crops Research*, 206, 11-20.
10. Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., & Li, Z. (2018). A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote sensing of environment*, 210, 35-47.

11. Cao, J., Zhang, Z., Luo, Y., Zhang, L., Zhang, J., Li, Z., & Tao, F. (2021). Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. *European Journal of Agronomy*, *123*, 126204.
12. Canata, T. F., Wei, M. C. F., Maldaner, L. F., & Molin, J. P. (2021). Sugarcane yield mapping using high-resolution imagery data and machine learning technique. *Remote Sensing*, *13*(2), 232.
13. Choudhary, K., Shi, W., Dong, Y., & Paringer, R. (2022). Random Forest for rice yield mapping and prediction using Sentinel-2 data with Google Earth Engine. *Advances in Space Research*, *70*(8), 2443-2457.
14. Chaudhary, Q. Z., & Rasul, G. (2004). Agro-climatic classification of Pakistan. *Sci Vis*, *9*(1-4), 59-66.
15. Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., ... & Peng, B. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and forest meteorology*, *274*, 144-159.
16. Dorosh, P., & Salam, A. (2008). Wheat markets and price stabilisation in Pakistan: An analysis of policy options. *The Pakistan Development Review*, 71-87.
17. Fu, Y., Yang, G., Wang, J., Song, X., & Feng, H. (2014). Winter wheat biomass estimation based on spectral indices, band depth analysis and partial least squares regression using hyperspectral measurements. *Computers and Electronics in Agriculture*, *100*, 51-59.
18. Feng, L., Wang, Y., Zhang, Z., & Du, Q. (2021). Geographically and temporally weighted neural network for winter wheat yield prediction. *Remote Sensing of Environment*, *262*, 112514.
19. Fahad, M., Ahmad, I., Rehman, M., Waqas, M. M., & Gul, F. (2019). Regional wheat yield estimation by integration of remotely sensed soil moisture into a crop model. *Canadian Journal of Remote Sensing*, *45*(6), 770-781.
20. Friedl, M. A., McIver, D. K., Hodges, J. C., Zhang, X. Y., Muchoney, D., Strahler, A. H., ... & Schaaf, C. (2002). Global land cover mapping from MODIS: algorithms and early results. *Remote sensing of Environment*, *83*(1-2), 287-302.
21. Faruquee, R., & Coleman, J. R. (1996). *Managing price risks in the Pakistan wheat market*. The World Bank.
22. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, *202*, 18-27.

23. Gumma, M. K., Kadiyala, M. D. M., Panjala, P., Ray, S. S., Akuraju, V. R., Dubey, S., ... & Whitbread, A. M. (2022). Assimilation of remote sensing data into crop growth model for yield estimation: A case study from India. *Journal of the Indian Society of Remote Sensing*, 50(2), 257-270.
24. Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., & Zhang, J. (2020). Prediction of winter wheat yield based on multi-source data and machine learning in China. *Remote Sensing*, 12(2), 236.
25. Hunt, M. L., Blackburn, G. A., Carrasco, L., Redhead, J. W., & Rowland, C. S. (2019). High resolution wheat yield mapping using Sentinel-2. *Remote Sensing of Environment*, 233, 111410.
26. IMRAN, A., & NOUREEN, K. Wheat Crop Development in Central Punjab (Faisalabad, 2020–21).
27. Iizumi, T., Shin, Y., Kim, W., Kim, M., & Choi, J. (2018). Global crop yield forecasting using seasonal climate information from a multi-model ensemble. *Climate Services*, 11, 13-23.
28. Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., & Lobell, D. B. (2019). Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sensing of Environment*, 228, 115-128.
29. Jovanović, D., Sabo, F., Govedarica, M., & Marinković, B. (2014). Crop yield estimation in 2014 for Vojvodina using methods of remote sensing. *Ratarstvo i povrtarstvo*, 51(3), 145-153.
30. Kayad, A., Sozzi, M., Gatto, S., Marinello, F., & Pirotti, F. (2019). Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques. *Remote Sensing*, 11(23), 2873
31. Khan, S. N., Li, D., & Maimaitijiang, M. (2022). A Geographically Weighted Random Forest Approach to Predict Corn Yield in the US Corn Belt. *Remote Sensing*, 14(12), 2843
32. Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., & Shearer, S. (2018). Integration of high-resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and electronics in agriculture*, 153, 213-225.
33. Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39, 261-283.
34. Kern, A., Barcza, Z., Marjanović, H., Árendás, T., Fodor, N., Bónis, P., ... & Lichtenberger, J. (2018). Statistical modelling of crop yield in Central Europe using

- climate data and remote sensing vegetation indices. *Agricultural and forest meteorology*, 260, 300-320.
35. Lee, J. H., Shin, J., & Realf, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering*, 114, 111-121.
 36. Leng, G., & Huang, M. (2017). Crop yield response to climate change varies with crop spatial distribution pattern. *Scientific Reports*, 7(1), 1463.
 37. Li, Y., Guan, K., Yu, A., Peng, B., Zhao, L., Li, B., & Peng, J. (2019). Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the US. *Field Crops Research*, 234, 55-65.
 38. Li, A., Liang, S., Wang, A., & Qin, J. (2007). Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. *Photogrammetric Engineering & Remote Sensing*, 73(10), 1149-1157.
 39. Liaqat, M. U., Cheema, M. J. M., Huang, W., Mahmood, T., Zaman, M., & Khan, M. M. (2017). Evaluation of MODIS and Landsat multiband vegetation indices used for wheat yield estimation in irrigated Indus Basin. *Computers and Electronics in Agriculture*, 138, 39-47.
 40. Li, Y., Guan, K., Yu, A., Peng, B., Zhao, L., Li, B., & Peng, J. (2019). Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the US. *Field Crops Research*, 234, 55-65.
 41. Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., & Foley, J. A. (2012). Closing yield gaps through nutrient and water management. *Nature*, 490(7419), 254-257.
 42. Monthly Climatic Normals of Pakistan, 1981-2010 (January 2013): Climate Data Processing Center, Pakistan meteorological Department, Karachi.
 43. Meng, L., Liu, H., L. Ustin, S., & Zhang, X. (2021). Predicting maize yield at the plot scale of different fertilizer systems by multi-source data and machine learning methods. *Remote Sensing*, 13(18), 3760.
 44. Ma, Y., Zhang, Z., Kang, Y., & Özdoğan, M. (2021). Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sensing of Environment*, 259, 112408.
 45. Mumtaz, R., Baig, S., & Fatima, I. (2017). Analysis of meteorological variations on wheat yield and its estimation using remotely sensed data. A case study of selected districts of Punjab Province, Pakistan (2001-14). *Italian Journal of Agronomy*, 12(3).

46. Olmos-Trujillo, E., González-Trinidad, J., Júnez-Ferreira, H., Pacheco-Guerrero, A., Bautista-Capetillo, C., Avila-Sandoval, C., & Galván-Tejada, E. (2020). Spatio-temporal response of vegetation indices to rainfall and temperature in a semiarid region. *Sustainability*, *12*(5), 1939
47. Petersen, L. K. (2018). Real-time prediction of crop yields from MODIS relative vegetation health: A continent-wide analysis of Africa. *Remote Sensing*, *10*(11), 1726.
48. Pede, T., Mountrakis, G., & Shaw, S. B. (2019). Improving corn yield prediction across the US Corn Belt by replacing air temperature with daily MODIS land surface temperature. *Agricultural and Forest Meteorology*, *276*, 107615.
49. Prasad, A. K., Chai, L., Singh, R. P., & Kafatos, M. (2006). Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied earth observation and geoinformation*, *8*(1), 26-33.
50. Robinson, N. P., Allred, B. W., Jones, M. O., Moreno, A., Kimball, J. S., Naugle, D. E., ... & Richardson, A. D. (2017). A dynamic Landsat derived normalized difference vegetation index (NDVI) product for the conterminous United States. *Remote sensing*, *9*(8), 863.
51. Roberts, M. J., Braun, N. O., Sinclair, T. R., Lobell, D. B., & Schlenker, W. (2017). Comparing and combining process-based crop models and statistical models with some implications for climate change. *Environmental Research Letters*, *12*(9), 095010.
52. Sharma, N., Sharma, R., & Jindal, N. (2021). Machine learning and deep learning applications-a vision. *Global Transitions Proceedings*, *2*(1), 24-28.
53. Shamshad, K. M. (1988). The meteorology of Pakistan. *Royal Book Company, Karachi*.
54. Shiu, Y. S., & Chuang, Y. C. (2019). Yield estimation of paddy rice based on satellite imagery: Comparison of global and local regression models. *Remote Sensing*, *11*(2), 111.
55. Skakun, S., Vermote, E., Roger, J. C., & Franch, B. (2017). Combined use of Landsat-8 and Sentinel-2A images for winter crop mapping and winter wheat yield assessment at regional scale. *AIMS geosciences*, *3*(GSFC-E-DAA-TN49944).
56. Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology*, *284*, 107886.

57. Shammi, S. A., & Meng, Q. (2021). Use time series NDVI and EVI to develop dynamic crop growth metrics for yield modeling. *Ecological Indicators*, *121*, 107124.
58. Sharifi, A. (2021). Yield prediction with machine learning algorithms and satellite images. *Journal of the Science of Food and Agriculture*, *101*(3), 891-896.
59. Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of neuroscience methods*, *220*(1), 85-91.
60. Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science*, *11*, 1120.
61. Shahhosseini, M., Martinez-Feria, R. A., Hu, G., & Archontoulis, S. V. (2019). Maize yield and nitrate loss prediction with machine learning algorithms. *Environmental Research Letters*, *14*(12), 124026.
62. Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology*, *284*, 107886.
63. Schafer, A., & Victor, D. G. (2000). The future mobility of the world population. *Transportation research part a: policy and practice*, *34*(3), 171-205.
64. Tuvdendorj, B., Wu, B., Zeng, H., Batdelger, G., & Nanzad, L. (2019). Determination of appropriate remote sensing indices for spring wheat yield estimation in Mongolia. *Remote Sensing*, *11*(21), 2568.
65. Vanli, Ö., Ahmad, I., & Ustundag, B. B. (2020). Area estimation and yield forecasting of wheat in southeastern turkey using a machine learning approach. *Journal of the Indian Society of Remote Sensing*, *48*, 1757-1766.
66. Wei, M. C. F., & Molin, J. P. (2020). Soybean yield estimation and its components: A linear regression approach. *Agriculture*, *10*(8), 348.
67. Wolanin, A., Mateo-García, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., ... & Guanter, L. (2020). Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environmental Research Letters*, *15*(2), 024019.
68. Yuan, W., Chen, Y., Xia, J., Dong, W., Magliulo, V., Moors, E., ... & Zhang, H. (2016). Estimating crop yield using a satellite-based light use efficiency model. *Ecological Indicators*, *60*, 702-709.
69. Zhang, Z., Song, X., Tao, F., Zhang, S., & Shi, W. (2016). Climate trends and crop production in China at county scale, 1980 to 2008. *Theoretical and Applied*

Climatology, 123, 291-302.

70. Zhou, W., Liu, Y., Ata-Ul-Karim, S. T., Ge, Q., Li, X., & Xiao, J. (2022). Integrating climate and satellite remote sensing data for predicting county-level wheat yield in China using machine learning methods. *International Journal of Applied Earth Observation and Geoinformation*, 111, 102861.

Appendix-1. Yield correlation with time series T2MAX (the maximum temperature at 2) (2019-20).

	yield	T2MAX_1	T2MAX_2	T2MAX_3	T2MAX_4	T2MAX_5	T2MAX_6	T2MAX_7	T2MAX_8	T2MAX_9	T2MAX_10
yield	1	0.14**	-0.01	0.04	0	0.07	0.08	0.06	0.22***	0.22***	0.24***
T2MAX_1	0.14**	1	0.94***	0.97***	0.95***	0.98***	0.97***	0.93***	0.97***	0.97***	0.97***
T2MAX_2	-0.01	0.94***	1	0.99***	0.99***	0.98***	0.97***	0.98***	0.94***	0.91***	0.92***
T2MAX_3	0.04	0.97***	0.99***	1	1.0***	1.0***	0.99***	0.98***	0.96***	0.95***	0.96***
T2MAX_4	0	0.95***	0.99***	1.0***	1	0.99***	0.98***	0.97***	0.95***	0.94***	0.94***
T2MAX_5	0.07	0.98***	0.98***	1.0***	0.99***	1	1.0***	0.96***	0.97***	0.97***	0.97***
T2MAX_6	0.08	0.97***	0.97***	0.99***	0.98***	1.0***	1	0.96***	0.97***	0.97***	0.97***
T2MAX_7	0.06	0.93***	0.98***	0.98***	0.97***	0.96***	0.96***	1	0.96***	0.91***	0.93***
T2MAX_8	0.22***	0.97***	0.94***	0.96***	0.95***	0.97***	0.97***	0.96***	1	0.97***	0.98***
T2MAX_9	0.22***	0.97***	0.91***	0.95***	0.94***	0.97***	0.97***	0.91***	0.97***	1	0.99***
T2MAX_10	0.24***	0.97***	0.92***	0.96***	0.94***	0.97***	0.97***	0.93***	0.98***	0.99***	1

Significance level *: p-value ≤ 0.1 , **: p-value ≤ 0.05 , ***: p-value ≤ 0.01

Appendix-2. Yield correlation with time series T2MAX (the maximum temperature at 2 meters) (2020-21).

	yield	T2MAX_1	T2MAX_2	T2MAX_3	T2MAX_4	T2MAX_5	T2MAX_6	T2MAX_7	T2MAX_8	T2MAX_9	T2MAX_10
yield	1	0.1	-0.04	0.01	-0.02	0.05	0.06	0.04	0.18***	0.19***	0.2***
T2MAX_1	0.1	1	0.95***	0.97***	0.95***	0.97***	0.97***	0.93***	0.97***	0.97***	0.98***
T2MAX_2	-0.04	0.95***	1	0.99***	0.98***	0.97***	0.96***	0.98***	0.95***	0.92***	0.93***
T2MAX_3	0.01	0.97***	0.99***	1	1.0***	1.0***	0.99***	0.98***	0.96***	0.96***	0.96***
T2MAX_4	-0.02	0.95***	0.98***	1.0***	1	0.99***	0.99***	0.98***	0.95***	0.94***	0.94***
T2MAX_5	0.05	0.97***	0.97***	1.0***	0.99***	1	1.0***	0.96***	0.97***	0.97***	0.96***
T2MAX_6	0.06	0.97***	0.96***	0.99***	0.99***	1.0***	1	0.96***	0.97***	0.97***	0.96***
T2MAX_7	0.04	0.93***	0.98***	0.98***	0.98***	0.96***	0.96***	1	0.96***	0.92***	0.93***
T2MAX_8	0.18***	0.97***	0.95***	0.96***	0.95***	0.97***	0.97***	0.96***	1	0.97***	0.98***
T2MAX_9	0.19***	0.97***	0.92***	0.96***	0.94***	0.97***	0.97***	0.92***	0.97***	1	0.99***
T2MAX_10	0.20***	0.98***	0.93***	0.96***	0.94***	0.96***	0.96***	0.93***	0.98***	0.99***	1

Significance level *: p-value \leq 0.1, **: p-value \leq 0.05, ***: p-value \leq 0.01

Appendix-3. Yield correlation with time series relative humidity (relative humidity at 2 meters) (2019-20).

	yield	RH2M_1	RH2M_2	RH2M_3	RH2M_4	RH2M_5	RH2M_6	RH2M_7	RH2M_8	RH2M_9	RH2M_10
yield	1	-0.12*	-0.11*	-0.13**	-0.01	0.01	-0.09	-0.13***	-0.16***	-0.14***	-0.11***
RH2M_1	-0.12*	1	0.93***	0.94***	0.96***	0.91***	0.95***	0.96***	0.96***	0.95***	0.94***
RH2M_2	-0.11*	0.93***	1	0.97***	0.97***	0.95***	0.99***	0.96***	0.97***	0.97***	0.97***
RH2M_3	-0.13**	0.94***	0.97***	1	0.96***	0.94***	0.96***	0.93***	0.96***	0.97***	0.96***
RH2M_4	-0.01	0.96***	0.97***	0.96***	1	0.95***	0.97***	0.96***	0.97***	0.94***	0.93***
RH2M_5	0.01	0.91***	0.95***	0.94***	0.95***	1	0.95***	0.88***	0.9***	0.91***	0.91***
RH2M_6	-0.09	0.95***	0.99***	0.96***	0.97***	0.95***	1	0.96***	0.97***	0.97***	0.97***
RH2M_7	-0.13***	0.96***	0.96***	0.93***	0.96***	0.88***	0.96***	1	0.99***	0.97***	0.96***
RH2M_8	-0.16***	0.96***	0.97***	0.96***	0.97***	0.9***	0.97***	0.99***	1	0.98***	0.97***
RH2M_9	-0.14***	0.95***	0.97***	0.97***	0.94***	0.91***	0.97***	0.97***	0.98***	1	1.0***
RH2M_10	-0.11***	0.94***	0.97***	0.96***	0.93***	0.91***	0.97***	0.96***	0.97***	1.0***	1

Significance level *: p-value ≤ 0.1 , **: p-value ≤ 0.05 , ***: p-value ≤ 0.01

Appendix-4. Yield Correlation with time series relative humidity (relative humidity at 2 meters) (2020-21).

	yield	RH2M_1	RH2M_2	RH2M_3	RH2M_4	RH2M_5	RH2M_6	RH2M_7	RH2M_8	RH2M_9	RH2M_10
yield	1	-0.13**	-0.12*	-0.14**	-0.02	0	-0.11	-0.19***	-0.18***	-0.25***	-0.25***
RH2M_1	-0.13**	1	0.94***	0.94***	0.97***	0.91***	0.95***	0.97***	0.96***	0.95***	0.94***
RH2M_2	-0.12*	0.94***	1	0.97***	0.97***	0.95***	0.99***	0.96***	0.97***	0.97***	0.97***
RH2M_3	-0.14**	0.94***	0.97***	1	0.96***	0.94***	0.96***	0.94***	0.96***	0.97***	0.96***
RH2M_4	-0.02	0.97***	0.97***	0.96***	1	0.95***	0.98***	0.96***	0.97***	0.94***	0.93***
RH2M_5	0	0.91***	0.95***	0.94***	0.95***	1	0.96***	0.88***	0.91***	0.91***	0.91***
RH2M_6	-0.11	0.95***	0.99***	0.96***	0.98***	0.96***	1	0.96***	0.97***	0.97***	0.97***
RH2M_7	-0.19***	0.97***	0.96***	0.94***	0.96***	0.88***	0.96***	1	0.99***	0.97***	0.96***
RH2M_8	-0.18***	0.96***	0.97***	0.96***	0.97***	0.91***	0.97***	0.99***	1	0.98***	0.98***
RH2M_9	-0.25***	0.95***	0.97***	0.97***	0.94***	0.91***	0.97***	0.97***	0.98***	1	1.0***
RH2M_10	-0.25***	0.94***	0.97***	0.96***	0.93***	0.91***	0.97***	0.96***	0.98***	1.0***	1

Significance level *, p-value <= 0.1, **, p-value <= 0.05, ***, p-value <= 0.01