Journal of the Saudi Society of Agricultural Sciences 22 (2023) 107–120

HOSTED BY

Contents lists available at ScienceDirect

Journal of the Saudi Society of Agricultural Sciences

journal homepage: www.sciencedirect.com

Full length article

Machine learning for yield prediction in Fergana valley, Central Asia

Mukesh Singh Boori [a,*], Komal Choudhary [a,b,*], Rustam Paringer [a,c], Alexander Kupriyanov [a,c]

[a] Scientific Research Laboratory of Automated Systems of Scientific Research (SRL-35), Samara National Research University, Samara, Russia
[b] Department of Land Surveying and Geo- Informatics, Smart Cities Research Institute, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[c] Image Processing Systems Institute of the RAS–Branch of the FSRC "Crystallography and Photonics", Samara, Russia

ARTICLE INFO

Article history:
Received 23 May 2022
Revised 19 July 2022
Accepted 27 July 2022
Available online 2 August 2022

Keywords:
Yield prediction
Regression
Sentinel-2
Spectral-indices
Machine learning
Phenology
Precision agriculture
Food security

ABSTRACT

Accurate yield prediction is essential for growers, researchers, governments, the farming industry, and policymakers for social peace, food safety, security, and sustainable development. The results of earlier techniques of data collecting and analysis for yield forecasts were typically delayed, expensive, time-consuming, site-specific, and riddled with errors and uncertainties. This study is a novel approach to using high-resolution satellite data in conjunction with environmental and topographic data to predict wheat yield variability at the farm scale using machine learning. In this research, winter wheat yield prediction was based on 36 indicators in machine learning using correlation and different regression models. Winter wheat yield was predicted using linear regression (LR), decision tree (DT), and random forest (RF) regression models with scikit-learn in machine learning. More than 10,000 data points from 45 farms were trained and validated in Fergana valley, Central Asia. Results indicate that at 10 m resolution using Sentinel-2 and other secondary data such as topographic, soil, environmental, and filed data can generate an accurate wheat yield prediction map. The accuracy of all regressions is lowest for LR (R^2:95, RMSE: 2.31), highest for RF (R^2:98, RMSE: 1.40), and intermediate for DT regression (R^2:97, RMSE: 1.85). Results also indicate that prediction in the early stage of the crop is less accurate in comparison to harvesting time as LR (R^2:85, RMSE: 2.66), DT (R^2:95, RMSE: 2.06), RF (R^2:97, RMSE: 1.54) have different R^2 and RMSE values. Appling the RF model, the winter wheat prediction is 3.29 to 4.30 t/ha therefore the total wheat production is approximately 100 t in the study area. Thus this study will demonstrate the capability of high-resolution satellite imagery and secondary data for highly accurate real-time crop yield prediction at the field scale, which can be used to assist precision agriculture and will provide a point of reference for crop area extraction, mapping, monitoring, and sustainable development with food security.

1. Introduction

Accurate crop yield observation and prediction are very important for food safety, security, and social peace, and they variate according to the soil, climate, and agriculture practices of a region (Ahn et al., 2022, Eklund et al. 2022, Ajami et al. 2020, Dokoohaki et al. 2015). It is especially important when population growth and urban sprawl are accelerating and putting undue strain on the agricultural sector (Obwocha et al., 2022, Azadi et al., 2021). Consequently, proper yield forecast aids in increasing the agriculture sector's productivity and toughness so that it can compete for market demand (Maftouh et al. 2022, Bhat and Huang 2021). It also supports farmers to know the actual effecting factors in agriculture and direct site-specific sustainable development plans for policymakers (Hamid et al., 2021, Yakupoğlu et al. 2022, Boori et al. 2020a, Norouzi et al. 2010). At high resolution, time series yield predictions suggest proper agricultural production management and the future scenario for food security (Kephe et al. 2021). For insurance and agribusiness choices, an accurate early crop production projection is essential (Ostaev et al. 2020). Many yield prediction techniques are currently available, based on local and national variables and requirements at various scales (Dubey et al. 2018, Fritz et al., 2019, Xie et al. 2020). These are based on vegetation indices statistics, radar data, net primary productivity estimation, crop height, biomass, and growth based (Chao et al., 2019). However these are focused on current crop conditions and never

* Corresponding authors.
E-mail addresses: boori.m@ssau.ru (M. Singh Boori), komal.kumari@connect.polyu.hk (K. Choudhary), RusParinger@ssau.ru (R. Paringer), akupr@ssau.ru (A. Kupriyanov).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jssas.2022.07.006
1658-077X/

estimate long time series at the field level, also neglect within-field yield variability, which is important for researchers to get the relationship between agriculture and environmental impact (Wolanin et al. 2020, Daniel et al., 2021, Boori et al. 2021).

There are currently many innovative approaches in the farming sector for mapping, monitoring, and crop variability observation at various crop growth stages (Khanal et al. 2020, Wolanin et al. 2020, Daniel et al., 2021). These precision farming are aiming to reduce waste and estimate accurate qualitative production (Ahmad et al. 2022, Choudhary et al. 2021). Since the beginning of satellite remote sensing, agriculture mapping, monitoring, management, and estimation have been critical activities in the agriculture sector (Pan et al. 2021, Poppiel et al. 2021, Choudhary et al. 2021). However, this technology also has some limitations such as data cost and frequency, pixel resolution, repeat cycle of the sensor, cloud cover, weather conditions, and technical limitations thus missing some key information during crop growth stages (Saiz-Rubio and Rovira-Más, 2020, Huang et al., 2018). In comparing Landsat and Sentinel in terms of resolution, cloud computing like google earth engine, sentinel-2 is a better option for time-series crop yield prediction (Liu et al., 2020, Pan et al. 2021, Poppiel et al. 2021, Taghizadeh-Mehrjardi et al. 2020, Tajik et al. 2020). However, the selection of appropriate data for a specific factor and method in agriculture mapping and monitoring is important (Feng et al., 2019, Boori et al. 2022). One of the most popular techniques for yield prediction is the use of vegetation indices (VIs) based on statistical analysis (Guo et al. 2020, Choudhary et al. 2019). This is based on vegetation reflectance values in the visible and near-infrared spectrum (Ali et al. 2019). Generally yield prediction based on selected indicators such as biomass, plant height, VIs, and environmental data (temperature, precipitation, evaporation, etc.), statistically, these approaches are broadly divided into two groups first conventional regression methods (eg. linear regression (LR) and stepwise regression (SR)) and second machine learning methods (decision tree (DT), random forest (RF) regression, support vector machine (SVM), artificial neural network (ANN)), etc. (Choudhary et al. 2022, Sishodia et al. 2020, Cai et al., 2019, Poppiel et al. 2021, Taghizadeh-Mehrjardi et al. 2020, Tajik et al. 2020).

Recently machine learning is a powerful tool with its large number and size of heterogeneous data-driven approaches at any scale (Azimi et al. 2020, Ye et al. 2018, Besalatpour et al. 2014, Zeraatpisheh et al. 2021, Naimi et al., 2021, Zeraatpishehet al. 2019). Machine learning makes nonlinear relationships between different sources of data and provides yield estimation and predictions (Palanivel and Surianarayanan 2019, Schulz et al. 2020). Machine learning is also facilitated by numerous inbuilt models, which can use a wide range of applications such as correlation, histogram, data description, classification, and so on (Seyedzadeh et al., 2020, Boori et al. 2020b, Choudhary et al. 2022, Sishodia et al. 2020, Cai et al., 2019, Poppiel et al. 2021, Taghizadeh-Mehrjardi et al. 2020, Tajik et al. 2020). In this study, 28 secondary data sets, including environmental, soil, topographic, and field data, as well as vegetation indices (VIs) based on Sentinel-2 (10 m resolution) images were evaluated using the scikit-learn package in machine learning using Python. The topographic attributes such as slope, aspect, and elevation are associated with soil moisture, organic matter, and nutrition content and it is variate according to height and slop, which significantly affect agriculture production (Ajami et al. 2020). As a result, satellite-generated DEM and soil characteristics have a significant impact on crop production because soil moisture and nutrition content are higher on low slopes or at the bottom, implying that crop production is more likely (Ajami et al. 2020). The newly developed machine learning model is helpful to access winter wheat yield variability at different crop growth stages at 10 m resolution within-field yield level. The model was trained and validated over 10,000 points collected from 45 fields in Fergana valley, Central Asia. Therefore the main aim of this research were.

1. To develop a yield estimation and perdition model based on machine learning.
2. The newly developed model assesses and contrasts the effectiveness of various types of regression, including DT, LR, and RF.
3. The created model is applicable to various stages of crop growth.
4. The developed model is applicable to any scale, including the field level.
5. The developed model assesses the yield variability of all multi-source heterogeneous spatial and temporal data.

2. Methods and materials

2.1. Study area

The Fergana valley is located in Central Asia and spreads across eastern Uzbekistan, southern Kyrgyzstan, northern Tajikistan, and the former part of the Soviet Union. The valley is ethnically diverse and in the early 21st century was the scene of conflict due to its fertile land in a very dry part of central Asia. The coordinates of the study area are 40.40 N and 71.30E (Fig. 1). This research work was conducted on 45 winter wheat farms in the Fergana valley. The maximum rain is from March to May months with an average of 12.7 mm and August is the least rain, while snow is in January and February months with an average of 1.3 in.. The average winter is from −3 °C to 3 °C (January), while the average summer temperature is from 20 °C to 35 °C (July). Mainly the irrigation water comes from the Syr Darya River which divides into two parts in the study area named as Naryn and Kara Darya River and provides around 70 % surface water (Savoskul et al. 2003). Fergana valley is the oldest, most important agricultural and most fertile land for agriculture due to its foothills location in Central Asia (Abdullaev et al. 2009). Around 70 % populations depend on agriculture sector-based income, which contributes around 24 % country's gross domestic products (Bichsel 2009). Earlier cotton was the major crop, which was sequentially appended by the winter wheat crop, with 5 t/ha wheat production in the entire Fergana valley. This agricultural productivity has made the valley the most densely populated area in Central Asia. The Central Asia population density is 19 people per km^2 while Fergana valley has 1000 people per km^2, which is the fastest-growing within Central Asia with a 32 percent growth rate. This fast-growing population put extreme pressure on the agriculture sector. Due to its center location, which allowed it to rule several empires from all four sides, it has historically been a very significant territory.

2.2. Data and pre-processing

Fig. 2 represents the methodological steps and data used in this research work. Broadly Sentinel-2 generated VIs, environmental data, soil data, field data, and topographic data were used for winter wheat yield observation and prediction (Boori et al. 2022). This was a huge heterogeneous data in terms of special, spectral, temporal resolution, scale, data type or format, etc. therefore first all data were converted at 10 m resolution in ArcGIS software by cubic resampling method. Since the field data was in vector format, it was first transformed to a raster-based pixel format. Other geometric, radiometric, and atmospheric errors were removed in ArcGIS software thus each pixel opens to its expected location on the globe with less than 10 % cloud cover. Following that, all data were standardized so that all indicators will get equal importance and any individual indicator will not affect the outcomes. Now, in the

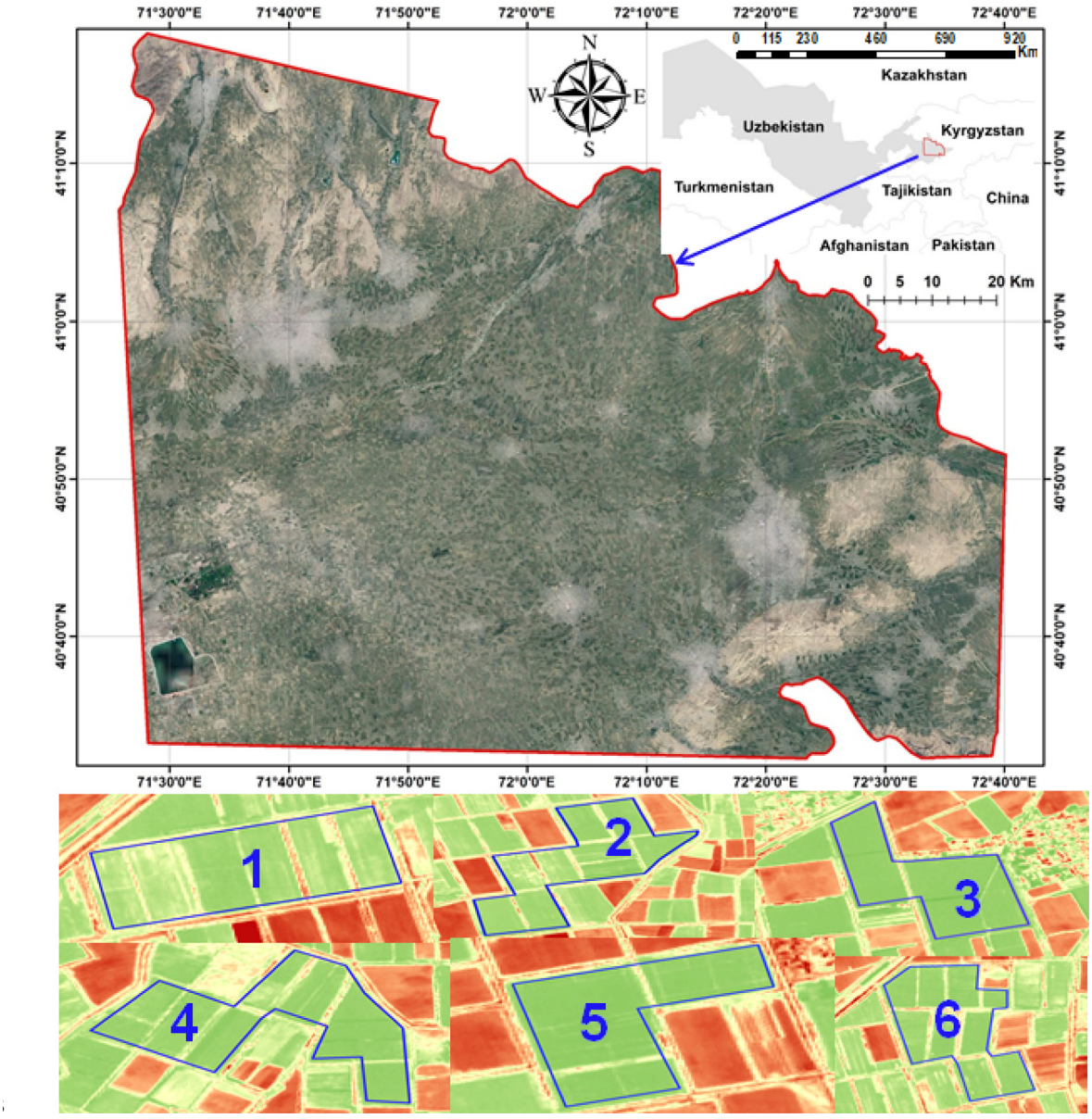

Fig. 1. The location map of the study area with six field experiment sites.

machine learning model, check the importance of each individual indicator by correlation with actual yield production. Finally, before running the model, all data were divided into training and test data sets with a 75:25 split ratio. Then linear regression (LR), decision tree (DT), and random forest (RF) regression models were run for yield prediction of winter wheat in the study area. All models were analyzed in the Jupiter notebook python programming language.

2.2.1. Wheat yield data

Crop yield data was obtained from the Fergana valley agriculture ministry, and the winter wheat crop was chosen as a study interest because it is the main crop that covers the most area in the study area (Löw et al., 2017). There are two wheat crops in a year winter and summer wheat crops. In this research, the winter wheat crop was used for prediction due to its high quality and quantitative production in the study area. Simple cleaning methods were used in machine learning to adjust all data points as some missing latitude/longitude and other information thus with the help of mean value all data points were corrected (Schulz et al. 2020).

2.2.2. Satellite data

Sentinel-2 at 10 m resolution data were used for land use/cover, crop classification, and vegetation indices (VIs) in the study area every month for the year 2021. The details of other collected data are in Table 1.

A<10 % cloud cover Sentinel-2 images were downloaded for the entire year of 2021from Copernicus open access hub and only 10 m resolution (band 2, 3, 4, 8) were selected for this research work.

2.2.2.1. LULC and crop classification. The following key classes were identified on the land use/land cover map (Fig. 3) using the supervised maximum likelihood classification technique: agricultural, forest, settlements, water, wetland, and others (Sisodia et al., 2014). Here other class covers bare land, and different types of seasonal crops such as orchards, maize, sunflower, sorghum, fruit plantation, etc. The reference fields were randomly selected

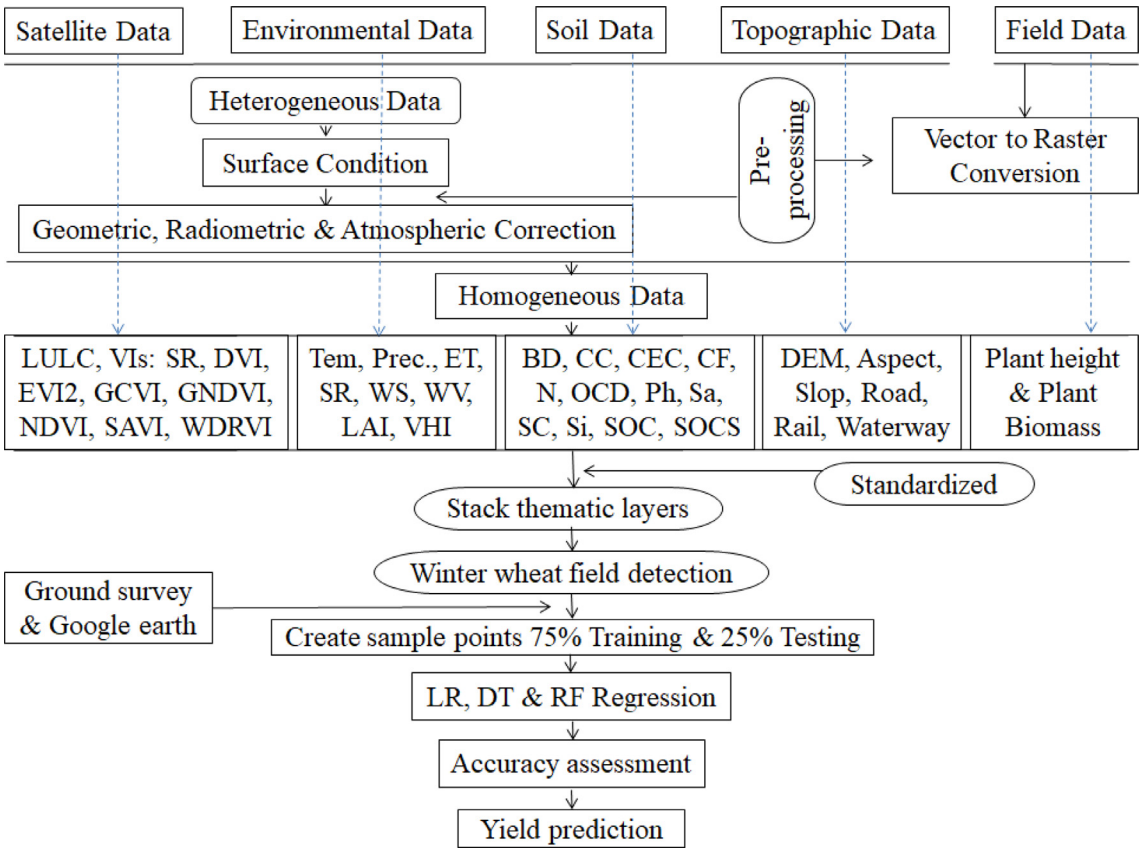

Fig. 2. Methodological steps data used in this research work.

Table 1
The details of the used data.

Data name	Attribute	Source
Sentinel-2 for VIs	05-Day temporal & 10 m spatial resolution	Copernicus Open Access Hub ()
MODIS 13Q1 NDVI	16-Day temporal & 250 m spatial resolution	NASA LAADS DAAC (https://ladsweb.modaps.eosdis.nasa.gov/search)
MODIS 16A2 ET data	8-Day temporal & 500 m spatial resolution	NASA LAADS DAAC (https://ladsweb.modaps.eosdis.nasa.gov/search)
MODIS 11A2 Temperature, Emissivity	8-Day temporal & 1 km spatial resolution	Earth-Explorer USGS (https://earthexplorer.usgs.gov/)
MODIS 15A2H LAI data	8-Day temporal & 500 m spatial resolution	Earth-Explorer USGS (https://earthexplorer.usgs.gov/)
MODIS 17A2H GPP data	8-Day temporal & 500 m spatial resolution	Earth-Explorer USGS (https://earthexplorer.usgs.gov/)
MODIS 12Q1 LULC	8-Day temporal & 500 m spatial resolution	NASA LAADS DAAC (https://ladsweb.modaps.eosdis.nasa.gov/search)
DEM/Elevation/Aspect	30 m spatial resolution	SRTM https://dwtkns.com/srtm30m/
AVHRR-NOAA VHI data	7-Day temporal & 1 km spatial resolution	NOAA https://www.star.nesdis.noaa.gov/smcd/emb/vci/VH/vh_ftp.php
NOAA-NCDC for Precipitation data	7-Day temporal & 1 km spatial resolution	https://www.ncdc.noaa.gov/cdo-web/
Road or topography data	shp	https://download.geofabrik.de/russia.html
Soil data	shp	https://soilgrids.org/
Socio-economic/ demographic data – Wind speed, Water vapor, Solar radiation	shp	Official website of Uzbekistan (https://www.gov.uz/en)
Field data – Plant height & Biomass	shp	field

throughout the entire study area for sampling. There were also some limitations in terms of data access and sampling due to limited access due to poor road or path conditions, broken bridges, not allowing access to specific sites, and so on, but all collected reference data were equally distributed in training and testing data in machine learning.

2.2.2.2. ***Vegetation indices*** *(VIs).* Following eight VIs were used in this research work, which was generated from Sentinel-2 data (Table 2). Generally, VIs highlight the vegetation part in the image and suppress other classes thus maximum VIs are generated with the Red (R) and near-infrared (NIR) band combinations as vegetation have the highest reflectance values in the R and NIR spectral range.

2.2.3. Environmental data

Temperature, Precipitation, evapotranspiration (ET), Solar Radiation (SR), Wind speed (WS), Water vapor (WV), leaf area index (LAI), and vegetation health index (VHI) were used under environmental data. The details of these data are described in Table 1 as spatial and temporal resolution, parental data, and source of the data. These indicators were chosen under environmental factors as they affect directly or indirectly crop production. As extreme temperature, ET, SR, and WS put a negative impact on vegetation

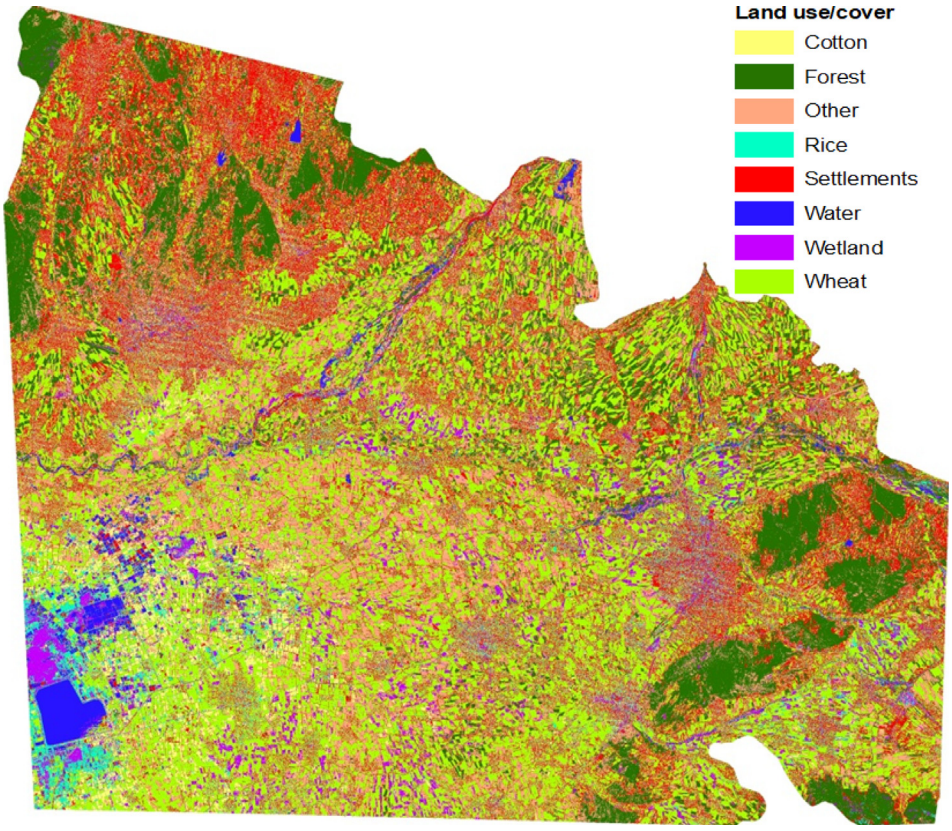

Fig. 3. LULC map of the study area.

by too much water extortion, while precipitation, WV, LIA, and VHI put a positive impact on vegetation.

2.2.4. Soil data

Soil data were obtained from soil grids (https://soilgrids.org/) from 0 to 5 cm in depth. Soil data has the following four major characteristics:

A) *Derived properties:* 1. Organic carbon density (OCD), 2. Soil organic carbon stock (SOCS).

B) *Physical soil properties:* 1. Bulk density (BD), 2. Clay content (CC), 3. Coarse fragment (CF), 4. Sand (Sa), Silt (Si).

C) *Chemical soil properties:* 1. Cation exchange capacity (CEC), 2. Nitrogen (N), 3. Soil organic carbon (SOC), 4. Ph.

D) *Soil classification:* based on world reference base (2006) soil groups.

2.2.5. Topographic data

In topography data, road, rail, and elevation-related information were collected. Road, rail, and waterway were collected from the open-street maps while height information such as elevation, aspect, and the slope was generated from the SRTM digital elevation model (DEM) at 30 m resolution, later on, converted into 10 m resolution to match the pixel size with other data.

2.2.6. Field data

Plant height and biomass data were collected in the selected fields during fieldwork. A real-time field cadaster data was digitized on high-resolution Google Earth images for the year 2021. Six wheat field sits were chosen for field data collection in the study area and collect plant height and biomass data. A total of 45 wheat farms were selected for analysis, which covers more than 10,000 data points. Later on, all data points were divided into test (25 %) and training (75 %) points. The test points were used to validate the training data set's results. A hand-held GPS with 2 m accuracy was used in the field and all obtained information during the fieldwork was used to verify the results.

2.2.7. Machine learning

A machine learning model was built for wheat yield prediction using the above-mentioned data. Machine learning is a powerful tool to handle numerous heterogeneous big data to make meaningful results. Once a module is prepared in machine learning later on with new input data related to crop and its effecting factors such as soil, weather, remote sensing, etc., new results can predict, or old results can update or modify at any scale based on input data. This type of machine learning model is based on the actual condition of the crop thus associated with the crop calendar or different crop growth stages. We can also change the baseline on the machine learning model such as in data cleaning, crop calendar dynamics based on time and location, variation in the user data range, and test, training data split ratio, etc. Therefore we used an improved scikit-learn based machine learning module in Jupyter notebook, run by python programming language from baseline in terms of reuse, scale, and flexibility, etc. Thus all sex selected case studies were run in above mentioned upgraded machine learning model. In this research work, the scikit-learn package with python language was used in machine learning due to its more user-friendly, easy to run, and generates different correlation and regression metrics (Azimi et al. 2020, Ye et al. 2018). First, a correlation metric was generated to know the importation of individual indicators with observed yield. Later on, with the help of multiple linear regressions (LR), decision tree (DT), and random forest (RF) regression models, yield observed and predicted values were generated.

Table 2
Vegetation indices calculations from sentinel-2 data at 10 m resolution, where R is red and NIR is a near-infrared band.

Vegetation Index	Abbreviation	Equation	Reference
Difference vegetation index	DVI	NIR - Red	Richardson and Weigand (1977)
Enhanced vegetation index 2	EVI2	2.5 * ((NIR – R) / (NIR + 2.4 * R + 1))	Jiang (2008), Huete (1988)
Green chlorophyll vegetation index	GCVI	(NIR/G) – 1	Gitelson et al. (2003)
Green normalized difference vegetation index	GNDVI	(NIR – G) / (NIR + G)	Gitelson et al. (1996)
Normalized difference vegetation index	NDVI	(NIR – R) / (NIR + R)	Rouse et al. (1973)
Soil-adjusted vegetation index	SAVI	((NIR – R) / (NIR + R + 0.5)) * (1 + 0.5)	Huete (1988)
Simple ratio	SR	NIR / R	Jordan (1969)
Wide dynamic range vegetation index	WDRVI	(0.2 * NIR - R) / (0.2 * NIR + R)	Gitelson (2004)

Table 3
LULC classes of the study area.

LULC	Cotton	Rice	Wheat	Forest	Settlement	Water	Wetland	Others
Area KmSq	444.09	5.57	2016.45	44.18	1342.38	98.33	305.57	1085.88
%	6.75	0.08	30.65	0.67	20.40	1.49	4.64	16.51

Table 4
R^2 and RMSE values of LR, DT, and RF regression models in all six sample fields at different growth stage/time.

		R^2				RMSE			
		Jan + Feb	Jan + Feb + Mar	Jan + Feb + Mar + Apr	Jan + Feb + Mar + Apr + May	Jan + Feb	Jan + Feb + Mar	Jan + Feb + Mar + Apr	Jan + Feb + Mar + Apr + May
Field 1	LR	0.66	0.75	0.85	0.92	5.34	4.59	3.58	2.66
	DTR	0.87	0.90	0.93	0.95	3.28	2.88	2.44	2.06
	RFR	0.93	0.94	0.96	0.97	2.36	2.18	1.80	1.54
Field 2	LR	0.32	0.73	0.80	0.94	7.28	4.60	3.90	2.13
	DTR	0.84	0.93	0.94	0.97	3.48	2.40	2.20	1.60
	RFR	0.91	0.96	0.96	0.98	2.66	1.85	1.66	1.18
Field 3	LR	0.44	0.71	0.82	0.96	3.13	2.24	1.76	0.87
	DTR	0.83	0.92	0.94	0.97	1.72	1.19	1.03	067
	RFR	0.90	0.96	0.97	0.98	1.29	0.87	0.73	0.51
Field 4	LR	0.31	0.62	0.68	0.94	7.74	5.76	5.23	2.29
	DTR	0.75	0.84	0.86	0.96	4.69	3.74	3.43	1.89
	RFR	0.86	0.91	0.92	0.98	3.53	2.82	2.59	1.37
Field 5	LR	0.29	0.74	0.80	0.95	4.97	3.01	2.63	1.33
	DTR	0.79	0.86	0.91	0.96	2.69	2.23	1.74	1.11
	RFR	0.86	0.92	0.95	0.98	2.17	1.64	1.28	0.88
Field 6	LR	0.26	0.65	0.72	0.95	8.82	6.06	5.41	2.31
	DTR	0.91	0.88	0.92	0.97	3.08	3.59	2.87	1.85
	RFR	0.94	0.94	0.96	0.98	2.42	2.49	2.14	1.40

2.2.8. Accuracy assessment

The above-generated model's accuracy was tested by the coefficient of determination (R^2) and root mean squared error (RMSE). R^2 is the measurement of the goodness of fit of a model, and how well the regression predictions approximate the real data points. Its value is from 0 to 1 range; where 1 indicates the best fit of the prediction model. RMSE is the standard deviation of the prediction errors, which use to measure the difference between, observed and predicted values, which means how far prediction values are from the regression line data points. RMSE measures the accuracy and is always non-negative and a 0 value is almost never possible. Normally a lower value in RMSE is better than higher values as it indicates less error.

3. Results

3.1. LULC and crop classes

Maximum likelihood supervised classification was used for land use /cover classification due to its better accuracy in comparison to other classification techniques as statistically it's based on probability to classify an unknown pixel. Broadly following nine major classes were identified in the study area: cotton, rice, wheat, forest, settlements, water, wetland, and others (Table 3). Wheat crop covered the most area at 2016 km^2 (30.65 %), followed by settlements at 1342 km^2 (20.40 %). Other class covers orchards, maize, sunflower, sorghum, fruit plantation, etc. It's a highly accurate LULC map with overall accuracy is 93 %. Some classes were misclassified such as water, wetland, and rice field due to water content in the field. Some parts of settlements are also misclassified due to greenness and hill terrain but still, it is a highly accurate LULC map with identifying individual crop fields, which was verified during field work.

3.2. Regression

Table 4 displays the resulted R^2 and RMSE values obtained from the linear, decision tree, and random forest regression from the beginning of the crop, mid or mature time, and in the last harvest time for yield prediction. The results demonstrate that R^2 values are initially quite low (0.26 to 0.66) and suggest lesser accuracy, but they gradually increase until harvest time and reach higher values (0.92 to 0.98) in all six sample fields, indicating higher accuracy. Here highest accuracy presents 1.008 t/ha wheat production

in the study area. RMSE values also show the same characteristics as R^2, in earlier prediction with higher error and in harvest time lowest error. When all models were compared, the RF model had the highest accuracy (higher R^2 and lower RMSE), the LR model had the lowest accuracy, and the DT model was in the middle (Table 4). Due to less accuracy in LR compared to RF, Fig. 4 show pinkish color in all field while RF show higher green color in all field. Overall, the RF model can be used to predict yield accurately based on the data used.

The validation of all models was based on test and training data. Each crop field has more than 10,000 data points thus 7500 points were used in training data sets and reaming 2500 points were used in testing data sets. These reaming data points were not used in training data thus accuracy is reliable as data were not overfittings in the models.

Here are some linear and random forest regression-derived histograms and their corresponding maps for the fields 1, 3, and 6 during the beginning and harvest stages. Within-field yield prediction is

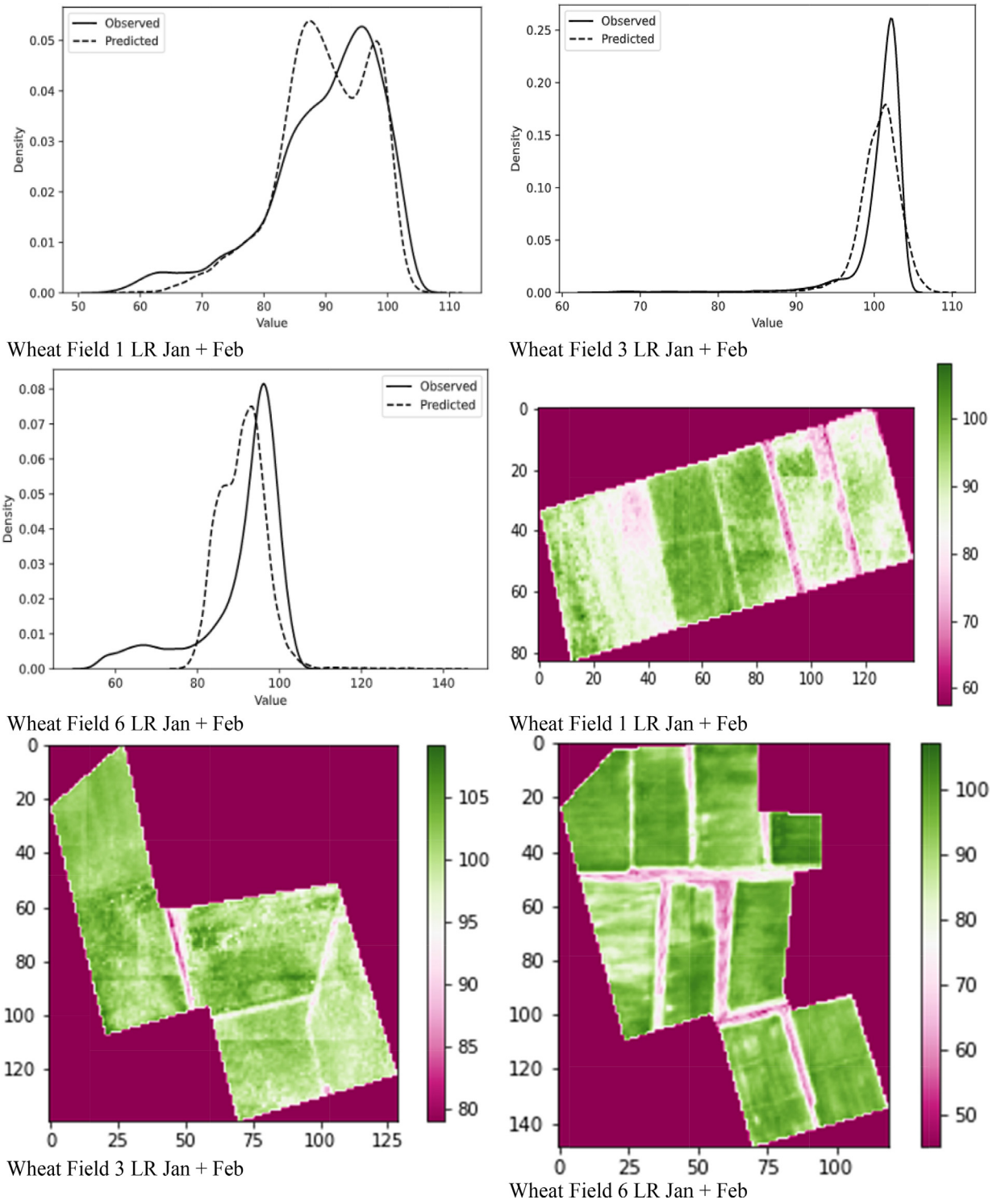

Fig. 4. Observed and predicted values from fields 1, 3, and 6 based on LR and RF regression models.

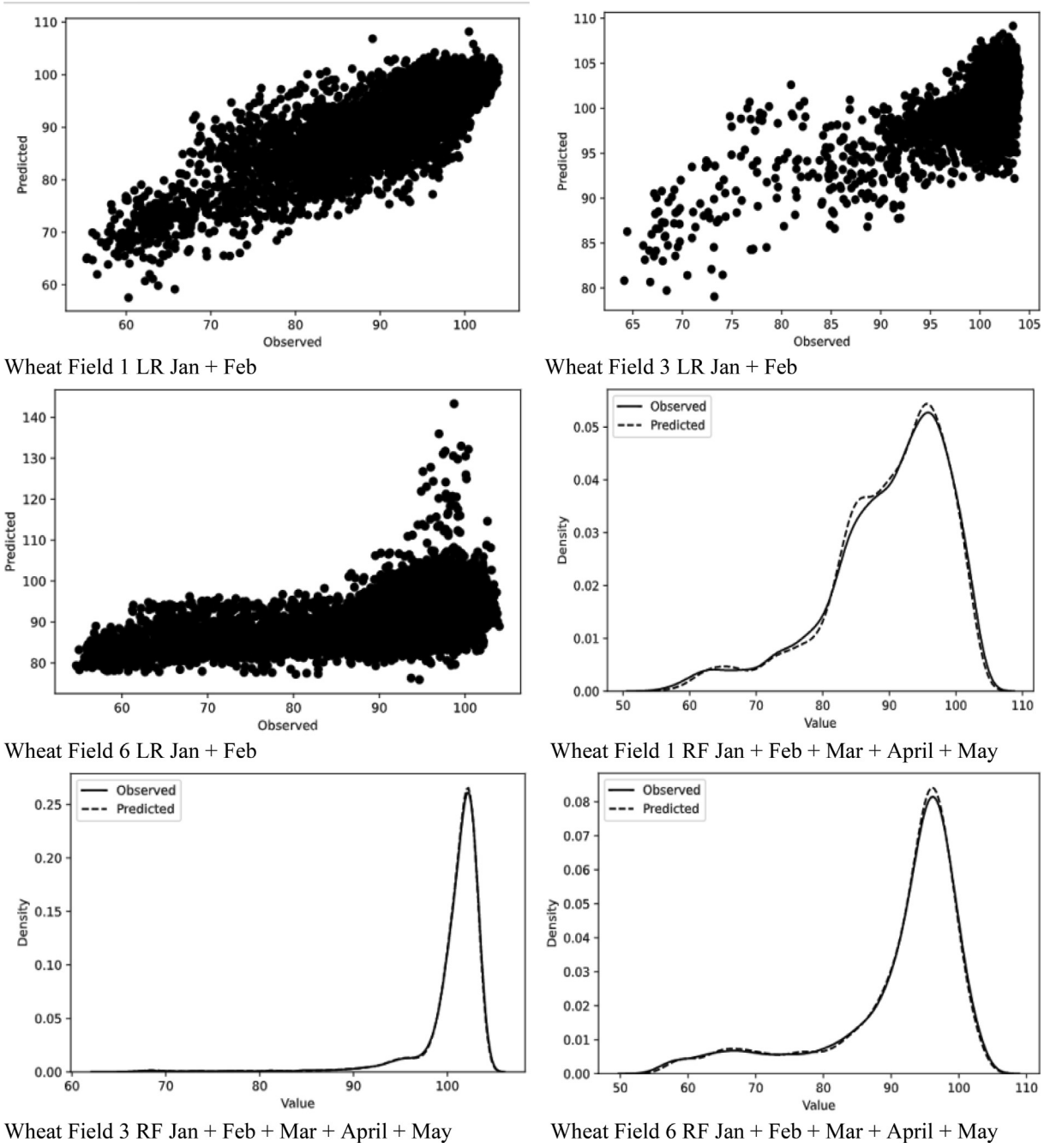

Fig. 4 (*continued*)

relatively accurate with an RMSE between 0.70 and 1 t/ha depending on the crop stage. This accuracy is variate at field level in predicated yield as in green color, when compare to observed and predicted yield. When comparing the frequency distribution between observed and predicted yield it varies in fields (Fig. 4), this is also conferred by frequency distribution trends in regression graphs.

By using some sample fields, high-resolution satellite data combined with environmental, soil, topographic, and field data can help predict yield in large landscape areas. Therefore in this research, a wide area in Fergana valley winter wheat was predicted. The wheat fields were identified by LULC maps of the study area. All six sample sites fall in training data points thus likelihood increase yield prediction accuracy. The high-resolution yield maps aid in understanding the difference and pattern of within-field and between-field yield differences. With the help of clustering, it is simple to identify highly productive and less productive fields in the same climate zone area on a colorful map. For example, in Fig. 4 field one central crop field has a dark green color which indicates higher yield prediction in comparison to other farms. The variation in this color clustering within a field must be studied further to know the exact cause of yield prediction variation. It may be due to different farming practices, used fertilizers, or environmental conditions. Aside from early yield prediction, it is useful to understand yield-limiting factors, farming practices, and management strategies to improve specific area yield production.

3.3. Machine learning

As this research work used different types of data from different sources thus machine learning support to handle this big data

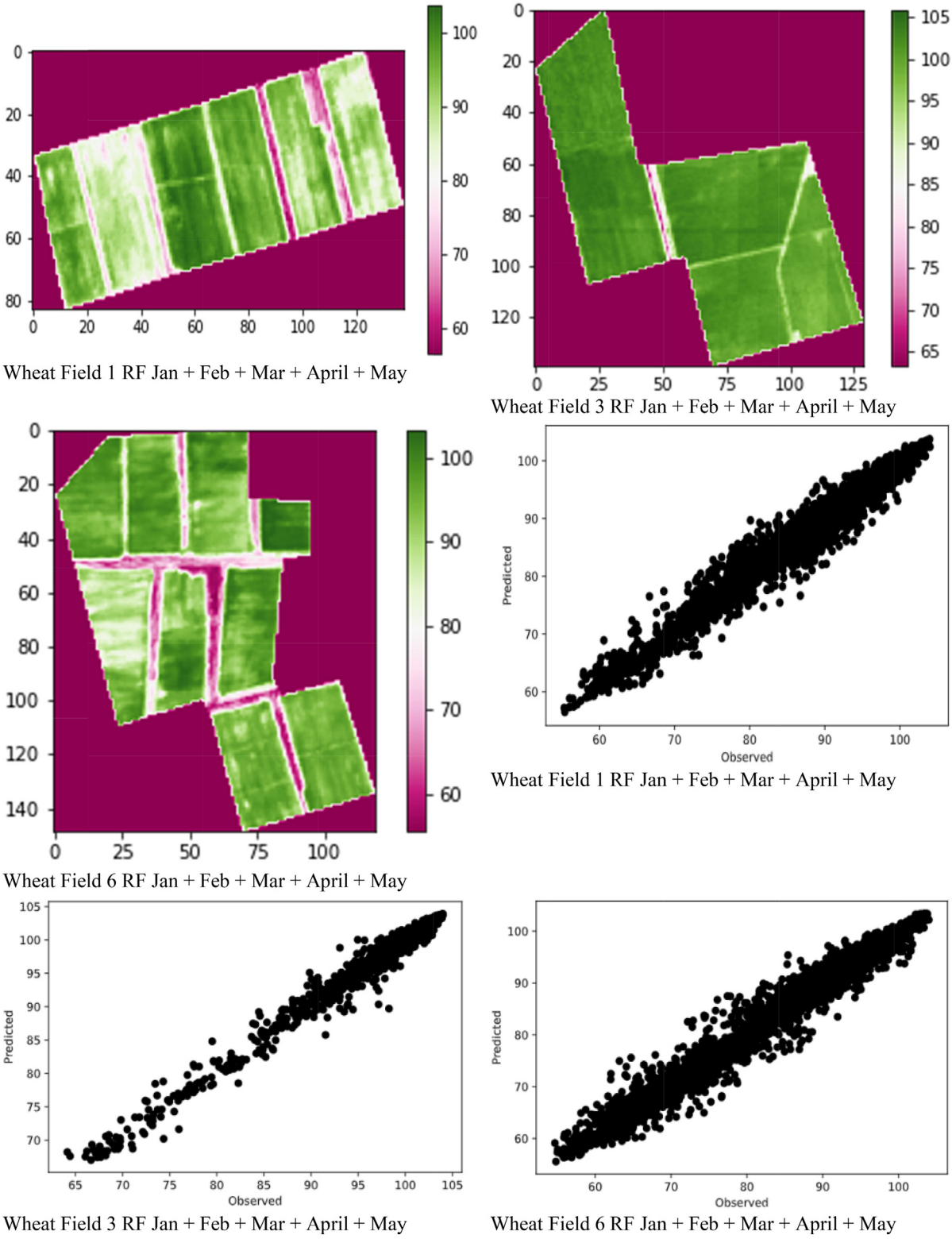

Fig. 4 (*continued*)

information in a single platform. The observed and predicted results already mentioned in the above section "regression" and the remaining few are as: In the first step machine learning provide complete details of the data with all specific indicators information such as the total number of points their maximum, minimum, mean, standard deviation values, and histogram, etc. Here histograms help to know the data values in a specific range as shown in Fig. 5.

Machine learning also supports knowing the correlation matrices of all individual indictors to predicted yield as shown in Table 5 and Fig. 6. Correlation results indicate the following metrics with observed yield.

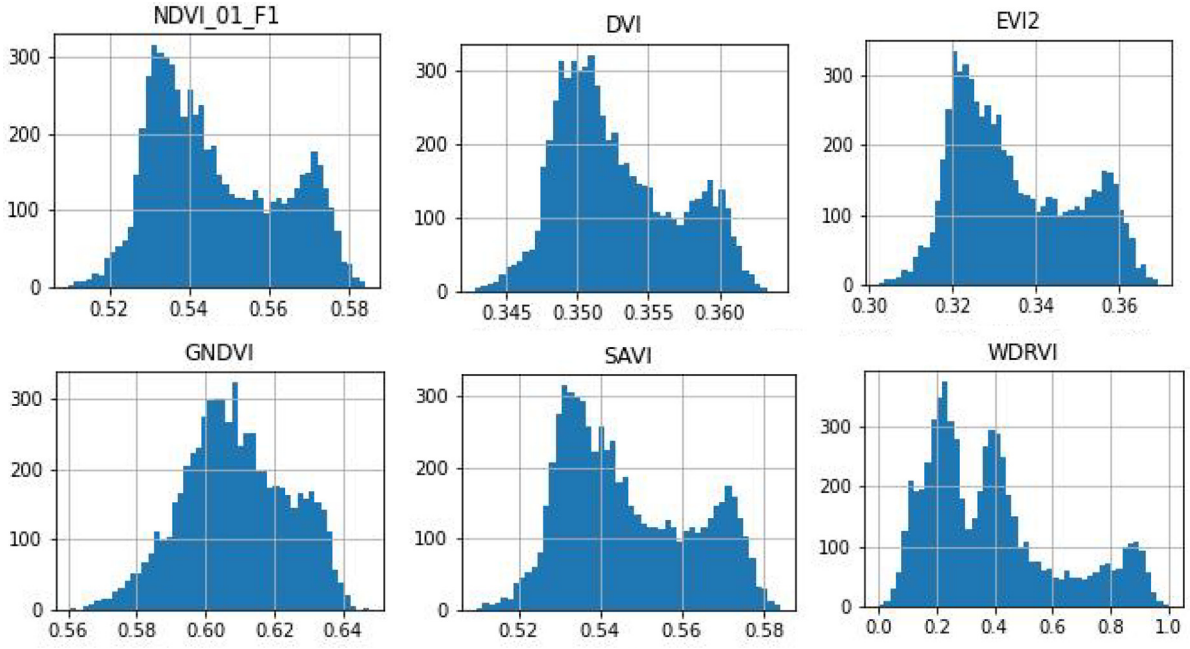

Fig. 5. Vegetation indices histograms in machine learning.

Table 5
Correlation matrices of individual indicators with yield.

NDVI	WDRVI	Soil	GNDVI	DVI	EVI2	SAVI	LAI	Water Vapor	Tem.	GPP	Solar radiation
0.99	0.98	0.86	0.82	0.80	0.78	0.76	0.75	0.73	0.71	−0.033	−0.122

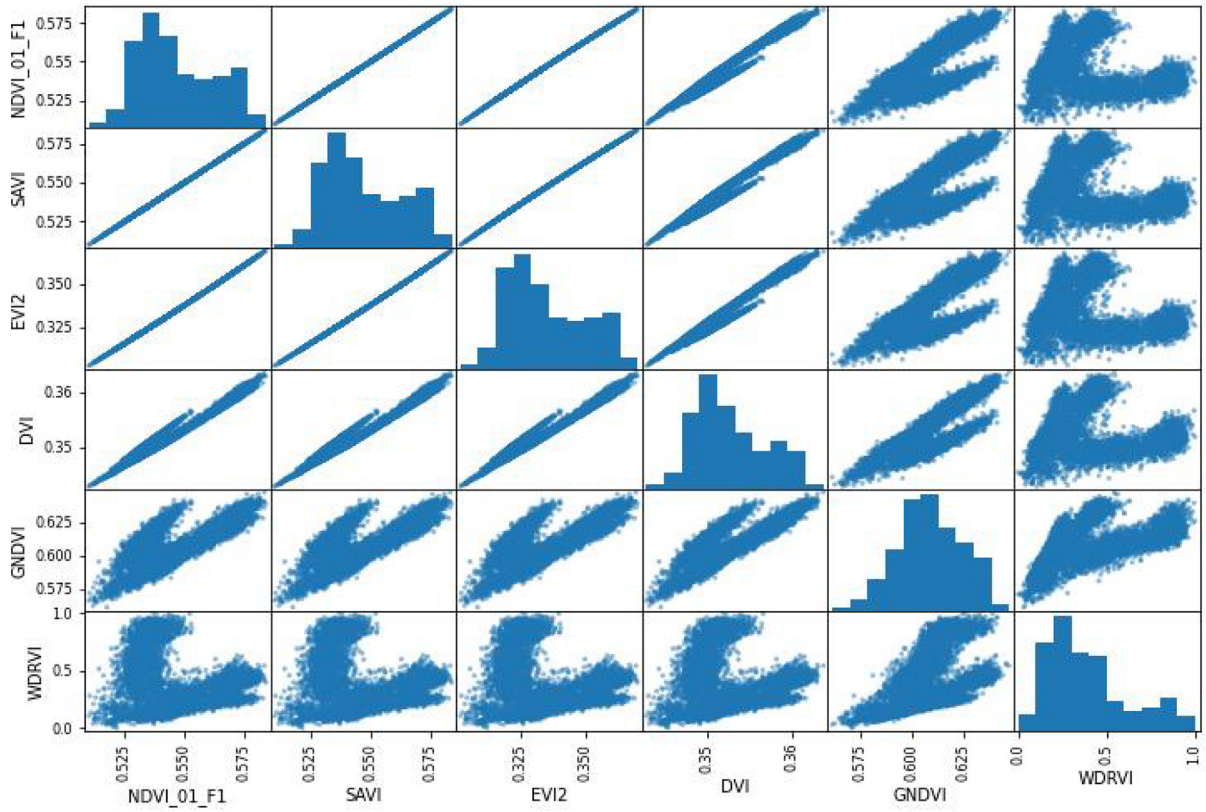

Fig. 6. Correlation in machine learning.

Table 6
Correlation matrices of individual indicators with yield.

VIs	Feature	Model	R^2	*p*
NDVI	Plant height	y = 149.81x + 3.0993	0.9628	<0.001
	Biomass	y = 1.2715x – 0.1388	0.9476	<0.001
EVI2	Plant height	y = 77.223x + 8.9022	0.9271	<0.001
	Biomass	y = 0.6749x – 0.1031	0.9675	<0.001
GNDVI	Plant height	y = 173.72x + 6.4608	0.9289	<0.001
	Biomass	y = 1.5064x – 0.1206	0.954	<0.001
SAVI	Plant height	y = 93.927x + 13.448	0.8582	<0.001
	Biomass	y = 0.8496x – 0.0783	0.9591	<0.001
DVI	Plant height	y = 0.0319x + 13.07	0.9028	<0.001
	Biomass	y = 0.0003x – 0.0737	0.9728	<0.001
GCVI	Plant height	y = 31.021x + 29.07	0.8286	<0.001
	Biomass	y = 0.2899x + 0.0529	0.9887	<0.001
SR	Plant height	y = 23.022x + 5.0181	0.8947	<0.001
	Biomass	y = 0.2073x – 0.1523	0.9914	<0.001
WDRVI	Plant height	y = -10.23x – 59.68	0.3847	<0.001
	Biomass	y = -0.0556x – 0.2979	0.155	<0.001

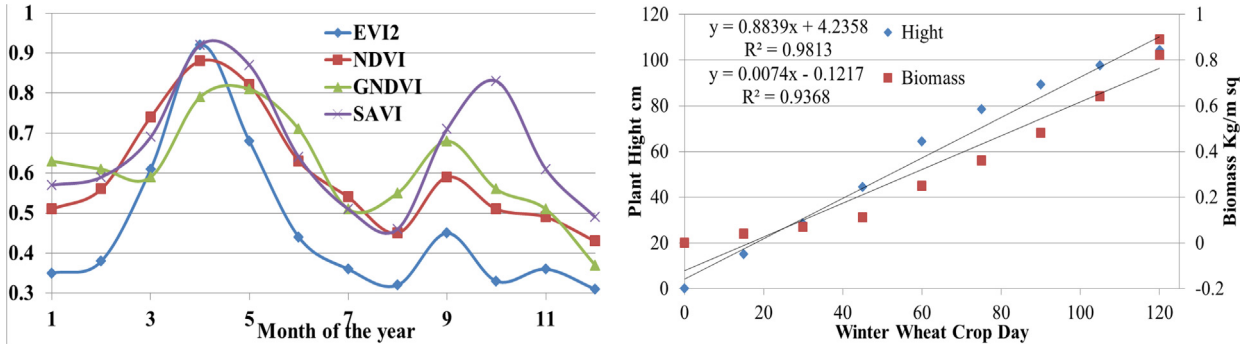

Fig. 7. VIs variation throughout the whole year (left) and linear relation in plant height, biomass, and day of the year for winter wheat crop (right).

According to Table 6, NDVI has the strongest association with yield, which means that yield increases as NDVI values rise. When temperature and VIs are compared, temperature has a lower correlation than VIs because increasing temperature reduces yield. GPP and solar radiation have a negative correlation which indicates adverse effects.

3.4. Accuracy

Instead of using only satellite data or its generated vegetation indices, the increasing number of secondary data sources improves accuracy. The accuracy of yield prediction increase with increasing crop growth and it was highest at harvest time as Table 4 show R^2 and RMSE values. The main cause of high accuracy with increasing growth stages is high amount of reflected energy to the sensor. The results were verified during field work and also take help from high resolution google earth images and earlier statistics results. The user accuracy, producer accuracy, overall accuracy and kappa statistics results show above 93 % accuracy.

4. Discussion

This study focused primarily on wheat crop height, observed biomass, VIs, and some secondary data for accurate yield prediction at various winter wheat crop growth stages. The main criteria was to analysis R^2 and RMSE values of VIs, plant height, biomass and other indicators, their relative importance in LR, DT and RF model for yield prediction (Van der Meij et al., 2017, Hoffmeister et al. 2016).

This research tries to make a combination of satellite data (Sentinel-2), vegetation indices, environmental data, soil data, and topographic and field data for yield prediction in machine learning through different regression models (Wang et al. 2021). Results indicated that RF has the highest accuracy in comparison to DT and LR regression. Results also show a correlation between VIs and field data such as plant height and biomass. Here remote sensing is a prestigious tool to develop biophysical parameters for accurate yield prediction for a large area at a field level (Wang et al. 2021).

In the comparison of all VIs, NDVI, GNDVI, SAVI, and EVI2 have the highest variation throughout the year. This variation was greater in the first half of the year (winter wheat crop) than in the second half (Impollonia et al. 2022). Winter wheat shows the highest VIs values around April-May and the lowest in August and December (Fig. 7 left). There is a great correlation between plant height and biomass with R^2 at 0.98 while with biomass R^2 is 0.93 (Fig. 7 right).

The sensitivity of all eight VIs with observed biomass or yield was variate according to vegetation density and area coverage. The linear regression R^2 values of all VIs with plant height and biomass are present in Table 6 and Fig. 8. All VIs except SR and WDRVI were positively correlated with observed biomass or yield. NDVI shows the highest R^2 value with plant height as 0.96, while WDRVI shows the lowest R^2 value for plant height as 0.38. For biomass, SR shows the highest R^2 at 0.99 and GCVI second highest with 0.98 while WDRVI has the lowest value of 0.15 respectively. Higher R^2 values indicate close to actual yield prediction or higher accuracy.

Finally, satellite data-based VIs and other secondary data combinations can be used in machine learning-based models for yield prediction in a very simple, interpretable, and acceptable manner (Zhang et al., 2022a, Tilly et al. 2015). Earlier studies indicated that plant growth stages play an important role in terms of sensitivity and model performance due to biophysical parameter variations

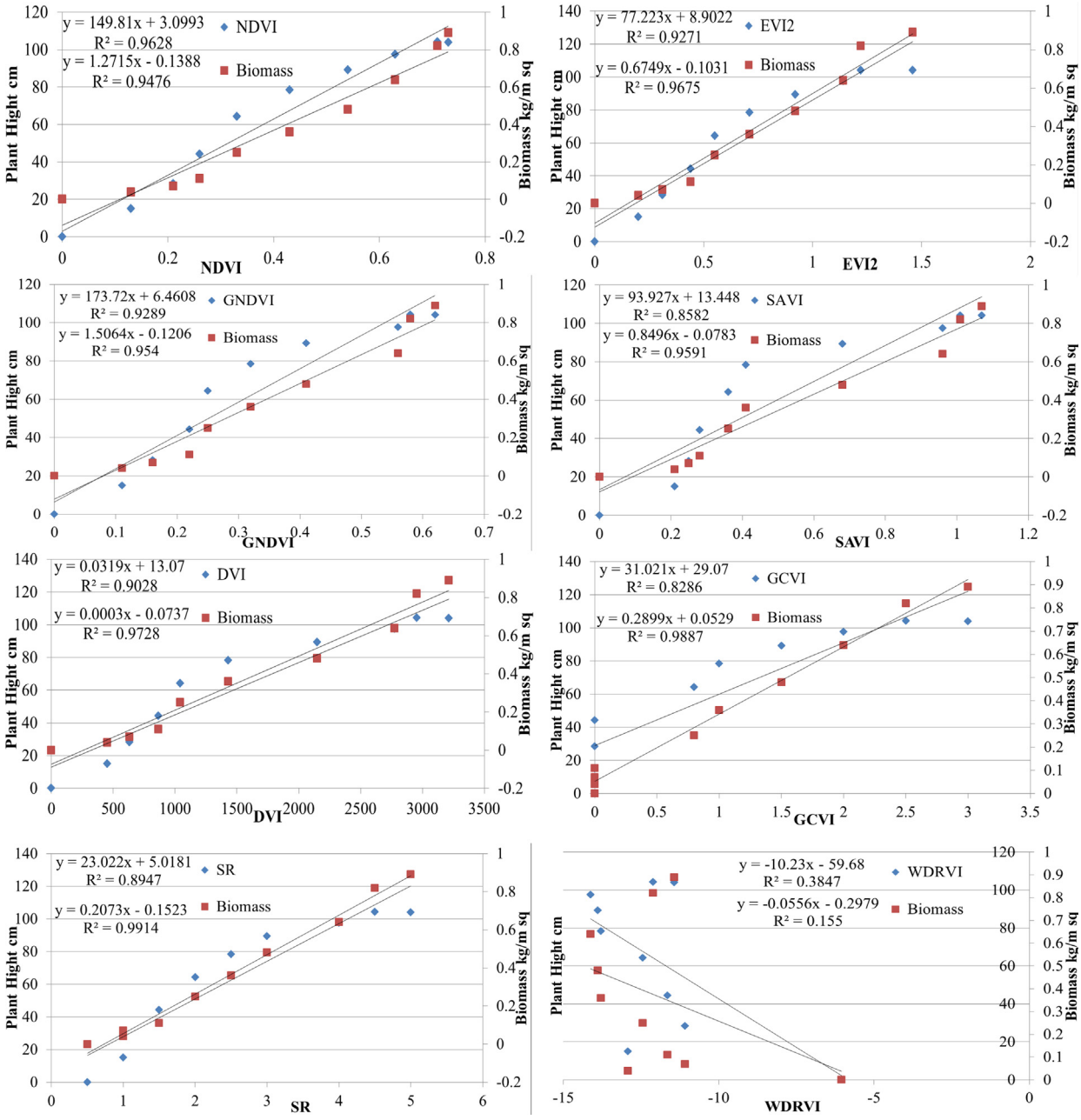

Fig. 8. Linear regression of different VIs with plant height and biomass.

(Zhang et al., 2022b). Therefore two things are important in yield prediction first, how to reduce soil effect in the early crop growth stage as its effect on reflectance. Second how to remove spectral indices saturation after canopy closure (Zhang et al., 2022b, Rivera et al. 2014). Thus this research used different VIs to overcome these problems for example soil adjusted vegetation index (SAVI) removes soil background and detects minor vegetation to make a good performance for yield prediction. Some narrow bandwidth-based VIs such as EVI2 also detect vegetation accurately even in adverse climate and soil affect and help for accurate prediction (Zhang et al., 2022a, Rusetska, 2014). Due to the wide spectrum and sensitivity of these VIs, this research also suggests that a large number of VIs with narrow bandwidth and more spectral bands aid in improving accuracy. In addition, crop height is an important parameter in VIs as the plant is growing vertically and thus has the main role in biomass prediction through VIs, especially in the late mature stage (Tilly et al. 2015, Rivera et al. 2014). In the last model, performance is based on a statistical relationship between indictors and biomass which depends on a number of fields therefore in this research work we used 45 crops farms for sampling to improve the model performance and accuracy. Moreover, yield or biomass prediction is a measuring of weight through spectral spectrum over a time thus it's also vitiated according to time therefore to improve the accuracy this parameter must be considered in future research.

5. Conclusion

This study demonstrates the successful use of machine learning on big data for yield prediction in the Fergana Valley of Central Asia. Finally, a crop prediction model was developed in which 45

crop farms at 6 sites were sampled to understand the sensitivity of VIs with plant height, biomass, and other indicators, and a significant correlation were observed in indictors and yield prediction ($P < 0.0001$). The increasing of input secondary data provides more accurate predictions. With an overall accuracy of 93 %, RF has the highest yield prediction accuracy when compared to LR, DT, and RF. The increasing numbers of field, spectral and spatial resolution also increase the performance of the model and affect accuracy. This research can be applied to other crops as well as other locations at various scales. This research demonstrates continuous mapping and monitoring of a sizable study area, finds regular crop variability throughout crop growth, and gives precise yield predictions that calls for food safety, security, sustainable development, and societal harmony. Further research is needed with more spectral features, data, and algorithms to improve the prediction accuracy which is required for precision agriculture.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant # 0777-2020-0017) and, was partially funded by RFBR, project number # 20-51-05008.

References

Abdullaev, I., Kazbekov, J., Jumaboev, K., Manthrithilake, H., 2009. Adoption of integrated water resources management principles and its impacts: lessons from Fergana Valley. Water Int. 34 (2), 230–241.

Ahmad, U., Nasirahmadi, A., Hensel, O., Marino, S., 2022. (2022), Technology and Data Fusion Methods to Enhance Site-Specific Crop Monitoring. Agronomy 12, 555. https://doi.org/10.3390/agronomy12030555.

Ahn, J., Briers, G., Baker, M., Price, E., Djebou, D.C.S., Strong, R., Piña, M., Kibriya, S., 2022. Food security and agricultural challenges in West-African rural communities: a machine learning analysis. Int. J. Food Prop. 25 (1), 827–844. https://doi.org/10.1080/10942912.2022.2066124.

Ajami, M. et al., 2020. Spatial Variability of Rainfed Wheat Production Under the Influence of Topography and Soil Properties in Loess-Derived Soils, Northern Iran. Int. J. Plant Prod., 1–12

Ali, A., Martelli, R., Lupia, F., Barbanti, L., 2019. (2019), Assessing Multiple Years' Spatial Variability of Crop Yields Using Satellite Vegetation Indices. Remote Sens. 11, 2384. https://doi.org/10.3390/rs11202384.

Azadi, H., Moghaddam, S.M., Burkart, S., Mahmoudi, H., Van Passel, S., Kurban, A., Lopez-Carr, D., 2021. Rethinking resilient agriculture: From Climate-Smart Agriculture to Vulnerable-Smart Agriculture. J. Cleaner Prod. 319,. https://doi.org/10.1016/j.jclepro.2021.128602 128602.

Azimi, M., Eslamlou, A.D., Pekcan, G., 2020. (2020), Data-Driven Structural Health Monitoring and Damage Detection through Deep Learning: State-of-the-Art Review. Sensors 20, 2778. https://doi.org/10.3390/s20102778.

Besalatpour, A.A., Ayoubi, S., Hajabbasi, M.A., Jazi, A.Y., Gharipour, A., 2014. Feature selection using parallel genetic algorithm for the prediction of geometric mean diameter of soil aggregates by machine learning methods. Arid Land Res. Manage. 28 (4), 383–394.

Bhat, S.A., Huang, N.-F., 2021. (2021), "Big Data and AI Revolution in Precision Agriculture: Survey and Challenges,". IEEE Access 9, 110209–110222. https://doi.org/10.1109/ACCESS.2021.3102227.

Bichsel, C., 2009. Conflict transformation in Central Asia. Irrigation Disputes in the Fergana Valley. Routledge, London (Central Asian studies series) 14.

Boori, M.S., Choudhary, K., Kupriyanov, A., 2020a. Crop growth monitoring through Sentinel and Landsat data based NDVI time-series. J. Comput. Opt. 44 (3), 409–418 https://doi.org/10.18287/2412-6179-CO-635.

Boori, M.S., Choudhary, K., Kupriyanov, A., 2020b. Detecting vegetation drought dynamic in European Russia. Geocarto Int. (Taylor & Francis). https://doi.org/10.1080/10106049.2020.1750063.

Boori, M.S., Choudhary, K., Paringer, R., Kupriyanov, A., 2021. (2021), Eco-environmental quality assessment based on pressure-state-response framework by remote sensing and GIS. Remote Sens. Appl.: Soc. Environ. 23,. https://doi.org/10.1016/j.rsase.2021.100530 100530.

Boori, M.S., Choudhary, K., Paringer, R., Kupriyanov, A., 2022. (2022), Using RS/GIS for spatiotemporal ecological vulnerability analysis based on DPSIR framework in the Republic of Tatarstan, Russia. Ecol. Informat. 67,. https://doi.org/10.1016/j.ecoinf.2021.101490 101490.

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Tianfang, X.u., Asseng, S., Zhang, Y., You, L., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. For. Meteorol. 274, 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010.

Chao, Z., Liu, N., Zhang, P., Ying, T., Song, K., 2019. Estimation methods developing with remote sensing information for energy crop biomass: A comparative review. Biomass Bioenergy 122, 414–425. https://doi.org/10.1016/j.biombioe.2019.02.002.

Choudhary, K., Shi, W., Singh, M., Corgne, S., 2019. Agriculture Phenology Monitoring Using NDVI Time Series Based on Remote Sensing Satellites: A Case Study of Guangdong, China. Opt. Memory Neural Netw. 28 (3), 204–214. https://doi.org/10.3103/S1060992X19030093.

Choudhary, K., Shi, W., Dong, Y., 2021. Rice growth vegetation index 2 for improving estimation of rice plant phenology in costal ecosystems. Computer Optics 45 (3), 438–448 https://doi.org/10.18287/2412-6179-CO-827.

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. Adv. Space Res. https://doi.org/10.1016/j.asr.2022.06.073.

dela Torre Daniel Marc, G., Gao, Jay, Macinnis-Ng, Cate, 2021. Remote sensing-based estimation of rice yields using various models: A critical review. Geo-spatial Informat. Sci. 24 (4), 580–603. https://doi.org/10.1080/10095020.2021.1936656.

Dokoohaki, H., Gheysari, M., Mehnatkesh, A.M., Ayoubi, S., 2015. Applying the CSM-CERES-Wheat model for rainfed wheat with specified soil characteristic in undulating area in Iran. Arch. Agron. Soil Sci. 61 (9), 1231–1245.

Dubey, S.K., Gavli, A.S., Yadav, S.K., et al., 2018. Remote Sensing-Based Yield Forecasting for Sugarcane (Saccharum officinarum L.) Crop in India. J. Indian Soc. Remote Sens. 46, 1823–1833. https://doi.org/10.1007/s12524-018-0839-2.

Eklund, L., Theisen, O.M., Baumann, M., Forø Tollefsen, A., Kuemmerle, T., Østergaard Nielsen, J., 2022. Societal drought vulnerability and the Syrian climate-conflict nexus are better explained by agriculture than meteorology. Commun Earth Environ 3 (1). https://doi.org/10.1038/s43247-022-00405-w.

Feng, P., Wang, B., Liu, D.L., Qiang, Y.u., 2019. Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia. Agric. Syst. 173, 303–316. https://doi.org/10.1016/j.agsy.2019.03.015.

Fritz, S., See, L., Bayas, J.C.L., Waldner, F., Jacques, D., Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., Crutchfield, J., Rembold, F., Rojas, O., Schucknecht, A., Van der Velde, M., Verdin, J., Bingfang, W.u., Yan, N., You, L., Gilliams, S., Mücher, S., Tetrault, R., Moorthy, I., McCallum, I., 2019. A comparison of global agricultural monitoring systems and current gaps. Agric. Syst. 168, 258–272. https://doi.org/10.1016/j.agsy.2018.05.010.

Gitelson, A.A., 2004. Wide dynamic range vegetation index for remote quantification of crop biophysical characteristics. J. Plant Physiol. 161, 165–173.

Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N., 1996. Use of a green channel in remote sensing of global vegetation from EOS- MODIS. Remote Sens. Environ. 58 (3), 289–298. https://doi.org/10.1016/S0034-4257(96)00072-7.

Gitelson, A.A., Viña, A., Arkebauer, T.J., Rundquist, D.C., Keydan, G., Leavitt, B., 2003. Remote estimation of leaf area index and green leaf biomass in maize canopies: REMOTE ESTIMATION OF LEAF AREA INDEX. Geophys. Res. Lett. 30 (5), n/a–n/a.

Guo, Y., Wang, H., Wu, Z., Wang, S., Sun, H., Senthilnath, J., Wang, J., Robin Bryant, C., Fu, Y., 2020. (2020), Modified Red Blue Vegetation Index for Chlorophyll Estimation and Yield Prediction of Maize from Visible Images Captured by UAV. Sensors 20, 5055. https://doi.org/10.3390/s20185055.

Hamid, F., Yazdanpanah, M., Baradaran, M., Khalilimoghadam, B., Azadi, H., 2021. Factors affecting farmers' behavior in using nitrogen fertilizers: society vs. farmers' valuation in southwest Iran. J. Environ. Plann. Manage. 64 (10), 1886–1908. https://doi.org/10.1080/09640568.2020.1851175.

Hoffmeister, D., Waldhoff, G., Korres, W., Curdt, C., Bareth, G., 2016. Crop height variability detection in a single field by multi-temporal terrestrial laser scanning. Precision Agric 17 (3), 296–312.

Huang, Y., Zhong-xin, C.H.E.N., Tao, Y.U., Xiang-zhi, H.U.A.N.G., Xing-fa, G.U., 2018. Agricultural remote sensing big data: Management and applications. J. Integrat. Agric. 17 (9), 1915–1931. https://doi.org/10.1016/S2095-3119(17)61859-8.

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25 (3), 295–309.

Impollonia, G., Croci, M., Ferrarini, A., Brook, J., Martani, E., Blandinières, H., Marcone, A., Awty-Carroll, D., Ashman, C., Kam, J., Kiesel, A., Trindade, L.M., Boschetti, M., Clifton-Brown, J., Amaducci, S., 2022. UAV Remote Sensing for High-Throughput Phenotyping and for Yield Prediction of Miscanthus by Machine Learning Techniques. Remote Sens. 14, 2927.

Jordan, C.F., 1969. Derivation of leaf-area index from quality of light on the forest floor. Ecology 50, 663–666. https://doi.org/10.2307/1936256.

Kephe, P.N., Ayisi, K.K., Petja, B.M., 2021. Challenges and opportunities in crop simulation modelling under seasonal and projected climate change scenarios for crop production in South Africa. Agric. Food Secur. 10, 10. https://doi.org/10.1186/s40066-020-00283-5.

Khanal, S., Kc, K., Fulton, J.P., Shearer, S., Ozkan, E., 2020. Remote Sensing in Agriculture—Accomplishments, Limitations, and Opportunities. Remote Sens. 12 (22), 3783.

Liu, L., Xiao, X., Qin, Y., Wang, J., Xinliang, X.u., Yueming, H.u., Qiao, Z., 2020. Mapping cropping intensity in China using time series Landsat and Sentinel-2 images and Google Earth Engine. Remote Sens. Environ. 239,. https://doi.org/10.1016/j.rse.2019.111624 111624.

Löw, F., Biradar, C., Fliemann, E., Lamers, J.P.A., Conrad, C., 2017. Assessing gaps in irrigated agricultural productivity through satellite earth observations—A case study of the Fergana Valley, Central Asia. Int. J. Appl. Earth Obs. Geoinf. 59 (2017), 118–134.

Maftouh, A., El Fatni, O., Fayiah, M., Liew, R.K., Lam, S.S., Bahaj, T., Butt, M.H., 2022. The application of water–energy nexus in the Middle East and North Africa (MENA) region: a structured review. Appl Water Sci 12 (5). https://doi.org/10.1007/s13201-022-01613-7.

Naimi, S. et al., 2021. Spatial Prediction of Soil Surface Properties in an Arid Region Using Synthetic Soil Image and Machine Learning. Geocarto Int., 1–22

Norouzi, M., Ayoubi, S., Jalalian, A., Khademi, H., Dehghani, A.A., 2010. Predicting rainfed wheat quality and quantity by artificial neural network using terrain and soil characteristics. Acta Agric. Scand. Sect. B-Soil Plant Sci. 60 (4), 341–352.

Obwocha, E.B., Ramisch, J.J., Duguma, L., Orero, L., 2022. The Relationship between Climate Change, Variability, and Food Security: Understanding the Impacts and Building Resilient Food Systems in West Pokot County, Kenya. Sustainability 14, 765. https://doi.org/10.3390/su14020765.

Ostaev, G., Shulus, A., Mironova, M., Smolin, Y., 2020. Accounting agricultural business from scratch: management accounting, decision making, analysis and monitoring of business processes. Amazonia Invest. 9 (27), 319–332 https://doi.org/10.34069/AI/2020.27.03.35.

Palanivel, Kodimalar, Surianarayanan, Chellammal, 2019. An Approach for Prediction of Crop Yield Using Machine Learning and Big Data Techniques (2019). Int. J. Comput. Eng. Technol. 10 (3), 110–118.

Pan, L., Xia, H., Zhao, X., Guo, Y., Qin, Y., 2021. (2021), Mapping Winter Crops Using a Phenology Algorithm, Time-Series Sentinel-2 and Landsat-7/8 Images, and Google Earth Engine. Remote Sens. 13, 2510. https://doi.org/10.3390/rs13132510.

Poppiel, R.R., Demattê, J.A.M., Rosin, N.A., Campos, L.R., Tayebi, M., Bonfatti, B.R., Ayoubi, S., Tajik, S., Afshar, F.A., Jafari, A., Hamzehpour, N., Taghizadeh-Mehrjardi, R., Ostovari, Y., Asgari, N., Naimi, S., Nabiollahi, K., Fathizad, H., Zeraatpisheh, M., Javaheri, F., Doustaky, M., Naderi, M., Dehghani, S., Atash, S., Farshadirad, A., Mirzaee, S., Shahriari, A., Ghorbani, M., Rahmati, M., 2021. High resolution middle eastern soil attributes mapping via open data and cloud computing. Geoderma 385, 114890.

Richardson, A.J., Weigand, C.L., 1977. Distinguishing vegetation from soil background information. Photogramm. Eng. Rem. Sens. 43 (12), 1541–1552.

Rivera, J.P., Verrelst, J., Delegido, J., Veroustraete, F., Moreno, J., 2014. On the Semi-Automatic Retrieval of Biophysical Parameters Based on Spectral Index Optimization. Remote Sens. 6, 4927–4951. https://doi.org/10.3390/rs6064927.

Rouse, J.W., Hass, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the great plains with ERTS. In: Third Earth Resour. Technol. Satell. Symp. 1. pp. 309–317 (doi:citeulike-article-id:12009708).

Rusetska, Uliana, 2014. How could agricultural trade between Ukraine and the EU benefit from institutional harmonization? A regression discontinuity approach. Second cycle, A2E. Uppsala: SLU, Dept. of Economics.

Saiz-Rubio, V., Rovira-Más, F., 2020. From Smart Farming towards Agriculture 5.0: A Review on Crop Data Management. Agronomy 10, 207. https://doi.org/10.3390/agronomy10020207.

Savoskul, O.S., Chevnina, E.V., Perziger, F.I., Vasilina, L.Y., Baburin, V.L., Danshin, A.I., Matyakubov, B., Murakaev, R.R., 2003. Water, climate food, and environment in the Syrdarya basin. Contribution to the project ADAPT, Adaptation strategies to changing environments.

Schulz, M.-A., Yeo, B.T.T., Vogelstein, J.T., Mourao-Miranada, J., Kather, J.N., Kording, K., Richards, B., Bzdok, D., 2020. (2020), Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. Nat. Commun. 11 (1). https://doi.org/10.1038/s41467-020-18037-z.

Seyedzadeh, S., Rahimian, F.P., Oliver, S., Rodriguez, S., Glesk, I., 2020. Machine learning modelling for predicting non-domestic buildings energy performance: A model to support deep energy retrofit decision-making. Appl. Energy 279,. https://doi.org/10.1016/j.apenergy.2020.115908 115908.

Sishodia, R.P., Ray, R.L., Singh, S.K., 2020. (2020), Applications of Remote Sensing in Precision Agriculture: A Review. Remote Sens. 12, 3136. https://doi.org/10.3390/rs12193136.

Sisodia, P.S., Tiwari, V., Kumar, A., 2014. Analysis of Supervised Maximum Likelihood Classification for remote sensing image. *International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)*, pp. 1-4. doi: 10.1109/ICRAIE.2014.6909319.

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., Scholten, T., 2020. Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. Remote Sens. 12 (7), 1095.

Tajik, S., Ayoubi, S., Zeraatpisheh, M., 2020. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. Geoderma Reg. 20, e00256.

Tilly, N., Aasen, H., Bareth, G., 2015. Fusion of Plant Height and Vegetation Indices for the Estimation of Barley Biomass. Remote Sens. 7, 11449–11480. https://doi.org/10.3390/rs70911449.

Van der Meij, B., Kooistra, L., Suomalainen, J., Barel, J.M., De Deyn, G.B., 2017. Remote sensing of plant trait responses to field-based plant–soil feedback using UAV-based optical sensors. Biogeosciences 14, 733–749. https://doi.org/10.5194/bg-14-733-2017.

Wang, J., Peng, J., Li, H., Yin, C., Liu, W., Wang, T., Zhang, H., 2021. Soil Salinity Mapping Using Machine Learning Algorithms with the Sentinel-2 MSI in Arid Areas, China. *Remote Sens.* 13, 305. https://doi.org/10.3390/rs13020305.

Wolanin, A., Mateo-García, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., Liangzhi, Y., Guanter, L., 2020. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. Environ. Res. Lett. 15 (2), 1748–9326. https://doi.org/10.1088/1748-9326/ab68ac.

Xie, H., Zhang, Y., Wu, Z., Lv, T., 2020. A Bibliometric Analysis on Land Degradation: Current Status, Development, and Future Directions. Land 9, 28. https://doi.org/10.3390/land9010028.

Yakupoğlu, T., Dindaroğlu, T., Rodrigo-comino, J., Cerdà, A., 2022. Stubble burning and wildfires in Turkey considering the Sustainable Development Goals of the United Nations. Eurasian J. Soil Sci. 11 (1), 66–76. https://doi.org/10.18393/ejss.993611.

Ye, H., Liang, L.e., Ye Li, G., Kim, JoonBeom, Lu, L.u., Wu, M., 2018. Machine Learning for Vehicular Networks: Recent Advances and Application Examples. IEEE Veh. Technol. Mag. 13 (2), 94–101.

Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., Finke, P., 2019. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma 338, 445–452.

Zeraatpisheh, M., Ayoubi, S., Mirbagheri, Z., Mosaddeghi, M.R., Xu, M., 2021. Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions using machine learning algorithms and environmental variables. Geoderma Reg. 27, e00440.

Zhang, J.M., Harman, M., Ma, L., Liu, Y., 2022a. Machine Learning Testing: Survey, Landscapes and Horizons. IIEEE Trans. Software Eng. 48 (1), 1–36.

Zhang, Y., Yin, P., Li, X., Niu, Q., Wang, Y., Cao, W., Huang, J., Chen, H., Yao, X., Le, Y.u., Li, B., 2022b. The divergent response of vegetation phenology to urbanization: A case study of Beijing city, China. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2021.150079, 803, (150079).