Available online at www.sciencedirect.com

ScienceDirect

Advances in Space Research 73 (2024) 4052–4066

ADVANCES IN SPACE RESEARCH (a COSPAR publication)

www.elsevier.com/locate/asr

Integrating the Sentinel-1, Sentinel-2 and topographic data into soybean yield modelling using machine learning

Khilola Amankulova [a,*], Nizom Farmonov [a], Khasan Omonov [b], Mokhigul Abdurakhimova [c], László Mucsi [a]

[a] Department of Geoinformatics, Physical and Environmental Geography, University of Szeged, Egyetem utca 2, Szeged 6722, Hungary
[b] Department of Land Resources, Cadastre and Geoinformatics, Karshi Institute of Irrigation and Agrotechnology "TIIAME" National Research University, Karshi, Uzbekistan
[c] Department of State Cadastres, Tashkent Institute of Irrigation and Agricultural Mechanization Engineers" National Research University, Tashkent, Uzbekistan

Received 4 May 2023; received in revised form 16 October 2023; accepted 20 January 2024
Available online 24 January 2024

Abstract

It is crucial to accurately and timely estimate crop yield within field variability for sustainable management and precision farming applications. Various Earth observation systems have been developed for crop monitoring and yield prediction. However, there is a need for further research that integrates multiplatform data, advances in satellite technology, and data processing to apply this knowledge to agricultural practices. The integration of satellite imagery and environmental data has been used increasingly in recent years to predict crop yields using machine learning techniques. In recent years, VIs derived from optical satellites, particularly Sentinel 2 (S2), have gained popularity, but their availability is affected by weather conditions. On the other hand, the backscatter data from Sentinel 1 (S1) is less commonly used in agriculture due to its complex interpretation and processing, but it is not influenced by the weather. This study aims to improve the accuracy of yield predictions by combining remote sensing data with environmental variables. The use of satellite data S1 and S2 was used to identify the optimal phenological period, and a training model was developed using four machine learning techniques, including Random Forest Regression (RF), K Nearest Neighbor (KNN), Multiple Linear Regression (MLR) and Decision Tree (DT). The results showed that RF provided the highest values among the four techniques. The validation process using RF demonstrated high accuracy rates, with R^2 ranging from 0.41 to 0.89, the mean square error of the root (RMSE) ranging from 0.122 to 0.224 t/ha, and the mean absolute error (MAE) ranging from 0.089 to 0.163 t/ha. The integration of satellite data S1 and S2 with topographical information may be useful for monitoring, mapping, and forecasting crop yields on small and fragmented farmlands. This approach can provide farmers, agricultural businesses, and policymakers with accurate and timely predictions of crop yield, which can facilitate decision making and provide early warnings for potential crop losses.

Keywords: Machine learning; Sentinels 1 and 2; Yield estimation; Crop phenology

1. Introduction

One of the most significant food crops in the world is soybean, which receives significant attention in the global food industry and is generally planted all over the world (Amherdt et al., 2022). Soybean crops are highly nutritious in farming systems, a source of raw materials from oil

* Corresponding author at: Egyetem utca 2, 6722 Szeged, Hungary.
E-mail address: amankulova.khilola@stud.u-szeged.hu (K. Amankulova).

https://doi.org/10.1016/j.asr.2024.01.040
0273-1177/

refineries, have a high protein content in their seeds, and have the potential to enrich the soil through symbiotic *fixation of N2,* which has significant economic benefits (Sinclair et al., 2014). Soybean crops are crucial to ensure national food security in many countries (She et al., 2020). In the case of Hungary, due to government support, both the area dedicated to soybean farming and the number of farmers has increased since 2015. Although the number of producers has increased to 5,000 ha, the production area has grown from 42,000 to 772,000 ha (Soós et al., 2022). Through technological improvements, a variety of instruments can now be installed on combine harvesters, such as a yield monitor that keeps records of the crop on a parcel using information collected by various sensors (Arslan and Colvin, 2002). Yield monitoring devices provide a novel and effective tool for zone management and field comparisons in the sector of precision agriculture (Pierce et al., 2015). Farmers can effectively plan their agricultural activities for the future growing season by using this to acquire new information by evaluating data for a specific field (Pejak et al., 2022). The recent implementation of the S2 (S2) satellite constellation by the European Space Agency (ESA) has the potential to improve the application of precision agriculture (PA) approaches, which present challenges for small and medium-sized farmers (Uribeetxebarria et al., 2023). Twin satellites (A and B) of the S2 series in particular were designed to satisfy the requirements of scientists and the agricultural industry (Segarra et al., 2020). These satellites' high-resolution images, 13 multispectral bands, and rapid revisit rates are all publicly available through the ESA's Copernicus program (accessed on March 13, 2023). Through various bands of the sensor, several VIs can be calculated. Remote sensing has been an important source of data for analyzing crop development and forecasting final yields in large regional circumstances since the 1980s. The basic point behind the correlation between VIs and yield is that canopy characteristics, such as biomass, chlorophyll content, and canopy structure, determine crop growth (Zhao et al., 2020). The vast majority of the research focused on the NDVI (Normalized Difference Vegetation Index) (Shang et al., 2015; Zhao et al., 2015). To accurately extract phenology, the Normalized Difference Vegetation Index (NDVI) is commonly implemented for monitoring crop growth conditions (Becker-Reshef et al., 2010; Saeed et al., 2017; Sehgal et al., 2011). Where the leaf area index is moderately high, the Green Normalized Difference Vegetation Index (GNDVI) is more effective in evaluating leaf chlorophyll variability (Gitelson et al., 1996). Given that it was less impacted by saturation, GNDVI provided a positive indication for a number of vegetation performance variables (Gianelle et al., 2009). To minimize the effect of spectral VIs by using red and near-infrared bands, the Soil Adjusted Vegetation Index (SAVI) is used (Qin et al., 2021). The variation in water content in plant leaves is evaluated using the Normalized Differential Water Index (NDWI) (Qin et al., 2021).

Because S2 is limited by cloud coverage, the amount of usable data available for certain areas and applications may be restricted (Uribeetxebarria et al., 2023). Using spaceborne microwave remote sensing, vegetation and soil conditions can be monitored on a range of scales. Synthetic aperture radars (SAR) produce observations with a high spatial resolution of tens of meters to monitor crops (Steele-Dunne et al., 2017). One of the key advantages of using S1 data for crop yield prediction is its ability to penetrate through clouds and obtain images regardless of weather conditions, allowing year-round monitoring of crop growth. In addition, SAR data can provide information on crop structural properties, such as canopy height, biomass, and density, which are essential factors to determine crop yield. A vertical transmit chain (V) and two parallel receive chains for the polarization of H and V (horizontal and vertical, respectively) are used by S1 C-band (5.405 GHz) SAR devices in Europe to facilitate the operation in dual polarization (VV+VH) over the land (Østergaard et al., 2011).

Machine learning (ML) techniques have become increasingly popular for yield prediction due to their ability to handle complex data and model non-linear relationships between predictor variables and crop yield. ML algorithms can learn from historical data and use that knowledge to make accurate predictions for future crop yields. With the use of these technologies, huge volumes of data collected from various sources, including satellite imaging, drones, and Internet of Things (IoT) sensors, can be processed and analyzed to produce precise and thorough predictions (Mishra et al., 2016). Supervised learning algorithms can be used to predict crop yields or identify patterns in crop growth. Other machine learning algorithms that have been used to predict yield include k closest neighbor (KNN), Decision Tree (DT), Random Forest (RF) and Multiple Linear Regression (MLR) (Obsie et al., 2020; Shao et al., 2015; Sharifi, 2021; Suominen et al., 2013).

So far, several studies have been developed to predict crop yield at different levels, for instance, Schwalbert et al., (2020) presented in their study highlights strengths, including the effective utilization of satellite and weather data, integration of multiple variables for improved forecasting, exploration of time-ordered data using Long short-term memory (LSTM), and high accuracy at the municipality level. Weaknesses include challenges in crop field detection, increased errors with early yield forecasts, data limitations that lead to squared bias, and potential regional applicability depending on data availability. Another study by Herrero-Huerta et al., (2020) developed two tree learning models, RF and eXtreme Gradient Boosting (XGBoost), for soybean yield prediction using unmanned aerial vehicle (UAV) based imagery. Strengths of ML models in this study include their accurate fitting of training data, quantitative assessment using various error metrics, and the superior performance of XGBoost compared to RF, particularly in handling overfitting. How-

ever, there is a risk of overfitting, and the models tend to exhibit underestimation at high yield values and overestimation at low values, which can be influenced by the data distribution and may require further refinement. Barbosa dos Santos et al. (2022) evaluated the response of soybeans in different irrigation supplementations and found that higher water supply resulted in increased dry matter and grain yield, leading to yield stability during the reproductive phases. The study effectively utilized thermal mapping to gain insight into how climate impacts different stages of soybean growth, enhancing the accuracy of predictive modeling. In particular, RF showed robust performance with a high R2 of 0.81 and a low RMSE, demonstrating its precision in forecasting yields. Additionally, the study's comparison of machine learning algorithms highlighted RF's superiority for similar forecasting tasks. Furthermore, the models successfully captured regional variations in yield, which is crucial for practical agricultural applications. On the other hand, the weaknesses observed include the tendency of models like SVM_RBF and SVM_POLY to underestimate yields in specific regions. The limited number of data points in certain areas may have affected the accuracy of predictions, particularly in years with extreme weather conditions. Furthermore, the study missed an opportunity to provide a broader perspective on model performance by not comparing its results with traditional forecasting methods.

Combining S1 and S2 datasets provides a more comprehensive view of the agricultural landscape, allowing better prediction of soybean yield at the pixel level. Additionally, the inclusion of light detection and classification (LiDAR) data can provide information on soil characteristics and topography, which can further improve the accuracy of yield prediction models. Machine learning techniques can be applied to these data sets to develop models that can accurately predict soybean yield, which can be used to inform agricultural management decisions and improve crop yields. Due to the importance of yield predictions in facilitating various decisions at different levels of the agroindustrial soybean value chain, and the necessity to revise yield predictions during the crop development phase. The main objective of this research was to examine the potential of SAR and multispectral satellite imagery, as well as elevation data, to predict soybean yield at the pixel level in Mezohegyes using ML techniques. To do this, the following objectives were set: (1) to develop predictive models that can determine the stage at which reliable predictions of harvest yield can be made based on soybean yield in two years (eg 2020–2021) at the pixel level; (2) to conduct a comprehensive and detailed investigation of multiple ML techniques, using data from two years to forecast soybean yield and determine the most appropriate algorithm to predict year 2022. To achieve this objective, the study will evaluate the effectiveness of various machine learning algorithms, including KNN, RF, MLR, and DT, in estimating soybean yield; (3) investigate the advantages of combining satellite imagery with elevation data in predicting soybean yield, and analyze how individual predictors impact model performance. Since yield predictions at the pixel level are most useful to farmers, this study also aimed to (4) assess how spatially averaged field-level results predict soybean yield. To achieve these objectives, remote sensing data from Copernicus S1 SAR and S2 multispectral satellites, which provide free of charge, high-resolution, and frequent imagery, were analyzed.

2. Materials and methods

2.1. Study area

The study area for this research is Mezhegyes, located in southeastern Hungary near the border with Romania (46° 19′ N, 20° 49′ E). All soybean parcels from 2020 to 2023 were included in the study area (Fig. 1). Specifically, parcels from 2020 and 2021 were used for training, while those from 2022 were used for testing, and their corresponding details are presented in Fig. 2. Mezhegyes is a town that spans 15,544 ha with a population of 4950 individuals. The soil in the meadows and lowlands is predominantly chernozem, a common soil type with high lime content that is particularly suitable for agriculture, particularly cereal and oil crops (Amankulova et al., 2021). The Mezhegyes experimental farm, operated by Mezhegyesi Ménesbirtok Zrt., is a significant contributor to agricultural activity in Mezhegyes and neighbouring communities.

2.2. Satellite imagery

S1 is equipped with a C-band radar instrument that allows it to capture images of the Earth's surface day and night, regardless of weather conditions. It uses the (SAR) to capture images, which allows it to penetrate through clouds, rain, and even vegetation. S1 provides images with a spatial resolution of up to 5 m and a revisit time of up to 12 days.

S2 is an MSI that provides high-resolution imagery of the Earth's surface. It has 13 spectral bands that allow observation of a wide range of features of land cover, including vegetation, water bodies, and urban areas. S2 provides images with a spatial resolution of up to 10 m and a revisit time of up to 5 days. Both S1 and S2 provide free and open access data, which can be accessed through various data portals such as the Copernicus Open Access Hub or the Sentinel Hub. The data from Sentinel 1 and 2 data were acquired during the soybean cultivation period between 1 April and 31 October in the years 2020, 2021, and 2022 (Table 1).

In this research, we used (SAR) data obtained from the S1 satellite in the Interferometric Wide (IW) mode of acquisition. The SAR images have a resolution of 5×20 m and a swath width of 250 km, with two polarization types (VV and VH) providing backscatter intensity information. These images were pre-processed at Level 1, resulting in complex data in the slant range that is geolo-

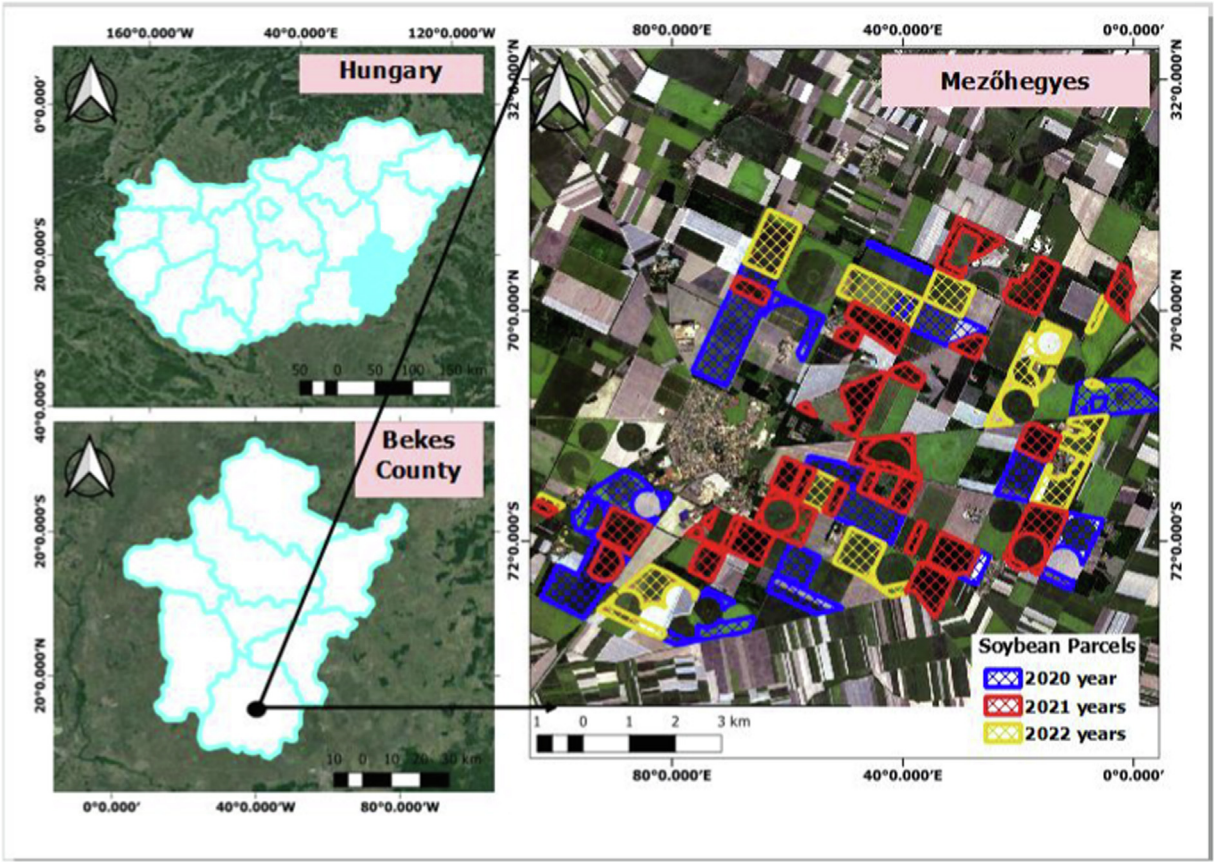

Fig. 1. The red, blue, and green colours indicate the year 2020, 2021, and 2022, respectively. The natural colour composite is based on S2 imagery, and the RGB bands used were 4, 3, and 2. The acquisition date for the image was 8 August 2021.

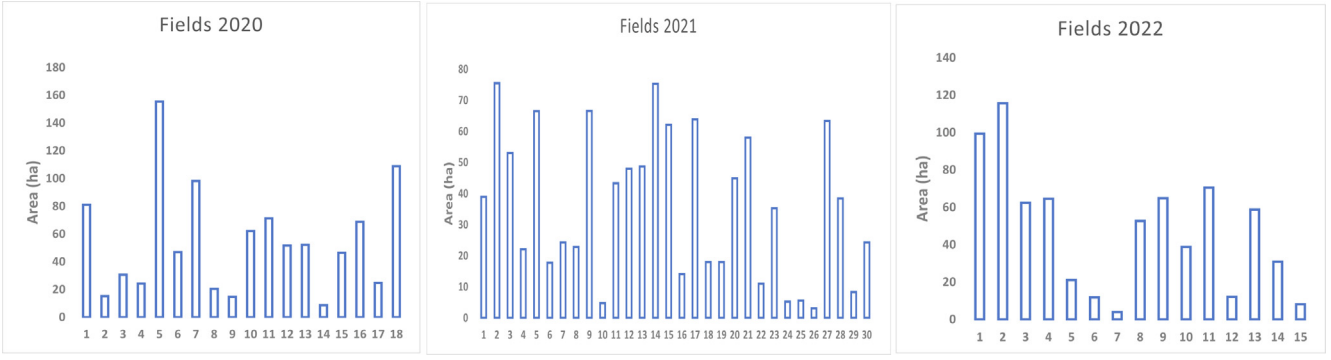

Fig. 2. Information about soybean fields for three years.

cated, radiometrically calibrated, and terrain-corrected. The images obtained were processed using Sentinel Application Platform (SNAP) version 8.0 software, developed by the European Space Agency (ESA), to make them suitable for further analysis. This involved adjusting the size of the image tiles to match the study area and obtaining precise orbit information by applying orbit files, since the metadata provided with the radar products are often insufficiently accurate. In addition, steps were taken to enhance image quality by eliminating thermal noise and radiometric artefacts from the edge edges of the image, calibrating the images for radiometrically calibrated backscatter, and removing the granular noise caused by backscatter from certain elements. The images were then assigned geographical coordinates and the backscatter values were converted to decibels in the final step. For S1, VV/VH was calculated (Veloso et al., 2017).

$$VV/VH = VV - VH \tag{1}$$

The S2 images used in this study were resampled at a 10 m resolution using the SNAP software after initially being obtained at varying pixel sizes. Fields in the study

Table 1
S1 and S2 imagery numbers for each growing season for three years.

Year	Month	Sentinel-1	Sentinel-2
2020	April	3	2
2021		4	3
2022		4	2
2020	May	4	4
2021		5	6
2022		5	5
2020	June	6	6
2021		6	5
2022		6	6
2020	July	4	5
2021		5	4
2022		5	6
2020	August	6	6
2021		6	5
2022		6	6
2020	September	4	4
2021		4	5
2022		3	4
2020	October	2	2
2021		2	2
2022		2	2

area were identified using an official crop plan map as a mask layer in QGIS 3.16. To identify the green peak soybean phenological stage, we generated averaged mosaics of S1 and S2 images for each month by computing their average values and a box plot was generated by computing minimum, maximum, mean, median, and standard deviation statistics. A box plot was then created to present the findings (Fig. 3). To obtain information about S2, the NDVI values were calculated (Tucker, 1979) and the statistical range from minimum to maximum was determined for each month (Fig. 4).

The analysis of the data revealed that August had a high indicator value in each table. This indicates that the best indicator to reflect the value of mosaic bands and indices, from minimum to maximum, is in August. The high indicator value in August suggests that it is the peak phenological period for soybeans and therefore the most suitable time for crop yield monitoring. Bolton and Friedl, 2013 demonstrates that considering crop phenology, especially the timing of peak vegetation index, enhances crop yield predictions. Identify specific days after greenup, varying for different crops like maize and soybeans, as optimal for yield prediction. This highlights the importance of timing the highest vegetation index values for accurate crop yield predictions. To demonstrate this, we conducted an experiment in which we used data from two satellites to make monthly predictions, and the results substantiated our approach (See Fig. 5).

Therefore, these results highlight the importance of using satellite imagery and machine learning techniques to monitor crop growth during the peak phenological period, especially in August, to ensure better crop yield and management. Specifically, we mosaicked the S1 and 2 images from each month to determine the phenological stage of soybeans. The process of selecting the greenest pixel composite is a technique used to create temporal mosaicking of satellite imagery. To incorporate temporal data and accommodate various stages of growth of soybean crops, we generated composite images by combining Sentinel-1 and Sentinel-2 data throughout the growing seasons from April to October. For Sentinel-1 data, we created monthly mosaics by averaging the images within each month (Shendryk et al., 2021). This method aims to choose the image captured under the least cloudy conditions and to reduce any discrepancies in vegetation phenology (Bey et al., 2020). To ensure that the spatial resolution of the S2 images matched that of the model development, a grid rectangle (eg polygon) was created at 10 × 10 m to extract pixel values. This involved combining multiple images to create a larger and more complete image of the soybean field at each stage of growth. In addition, we calculated environmental data to include in our model, such as aspect, slope, and TWI using QGIS 3.16 software. Overview of the methodology adopted for the soybean yield prediction procedures given in workflow form (Fig. 6).

2.3. Environmental data

The study area was mapped using LiDAR technology to create a high-resolution digital terrain model (DTM) with a spatial resolution of 5 cm. The DTM was derived from radar data collected during the airborne campaign in 2019. To ensure compatibility with S2′s spatial resolution, the data were rescaled to 10 m using the cubic convolution method in ERDAS IMAGINE 2020 software. The rescaled data was used to calculate slope and aspect, secondary variables used as input parameters in the estimation model (Farmonov et al., 2023). The Topographic Wetness Index (TWI) is a measure of topographic control of the water flow and the water storage potential in a terrain. It is a function of the accumulation of slope and flow, and it characterizes the degree of topographic convergence or divergence in a given area. TWI is a useful tool for predicting hydrological processes, such as soil moisture, groundwater recharge, and runoff generation. TWI is a topographic index that characterizes the pattern of water accumulation pattern across the landscape (Qin et al., 2011; Silva and Alexandre, 2005) and is known to be correlated with crop yield (Maestrini and Basso, 2018; Silva and Alexandre, 2005). In the case of LIDAR data, the TWI can be calculated by first generating a Digital Elevation Model (DEM) from the LIDAR data, then computing the flow direction and flow accumulation grids from the DEM using a hydrological model, and finally applying the TWI equation, which involves dividing the natural logarithm of flow accumulation by the slope of the terrain.

2.4. Field data

High-resolution soybean yield data for three years (2020, 2021 and 2022) were collected using a GPS-

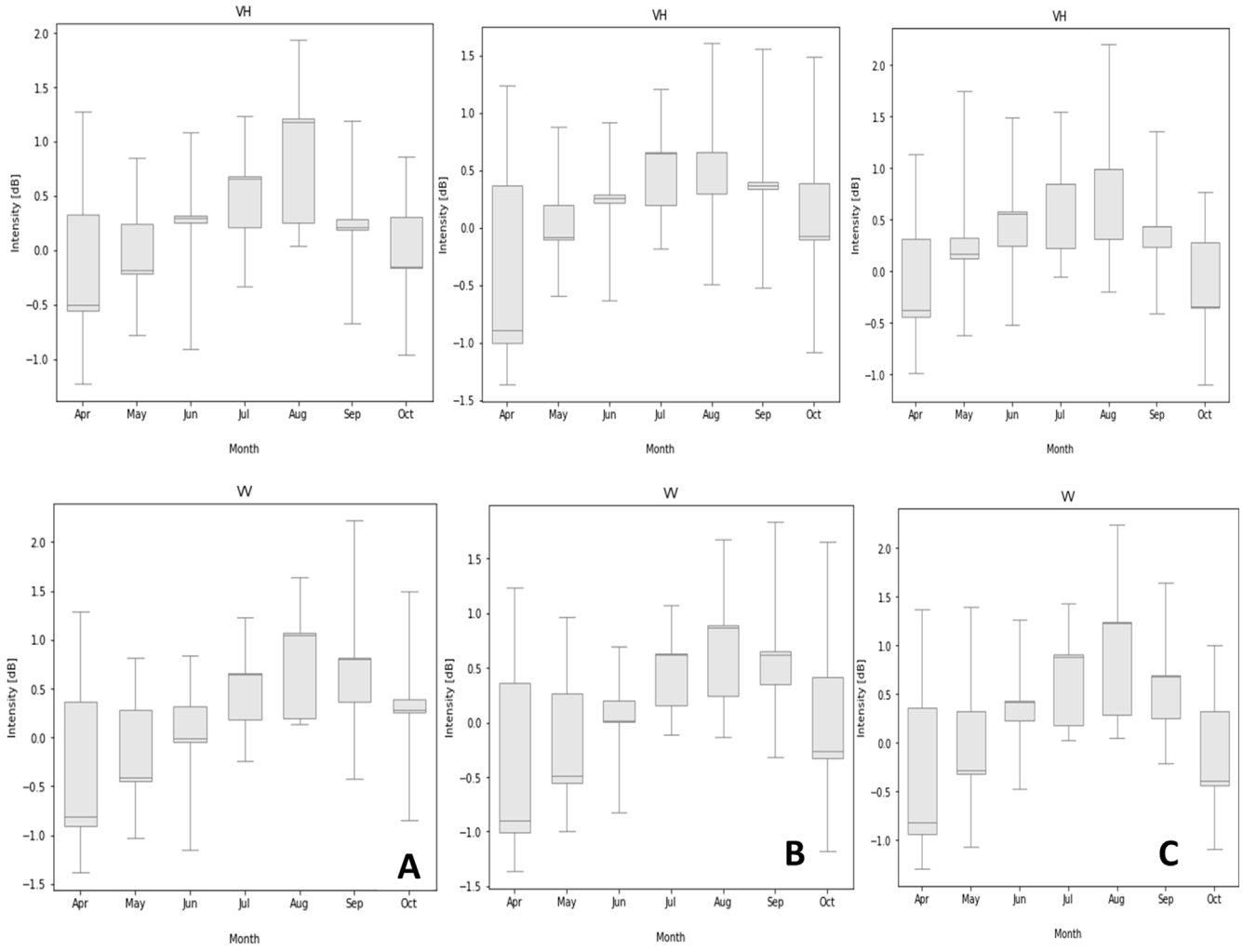

Fig. 3. Boxplots were generated for each month in 2020 (A), 2021 (B), and 2022 (C). The data is sourced from S1 mosaics, and VH values are represented in the upper layer, while VV values are in the lower layer.

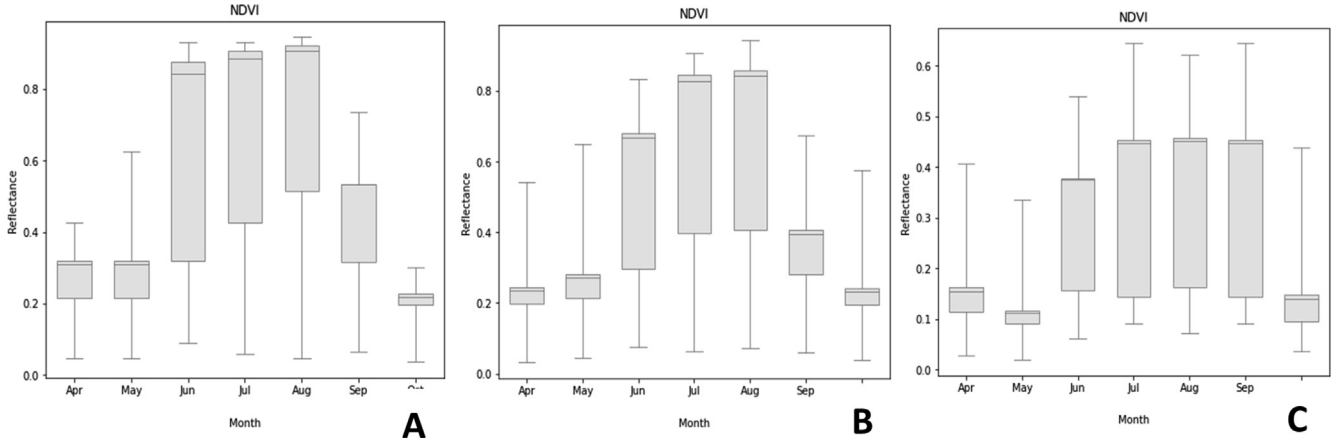

Fig. 4. Boxplot displaying NDVI values from April to October for the years 2020, 2021, and 2022 was created using S2 mosaic imagery.

equipped combine harvester. In Hungary, soybeans are typically planted in April and harvested between September and October. To eliminate biases caused by combine harvester dynamics and positioning data inaccuracy, raw yield data were cleaned according to the method proposed by Lyle et al. (2014). Crop yield data were adjusted and filtered to remove incorrect data caused by overlapping crop rows, resulting in a linear sequence of near-zero productiv-

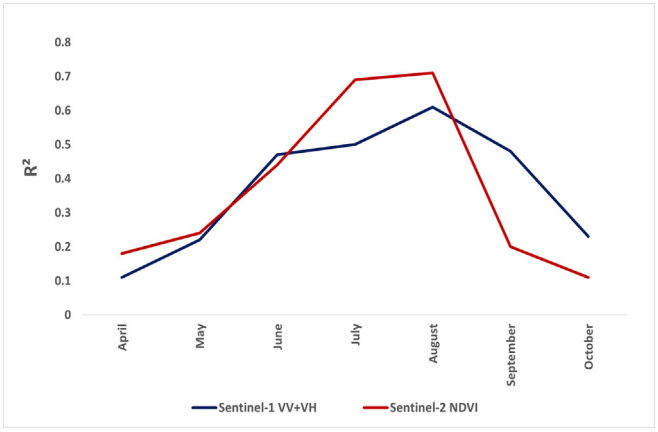

Fig. 5. Seven-Month Yield Prediction Time-Series Analysis Using Sentinel-1 VV+VH and Sentinel-2 NDVI.

ity areas. The company involved in agriculture in the study area provided the yield data, which were adjusted to match the head dimensions of the harvester (2 m × 6 m) and converted to raster format using the inverse distance-weighted (IDW) interpolation method of QGIS v.3.16 with 10 m × 10 m pixels to match the resolution of the satellite images. Response variables for yield prediction models were obtained using RS-derived VIs, VV/VH bands, LIDAR data and their combinations. Fishnet grid polygons with dimensions of 60 × 30 m were created to accurately predict yield (Fig. 7), which contains S1 and 2 pixels. The average S1 bands, VIs, LIDAR data, and crop yield values were calculated for the corresponding grids.

2.5. VIs

VIs are commonly used to assess vegetation health and productivity (Table 2). In this study, we used four different VIs. The indices were chosen on the basis of their ability and potential to capture crop growth dynamics. Although there are various VIs available, the Normalized Difference Vegetation Index (NDVI) has been the focus of many studies (Zhao et al., 2015; Shang et al., 2015). The Green Nor-

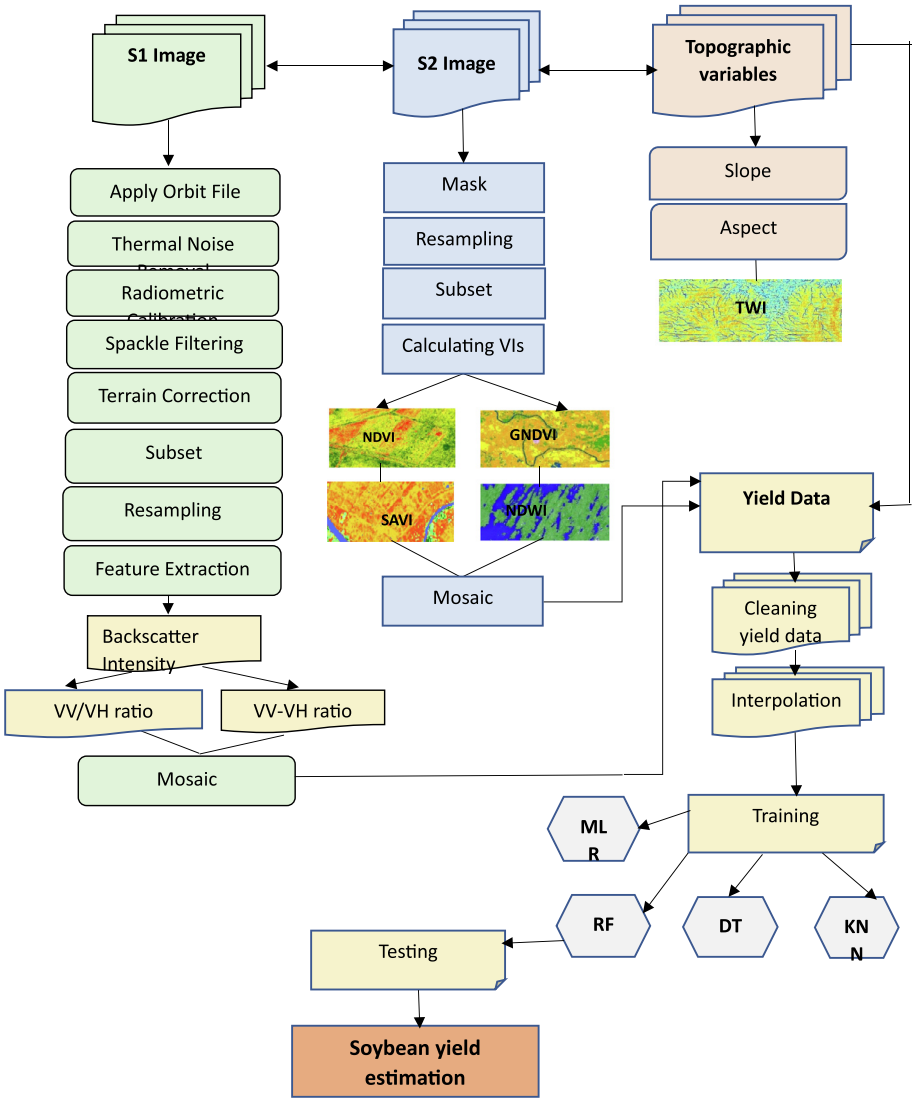

Fig. 6. Schematic diagram of workflow in this study.

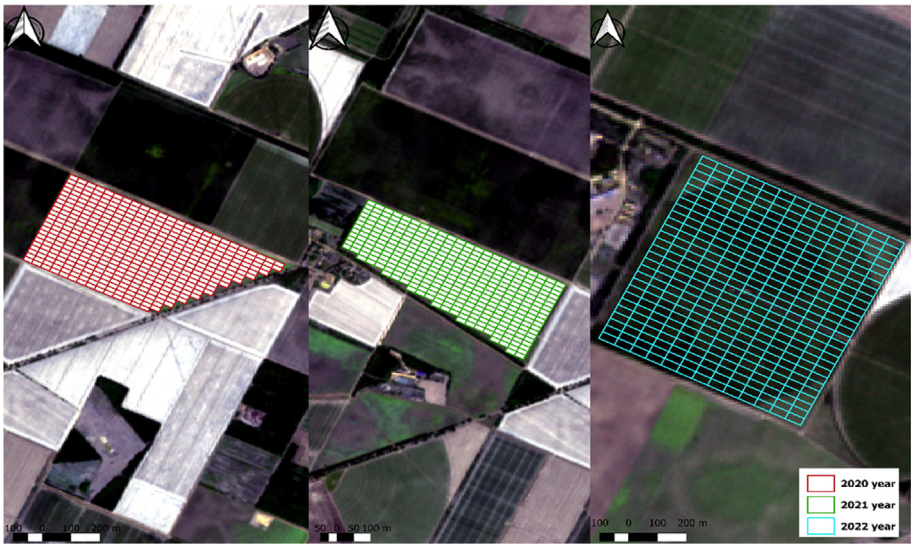

Fig. 7. Fishnet polygons were created to define the field boundaries at the pixel level for predicting crop yield for three years.

Table 2
Definition of the vegetation indices used in the study.

Index	Equation	Reference
Normalized difference vegetation index (NDVI)	$\frac{NIR-Red}{NIR+Red}$	(Rouse et al., 1973)
Green normalized difference vegetation index (GNDVI)	$\frac{NIR-Green}{NIR+Green}$	(Gitelson et al., 1996)
Soil-adjusted vegetation index (SAVI)	$(1+L)\frac{(NIR-Red)}{(NIR+Red+L)}$	(Huete, 1988)
Normalized difference water index (NDWI)	$\frac{NIR-SWIR}{NIR+SWIR}$	(McFEETERS, 1996)

malized Difference Vegetation Index (GNDVI) is a modified version of NDVI that replaces the red band with the green band (Gitelson et al., 1996). This change may be more advantageous in evaluating changes in green biomass at the canopy level. To account for soil background effects that can affect the reflectance of crop canopies, the Soil Adjusted Vegetation Index (SAVI) was developed (Huete, 1988). SAVI has been applied for the prediction of total biomass and crop yield (Elwadie et al., 2005; Panda et al., 2010). It involves an adjustment factor (L) in the NDVI equation that removes soil noise, the value of L being dependent on the density of the vegetation. NDWI was developed to detect water content in vegetation and is more sensitive to water stress in plants, making it more effective in capturing the impact of drought on crop yields (Gu et al., 2008, 2007).

2.6. Machine learning algorithms

The ML algorithms used in this study are RF, MLR, DT, and KNN. The study used the Scikit-learn library (Pedregosa et al., 2011) to search for optimal machine learning pipelines for each response variable, using randomly generated pipelines. The study also employed an ensemble algorithm called Random Forest (RF), which uses multiple decision trees to make predictions. RF works by creating a large number of decision trees and combining their predictions through methods like averaging or majority voting (Breiman, 2001). This approach helps to reduce overfitting and variance issues commonly associated with single decision tree models. RF is capable of handling high-dimensional and correlated features and can be used for classification and regression tasks (Tin Kam Ho, 1995). It also provides an estimate of feature importance, which is beneficial for feature selection and understanding of the underlying relationships in the data. Optimizing the number of regression trees (ntree) and the selection of different predictors at each leaf node (mtry) is necessary for the implementation of the RF algorithm (Dewi et al., 2019). This study performed a grid search optimization of these parameters using Python 3.11.3 version with the Scikit-learn (sklearn) package. The ntree values were tested from 50 to 500 at intervals of 50, while the mean values were tested from 5 to 100. The optimal result was achieved by setting the value at 500 and selecting the default value of mtry, which is calculated as the total number of predictors divided by 3, as the number of variables tried at each split (Amankulova et al., 2023).

MLR has been widely used across diverse fields as a preferred linear regression technique. Considering that a phenomenon is often associated with multiple influencing factors, using multiple independent variables in MLR has proven to be more effective and realistic than the use of a single independent variable alone, as suggested by Sousa et al. (2007). Therefore, MLR is considered more practical than single linear regression and is commonly utilized to

model linear relationships between a set of multiple independent variables and a dependent variable, as pointed out by Aiken et al. (2012).

KNN is a machine learning method for regression and classification problems. It uses a distance function such as Manhattan or Euclidean to calculate the target value for new samples based on the nearest neighbours of k. K is directly proportional to the prediction, with a smaller K indicating high variance and low bias and a larger K indicating low variance and high bias. The advantage of KNN is that it does not require training or optimization, but has higher complexity and time consumption, as it uses past datasets to predict new ones (Medar et al., 2019).

The DTR method makes predictions for the target variable by building a tree with nodes representing each feature based on the training data. This method can be used for both classification and regression problems and has the advantage of providing easily interpretable results in a tree structure. The algorithm uses binary splits to separate the data into two parts and minimize the sum of squared deviations from the mean in each part until a minimum node size specified by the user is reached (Millán-Castillo et al., 2020; Xu et al., 2005).

The performance of the yield prediction model was assessed by calculating the coefficients of determination (R2), the root mean square error (RMSE) and the mean absolute error (MAE) accuracy metrics. These metrics can be calculated using the following equations: (2)–(4)

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)\left(f_1 - \bar{f}_i\right)\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2 \sum_{i=1}^{n}\left(f_1 - \bar{f}_i\right)^2} \quad (2)$$

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{\left(y_i - f_i\right)^2}{n}} \quad (3)$$

$$MAE = \frac{\sum_{i-1}^{n}|y_i - f_i|}{n} \quad (4)$$

In these equations, n (i = 1, 2, . . ., n) represents the number of samples used to test the ML model, y_i represents the observed yield, $\bar{y}_i$ represents the corresponding mean value, f_i represents the predicted yield, and $\bar{f}_i$ represents the corresponding mean value. A high value of R^2 indicates a better performance of the model in predicting the yield. Lower RMSE values indicate less discrepancy between the predicted and observed yield.

2.7. Model development

The model was developed through a thorough process of analyzing and testing the data, which spanned two years. Various techniques were used to build the machine learning model, including merging the data from 2020 and 2021 to create a more comprehensive data set. To ensure accurate results, each band of S1 and each vegetation index, as well as environmental data, were individually calculated and tested in different combinations. The aim was to determine the optimal combination of features that would yield the most accurate predictions. The metrics were computed separately for each model and the outcomes were compared to identify the best model. Analysis was carried out in August, which is the peak phenological period of soybeans, to ensure that the results were a true representation of the actual yield during this period.

2.8. Model training

In this study, we combined two years (2020 and 2021) of crop data to create a model that was used for training. To test the model, we divided the data into two parts, 70 % used for training and 30 % for testing. Four machine learning techniques, namely RF, KNN, MLR, and DTR, were used to check the model, and three metric values, namely R-squared, RMSE, and MAE, were calculated from the results. The calculations were carried out separately for each of the S1 and VI data and their combination, followed by the topographic data. Finally, all data were combined and the regression was calculated (Fig. 8). The results showed that the R^2 values for S1 ranged from 0.2 to 0.5, for VIs from 0.54 to 0.90, and for the combination of S1 and VIs from 0.32 to 0.90. When combined with topographic data (ie, aspect, slope, and TWI), the R2 values increased from 0.85 to 0.91. The RMSE and MAE values had similar indicators, with separate calculations for S1 resulting in RMSE values ranging from 0.143 to 0.192 t/ha, for VIs from 0.119 to 0.132 t/ha, and for their combination from 0.105 to 0.130 t/ha. The MAE values ranged from 0.126 to 0.151 t/ha for S1, from 0.116 to 0.141 t/ha for VI, and from 0.089 to 0.110 t/ha for their combination. In general, these findings represent that the combination of S1, VIs and topographic data could potentially improve the prediction of crop yields.

3. Results

3.1. Future selection

We examine correlation-based feature selection (CFS), a popular technique for selecting the most relevant features in a dataset. CFS evaluates the correlation between each feature and the target variable and selects the features with the highest correlation. We used Python 3.11.3 software, several libraries provide CFS functionality. One of the most commonly used libraries is scikit-learn. Scikit-learn provides a SelectKBest function, which can be used to select the K-highest features based on a scoring function. The scoring function can be set to correlation, which will select the features with the highest correlation with the target variable. The analysis revealed that GNDVI, NDVI and SAVI were the most significant features in predicting crop yield, with correlation coefficients (r) of almost 1, 0.95, and 0.85, respectively (Fig. 9). These results suggest that these VIs are highly indicative of crop productivity.

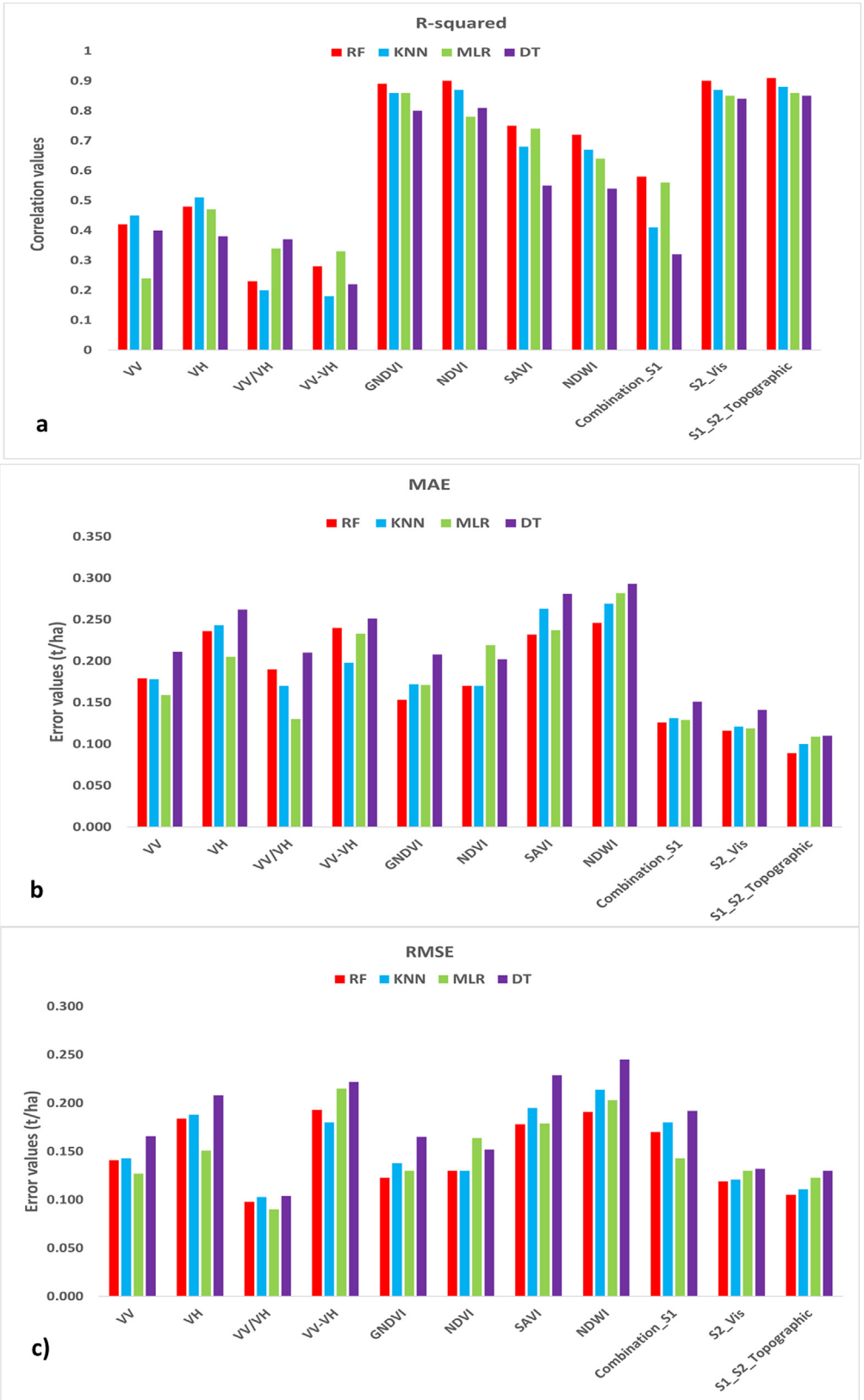

Fig. 8. The training of a predictive model using various ML techniques, and provides information on its performance metrics, specifically the a) R^2, b) MAE, and c) RMSE values.

Furthermore, the *r* values for the polarization types HH and HV were 0.55 and 0.5, respectively, indicating that they have moderate relevance to predict crop yield. Furthermore, topographic factors such as aspect, slope, and TWI were found to have the lowest impact, ranging from 0.1 to 0.2 for productivity prediction. These findings suggest that the combination of VIs, polarization types, and topographic data in crop productivity models can improve the accuracy of prediction.

3.2. Model validation

After creating and testing the model using four different machine learning techniques and various metric values, we found that RF consistently performed the best in predicting soybean yield. Therefore, we chose RF as the preferred ML technique and used it for model validation. We validated the two-year RF training model on independent soybean yield from 2022. We calculated S1, VIs and topographic values for the 2022 yield and divided the plots into fishnet sections for analysis. To demonstrate the results, we present an example of six parcels. The R^2 values in fields 2 and 5 were found to be the lowest with values ranging from 0.41 to 0.77, while fields 1 and 3 showed average values of 0.82 to 0.81, and fields 4 and 6 presented the best value of 0.89. It is important to note that low values of RMSE and MAE are generally observed in areas where R^2 is high, while high RMSE and MAE are associated with lower R^2 values. The MAE values in fields 2 and 5 were calculated as 0.089 and 0.117 t/ha, respectively, while fields 1 and 3 had values of 0.163 and 0.129 t/ha, and fields 4 and 6 yielded values of 0.103 and 0.126 t/ha. Similarly, RMSE values in fields 2 and 5 were found to be 0.122 and 0.153 t/ha, respectively, while fields 1 and 3 had values of 0.224 and 0.171 t/ha, and fields 4 and 6 produced values of 0.138 and 0.165 t/ha. We employed a boxplot to visually represent the RMSE and MAE values in the context of accuracy metrics.

4. Discussion

4.1. Analyzing the vegetative period at peak

The S1 and S2 data was utilized by mosaicing each month to identify the optimal growing season. Minimum, maximum, mean, median, and standard deviation were calculated for each month over three years. The VV and VH bands were calculated for S1 (Fig. 3), while the NDVI index was calculated for S2 (Fig. 4). The study revealed that the peak period of soybean harvest was August in both sensors and all three years, which corresponds to the beginning pod, the phenological period of the entire pod of soybean. Numerous research studies have shown that peak phenological stages can yield superior outcomes. Bai et al., (2019) employed the optimal phenological stage to determine the best time for yield estimation, relying primarily on phenological stages and VIs derived from Landsat 8 imagery, and utilized the month of peak phenology to forecast crop yield. Using the peak phenological period, we based our selection on previous results (Amankulova et al., 2023), monitored the growth period of sunflowers through VIs acquired from S2, and developed a yield prediction model using the highest phenological stage.

4.2. The benefits of using RF compared to other machine learning techniques

In our study, we evaluated the performance of four different machine learning techniques to predict soybean yield. After testing each method in the training dataset,

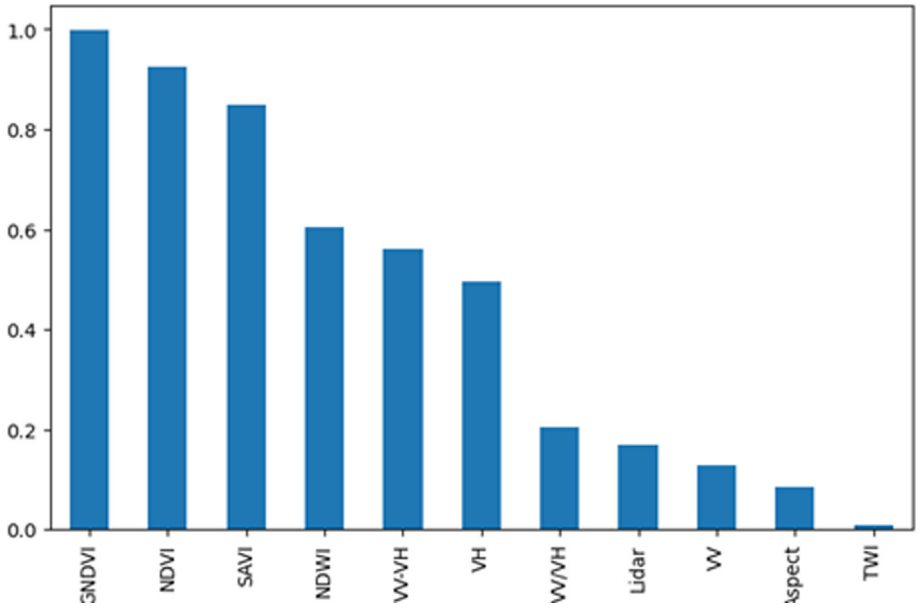

Fig. 9. Correlation-based Feature Selection results.

we found that RF produced the most accurate predictions. Fig. 8 shows the comparison of R^2 values between the four ML techniques, where RF had the highest value of 0.91 than other techniques KNN 0.88, MLR 0.86, DT 0.85. This indicates that RF outperformed the other methods in terms of predicting soybean yield using combined satellite and topographic data. In addition, high R^2 values indicate low RMSE and MAE values, which are important metrics for evaluating prediction accuracy. The RF model demonstrated the lowest RMSE and MAE values of 0.105 and 0.089 t/ha, respectively. After considering the results of other relevant studies, Kumar et al., (2018) concluded that the analysis of S1A SAR was utilized to estimate the growth parameters of winter wheat crops in Varanasi district, India. In general, RF was the most precise algorithm for estimating winter wheat parameters, followed by the SVR, ANNR and LR algorithms. Pang et al., (2022) used RF models on satellite imagery to predict wheat yields in three south-east Australian paddocks. The RF composite region-wide RF model had an R^2 of 0.86 and an RMSE of 0.18 t ha^{-1}, while individual paddocks in Victoria and New South Wales performed well with R^2 values of 0.89 and 0.87 and low RMSE values of 0.15 and 0.07 t ha^{-1}. However, the South Australia model had moderate performance with an R^2 of 0.45 and an RMSE of 0.25 t ha^{-1}. The study highlights the potential of using RF models on satellite imagery for regional- and local-scale yield prediction. In our previous article (Amankulova et al., 2023), we conducted a study to investigate the feasibility of using remote sensing data to monitor crop phenology and predict sunflower crop yield at the field scale. Multiple linear regression and two machine learning approaches were used to predict sunflower crop yield using remote sensing data. The best performing model was found to be the RF with an R^2 of approximately 0.6 and an RMSE of 0.284–0.473 t/ha.

4.3. Importance of future combination S1, S2, and topographical data for soybean yield prediction

The S1 data provide important information on soil moisture, which is a key factor in crop growth and yield, while the S2 data provide high-resolution multispectral imagery, allowing a detailed analysis of crop health and growth patterns. We also investigated VIs derived from S2 images to predict soybean yield. The use of VIs in the prediction of crop yield is an important area of research and several studies have examined its significance. VIs are indicators of crop health and can provide information on vegetation density, photosynthetic activity, and other plant characteristics (Joshi et al., 2023). However, according to other studies, yield accuracy estimates could not be improved by calculating independent VIs (Hunt et al., 2019). This would imply that RF can obtain important information for the estimation of the yield from the specific satellite bands themselves, which is often provided by VIs. When developing a training model, it was observed that the use of only S1 RMSE = 0.180 t / ha or only VIs RMSE = 0.119 t/ha did not produce satisfactory results.

The generated model outperformed the previously established models when the environmental data was integrated with the S1, S2 and topographical data. According to several studies (Burt, 2012; Hunt et al., 2019; Schwalbert et al., 2020), combining environmental data with satellite data to improve crop production assessment produced superior results. Consequently, a decision was made to combine these two satellite images. The integration of topographic data with satellite images led to a significant improvement in the performance of the model. This combination increased the accuracy of the estimate. Specifically, the RMSE value decreased to 0.105, while the MAE was reduced to 0.089 t/ha (Fig. 8) for the random forest regression model. These findings highlight the importance of using a combination of satellite images and topographic data for accurate yield prediction. In the results section, we demonstrate that the combined data approach was effective by conducting a validation for 2022. We applied the combined data approach to all six parcels and observed favorable results with R2 values ranging from 0.41 to 0.89, RMSE values ranging from 0.122 to 0.224 t / ha and MAE values ranging from 0.089 to 0.163 t/ha (Fig. 10). These results highlight the potential of using the combined data approach for soybean yield prediction, as it offers a more comprehensive and accurate assessment of crop conditions and provides valuable insights for farmers and decision makers in the agricultural industry (See Fig. 11).

4.4. Limitations of the study

The main limitation of using combine harvester yield data for yield mapping and monitoring. Inaccurate data could have a variety of causes. These include operating multiple machines with various calibrations, choosing the wrong header height and cut width settings, and making mistakes with speed and travel distance. This can make it difficult to accurately capture within-field variations in crop yield, which can be influenced by a variety of factors, such as soil type, topography, and plant health. Additionally, combine harvester yield data are only available after harvest, which limits their usefulness for making in-season management decisions. Finally, yield data from combine harvesters may be affected by factors such as machine calibration, crop lodging, and operator variability, which can introduce errors and uncertainty into yield estimates (Thylén and Murphy, 1996).

5. Conclusions

This study has demonstrated the potential of using a combination of S1 and S2 satellite imagery along with other geographic attributes to forecast soybean yield in the field at an early stage. The study first determined the optimal phenological stage for soybean harvest in August by analyzing satellite images. Two years were used for

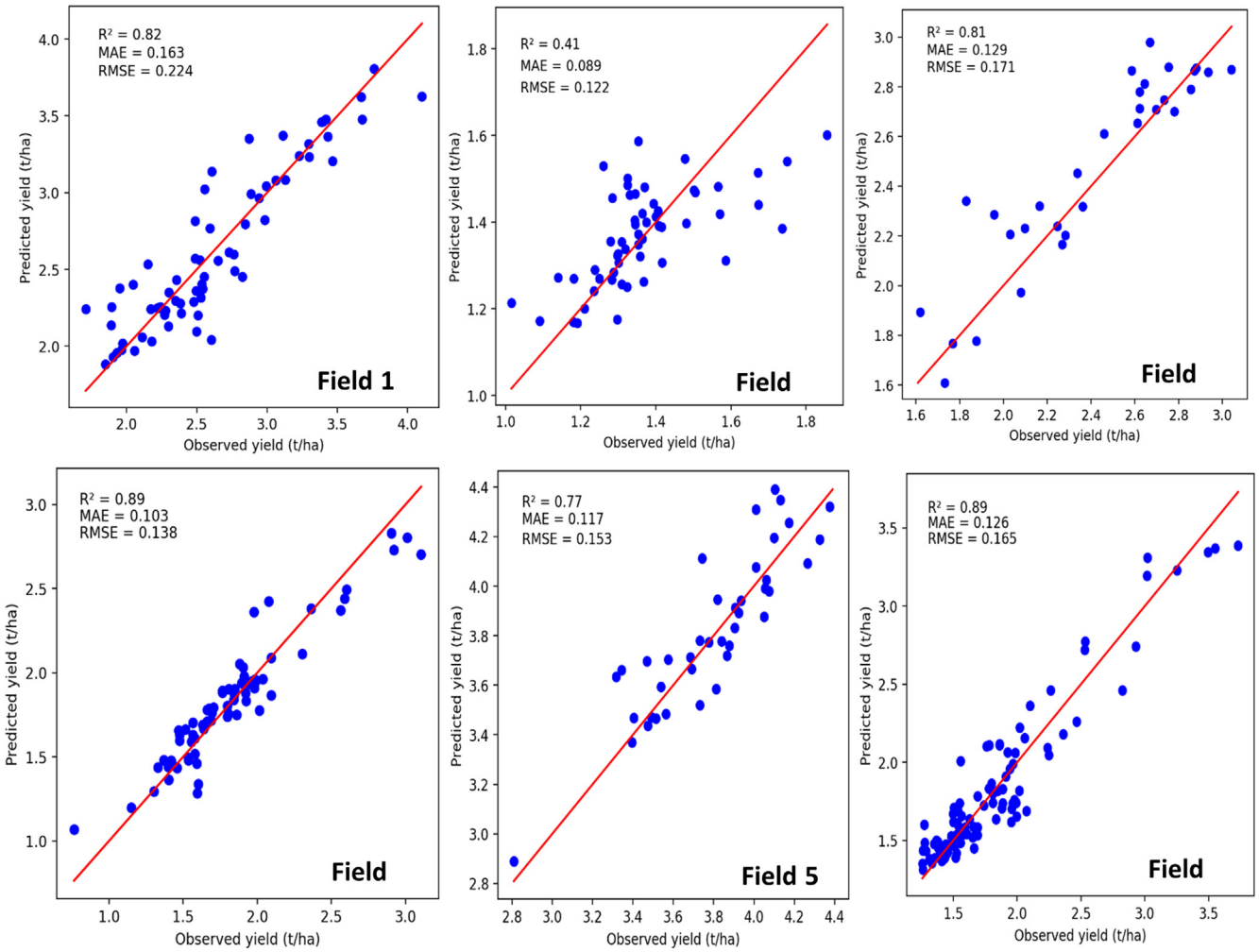

Fig. 10. Observed and predicted soybean yield data from validation for 2022 using combined predictor variables (i.e. satellite imagery, environmental data) extracted from August soybean yield.

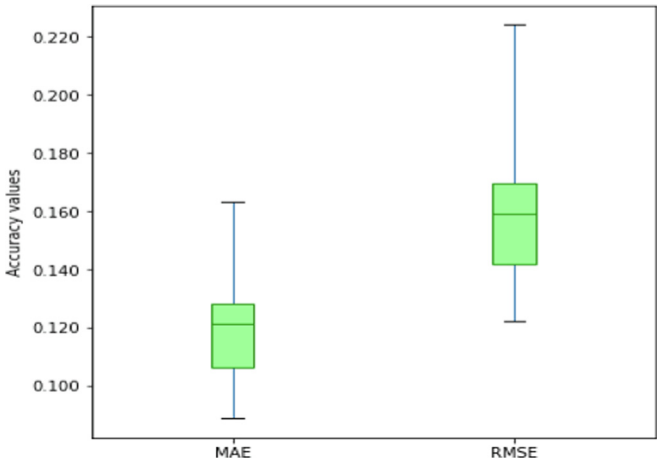

Fig. 11. Box Plots illustrating a summary of accuracy metrics inluding RMSE and MAE for validation datasets accross all fields.

model training and testing, while the model was validated on the data set of 2022 years, utilizing S1 bands and VIs obtained from monthly mosaiced images, as well as topographic data. The individual calculations were subsequently combined and it was determined that the RF regression algorithm was the most effective machine learning technique. The combination data was then calculated using RF in the validation process, resulting in high accuracy rates, with R^2 ranging from 0.41 to 0.89 in parcel sections, RMSE ranging from 0.122 to 0.224 t/ha, and MAE ranging from 0.089 to 0.163 t/ha. The results of this study indicate that the integration of satellite data S1 and 2 with topographical information can facilitate the monitoring, mapping and forecasting of crop yields on small and fragmented farmlands, thus aiding agricultural decision-making and allowing early warnings.

By combining data from S1 and S2, the outcomes were found to be more effective than using data only from S2. However, further research is needed to improve our understanding of the relationship between backscattering and crop yield. In future studies, it would be useful to consider high-resolution meteorological and soil variables such as temperature, precipitation, and soil moisture to gain a better understanding of the factors affecting crop yield.

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.

Acknowledgments

This research was supported by the [University of Szeged Open Access Fund] under Grant [number 6784].

References

Amankulova, K., Farmonov, N., Mukhtorov, U., Mucsi, L., 2023. Sunflower crop yield prediction by advanced statistical modeling using satellite-derived vegetation indices and crop phenology. Geocarto Int. 38, 2197509. https://doi.org/10.1080/10106049.2023.2197509.

Amherdt, S., Di Leo, N.C., Pereira, A., Cornero, C., Pacino, M.C., 2022. Assessment of interferometric coherence contribution to corn and soybean mapping with Sentinel-1 data time series. Geocarto Int. 1–22. https://doi.org/10.1080/10106049.2022.2144472.

Arslan, S., Colvin, T.S., 2002. No title found. Precis. Agric. 3, 135–154. https://doi.org/10.1023/A:1013819502827.

Bai, T., Zhang, N., Mercatoris, B., Chen, Y., 2019. Jujube yield prediction method combining Landsat 8 Vegetation Index and the phenological length. Comput. Electron. Agric. 162, 1011–1027. https://doi.org/10.1016/j.compag.2019.05.035.

Becker-Reshef, I., Vermote, E., Lindeman, M., Justice, C., 2010. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sens. Environ. 114, 1312–1323. https://doi.org/10.1016/j.rse.2010.01.010.

Bolton, D.K., Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agric. For. Meteorol. 173, 74–84. https://doi.org/10.1016/j.agrformet.2013.01.007.

Breiman, L., 2001. No title found. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Elwadie, M.E., Pierce, F.J., Qi, J., 2005. Remote sensing of canopy dynamics and biophysical variables estimation of corn in Michigan. Agron. J. 97, 99–105. https://doi.org/10.2134/agronj2005.0099.

Farmonov, N., Amankulova, K., Szatmári, J., Urinov, J., Narmanov, Z., Nosirov, J., Mucsi, L., 2023. Combining PlanetScope and Sentinel-2 images with environmental data for improved wheat yield estimation. Int. J. Digital Earth 16, 847–867. https://doi.org/10.1080/17538947.2023.2186505.

Gianelle, D., Vescovo, L., Marcolla, B., Manca, G., Cescatti, A., 2009. Ecosystem carbon fluxes and canopy spectral reflectance of a mountain meadow. Int. J. Remote Sens. 30, 435–449. https://doi.org/10.1080/01431160802314855.

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, 289–298. https://doi.org/10.1016/S0034-4257(96)00072-7.

Gu, Y., Brown, J.F., Verdin, J.P., Wardlow, B., 2007. A five-year analysis of MODIS NDVI and NDWI for grassland drought assessment over the central Great Plains of the United States. Geophys. Res. Lett. 34, L06407. https://doi.org/10.1029/2006GL029127.

Gu, Y., Hunt, E., Wardlow, B., Basara, J.B., Brown, J.F., Verdin, J.P., 2008. Evaluation of MODIS NDVI and NDWI for vegetation drought monitoring using Oklahoma Mesonet soil moisture data. Geophys. Res. Lett. 35, L22401. https://doi.org/10.1029/2008GL035772.

Herrero-Huerta, M., Rodriguez-Gonzalvez, P., Rainey, K.M., 2020. Yield prediction by machine learning from UAS-based multi-sensor data fusion in soybean. Plant Methods 16, 78. https://doi.org/10.1186/s13007-020-00620-6.

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25, 295–309. https://doi.org/10.1016/0034-4257(88)90106-X.

Kumar, P., Prasad, R., Gupta, D.K., Mishra, V.N., Vishwakarma, A.K., Yadav, V.P., Bala, R., Choudhary, A., Avtar, R., 2018. Estimation of winter wheat crop growth parameters using time series Sentinel-1A SAR data. Geocarto Int. 33, 942–956. https://doi.org/10.1080/10106049.2017.1316781.

Maestrini, B., Basso, B., 2018. Drivers of within-field spatial and temporal variability of crop yield across the US Midwest. Sci Rep 8, 14833. https://doi.org/10.1038/s41598-018-32779-3.

McFEETERS, S.K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. Int. J. Remote Sens. 17, 1425–1432. https://doi.org/10.1080/01431169608948714.

Medar, R., Rajpurohit, V.S., Shweta, S., 2019. Crop Yield Prediction using Machine Learning Techniques. In: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT). Presented at the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), IEEE, Bombay, India, pp. 1–5. https://doi.org/10.1109/I2CT45611.2019.9033611.

Mishra, S., Mishra, D., Santra, G.H., 2016. Applications of machine learning techniques in agricultural crop production: a review paper. Indian J. Sci. Technol., 9 https://doi.org/10.17485/ijst/2016/v9i38/95032.

Obsie, E.Y., Qu, H., Drummond, F., 2020. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. Comput. Electron. Agric. 178. https://doi.org/10.1016/j.compag.2020.105778 105778.

Panda, S.S., Ames, D.P., Panigrahi, S., 2010. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. Remote Sens. (Basel) 2, 673–696. https://doi.org/10.3390/rs2030673.

Pang, A., Chang, M.W.L., Chen, Y., 2022. Evaluation of Random Forests (RF) for regional and local-scale wheat yield prediction in Southeast Australia. Sensors 22, 717. https://doi.org/10.3390/s22030717.

Pejak, B., Lugonja, P., Antić, A., Panić, M., Pandžić, M., Alexakis, E., Mavrepis, P., Zhou, N., Marko, O., Crnojević, V., 2022. Soya yield prediction on a within-field scale using machine learning models trained on Sentinel-2 and soil data. Remote Sens. (Basel) 14, 2256. https://doi.org/10.3390/rs14092256.

Pierce, F.J., Anderson, N.W., Colvin, T.S., Schueller, J.K., Humburg, D. S., McLaughlin, N.B., 2015. Yield Mapping. In: Pierce, F.J., Sadler, E. J. (Eds.), ASA, CSSA, and SSSA Books. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI, USA, pp. 211–243. https://doi.org/10.2134/1997.stateofsitespecific.c11.

Qin, Q., Xu, D., Hou, L., Shen, B., Xin, X., 2021. Comparing vegetation indices from Sentinel-2 and Landsat 8 under different vegetation gradients based on a controlled grazing experiment. Ecol. Ind. 133. https://doi.org/10.1016/j.ecolind.2021.108363 108363.

Qin, C.-Z., Zhu, A.-X., Pei, T., Li, B.-L., Scholten, T., Behrens, T., Zhou, C.-H., 2011. An approach to computing topographic wetness index based on maximum downslope gradient. Precision Agric. 12, 32–43. https://doi.org/10.1007/s11119-009-9152-y.

Saeed, U., Dempewolf, J., Becker-Reshef, I., Khan, A., Ahmad, A., Wajid, S.A., 2017. Forecasting wheat yield from weather data and MODIS NDVI using Random Forests for Punjab province, Pakistan. Int. J. Remote Sens. 38, 4831–4854. https://doi.org/10.1080/01431161.2017.1323282.

Segarra, J., Buchaillot, M.L., Araus, J.L., Kefauver, S.C., 2020. Remote sensing for precision agriculture: Sentinel-2 improved features and applications. Agronomy 10, 641. https://doi.org/10.3390/agronomy10050641.

Sehgal, V.K., Jain, S., Aggarwal, P.K., Jha, S., 2011. Deriving crop phenology metrics and their trends using times series NOAA-AVHRR NDVI data. J. Indian Soc. Remote Sens. 39, 373–381. https://doi.org/10.1007/s12524-011-0125-z.

Shang, J., Liu, J., Ma, B., Zhao, T., Jiao, X., Geng, X., Huffman, T., Kovacs, J.M., Walters, D., 2015. Mapping spatial variability of crop growth conditions using RapidEye data in Northern Ontario, Canada. Remote Sens. Environ. 168, 113–125. https://doi.org/10.1016/j.rse.2015.06.024.

Shao, Y., Campbell, J.B., Taff, G.N., Zheng, B., 2015. An analysis of cropland mask choice and ancillary data for annual corn yield forecasting using MODIS data. Int. J. Appl. Earth Obs. Geoinf. 38, 78–87. https://doi.org/10.1016/j.jag.2014.12.017.

Sharifi, A., 2021. Yield prediction with machine learning algorithms and satellite images. J. Sci. Food Agric. 101, 891–896. https://doi.org/10.1002/jsfa.10696.

She, B., Yang, Y., Zhao, Z., Huang, L., Liang, D., Zhang, D., 1. School of Geomatics, Anhui University of Science &Technology, Huainan 232001, Anhui, China, 2. National Engineering Research Center for Agro-Ecological Big Data Analysis & Application, Anhui University, Hefei 230601, China, 2020. Identification and mapping of soybean and maize crops based on Sentinel-2 data. Int. J. Agric. Biol. Eng. 13, 171–182. https://doi.org/10.25165/j.ijabe.20201306.6183.

Shendryk, Y., Davy, R., Thorburn, P., 2021. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. Field Crop Res 260. https://doi.org/10.1016/j.fcr.2020.107984 107984.

Silva, J.R.M.D., Alexandre, C., 2005. Spatial variability of irrigated corn yield in relation to field topography and soil chemical characteristics. Precision Agric. 6, 453–466. https://doi.org/10.1007/s11119-005-3679-3.

Sinclair, T.R., Marrou, H., Soltani, A., Vadez, V., Chandolu, K.C., 2014. Soybean production potential in Africa. Glob. Food Sec. 3, 31–40. https://doi.org/10.1016/j.gfs.2013.12.001.

Steele-Dunne, S.C., McNairn, H., Monsivais-Huertero, A., Judge, J., Liu, P.-W., Papathanassiou, K., 2017. Radar remote sensing of agricultural canopies: a review. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. 10, 2249–2273. https://doi.org/10.1109/JSTARS.2016.2639043.

Suominen, L., Ruokolainen, K., Tuomisto, H., Llerena, N., Higgins, M. A., 2013. Predicting soil properties from floristic composition in western Amazonian rain forests: performance of k -nearest neighbour estimation and weighted averaging calibration. J Appl Ecol 50, 1441–1449. https://doi.org/10.1111/1365-2664.12131.

Thylén, L., Murphy, D.P.L., 1996. The control of errors in momentary yield data from combine harvesters. J. Agric. Eng. Res. 64, 271–278. https://doi.org/10.1006/jaer.1996.0068.

Ho, T.K., 1995. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. Presented at the 3rd International Conference on Document Analysis and Recognition, IEEE Comput. Soc. Press, Montreal, Que., Canada, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994.

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8, 127–150. https://doi.org/10.1016/0034-4257(79)90013-0.

Uribeetxebarria, A., Castellón, A., Aizpurua, A., 2023. Optimizing wheat yield prediction integrating data from Sentinel-1 and Sentinel-2 with CatBoost algorithm. Remote Sens. (Basel) 15, 1640. https://doi.org/10.3390/rs15061640.

Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.-F., Ceschia, E., 2017. Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. Remote Sens. Environ. 199, 415–426. https://doi.org/10.1016/j.rse.2017.07.015.

Xu, M., Watanachaturaporn, P., Varshney, P., Arora, M., 2005. Decision tree regression for soft classification of remote sensing data. Remote Sens. Environ. 97, 322–336. https://doi.org/10.1016/j.rse.2005.05.008.

Zhao, Y., Chen, X., Cui, Z., Lobell, D.B., 2015. Using satellite remote sensing to understand maize yield gaps in the North China Plain. Field Crop Res 183, 31–42. https://doi.org/10.1016/j.fcr.2015.07.004.

Zhao, Y., Potgieter, A.B., Zhang, M., Wu, B., Hammer, G.L., 2020. Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 Satellite imagery and crop modelling. Remote Sens. (Basel) 12, 1024. https://doi.org/10.3390/rs12061024.