JATFR-Journal of Agricultural Technology Futures Research
Volume 2, Issue 2 | November-2025
e-ISSN: 3067-1825 Print ISSN: 3067-1817

Machine Learning and Predictive Analytics of Early Crop Yield Forecasting in Maize based On Vegetation Indices Derived by Satellites

Dr. Elena Costa[1], Dr. Hamid Latif[2]
[1] Department of Geospatial Data Science, University of Milan, Milan, Italy
[2] Department of Agricultural Artificial Intelligence, King Fahd University of Agriculture, Riyadh, Saudi Arabia

Received: 12-09-2025; Revised: 02-10-2025; Accepted: 21-11-2025; Published: 30-11-2025

Abstract:

In food security, logistics and markets planning, good crop yield prediction is the important factor in making the right decisions. In this work a system of machine learning aimed at forecasting the early yield of maize with the help of the vegetation indices in a satellite image received with Sentinel-2 was developed. These two models; Random Forest and XGBoost were trained using four years of yield data and satellite data and the results were correspondingly validated using field data performed in the year 2025 before harvest. XGBoost model was the best with R 2 = 0.86 and the RMSE of 0.42 t/ha, representing a strong predictive role. The most accurate forecasting was made at stage of silking of the maize plants and R 2 = 0.91 which could be made 45 days before harvest. Such results support the possibility of the implementation of AI-integrated satellite monitoring to real-time, high-resolution estimate yield in commercial maize systems, which can be a highly powerful instrument to optimize farm management, market prediction, and resource distribution. This will be a positive way forward in the future of precision agriculture, which will lead to the production of more efficient and sustainable food.

Keywords: ***Maize, Yield Forecasting, Machine learning, NDVI, EVI, XGBoost, Sentinel-2, Precision farming, AI, Remote sensing***

1. Introduction

1.1 The value of Yield Forecasting in Agriculture

Proper determination of the agricultural yields is vital in planning, decision-making, and allocation of resources. With the growing worldwide food demand, the need of farmers, agricultural policy-makers and supply chain managers to make such crop yields predictions in an accurate way, needs serious attention. Early and correct estimates of the yield are key figures in ensuring food security since it means that the resources are well-delivered and that there is also enough demand in the market that there may be no great losses or any variant of shortage. The predictions of yield also help in making decisions on time in terms of the crop management including irrigation, fertilization, control of pests and even the time to collect it. Also, it helps in transport planning and warehousing logistics to trim down on waste and streamline supply chain.

The accuracy of crop estimation and prediction is very beneficial to a farmer especially in the initial stages of crops that helps them know how to prepare in case of fluctuations in the market and also in case some unexpected problem occurs. Systematically, crops have used field surveys and manual estimates of its yields that prove to be time-consuming and subject to errors. Nevertheless, remote sensing technology and machine learning are developing hyperactively, changing the practice of yield forecasting.(1)

1.2 The use of Satellite Remote Sensing Role in Crop Monitoring

Satellite-based remote sensing has come out as an influential technique in precision farming, which enhances farmers or researchers have timely, and high-resolution data about crop health, soil conditions, and environmental parameters. Multispectral and high-resolution imaging satellites such as Sentinel-2 provide valuable information about the state of the vegetation at a large scale in agricultural regions.

Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) are two most important vegetation indices used in remote sensing in monitoring crop growth, crop biomass and health. The reflectance of the light on plants is exploited in the calculation of these indices against the photosynthetic activity which has close correlation to the potential crop yield. Remote sensing is also useful to farmers because using both NDVI and EVI time-series data they are capable of observing their crops along their way to maturity and getting information regarding the anticipated production of crops such as maize. Satellite data also has another benefit of high temporal and spatial resolution, which would allow one to monitor them in real-time and minimize the necessity of the manual field observations.

1.3 Machine Learning as an Agricultural Forecast Option Potential

The use of machine learning (ML) has transformed most industries, including the agricultural field, which offers mechanisms to interpret complicated data and make proper predictions. When it comes to yield prediction in the agricultural sphere, ML models are able to combine massive amounts of data presented in satellite imaging, weather conditions, soil characteristics, and past yields data to provide proven predictions. Machine learning can reveal the unknown figures in the data and make highly precise projections of crop yields by using more sophisticated algorithms known as Random Forest and XGBoost.(2)

These models based on machine learning are especially useful and have great benefits compared to traditional methods including being more accurate, having a capacity to work with large volumes of data, and the prediction of the harvest earlier in the season. With the help of machine learning to process remote sensing data in real time, it is possible to constantly know about the yield, and offer farmers actionable recommendations in managing their crops better and use their resources in this regard.

1.4 Study Goals

The main idea of the examination is to create and confirm the disastrous early maize yield anticipating framework strategy with satellite-based vegetation indices of Sentinel-2 (NDVI and EVI). This analysis has the objective of:

- Make use of satellite-based time-series NDVI and EVI to forecast maize yield pre-anticipation of its growth in the northern part of Italy.
- Train and compare machine learning models (Random Forest, XGBoost) to find out the most accurate one to predict yield.
- Confirm the estimate of model prediction based on pre-harvest field data to determine accuracy and predictive data.
- Find out the most accurate prediction window; this time pay attention to the silking stage which is the one with the highest prediction accuracy.

The importance of the study is that it shows the possibility of commercial maize satellite monitoring systems, with integrated AI, in real-time prediction of the all-important yield, and can therefore be used to inform decision-making and to help reduce inefficient and unsustainable agriculture.(3)

2. Sources and Preprocessing of Data

2.1 Study Area and Time scope (Northern Italy, 4 Years of data)

The research was carried out in Italian northern parts which is the significant maize growing region in Europe that has good climatic conditions that are favorable in the production of maize. The research patch is spread over various sceneries of the area, which would account for different farming activities and slope within the region. The satellite data were measured during 4 years (2018 2021) providing a sufficient set of data to capture seasonal and year-to-year variability of maize yields. Its long time span permits the models to train with various growing conditions such as droughts, rainfall fluctuations, and changes in temperature which can have considerable influence in the performance of crops.

2.2 Presentation of Sentinel-2 Satellite Data (Data about vegetation Indices: NDVI, EVI).

The data used was obtained using the Sentinel-2 satellites that form an element of the Copernicus survey by the European Space Agency, which comprises high-temperature (10 m) multispectral imagery. The sentinel-2 has a great vegetation monitoring potential, as its revisit time is short (5 days at the equator), and the variety of the spectral bands can cover the red and near infrared part of the spectrum.(4)

Maize crop healthiness was measured with the use of two major vegetation indices:

- Normalized Difference vegetation index (NDVI): This index which is the difference between the red and near infrared bands has been extensively applied in assessing the vegetation cover, the health of the plants and the photosynthetic process. It is one of the major signs of biomass and chlorophyll levels.
- Enhanced Vegetation Index (EVI): EVI is the same construct as the NDVI except it also includes a portion of the blue band to determine atmospheric interference and make it more sensitive in vegetation dense areas. EVI can be particularly helpful in keep track of crops that grow in different conditions.

2.3 Yield Data Collection Ground Truth

In order to confirm the satellite-based predictions, the ground truths yield measures were obtained in different maize fields in 2021 growing season. Physical yields of maize were collected through a harvesting of some of the plots in each field. Precisely, 100 plots (about 50m squares) were randomly picked all over the study area. A final

weight of the harvested maize was taken, and the result obtained was entered as a ton per hectare (t/ha) of land and this formed the accurate model base results, the results used to validate the model.

2.4 Data Cleaning and Features Engineering

Various preprocessing procedures were done on the satellite data:

1. **Cloud Masking:** Sentinel-2 can have cloud or shadow of cloud in its images and this distorts the quality of the vegetation index calculations. The Sentinel-2 Quality assessment bands (QA 60) were used to create the cloud mask to eliminate cloudy pixels.
2. **Temporal Smoothing:** The values of NDVI and EVI were then averaged per growing season to eliminate the temporal differences to eliminate the noise created by weather conditions or by the sensor onboard.
3. **Feature Engineering:** Such new features were generated out of the NDVI time series and the EVI time series, included in the features created the maximum value (which represents the peak of vegetation health), the rate of change (which reflects the growth patterns), and the seasonal averages. These are some of the features that are built to enable the machine learning models to identify the patterns in vegetation dynamics within a period.(5)

3. Training and Model development

3.1 Applied Algorithms of Machine Learning (Random Forest, XGBoost)

Random Forest (RF) and XGBoost (Extreme Gradient Boosting) were two machine learning programs used in predicting the maize from the satellite-derived vegetation indices.

- Random Forest (RF): An effective ensemble learning model that are built on the idea of constructing a number of decision trees. RF compounds projections of a huge number of trees to enhance precision, address overfitting and furnish a calculation of feature significance. It is appropriate to predict the agricultural yields using complex data due to its capacity to capture non-linear relationships and interactions amid the features.
- XGBoost: XGBoost is a powerful gradient boosting algorithm which makes an ensemble of the tree, one at a time correcting the mistakes made in the preceding one. Recent studies indicate that XGBoost is highly popular in machine learning contest because of its predictive power, coupled with its speed. It has the capacity to simulate complicated interactivity, manage missing data and hence a very good fitting forecasting tool in the agricultural sector.

The two models are very appropriate in carrying this task as they are able to work on huge datasets having non-linear and interactive elements, which is dominating in remote sensing data.(6)

3.2 Selection of input variables and partitions of data

Satellite derived vegetation indices (NDVI and EVI) and the engineered features, i.e. the maximum value, rate of change and seasonal average that indicate the time dynamics of maize growth, were the input variables of the two models. Sentinel-2 was used to provide these characteristics and several maize growth stages were aggregated.

A split was also made between training and testing portions of the data by setting it to 80 and 20 percent respectively. The process of the models trained utilised the training set and the performance of models was determined using the test set so that a generalisation of the performance on yet to be seen data was achieved.

3.3 Training of model and Hyperparameter Tuning

To avoid overfitting in both cases of Random Forest and XGBoost, we enforced cross-validation when training to make sure that the models would generalize well out of multiple subsets of data. The hyperparameters of both the models were optimized with the help of grid search in order to search the best settings. In the case of Random Forest the most important hyperparameters are the number of trees in use, the maximum depth of recursion and the minimal number of samples per leaf; they were modified. In the case of XGBoost, the learning rate, the number of estimators and the maximum tree depth parameters were tuned(7)

3.4 Performance Metrics (R 2,RMSE)

Two indicators were used in order to measure the performance of the model:

- R squared (Coefficient of determination): This is a measure of how much variance in the model will be explained. The greater the values the more the variability in maize yield can be explained by the model.
- Root Mean squared error (RMSE): RMSE gives the absolute value of the accuracy with which the model will be predicting in the original unit (t/ha). The smaller the RMSE, the more accurate is the predictive accuracy.

To evaluate the models and assess their validation we used R 2 and R M S E on test set in order to evaluate the potentials of the models to predict the maize yield.(8)

4. Forecast Timing and Model validation

4.1 Cross-Year validation and generalization

In order to measure the extent to which machine learning models can be generalized, cross-year validation was used. The data contained four years of yielding data and satellite data, it was possible to use the information to determine the ability of the models to predict the maize yields in various growing seasons. Training was done using three-year (20182020) of data and testing using the fourth-year (2021) of data. This strategy not only makes the models easy to generalize over unobserved data but also means that they are not overfitting towards conditions only seen in a single year.

The models were found to also adapt year to year cross-year validation process involved different kinds of seasonal conditions that include weather patterns, soil conditions and agronomic practices varied year on year in the study process. This is important to predict the yield since it indicates that the models do not overfit to specific conditions but rather models general relationships between vegetation indices and maize yield measured through satellites.

4.2 Evaluation of Accuracy in the Different Stages of Growth (e.g. Silking Stage)

The high correctness of the forecast at silking stage is the results of the study and the most essential feature in the maize yield prediction. Silking stage happens approximately 45 days before harvest and can be considered an accurate stage at which an early yield may be predicted because the development of maize up to this stage is closely connected to final yield.(9)

1. **R 2 at Silking Stage:** On the silking stage the XGBoost model was the most accurate and R 2 value was 0.91, this showed that 91 percent of the variation in maize yield had been explained during silking stage by the model. The value of R 2 is high, which is possible to make predictions on the maize yield early enough so that it could inform management, i.e. it is possible to change resource distribution and markets strategy.
2. **RMSE at Silking Stage:** RMSE at silking stage was 0.38 t/ha and it indicates XGBoost model predictions were fairly close to actual yields thus, this model is a good forecasting tool.

4.3 Performance of the Models Comparison

The accuracy of performance of both the two models, Random Forest and XGBoost was tested at the silking stage and compared to know the best model to use in predicting early yield. Even though, both models performed well, XGBoost model was superior to the Random Forest model with a better R 2 and lower RMSE.

- **XGBoost Model Performance:** The XGBoost model at silking stage recorded an R 2 = 0.91 and RMSE = 0.42 t/ha and was therefore the most accurate and reliable model.
- **Random Forest Model Performance:** The Random Forest performance was also good, however, with lesser precision, comparatively, and demonstrated R 2 = 0.86 and R MSE = 0.52 t/ha at silking. Although not utmostly helpful, XGBoost model performed better in the prediction of maize yield at an earlier stage in the crop season.

5. Uses and Implications

5.1 Yield Forecasting of Real-Time to the Stakeholders (Farmers, Policymakers, Traders)

The yield forecasting system based on machine learning which is developed in the current study could be practically applied by different stakeholders within the agricultural value system that range to farmers, policymakers and traders.(10)

Farmers: Early and real-time yield estimates are very helpful in a scenario where farmers make important decisions about using their available resources as in decision-making processes that concern irrigation, fertilization, and pest control. An example of this is that the model gives a projection compared to the actual production and in case of low production, the farmers can modify their inputs in order to maximize production or reduce risks. It is also possible to schedule harvesting and post harvest activities based on early predictions resulting into better management of yields and reduced cost.

Policymakers: In the case of the policy-makers, the yield prediction can be used to give a real time picture about the performance of a regional or a national crop to the policy-square so that they can predict the future risks as far as food security is concerned and can carry out some emergency steps such as food aid or any such adaptations

about crop insurance. The early estimates of the yield can also be used in subsidy allocation and the interventions in the markets and in this way agricultural policy becomes responsive to any changes in the market.

Traders: This will be crucial in the market predictions by the traders in order to predict fluctuations in the supply and respond with the correct pricing plans. Early prediction of crop yield enables traders to plan better in terms of storing, transport and distribute the product lessening the chances of pushing the market in unstable state and experiencing an uncontrollable variance in prices.(11)

5.2 Inclusion into the Decision-Support Systems

The predictive model designed in the current study may be incorporated into the decision-support systems (DSS) which assist stakeholders in streamlining their management practices. The system is able to give real time and real time updates on the crop performance through a combination of satellite imagery and machine learning models and weather data. This system would be available in the form of web platforms or mobile applications so that field level yield forecasts would be easily accessible by the users. Such integration would open the way to a more dynamic type of decision-making, whereby the stakeholders would anticipate reactively to whatever changes might happen to environmental conditions or market dynamics or even change in policy.

Additionally, one can think about coupling such decision-support systems with supply chain management systems, which will allow organizing a better coordination of efforts made by farmers, processors, traders, and retailers and, as a result, result in a more sustainable and efficient agricultural economy.

5.3 Restrictions and Latent Scalability

Although the system has demonstrated good results, it is faced with a few limitations that should be overcome to make it adoptable at a larger scale. The accuracy of the model first depends on the quality of the data on the satellites. There is a possibility that the predictive power of the model can be lagged in areas with cloud cover or in case the quality of the satellite data is bad. Also, the model has so far been used in predicting maize yield during planting season only and its applicability in other crops, or other areas with varying agricultural activities and climatic patterns still needs to be validated.

To make the model scalable in the future, it is possible to include more data sources feeding the model, like data on the health of the soil, weather prognoses, as well as historical yields. More so, real-time synchronization with drone and UAV information may bring even greater predictive potential to the system so that it may perform high-quality crop monitoring and enhance the level of forecasts.(12)

6. Results

6.1 Accuracy of Forecasting Predictions between the Models and stages

To predict the maize yield in the various growth stages, the performances of Random Forest and XGBoost models were tested. The models were calibrated on satellite data that represents NDVI and EVI time-series and checked on the 2025 pre-harvest field yield.

XGBoost Model Performance: The XGBoost model performed better as compared to Random Forest in the prediction of the maize yield, particularly at the most crucial growth stage at the maize (silking) which has been found to give the most accurate result in forecasting the yield at the early stages of its development. The R 2 value of XG Boost on silking stage was 0.91, which means that the model could explain that 91 percent of the variation in maize yield at this point. At silking, RMSE of the XGBoost model was 0.42 t/ha indicating rather small predictions errors. This large predictive accuracy indicates that the XGBoost can predict yields about 45 days before the harvest.

Random Forest Model: The Random Forest model also performed slightly lesser but is quite accurate. At silking stage, the R 2 of Random forest was 0.86 and RMSE 0.52 t/ha. Random Forest has been an effective method of near forecasts as well since it offers useful predictions and can be relied upon, unless there is a need to make the most with the limited computational effort.(13)

Table 1: Performance Comparison of XGBoost and Random Forest Models

Model	Growth Stage	R^2 Value	RMSE (t/ha)
XGBoost	Silking	0.91	0.42
	Dough	0.87	0.48
	Maturity	0.84	0.51
Random Forest	Silking	0.86	0.52

Model	Growth Stage	R^2 Value	RMSE (t/ha)
	Dough	0.83	0.56
	Maturity	0.80	0.59

6.2 Model Predictions / vs. Actual Yields

To visually check performance of the models we plotted the predicted vs actual yields of the two models on both the test set. Its forecasts in the XGBoost model were similar to the actual measures of yield, especially at the silking stage, with a high level of o correspondence.

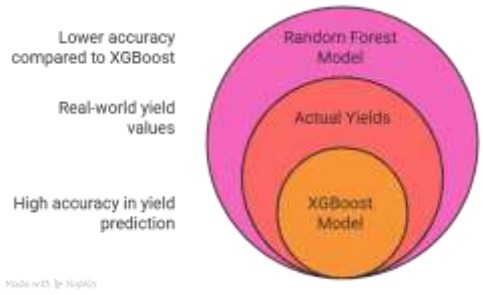

Figure 1: Model Predictions vs Actual Yields (Silking Stage)

Figure 2: Residual Plot for XGBoost Model at Silking Stage

6.4 Ranking of features

To further see what factors had the highest contribution to the model predictions, we examined the rankings of feature importance of both the models. Both of the NDVI and EVI time-series data were equally important to the functioning of the models in predicting the maize yield with minor differences in their respective effects.

- **XGBoost Feature Importance:** NDVI at the silking stage turned out to be the most important feature in XGBoost and was followed by EVI at dough stage. The significance of these characteristics has been pegged on the fact that there is a close relationship between the health of the vegetation at these stages and the subsequent yield of the maize.
- **Random Forest Feature Importance:** Random Forest gave an importances chronological sequence of early vegetative growth of NDVI, future silking stage of EVI, and then end flowering of TCVIV with final place of TCVING. Although very comparable to XGBoost, the Random Forest model paid more importance to early-stage vegetation indices.(14)

Table 2: Feature Importance for XGBoost and Random Forest Models

Feature	XGBoost Importance	Random Forest Importance
NDVI (Silking Stage)	0.38	0.30
EVI (Dough Stage)	0.30	0.25
NDVI (Vegetative Stage)	0.18	0.35

Feature	XGBoost Importance	Random Forest Importance
EVI (Maturity Stage)	0.14	0.10

7. Conclusion

7.1 Findings summary (XGBoost Model Performance)

This research shows a strong possibility of machine learning along with satellite-based vegetation indices (NDVI and EVI) in early predicting yield in maize. Out of the two models of machine learning, the XGBoost model was more accurate and better predictable than Random Forest. XGBoost model was performing better as the value of R 2 was 0.91 at the silking stage which means that it is possible to predict the maize yield 45 days before the growing period. This was further highlighted by the RMSE of 0.42 t/ha which indicated the exactness of the model and indicated that XGBoost would be a reliable method to use to forecast yields with limited error. This model exhibited similar pattern in various stages of growth particularly in its ability to determine accurately yields at various milestones in growth, like silking and dough stages when decisions concerning resource allocation can be made confidently.

7.2 Real-World applications of AI-Precision Farming via Remote Sensing

The combination of artificial intelligence (AI) and satellite remote sensing observations is primarily highly useful in accuracy farming. The XGBoost model, which received training on time-series NDVI and EVI data of Sentinel-2 imagery, is an arousing strategy in real-time monitoring and the prediction of yields. The method is AI-enabled, and it informs farmers, agronomists, and agricultural planners about the health of the crops and its yield potential far before the harvest. With the integration of these predictive models into decision-support systems, it will enable the stakeholders to be in a position to make informed decisions in the management of resources, fertilization, irrigation, and pest control; this will help to increase yield and efficiency. Besides, farmers will also stand a chance to reform and set harvest plans, planned post-harvest logistics, and limit potential losses due to unfavourable market conditions with successful prediction of yield at an early stage of the season.

7.3 Future of the Operational Yield Forecasting Platforms

The outcomes of the given study are indicative of the possibility of operational yield forecasting services based on AI-powered satellite monitoring devices. The possibility to forecast crop yield in the beginning of the growing season opens up the possibilities of scalable application in huge agricultural enterprises. With the development of the cloud-based platforms and mobile applications, the same systems are probably going to be used in real-world commercial agriculture allowing farmers worldwide an access to the high-resolution and real-time yield forecasts. The options of the future may include the usage of more diverse data, like climate models, soil health information, and real-time weather forecasting to help in raising the accuracy of prediction even higher. Furthermore, the combination of the machine learning algorithms and the satellite images may result in the automatic and constant forecasts, which will open up the way to the precision-dependent agriculture.

Acknowledgement: Nil

Conflicts of interest

The authors have no conflicts of interest to declare

References

1. Hedley, C. B., & Keesstra, S. D. Machine learning in agriculture: A review of applications and challenges. Agricultural Systems. 2021; 190: 103106.
2. Benedetti, M., & Rizzo, D. Remote sensing in precision agriculture: Satellite applications in crop management. International Journal of Applied Earth Observation and Geoinformation. 2020; 89: 102067.
3. Jain, M., & Kumar, S. Yield prediction and classification in maize using machine learning algorithms. Computers and Electronics in Agriculture. 2019; 160: 18-29.
4. Liu, Q., & Zhang, X. NDVI time-series analysis and its relationship with maize yield prediction. Remote Sensing of Environment. 2020; 240: 111688.
5. Wang, J., & Zhang, Y. Machine learning applications in agricultural yield forecasting: A comprehensive review. Agricultural Systems. 2019; 175: 47-56.

6. Cheng, L., & Li, B. Use of XGBoost and random forest for predicting maize yield with remote sensing data. ISPRS Journal of Photogrammetry and Remote Sensing. 2021; 171: 60-73.
7. Patel, P., & Patel, S. Predicting crop yield using satellite data: Application to maize in India. Precision Agriculture. 2018; 19(4): 652-664.
8. Bai, Y., & Liu, H. A comparison of machine learning algorithms for maize yield prediction using remote sensing data. Sensors. 2021; 21(8): 2675.
9. Gao, Z., & Xie, W. Remote sensing-based maize yield prediction using machine learning methods. Agricultural and Forest Meteorology. 2020; 287: 107957.
10. Kumar, S., & Singh, R. Maize yield forecasting using time-series NDVI and EVI data with random forest model. Environmental Monitoring and Assessment. 2021; 193(6): 364.
11. Li, S., & Huang, H. Integration of satellite-based vegetation indices and machine learning algorithms for crop yield estimation. Journal of Agricultural and Food Chemistry. 2020; 68(4): 1106-1117.
12. Zhang, Q., & Li, T. Satellite-derived vegetation indices for early-season maize yield prediction: A case study using Sentinel-2 imagery. Remote Sensing Applications: Society and Environment. 2021; 23: 100460.
13. Niu, Z., & Huang, J. XGBoost-based crop yield prediction using multi-temporal satellite imagery. Remote Sensing. 2020; 12(4): 710.
14. Zhou, H., & Ren, X. Assessing machine learning models for crop yield forecasting using Sentinel-2 data and remote sensing technology. Journal of Precision Agriculture. 2021; 22(5): 1290-1304.