*environmental sciences proceedings*

MDPI

*Proceeding Paper*

# Enhancing Winter Wheat Yield Estimation Using Machine Learning and Fusion of Radar and Optical Satellite Imagery †

**Shabnam Asgari [1], Mahdi Hasanlou [1,*] and Saeid Homayouni [2]**

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran; asgari.shabnam@ut.ac.ir
[2] Centre Eau Terre Environnement, Institut National de la Recherche Scientifique, 490 Rue de la Couronne, Quebec City, QC G1K 9A9, Canada; saeid.homayouni@ete.inrs.ca
* Correspondence: hasanlou@ut.ac.ir
† Presented at the 5th International Electronic Conference on Remote Sensing, 7–21 November 2023; Available online: https://ecrs2023.sciforum.net/

**Abstract:** Accurate crop yield Mapping is paramount in agricultural monitoring and food security. In this study, we present a comprehensive investigation into estimating winter wheat yield in the Qazvin plane of Iran, leveraging the synergy between machine learning algorithms and the fusion of remote sensing data from radar and optical satellite sensors. The research is based on the availability of high-quality in situ yield data gathered by the Ministry of Agriculture in collaboration with the Food and Agriculture Organization (FAO), collected during the 2019-2020 crop year. The study area encompasses the Qazvin plane, an agriculturally significant region renowned for winter wheat production in Iran. In-situ data from various agricultural fields and seed types as reference measurements enabled us to conduct rigorous validation of the performance of machine learning algorithms and the effectiveness of the fused remote sensing data. The primary objective of this study is to assess and compare the performance of seven prominent machine learning algorithms for accurate estimation of the annual winter wheat yields. Furthermore, we investigate the individual and synergistic capabilities of radar and optical satellite sensors in estimating winter wheat yield. Through rigorous analysis of the pixel-level confusion matrices, we identify the most effective model for yield estimation, evaluating the complementarity and information redundancy between the two types of remote sensing data. In this study, we conducted an extensive comparison of various machine learning algorithms for winter wheat crop yield estimation in the Qazvin plane of Iran. Among the four best-performing algorithms examined, namely polynomial regression (RMSE = 0.5657 $t/ha^{-1}$), random forest (RMSE = 0.1632 $t/ha^{-1}$), XGBoost (RMSE = 0.3153 $t/ha^{-1}$), and the proposed Multi-Layer Perceptron (MLP) (RMSE = 0.1324 $t/ha^{-1}$), the MLP demonstrated superior performance. The MLP's yield estimation exceeded the total yearly agricultural statistics of Qazvin by 0.19 percent. However, this discrepancy can be attributed to various factors, including errors in wheat and barley field mapping, miscalculation in cumulative statistics, and the inherent limitations of yield estimation algorithms in capturing the dynamic nature of agricultural systems. The findings of this research provide valuable insights into the potential of machine learning algorithms and remote sensing data fusion for accurate crop yield estimation, paving the way for enhanced agricultural monitoring and decision-making processes in the region.

**Keywords:** winter wheat; crop yield estimation; machine learning algorithms; remote sensing data fusion; Sentinel 1/2

**Citation:** To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Published: date



## 1. Introduction

In recent years, agriculture has witnessed a paradigm shift closer to harnessing remote sensing instruments and superior technology to address meal protection and aid control challenges. Remote sensing technology provides real-time insights into crop

health, environmental situations, and land attributes, underpinning the concept of precision agriculture for optimized resource use and minimized environmental impact [1]. With the appearance of system studying and deep mastering, these facts are leveraged for certain insights into crop boom styles, ailment detection, and yield estimation [2,3]. In the endeavor to develop a winter wheat yield estimation model, this has a look at integrating the evolution of remote sensing generation and precision agriculture principles to drive the sphere closer to bloom accuracy and efficacy.

The launch of ESA's Sentinel-1 and Sentinel-2 remote sensing satellites has been a new generation in precision agriculture by way of imparting high-resolution, multi-spectral data crucial for tracking and coping with agricultural landscapes [2,4]. Sentinel-1, ready with Synthetic Aperture Radar (SAR) generation, gives a transformative perspective with its specific skills. Its all-weather imaging prowess allows the assessment of crop water content, soil moisture, and area conditions even in tough weather scenarios, bearing in mind non-stop monitoring and decreased monitoring gaps [4]. This exceptional potential to penetrate cloud cover and bring steady imagery enhances the reliability of plant tracking. In contrast, Sentinel-2's multispectral abilities complement Sentinel-1's strengths by delving into the intricacies of crop health, vegetation density, and land cover dynamics [5]. This synergy equips agricultural practitioners and researchers with a complete toolkit for knowledge and optimizing crop growth throughout numerous situations. By fusing information from these satellites, this observation harnesses the inherent strengths of Sentinel-1 and Sentinel-2 to raise winter wheat yield estimation, thereby advancing agricultural practices.

While the potential of Sentinel-1 and Sentinel-2 brings strengths such as all-weather imaging and comprehensive data, it also introduces limitations. Sentinel-1's limited spectral information challenges differentiation between vegetation types, even with its cloud-penetrating capabilities [6]. Conversely, while Sentinel-2 excels in assessing vegetation health and land cover dynamics, cloud cover may hinder the availability of high-resolution images, affecting data availability for analysis [5]. Therefore, a nuanced approach to combining the strengths of these sensors while navigating their limitations becomes paramount for realizing their full potential in enhancing yield estimation.

Recent years have seen the exploration and application of diverse machine-learning algorithms for wheat yield estimation. Algorithms such as Support Vector Machines (SVM) [5], Convolutional Neural Networks (CNN) [7], and Random Forest [8] leverage data-driven approaches to analyze intricate relationships [9] and forecast crop yield. The integration of these algorithms within a synergistic framework, alongside satellite imagery fusion, holds promise for heightening the accuracy of wheat yield estimation. Through the fusion of machine learning methodologies and remote sensing data, this study endeavors to propel yield estimation to new levels of precision, paving the way for advanced agricultural monitoring and decision-making processes.

To achieve the outlined principles, this study follows a comprehensive methodology. Firstly, it harnesses the complementary strengths of optical and radar satellite data to mitigate weather effects on optical images and enhance the quality of radar images. The fusion of these data sets enhances the accuracy and reliability of winter wheat yield estimation. In addition, in-season satellite image time series are used to capture the dynamic growth patterns of winter wheat, enabling accurate and timely yield predictions. To attain the most accurate yield estimation, eight prominent machine learning algorithms are systematically examined, evaluating their performance in capturing the intricate relationships between various parameters and crop yield. The accuracy of the proposed methodologies is rigorously assessed using in-situ field yield measurements provided by the Ministry of Agriculture of Iran, ensuring the reliability of the findings. Furthermore, the study's conclusions are bolstered by comparison with the annual statistical book of 2019-2020, providing a comprehensive validation of the yield estimation results.

In summary, this study bridges the historical principles of precise agriculture with modern technological advancements in remote sensing and machine learning. By fusing

optical and radar satellite data, employing in-season satellite imagery, and evaluating a diverse array of machine learning algorithms, this research endeavors to enhance winter wheat yield estimation. The application of these principles contributes to the advancement of agricultural practices, empowering decision-makers and stakeholders with accurate and reliable insights for informed choices in the pursuit of global food security.

## 2. Materials and Methods

### *2.1. Study Area*

The study area is Qazvin Province, located approximately 45 kilometers west of Tehran, Iran, with a land area of around 15,830 square kilometers. Geographically, it spans from 48 degrees and 45 minutes to 50 degrees and 50 minutes east longitude and 35 degrees and 37 minutes north latitude. Qazvin's climate varies from semi-humid temperate to very humid cold and cold-dry. According to the 2018-2019 Agricultural Census, Qazvin Province contributes 2.4% of Iran's total cultivated area, with an estimated 145,000 hectares under cultivation. Of this, 67% is rain-fed, while 33% is irrigated. The wheat planting season in the Qazvin agricultural plain begins in the fall, leading to harvest in the subsequent spring and summer so we only used satellite images during the winter wheat growing cycle (September -June).

### *2.2. In-Situ Data*

While the In-situ dataset served as the bedrock of our analysis, our research cast a wider net by also incorporating the annual provincial statistics released by the Statistical Center of Iran for the same crop season. These statistics, while invaluable in their own right, served as supplementary resources that augmented our understanding of crop yields in the Qazvin region. It's important to emphasize that our primary focus was on the in-situ data. We used approximately 50% of this data for training and the rest for testing, with a dataset of around 38,000 10-meter wheat pixels gathered from 105 winter wheat fields in area. This division of the in-situ data into 70% for training ensuring that model performs well on new, unseen data and doesn't overfit to the training data[10]. The reason for unwavering attention is twofold. Firstly, the Ministry's highly detailed data, encompassing both final crop yield measurements and unique field characteristics, provides a valuable resource for machine learning algorithms to uncover intricate insights into the factors affecting crop yields. Secondly, the in-situ dataset was crucial for training our machine-learning algorithms to match the specific dynamics of Qazvin's agriculture.

### *2.3. Satellite Images*

During the course of our research, we harnessed the power of satellite imagery over 10 critical months, encompassing the entire wheat growth cycle. To elevate the precision of our findings, we employed a dual-pronged approach, integrating both optical and radar imagery. This fusion of data sources allowed us to leverage the unique strengths of each technology, resulting in a comprehensive and highly accurate assessment of the evolving wheat fields.

#### 2.3.1. Sentinel-1

The Sentinel-1 satellite utilizes a C-band radar sensor acquiring data in dual polarization modes, VV and VH, enabling the capture of electromagnetic waves in various configurations. Data can be acquired from both ascending and descending orbits, with a notable difference being the more pronounced radar backscatter values in ascending orbits due to steeper incidence angles. The preprocessing of Sentinel-1 GRD (Ground Range Detected) data has been done in several key stages. Initially, data filtering is applied based on parameters such as date (2019-2020), polarization (VV and VH), and incidence angle, tailored to the study's specific requirements, including temporal alignment with the wheat growth period and spatial coverage of Qazvin Province. Subsequently, the region-of-

interest clipping isolates relevant satellite data to reduce volume and maintain spatial precision, with the administrative boundary of Qazvin Province serving as the clipping layer, one image per month was chosen. Calibration is the next step, converting digital units to physical units through adjustments of historical calibration constants and value conversion to dB (decibels). Finally, speckle filtering is applied to mitigate speckle noise inherent in radar images, preserving essential details while removing interference patterns. The Goldstein algorithm is utilized for this purpose. In the context of our study, these four essential preprocessing steps are sufficient for radar data, making the images ready for subsequent analysis and utilization as outputs.

#### 2.3.2. Sentinel-2

In the selection of Sentinel-2 images for our study, we adopted a meticulous approach to ensure the highest data quality and minimal cloud cover. From the available images for each month, we carefully handpicked pixels with superior quality and the lowest cloud coverage to represent that specific month. This selection process was crucial to maintain data quality and precision throughout our analysis.

In the data preprocessing stage for Sentinel-2 data, we began by getting 13-band images from the Sentinel-2 Level-1C archive through the Google Earth Engine interface. Subsequently, we filtered all images based on the date (September 2019 - June 2020) and the study area encompassing Qazvin Province. Further refinement involved selecting images with cloud coverage of less than 10%. We then employed an algorithm for cloud masking within the Sentinel images, reducing multiple monthly images into a single image using image collection reduction tools. In this final step, we selected the pixel with the least cloud cover from all images in a given month to ensure the highest-quality image for each specific month.

### *2.4. Regression Models*

In our study, we employed a diverse set of eight regression models to predict and analyze winter wheat yield. The first set of models included traditional linear and polynomial regressions, which provided a baseline understanding of the relationship between input features and wheat yield, considering both linear and nonlinear patterns. To address potential issues of multicollinearity and overfitting, we incorporated regularized regression techniques like Ridge and Lasso regression. These methods helped us strike a balance between model complexity and accuracy by penalizing large coefficients, ultimately enhancing the robustness of our predictions. Moving beyond linear models, we explored Bayesian regression, which incorporates probabilistic principles to capture uncertainty in our predictions. Additionally, ensemble methods such as Random Forest Regression and Gradient Boosting Regression were utilized to harness the collective predictive power of multiple decision trees. Lastly, we leveraged Multi-Layer Perceptron (MLP) regression, a type of artificial neural network, to model complex, nonlinear relationships within the data, allowing for highly flexible and adaptable predictions. These diverse regression techniques collectively contributed to a comprehensive and nuanced analysis of winter wheat yield prediction, enabling us to extract valuable insights and improve model performance.

In order to optimize the performance of our regression models, we employed a meticulous approach to select hyperparameters and conduct cross-validation. We adopted two distinct strategies: grid search and genetic algorithms. Grid search systematically explored a predefined range of hyperparameters for each model, exhaustively evaluating their combinations to identify the set that yielded the best performance. This method provided a comprehensive view of the hyperparameter landscape but could be computationally expensive. In parallel, genetic algorithms were utilized as a more heuristic approach, emulating the process of natural selection to iteratively evolve and adapt hyperparameter configurations over multiple generations. This allowed us to strike a balance between

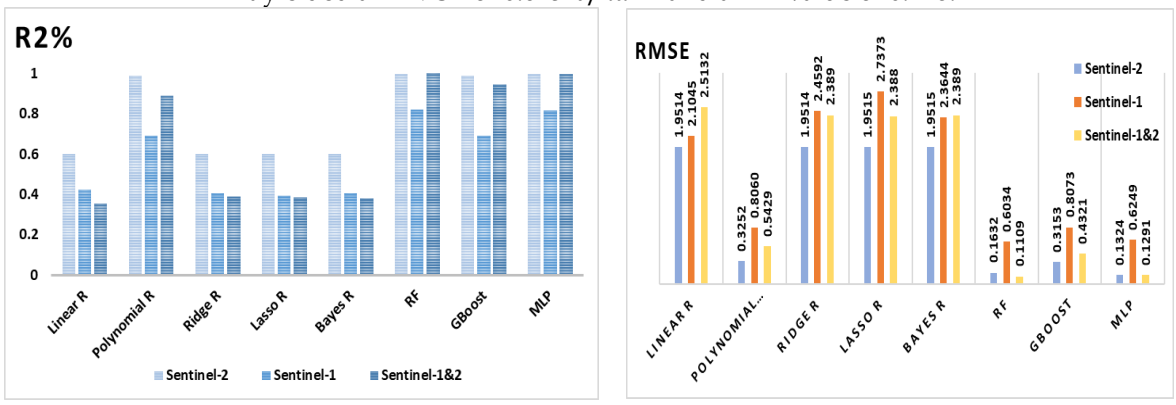
exploration and exploitation, seeking optimal hyperparameters while mitigating computational costs. The combination of these approaches ensured that our regression models were fine-tuned to deliver robust and accurate predictions for winter wheat yield estimation.

To assess the performance of the diverse regression models employed in our study, we utilized two widely recognized metrics, $R^2$ (Coefficient of Determination) and RMSE (Root Mean Square Error). $R^2$ provides insight into the proportion of the variance in the dependent variable that is explained by the model, giving a measure of goodness-of-fit. It ranges from 0 to 1, with higher values indicating a better fit. On the other hand, RMSE quantifies the average prediction error, providing a measure of model accuracy. Lower RMSE values signify better predictive performance. By applying these metrics, we were able to quantitatively evaluate and compare the models' abilities to predict winter wheat yield, ensuring that our analysis was not only comprehensive but also grounded in rigorous statistical assessment.

## 3. Results

### *3.1. Pixel-Based Results*

In this study, we investigated the utilization of machine learning and the fusion of radar and optical satellite imagery to enhance the estimation of winter wheat yields. Our research employed time series satellite images of winter wheat within a designated study area in Iran. To achieve this, we trained eight machine-learning algorithms on the satellite images, utilizing both optical and radar sensors. Subsequently, we assessed the models' performance by comparing the predicted yields with the actual yields. The results clearly indicate that models incorporating both optical and radar images exhibited lower Root Mean Square Error (RMSE) values and higher R-squared ($R^2$) values compared to models relying solely on optical images. Notably, the top-performing model was the random forest model, which achieved an RMSE of 0.559 $t/ha^{-1}$ and an $R^2$ value of 0.841 when utilizing both optical and radar images. In contrast, when exclusively using optical images, it yielded an RMSE of 0.625 $t/ha^{-1}$ and an $R^2$ value of 0.778.




(**a**) (**b**)

**Figure 1.** This figure shows the results of the eight machine learning algorithms when using only optical images, only radar images, and both optical and radar images. (**a**) The $R^2$ values are expressed as percentages for the different models. (**b**) The RMSE values are expressed in tons per hectare for the different models.

These findings underscore the potential of radar and optical image fusion in improving the accuracy of winter wheat yield estimation. This enhancement is achieved through the reduction of errors between predicted and actual yields, as well as improved model fitting to the data. The benefits of employing multiple satellite types are multifaceted. Firstly, it aids in overcoming challenges such as low image quality and cloud cover during Iran's snowy seasons (January and February), ensuring more comprehensive in-season

time series data. Secondly, each sensor type provides distinct data, and when harnessed alongside sophisticated modeling techniques, as evidenced in Figure 1, the fusion of data contributes to heightened accuracy. Notably, algorithms like linear, Bayes, Lasso, and Ridge regression, grounded in basic mathematical computations, demonstrated weaker performance when contrasted with more advanced and intricate algorithms such as Random Forest, MLP, Gradient Boosting, and Polynomial Regression. The preeminent choice, the random forest model, excelled in yield prediction, primarily due to its non-parametric nature, enabling it to capture intricate relationships between features and the target variable.

The limitations of this study include the field size samples and the fact that the study was conducted in a single location. Future studies should be conducted with pixel-size in-situ data by using more accurate sampling methods and in multiple locations to confirm the results of this study.

### *3.2. Region-Based Results*

To calculate the total estimated crop yield in the agricultural region of Qazvin province, we initially imported the shapefile layer of Qazvin province's borders and the classified field data as masks for wheat and barley fields into Google Earth Engine. Subsequently, we extracted image layers for the masked fields in Qazvin province and inputted the output data into eight regression algorithms. The output results, measured in tons per hectare, were then multiplied by the area of each pixel (100 square meters = 0.01 hectares), and the totals for wheat and barley pixels were summed.

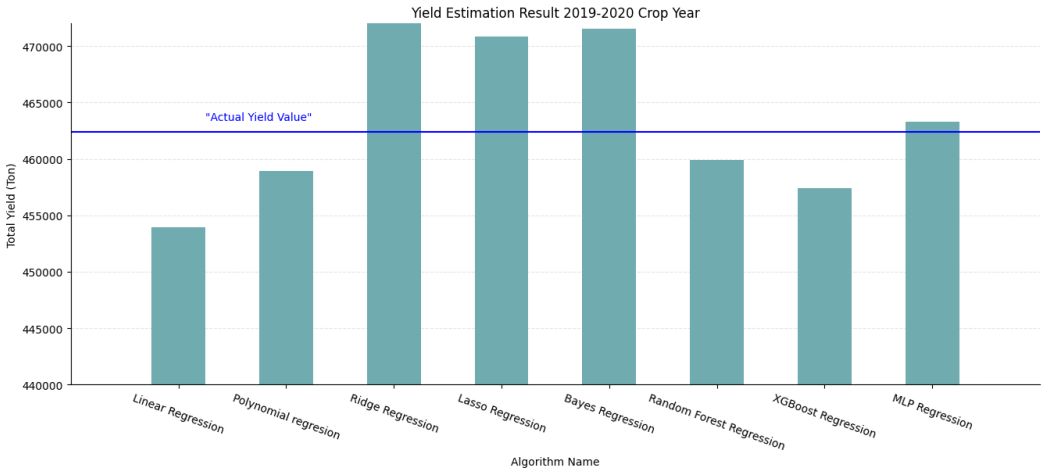



**Figure 2.** Illustrates the comparison between the results of the eight machine learning algorithms and the actual yearly yield statistics for the region.

As depicted in Figure 2., the closest estimate to the annual statistical yield reports was achieved by the MLP algorithm, given its complexities and more accurate responses. Another noteworthy observation is the estimated values produced by the Bayesian, Lasso, and Ridge algorithms, which rely on statistical methods. All three algorithms estimated values higher than the actual crop yield, attributed to their similarity in the underlying statistical approaches. The estimated values by Random Forest and Polynomial Regression algorithms were also close to the actual values. Linear Regression yielded the lowest crop yield estimate, while Ridge Regression provided the highest estimate.

## 4. Results and Discussion

In conclusion, our study underscores the substantial potential for improving winter wheat yield estimation by fusing radar and optical satellite imagery. Among the range of regression models explored the random forest algorithm emerged as the most effective

choice for yield prediction, offering not only strong performance but also ease of training. Incorporating the results of our regression models, it is evident that these models consistently produced yield predictions that closely aligned with yearly statistical data. Specifically, our linear regression model achieved a minimal deviation of only 1.837% lower than the actual statistics, while the polynomial regression model demonstrated a similarly impressive performance with just a 0.756% deviation. Notably, the random forest model showcased exceptional accuracy, with predictions nearly identical to the yearly statistics, deviating by a mere 0.545%. However, the XGBoost model displayed predictions that were 1.076% lower than the yearly statistics. In contrast, the Bayesian regression model yielded predictions that were 1.970% higher, Ridge 2.256% higher, Lasso 1.823% higher, and the Multi-Layer Perceptron (MLP) regression model exhibited a minimal deviation of only 0.190% higher than the yearly statistics. These results underscore the precision of our models in capturing the complexities of winter wheat yield estimation, further highlighting the potential for improving agricultural management and food security on a global scale.

## References

1. Krishnan, P. Using Satellite Imagery Datasets in the Agriculture Sector. *Medium* 2022.
2. Remote Sensing | Free Full-Text | Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China Available online: https://www.mdpi.com/2072-4292/12/2/236 (accessed on 29 August 2023).
3. Cheng, E.; Zhang, B.; Peng, D.; Zhong, L.; Yu, L.; Liu, Y.; Xiao, C.; Li, C.; Li, X.; Chen, Y.; et al. Wheat Yield Estimation Using Remote Sensing Data Based on Machine Learning Approaches. *Frontiers in Plant Science* **2022**, *13*.
4. Ahmadian, N. Integrating Satellite Remote Sensing and In-Situ Measurements to Estimate the Biophysical Parameters of Agricultural Crop Using Multispectral and Radar Data. **2017**.
5. Bogdanovski, O.P.; Svenningsson, C.; Månsson, S.; Oxenstierna, A.; Sopasakis, A. Yield Prediction for Winter Wheat with Machine Learning Models Using Sentinel-1, Topography, and Weather Data. *Agriculture* **2023**, *13*, 813, doi:10.3390/agriculture13040813.
6. Castro Gómez, M.G. Joint Use of Sentinel-1 and Sentinel-2 for Land Cover Classification: A Machine Learning Approach Available online: http://essay.utwente.nl/83458/ (accessed on 29 August 2023).
7. Sharifi, A. Yield Prediction with Machine Learning Algorithms and Satellite Images. *J Sci Food Agric* **2021**, *101*, 891–896, doi:10.1002/jsfa.10696.
8. Asgari, S.; Hasanlou, M. A COMPARATIVE STUDY OF MACHINE LEARNING CLASSIFIERS FOR CROP TYPE MAPPING USING VEGETATION INDICES. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **2023**, *X-4-W1-2022*, 79–85, doi:10.5194/isprs-annals-X-4-W1-2022-79-2023.
9. Rostami, A.; Akhoondzadeh, M.; Amani, M. A Fuzzy-Based Flood Warning System Using 19-Year Remote Sensing Time Series Data in the Google Earth Engine Cloud Platform. *Advances in Space Research* **2022**.
10. Aghdami-Nia, M.; Shah-Hosseini, R.; Rostami, A.; Homayouni, S. Automatic Coastline Extraction through Enhanced Sea-Land Segmentation by Modifying Standard U-Net. *International Journal of Applied Earth Observation and Geoinformation* **2022**, *109*, 102785.