remote sensing

MDPI

Article

Exploring the Use of High-Resolution Satellite Images to Estimate Corn Silage Yield Within Field

Srinivasagan N. Subhashree [1], Manuel Marcaida III [1], Shajahan Sunoj [1], Daniel R. Kindred [2], Laura J. Thompson [3] and Quirine M. Ketterings [1,*]

[1] Nutrient Management Spear Program, Department of Animal Science, Cornell University, Ithaca, NY 14853, USA; sn558@cornell.edu (S.N.S.); mmarcaida@cornell.edu (M.M.III); sunoj@illinois.edu (S.S.)
[2] ARC Agronomic Sciences Ltd., Woodbridge IP13 9PG, UK; daniel@arcagsci.com
[3] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583, USA; laura.thompson@unl.edu
* Correspondence: qmk2@cornell.edu; Tel.: +1-607-255-3061

Abstract: Corn (*Zea mays* L.) silage yield monitor data offer crucial insights into spatial and temporal yield variability. However, equipment's sensor malfunctioning can result in data loss, and yield sensor systems are expensive to purchase and maintain. In this study, we analyzed corn silage yield data from two fields and three years each for two dairy farms (Farm A and B). We aimed to explore the potential of integrating high-resolution satellite data, topography, and climate data with machine learning models to estimate missing yield data for a field or a year. Our objectives were to identify key yield-explaining features and assess the accuracy of different machine learning models in estimating silage yield. Results showed that the features differed among farms with a Two-Band Enhanced Vegetation Index, EVI2 (Farm A), and elevation (Farm B) emerging as the most prominent predictors. Ensemble-based models like XGBoost, Random Forest, and Extra Tree regressors exhibited superior predictive performance. However, XGBoost performed poorly when applied to unseen fields or years, whereas Extra Tree regressor, followed closely by Random Forest, emerged as a more reliable model for predicting missing data. Despite achieving reasonable accuracy, the best performance for estimating data for a missing field (6.46 Mg/ha) and year (5.51 Mg/ha) fell short of the acceptable error threshold of 4.9 Mg/ha currently used in state policy to evaluate if a management change resulted in a yield increase. These findings emphasize the need for higher-resolution data and extended years of yield records to better capture the trends in farm-scale yield applications.

Keywords: corn yield; feature selection; machine learning; remote sensing; yield estimation

check for updates

Citation: Subhashree, S.N.; Marcaida, M., III; Sunoj, S.; Kindred, D.R.; Thompson, L.J.; Ketterings, Q.M. Exploring the Use of High-Resolution Satellite Images to Estimate Corn Silage Yield Within Field. *Remote Sens.* **2024**, *16*, 4081. https://doi.org/10.3390/rs16214081

Academic Editor: Jianxi Huang

Received: 30 August 2024
Revised: 15 October 2024
Accepted: 29 October 2024
Published: 1 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Accurate determination of corn (*Zea mays* L.) silage yield is important in New York's agricultural landscape, given its status as a dairy state. Corn silage is the primary feed source for dairy cattle, directly influencing milk production. Therefore, it is crucial for ensuring a steady supply of feedstock for sustainable production. The number of farms that harvest with forage yield monitor systems capable of recording spatial yield points at one-second intervals during harvest has grown over the past 15 years [1]. With accurate yield maps, farmers can better target crop management inputs such as seed and fertilizer. Previous studies have leveraged yield monitor data to analyze yield trends, identifying spatial and temporal patterns to delineate areas with high, low, and variable yields within fields [2–4]. These data have also been crucial for setting field-specific yield goals and adaptive nitrogen management, where precise yield measurement is a key component for setting and evaluating crop nitrogen application rates [5].

Despite important technology advances, yield monitor data may not always provide accurate measurements stemming from factors such as sensor calibration errors and equipment malfunctioning, creating a risk of loss of important information for not just one field

but likely for more fields on a farm [6]. As harvest cannot be undone, this causes a loss of valuable information across the farm in a given year. Inaccurate or missing field data can lead to resource misallocation, disruption in feed planning, inefficient farm management, and gaps in long-term yield insights. In addition, post-processing of raw yield data is essential to rectify sensor delays and ensure accurate calibration [7], and this is a time-consuming step. Research is needed to evaluate other approaches that can be used to estimate missing field data.

Recent studies have showcased the potential benefit of topography, climate variables, and raw bands, as well as soil and vegetation indices derived from remote sensing data, for estimating corn yield [8–10]. Topography influences spatial variation as it can regulate water drainage and retention across the landscape. Various topographic variables, such as elevation, slope, landform, topographic wetness, and topographic position index, can be used to evaluate the impact of topography on yield. A study employing high-resolution drone imagery (0.3 m) to estimate corn yield and soil properties discovered that among the seven topographic factors examined, elevation exhibited a significant correlation with raw bands like green and blue, as well as soil indices such as the redness index (RI), indicating the potential for elevation data for accurately explaining and estimating spatial variability of soil properties and yield [11]. Furthermore, a field-scale study estimating corn yield using PlanetScope satellite data (3 m) demonstrated substantial accuracy by integrating topographic wetness index, topographic position index, and raw spectral bands within a stacked ensemble model [9].

Climate variables also play a crucial role in corn yield prediction in rainfed agriculture, due to their direct and indirect impacts on crop growth and yield. A study estimating corn and soybean yield based on Moderate Resolution Imaging Spectroradiometer (MODIS) data (232 m) showed that the inclusion of weather-related variables along with remote sensing data improved the accuracy of the model for data collected from rainfed areas [12]. Similarly, another study spanning twelve states within the US Corn Belt using the MODIS (500 m) dataset showcased enhanced corn yield estimation and enabled uncertainty analysis by integrating key climate variables such as temperature, precipitation, and vapor pressure [13]. Actual spectral reflectance or raw bands from remote sensing images provide valuable information by themselves. A comparison study used spectral reflectance from four satellite platforms (3–30 m) and found that the higher spatial resolution captured most of the within-field variability of corn and soybean yield, where raw spectral bands such as green, red-edge, and near-infrared were the most significant in explaining the variability [14].

Vegetation indices (VIs) have been introduced using raw reflectance bands for monitoring crop growth, health, and yield. Some of the prominent VIs for corn yield estimation are the Normalized Difference Vegetation Index (NDVI), Green NDVI (GNDVI), Green Chlorophyll Vegetation Index (GCVI), Simple Ratio (SR), and the Enhanced Vegetation Index (EVI). For several decades, NDVI has been used as an important VI for estimation and early season prediction of corn yield. Recent studies with satellite and drone imagery have shown the rise of GCVI and GNDVI in explaining corn growth [15–19], given their sensitivity to chlorophyll content and vegetation greenness, both of which are closely related to biomass accumulation.

In addition to the use of indices for monitoring vegetation growth, soil indices can be used to monitor soil physical and chemical properties of importance for agricultural management and land use planning. Several soil indices such as hue, saturation, and brightness, among others, have been used for corn yield prediction [11,13,20].

The interplay between soil, climate, and topographical information is crucial for accurately estimating yield. However, traditional statistical models may struggle with the high-dimensional and non-linear nature of the data. In contrast, machine learning (ML) algorithms have demonstrated promising performance in yield estimation and prediction, particularly in handling complex data within intricate agricultural systems. Decision tree-based models have recently gained popularity due to their predictive capabilities with crop

yield. A study on corn grain yield estimation, which examined seven ML regression models using 33 vegetation indices extracted from multispectral UAV imagery, concluded that Random Forest outperformed other models, especially when trained with all VIs and the top three VIs [21]. Similarly, another study comparing various statistical and ML methods for estimating soybean (*Glycine max* (L.) Merr) and corn yield based on soil and topographic characteristics using high-resolution field-scale data found that Random Forest was the most effective in estimating corn grain yield, while multiple linear regression was the least effective [22]. While ML has been increasingly applied to estimate corn yield, most studies have concentrated on corn grain yield. Unlike grain yield, corn silage involves the harvest of the entire plant biomass. Remote sensing may be a more suitable method for silage yield predictions as it directly captures the total biomass visible to satellite platforms, while corn grain yields can only be partially assessed through remote sensing of vegetative growth stages. Given the challenges associated with the accuracy and reliability of harvester-mounted yield monitor systems, this study seeks to investigate satellite platforms as a viable alternative for on-farm corn silage yield estimation. Our unique focus is on building ML models by integrating high-resolution satellite imagery with topographic and climate data for individual fields and years and evaluating their accuracy for estimating corn silage yield for two practical scenarios: (1) a year in which yield data were lost for a nearby field (missing field); and (2) a field for which we have a year of data but additional years are missing (missing year).

The specific objectives of this study are the following: (i) identify important features to explain corn silage yield; (ii) evaluate and select the most effective ML model for accurately estimating corn silage yields; (iii) assess the effectiveness of the ML model developed using nearby fields' data to predict the missing field's yield for the same year (farm-wise fit); and (iv) evaluate the ML model's performance, built using data from multiple years, for predicting a missing year's yield for the same field (year-wise fit).

2. Materials and Methods

2.1. Experimental Fields

This study utilized data from four commercial corn fields of two dairy farms in northern New York (Table 1). These farms were selected based on the availability of yield monitor data and a minimum of three consecutive years of data to ensure reliability in data analysis and examine trends over time.

Table 1. Distribution of yield data from silage fields within Farms A and B.

Year	Farm	Field	Area (ha)	Yield (Mg/ha)					
				Min	Max	Mean	Q1	Q2	Q3
1	A	A1	45.6	27.20	79.49	52.90	48.01	53.27	58.23
		A2	30.4	8.43	78.14	45.82	38.01	48.96	55.27
	B	B1	65.8	28.50	91.77	52.13	47.67	51.11	55.58
		B2	41.8	42.89	96.21	67.04	63.75	67.7	71.17
2	A	A1	45.6	32.67	86.17	60.43	55.91	60.84	65.26
		A2	30.4	19.50	78.4	49.12	43.98	50.16	55.57
	B	B1	65.8	28.91	87.64	60.19	57.15	60.2	63.48
		B2	41.8	30.35	94.66	63.93	60.43	64.11	67.69
3	A	A1	45.6	7.16	82.05	47.00	40.37	49.18	54.63
		A2	30.4	5.93	70.47	36.98	24.70	39.10	49.20
	B	B1	65.8	24.10	76.46	53.94	50.48	53.77	57.40
		B2	41.8	40.34	93.17	61.17	57.96	61.23	64.48

Q1—first quartile (25th percentile); Q2—second quartile (50th percentile); and Q3—third quartile (75th percentile).

Each farm supplied data for two nearby silage fields for three years. The size of the fields ranged from 30.4 to 65.8 ha. Within each farm, one of the fields (designated A1 and B1) was 1.5 times larger than the other (designated A2 and B2). Despite some minor

variations, similar patterns in yield were noted across both farms and fields. While the distribution of yield fluctuated from year to year, no consistent trend in yield between the fields was evident, as observed in New York silage fields [3]. The average annual rainfall for Farm A ranges between 12 and 25 mm, and for Farm B between 9 and 17 mm. Neither farm used irrigation, consistent with the way corn is grown in New York. The average daily temperature was 19–20 °C at Farm A and 16–18 °C at Farm B. Elevations for Farm A spanned 75 to 80 m above mean sea level, while Farm B ranged from 92 to 102 m. The soil type for Farm A consisted of silty clay loam, whereas Farm B had loamy fine sand.

2.2. Data Collection

2.2.1. Yield Monitor Dataset for Ground-Truth

Yield data were acquired from the forage choppers fitted with yield monitor systems. The raw yield monitor data were calibrated and processed for sensor delays [7]. The cleaned yield data were rasterized using the Matern kriging approach, selected for its reliability and minimal error among other spatial interpolation techniques [23]. Kriging was employed for the smoothness of the estimates of yield across the entire field by estimating values at the points where yield monitor data was not available. The resulting kriged data (2 × 2 m pixel size) served as ground-truth data.

2.2.2. Satellite Imagery

Satellite imagery from the PlanetScope constellation (Dove satellite) was acquired through PlanetLabs PBC using a research and education license [24]. The data was accessed and downloaded through the PlanetExplorer plugin in QGIS (version 3.4.11). The plugin facilitated the specification of search parameters such as cloud cover percentage and date range. Cloud cover was restricted to <15%, and the search period was defined from planting to harvest. The resulting satellite images consisted of four spectral bands: blue (455–515 nm), green (500–590 nm), red (590–670 nm), and near-infrared (NIR, 780–860 nm). Each image had a resolution of 3.125 m/pixel. Over the course of each year and field, 15–20 satellite images were downloaded during the search period (April–October). The NDVI was calculated using the red and NIR bands to determine the peak canopy period, which was chosen as it reflects the time of maximum vegetation cover, typically corresponding to potential yield. Subsequently, three peak canopy dates, which could vary by year due to climatic changes or growing conditions, were selected for further analysis, including feature selection and machine learning model building. In addition to the peak canopy images, bare soil images of the fields were also collected by changing the search period to one or two days prior to planting.

2.2.3. Climate Data

Spatial climate data from NASA POWER were accessed from the website http://power.larc.nasa.gov/ (accessed on 22 June 2023) and provided detailed information about various climate parameters across different geographical locations. These datasets typically include information such as temperature, precipitation, humidity, solar radiation, and wind speed, among others, at daily frequency. NASA POWER collects and processes data from multiple sources, such as satellite observations, weather stations, and climate models, and provides these datasets at a coarse resolution (~50 km). The spatial climate data source allows access to historical data as well as real-time and forecast data depending on the availability. In this study, 14 spatial climate datasets sourced from NASA POWER were incorporated as additional data layers to test and enhance the estimation of corn yield. Due to the coarser resolution, within-field weather variability was not captured. Rather, each weather variable represented a single daily average observation at the three peak canopy dates for each field.

2.2.4. Topographic Information

Elevation and landform topographic data were extracted from the digital elevation map (DEM). The DEM for each field was acquired from the New York State Geographic Information System Service (https://gis.ny.gov/ (accessed on 22 June 2023)). Elevation information was directly obtained from the DEM images at a spatial resolution of 1 m/pixel. The same DEM images were utilized to derive landform information, although they were resampled to a 5 m pixel resolution, which was deemed suitable for characterizing landforms in vegetated landscapes [25]. The landform analysis involved the calculation of the topographic position index (TPI), which compares the elevation of a specific location to that of its neighboring cells within a defined window or pixel (5 m). Subsequently, each pixel was categorized into either of the six classes: valley, lower slope, middle slope, upper slope, flat area, and ridge [26]. Landform was the only categorical variable used in this study.

2.3. Data Extraction

A custom Python (version 3.9) program pipeline was employed for data extraction. The spatial resolution of the collected data layers varied: kriged yield monitor data (2×2 m), remote sensing data encompassing surface reflectance bands, VIs, and soil indices (3×3 m), topographic information including elevation (1×1 m) and landform (5×5 m), and climate variables. To standardize the resolution across all layers, the kriged yield data served as a reference layer. This involved converting the kriged yield data into a vector layer of a 2×2 m vector grid. Subsequently, all other raster layers were resampled and superimposed on this vector grid for value extraction. Zonal statistics were applied to each grid cell to compute average values within a 2 m vector grid for continuous variables, including spectral indices, elevation, and climate variables. For the landform layer, which had a resolution of 5×5 m and was categorical, point sampling was conducted for each 2×2 m grid cell to extract the corresponding landform category values. This approach provided an accurate and standardized dataset for subsequent model development and analysis.

2.3.1. Surface Reflectance Bands

The reflectance bands alone can serve as good indicators of crop yield. Therefore, the reflectance bands of the satellite image, including red, green, blue, and NIR, were incorporated to investigate their potential correlation with yield.

2.3.2. Soil Indices

Soil indices were derived from the bare soil imagery captured before planting. Five specific soil indices were chosen based on their capacity to serve as proxies for soil quality and characteristics [27]. These indices represent the visual reflectance of the soil, with brightness indicating the degree of lightness or darkness, potentially reflecting soil moisture and organic matter content. The brightness index (BI) measures the global reflectance of the soil surface capturing the brightness and intensity by calculating Euclidean distance in the spectral space that helps distinguish soil types. The saturation index (SI) measures the intensity and vividness of soil color and offers insights into soil texture. The hue index (HI) delineates the predominant color tones present in the soil, providing information about soil composition. Additionally, the coloration index (CI) combines brightness, saturation, and hue, providing a comprehensive understanding of soil properties and their effect on corn silage yield. Lastly, the redness index (RI) emphasizes the redness in the soil which can be indicative of iron oxide concentration that could impact soil fertility. Each soil index was derived as follows:

$$\mathrm{BI} = \sqrt{\frac{\mathrm{Red}^2 + \mathrm{Green}^2 + \mathrm{Blue}^2}{3}} \quad (1)$$

$$\mathrm{SI} = \frac{\mathrm{Red} - \mathrm{Blue}}{\mathrm{Red} + \mathrm{Blue}} \quad (2)$$

$$\mathrm{HI} = \frac{2 \times \mathrm{Red} - \mathrm{Green} - \mathrm{Blue}}{\mathrm{Green} - \mathrm{Blue}} \quad (3)$$

$$\mathrm{CI} = \frac{\mathrm{Red} - \mathrm{Green}}{\mathrm{Red} + \mathrm{Green}} \quad (4)$$

$$\mathrm{RI} = \frac{\mathrm{Red}^2}{\mathrm{Blue} \times \mathrm{Green}^3} \quad (5)$$

2.3.3. Vegetation Indices

Seven VIs were computed from the extracted surface reflectance bands based on their effectiveness in estimating crop yield. The NDVI [28]) is the most widely used VI. Others included were the GNDVI [29], the Two-Band Enhanced Vegetation Index (EVI2, [30]), GCVI [31], and SR [32], all of which incorporate NIR and visible bands. Indices like the Excess Green Index (ExG, [33]) and the Triangular Greenness Index (TGI, [34]), which only utilize visible band wavelengths, were also included.

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red}} \quad (6)$$

$$\mathrm{GNDVI} = \frac{\mathrm{NIR} - \mathrm{Green}}{\mathrm{NIR} + \mathrm{Green}} \quad (7)$$

$$\mathrm{EVI2} = \frac{2.5 \times (\mathrm{NIR} - \mathrm{Red})}{\mathrm{NIR} + 2.4 \times (\mathrm{Red} + 1)} \quad (8)$$

$$\mathrm{SR} = \frac{\mathrm{NIR}}{\mathrm{Red}} \quad (9)$$

$$\mathrm{GCVI} = \frac{\mathrm{NIR}}{\mathrm{Green}} - 1 \quad (10)$$

$$\mathrm{ExG} = 2 \times \mathrm{Green} - \mathrm{Red} - \mathrm{Blue} \quad (11)$$

$$\mathrm{TGI} = \mathrm{Green} - 0.39 \times \mathrm{Red} - 0.61 \times \mathrm{Blue} \quad (12)$$

2.4. Data Processing

The data extracted from various layers in the satellite imagery underwent data cleaning, involving the identification of valid data points and removal of missing values. This process was based on three criteria: missing values, replacing values, and consistency in pixels across years. The missing values were particularly noticed at the field edges. The yield grid was at 2 × 2 resolution, while all the other layers were at a lower resolution, except the DEM layer. Overlaying the 2-m grid over these layers resulted in missing values at the field edges as the pixels were misaligned. These values were removed using the 'dropna()' function. Landform was the sole categorical variable among all the features. To address the dimensionality issue that this generates, and to capture the relationship between the category and the target variable (corn silage yield in this context), the categorical landform data were converted into a numeric format using the target encoding method, also known as mean encoding or likelihood encoding [35]. For each category in the landform, the average value of corn silage yield was calculated, and each category was then replaced with its corresponding average corn silage yield value.

2.5. Data Analysis and Modeling

The Google Colab platform was used for feature selection and training ML models due to its capability to fulfill the demand for hardware and Graphic Processing Unit (GPU) resources. RAPIDS, built on top of NVIDIA's Compute Unified Device Architecture (CUDA) platform, offers a parallel computing framework for GPU acceleration, facilitating significantly faster processing of large datasets compared to traditional Central Processing Unit (CPU)-based approaches. Datasets were uploaded to Google Drive, and the outputs were stored there as well, simplifying access from within Google Colab. Feature selection

and ML model building were performed on the larger fields (A1, B1) incorporating data from all three years within a farm. The smaller fields (A2 and B2, unseen data) were utilized to validate the developed models.

2.5.1. Feature Selection

In this study, the features from all the data layers resulted in a total of 32 features. Building ML models with all the features would be computationally inefficient. Therefore, we used Recursive Feature Elimination to identify and select the top 10 important features. RFE is a wrapper-based algorithm that utilizes the Random Forest model (ML) to identify the optimal features. In Python, the RFE feature selection process was conducted using the 'rfe()' function within the RFE class from 'sk.feature_selection' module. Initially, all features were utilized, and the RFE method iteratively eliminated the feature with the lowest Root Mean Square Error (RMSE) calculated from the out-of-bag (OOB) data. Subsequently, a new Random Forest model was developed using the remaining features. This process was iteratively applied using 10-fold cross-validation (repeated 5 times) to refine the variable selection process and identify the most crucial features. At each iteration, the Random Forest model with the selected subset of features yielding the lowest RMSE was considered optimal. However, the feature rankings were updated if another model with a different subset of features resulted in a lower RMSE. Through this iterative process, RFE determined the best subset of features that minimized the RMSE value. Feature selection was performed individually for data pertaining to three peak canopy dates in a year, and the rankings were aggregated to obtain an overall ranking for that particular year. Following that, the top ten features were selected for training the ML models.

2.5.2. Model Selection and Training

The new dataset contained only the top ten features identified as important according to the RFE results. Subsequently, this filtered dataset was partitioned into training and testing sets, with 80% for training and 20% for testing. The training data was utilized to develop ML models. PyCaret, a Python library, was employed to explore various ML regression models [36]. PyCaret integrates with RAPIDS to leverage GPU acceleration, significantly boosting computational efficiency for ML models. It simplifies regression tasks by automating workflows such as cross-validation and model interpretation while offering a range of performance metrics (e.g., MAE, RMSE, R-squared) and visualizations for easy comparison of different models.

The ML models were ranked based on their performance metrics, and among the 20 models tested, the top three models were selected to construct individual models. The selected models were trained using the data from the larger fields (A1 and B1) across all three years from Farms A and B and were used for hyperparameter tuning.

2.6. Model Performance Assessment

2.6.1. Hyperparameter Tuning

The top three ML models were subjected to hyperparameter tuning. The randomized search method was employed for efficient and high-dimensional parameter search as opposed to a conventional grid search, which can be exhaustive because it searches from all possible combinations. By evaluating a subset of parameter combinations, the randomized search is effective in finding good hyperparameter settings in less time [37].

2.6.2. Performance Metrics

The performance of the top three models was assessed using a 10-fold cross-validation technique, repeated 5 times. The evaluation was based on two commonly used regression performance metrics: RMSE and MAE). The RMSE quantifies the average deviation between predicted and actual values, considering the magnitude of errors. Conversely, MAE calculates the average absolute deviation between predicted and actual values. In

general, MAE is less influenced by outliers than RMSE, offering a more robust measure of the average prediction error.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (13)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (14)$$

where n = total number of observations, y_i is the actual yield, $\hat{y}_i$ is the estimated yield.

2.6.3. Validation of the Developed Model

Farm-level fitting ("Farm-wise fit") involved assessing whether missing yield monitor data can be substituted using a model trained on a nearby field. Specifically, the model trained using data from fields A1 and B1 were utilized to estimate yield values for nearby fields A2 and B2 across each year. All three hyper-tuned models were employed for this estimation task. The field and year yielding the lowest RMSE, indicating more accurate estimation, was then selected for visualizing the estimated yield map in comparison to the actual yield map.

Year-level fitting ("Year-wise fit") aimed to determine if missing yield monitor data for a specific year within a field could be estimated accurately using a model developed from other years. To achieve this, all the three models were trained using data from one year in fields A1 and B1 and validated with the remaining two years. Subsequently, the best-performing year and field combination was identified and visually depicted and compared to both actual and estimated yields for the trained year and remaining two validation years.

3. Results

3.1. Feature Selection Outcomes

Some of the common features that appeared across all years in both farms included elevation, NIR, BI, SI, and CI. However, in Farm A, across all years, elevation and EVI2 consistently emerged as an important feature (top 5) indicating the robustness of the prediction power of EVI2 across all seasons (Figure 1). At Farm B, elevation was the most important feature across all three years, while EVI2 was not among the top features (Figure 1). Farm A's importance ranking was consistent across all years, which means the same features appeared as important but in a different order. Results for Farm B displayed more variability with the soil indices and NIR fluctuating in importance across different years. Climate variables were considered significant for Farm B solely in year 1 and did not exhibit the same level of importance in subsequent years or on any other farm. These findings highlight the importance of farm-specific and year-specific model development.

3.2. Model Comparison Using PyCaret

Among the 20 regression models tested, ensemble-based decision tree models, Extra Trees, Random Forest, and XGBoost emerged as the top-performing models with Extra Tree at the top (RMSE = 2.80 Mg/ha) followed by Random Forest (RMSE = 3.13 Mg/ha) and XGBoost (RMSE = 3.58 Mg/ha) (Figure 2).

All three models can handle nonlinear relationships, making them suitable for agriculture applications. On the other hand, models such as Bayesian Ridge, Least Angle, and Passive-Aggressive exhibited subpar performance for the data (RMSE 3.91 to 42.03 Mg/ha).

3.3. Hyperparameter Tuning Results

The resulting hyperparameter values for XGBoost values demonstrated consistency across all years and farms, with parameters set to a learning rate of 0.3, 200 estimators, maximum depth of 9, subsample of 0.9, and colsample_bytree of 0.9. Additionally, other parameters such as objective, booster, device, gamma, missing, and n_jobs were determined, ensuring uniform model configuration across tuning. The RMSE was employed as the evaluation metric, with 'gpu_hist' utilized as the tree method for efficient GPU-accelerated computation.

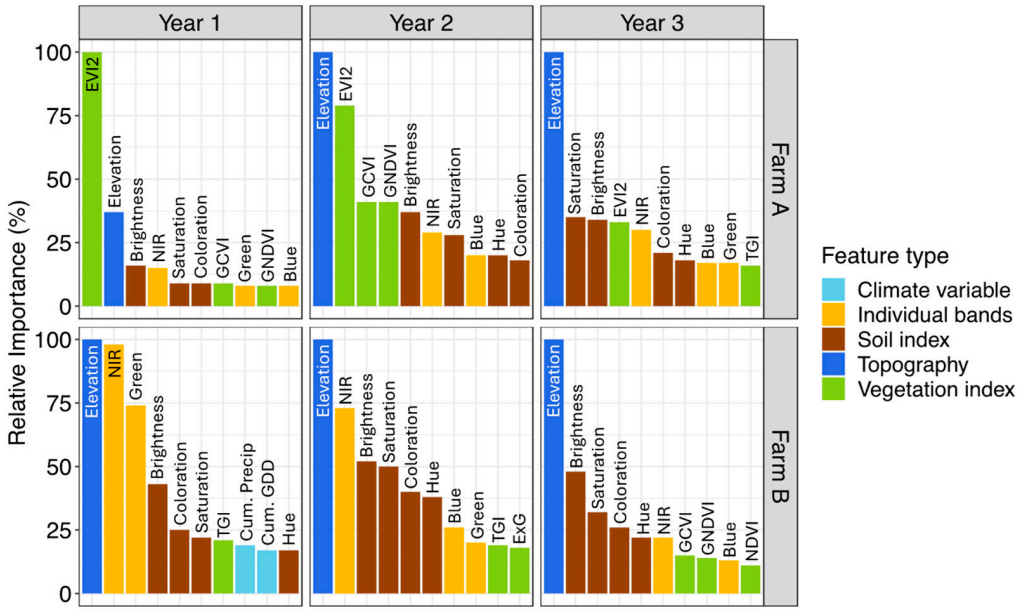


Figure 1. Feature selection ranking across three years for Farm A (field A1) and Farm B (field B1). NIR—near infrared, GCVI—Green Chlorophyll Vegetation Index, NDVI—Normalized Difference Vegetation Index, EVI2—Two-Band Enhanced Vegetation Index, TGI—Triangular Greenness Index, Cum.Precip.—cumulative precipitation, Cum.GDD—cumulative growing degree days, Exgod—excess green, GNDVI—green NDVI.

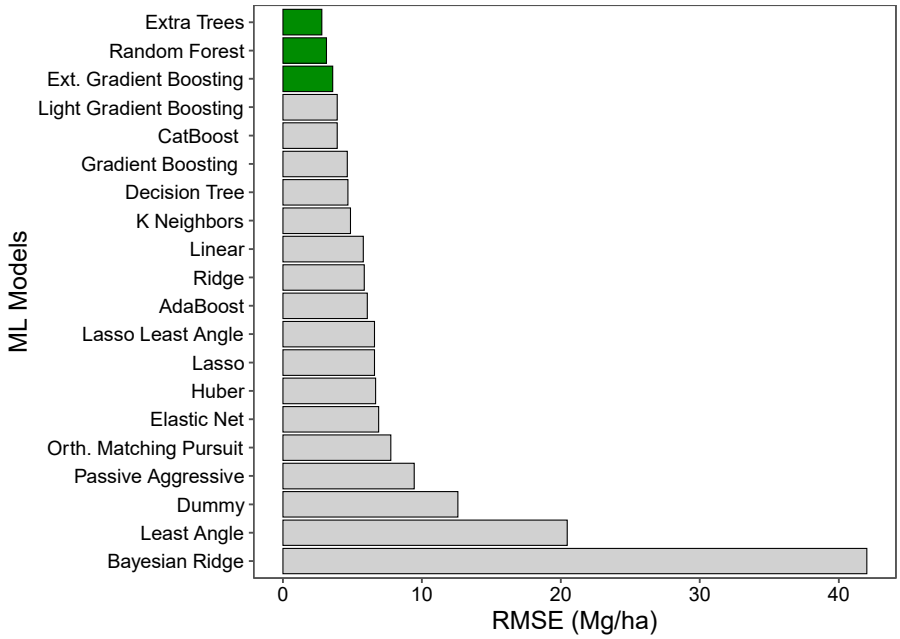


Figure 2. Performance comparison of various machine learning (ML) regression models in predicting corn silage yield using data from larger fields A1 and B1 including all three years of data within Farms A and B.

The hyperparameter tuning results for Random Forest models and Extra Trees revealed consistent configurations across different years and farms, indicating a uniform approach to model training. For Random Forest models, these were trained with bootstrap enabled, no cost complexity pruning (ccp_alpha), and the squared error criterion. The models were constrained to a maximum depth of 9, considered a maximum of 10 features for splitting, and required a minimum of 2 samples per leaf and 5 samples per split. Typically,

300 estimators were used, except for year 3 in Farm B, which utilized 200 estimators. These models did not use OOB samples and did not implement warm starting.

Similarly, the Extra Trees models were configured with bootstrap enabled, ccp_alpha, and used the squared error criterion for evaluation. These models also had a maximum depth of 9 and considered up to 10 features for splitting. Additionally, they required a minimum of one sample per leaf and five samples per split, with no weight fraction for leaf samples. These settings were consistently applied to 200 estimators and executed in parallel using all available CPUs, further highlighting the consistency in the approach to hyperparameter tuning across various conditions and datasets.

3.4. Selected Model Performance

After hyperparameter tuning, across both farms and all years, XGBoost generally performed better in terms of RMSE and MAE compared to Random Forest and Extra Trees (Figure 2). The RMSE of Farm A was, on average, 1.4 times the RMSE of Farm B. The prediction performance of all models was lower with higher RMSE values in year 3 for Farm A and the prediction performance in Farm B was lower for year 1 compared to other years. The RMSE for the testing ranged between 3.52 and 9.32 Mg/ha while the MAE ranged between 2.57 and 7.17 Mg/ha (Figure 3).

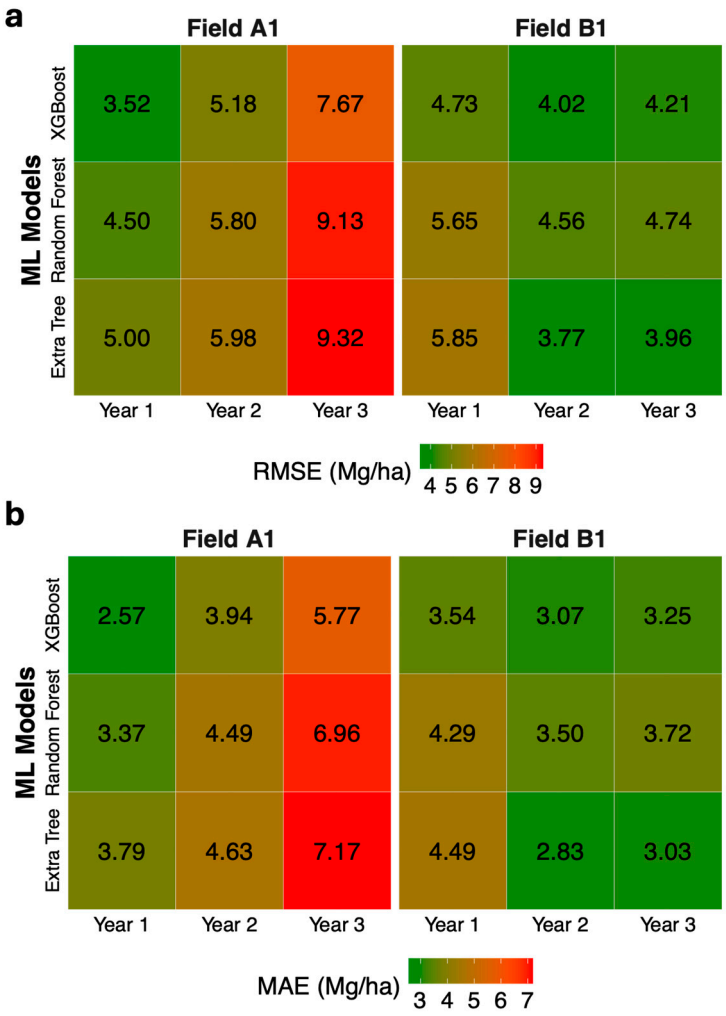


Figure 3. Heatmap comparison of ensemble model performance across three years for fields A1 and B1 using (**a**) Root Mean Square Error (RMSE), and (**b**) Mean Absolute Error (MAE).

3.5. Validation of the Best Models

3.5.1. Farm-Wise Fit

In general, models trained on the same field and tested on the same field had better predictive performance (Figure 4). However, poor prediction performance was obtained for validation using an unseen field. Also, the variation of RMSE values varies across different years, based on the RMSE values the performance of models in year 2 decreased by 55 and 48% compared to year 1 and year 3, respectively. Similar to model prediction results, Farm B had better validation results than Farm A. Even though prediction results show that XG-Boost was a better model, it was the least-performing model during validation. Extra Tree regressor emerged as the best followed by the Random Forest model. The best validation result was observed for Farm B, year 2 using the Extra Tree regressor to build the model from field B1 (RMSE = 3.77 Mg/ha) and validate using field B2 (RMSE = 6.46 Mg/ha).

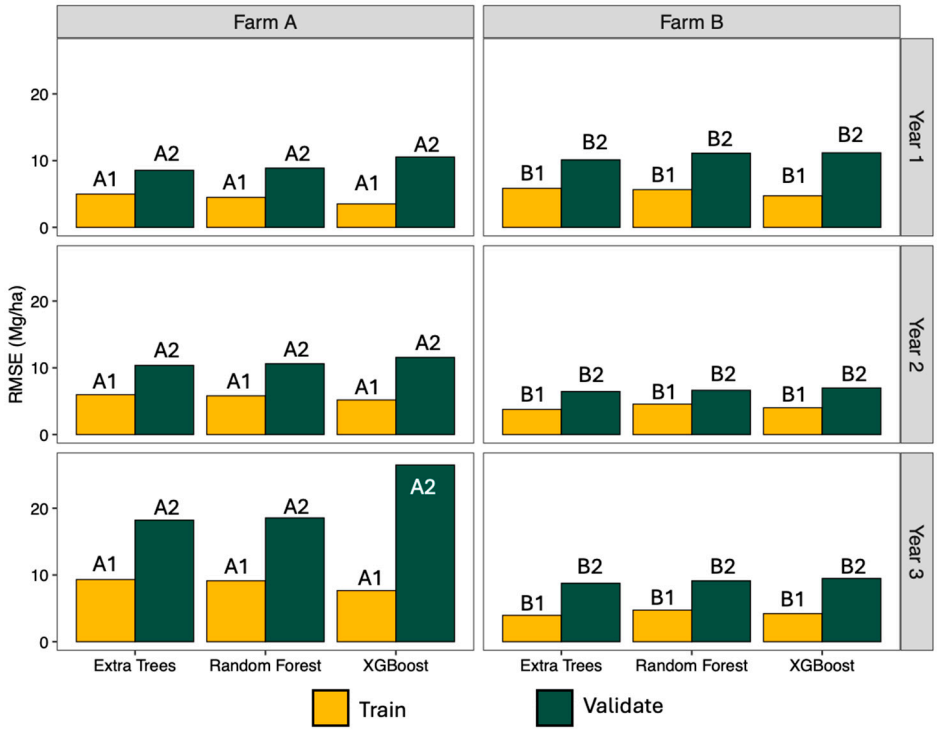


Figure 4. Comparison of Root Mean Square Error (RMSE) between training and testing fields predicting actual yield (Mg/ha) for Farms A and B over three years using three selected machine learning models for farm-wise fit.

Yield maps from the model trained on the same field closely represented the actual data (compare Figure 5a,b). However, the yield map generated for the unseen field, using the predicted data from the best model, significantly underestimated the actual yield values (compare Figure 5c,d).

3.5.2. Year-Wise Fit

Similar to farm-wise prediction results, generally models trained using one year of data and tested using the same year had lower RMSE values indicating better prediction (Figure 6). In general, the model performed better in predicting the prior years and had lower performance in predicting the future. As observed with the earlier results, the model for Farm B had better prediction performance than the model for Farm A, emphasizing the importance of farm-specific models. Although XGBoost performed better when predicting within the same year, it performed worse when predicting validation data from unseen years. The Extra Tree regressor again emerged as the best model in both farms across all years. The consistency in performance for Extra Tree regression suggests that the model

may be robust across different seasons and conditions. The highest validation results were achieved for Farm B year 1 using the Extra Tree regressor model, where the prediction of unseen data resulted in an RSME of 5.51 Mg/ha for year 2 and 7.88 Mg/ha for year 3 (Figure 6).

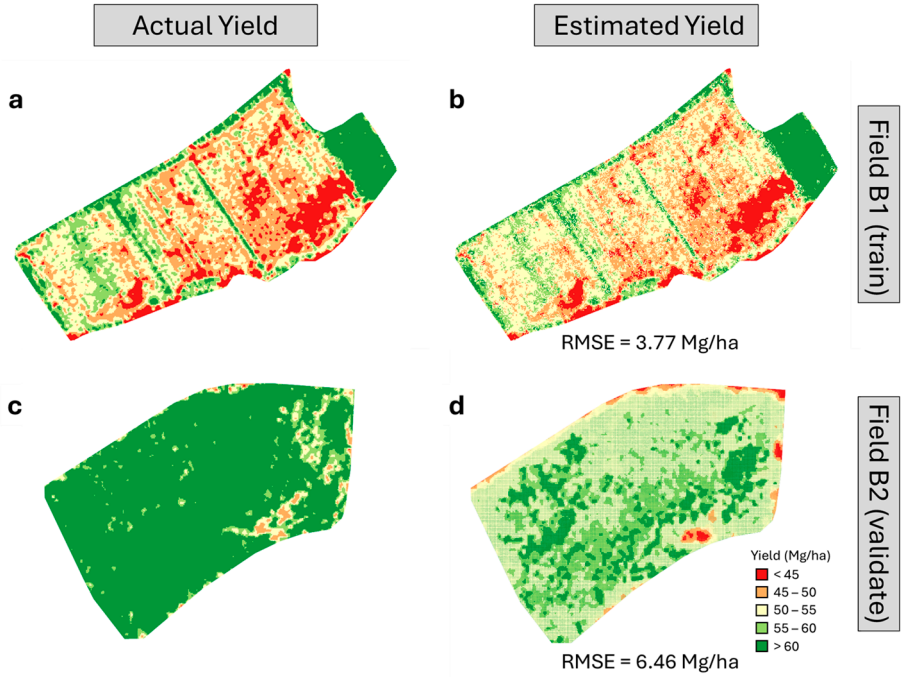


Figure 5. Farm-wise validation—for year 2, trained field B1 model validated by estimating the unseen yield in neighboring field B2. (**a**,**c**) show actual yield values while (**b**) is estimated yield of the trained field and (**d**) shows estimated yield of the unseen validated field.

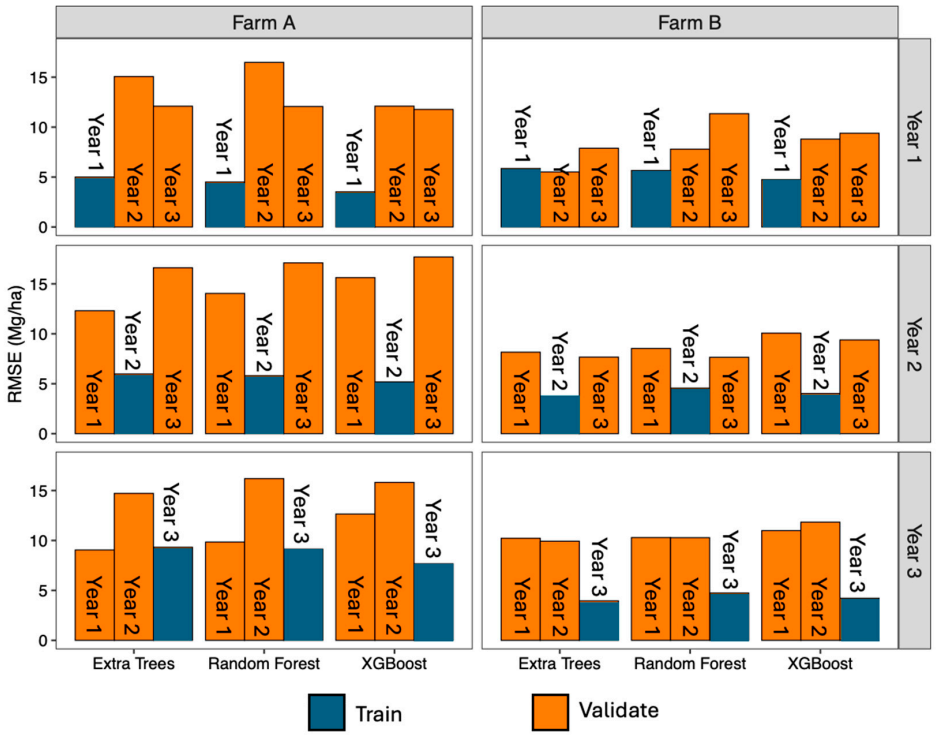


Figure 6. Comparison of Root Mean Square Error (RMSE) between training year and validating unseen years (2 years) for estimating silage yield (Mg/ha) for Farms A and B using three different machine learning models for year-wise fit.

Yield maps generated from the model developed and tested in the same year were most representative of the actual yield maps (compare Figure 7a,b). The yield map for year 2 generated from the model developed using year 1, captured the high-yielding spots (>60 Mg/ha) better than other classes (compare Figure 7c,d). The year 3 yield map showed that the low-yielding spots (<50 Mg/ha) were predicting better than the other classes (compare Figure 7e,f).

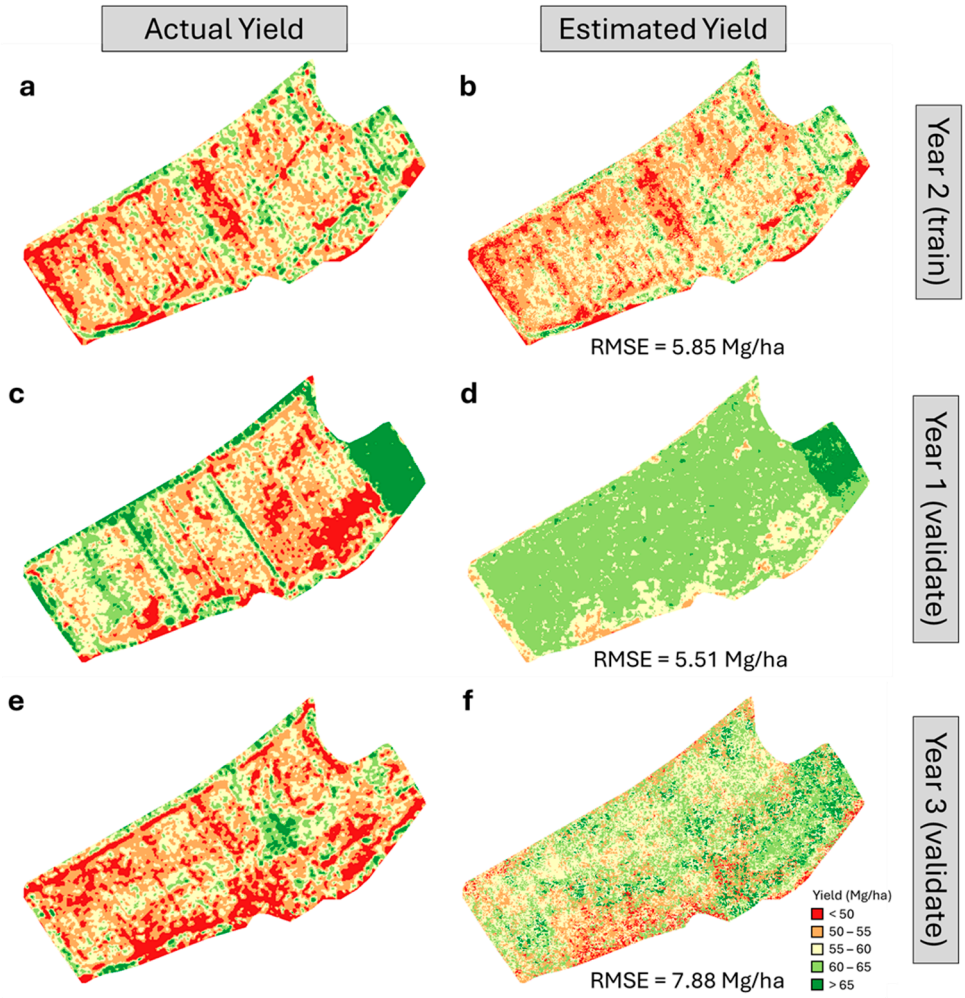


Figure 7. Year-wise validation—trained model with field B1 in year 2 and validated by predicting unseen field B1 in year 1 and year 3. (**a**,**c**,**e**) show actual yield values while (**b**,**d**,**f**) show estimated yield values.

4. Discussion

4.1. Feature Selection Implications

The consistent emergence of EVI2 across all years in Farm A reveals its reliability in capturing variation in crop health and productivity across years. This consistency suggests the potential of EVI2 to serve as a good indicator for monitoring crop performance in Farm A. Earlier studies have also reported the importance of EVI2 in predicting crop yield [19,38,39]. In contrast, Farm B's results indicate that elevation plays a more significant role, reflecting the impact of topographical factors on yield prediction. Interestingly, elevation also ranked high in Farm A, demonstrating its robustness across different farms. Previous studies show that elevation and other topographical features are crucial in increasing the spatial accuracy of yield prediction [9,40]. Unlike EVI2, which needs to be collected multiple times within a season, elevation is stable over time and can be gathered once for ongoing use. This makes elevation a more practical and cost-effective feature for

yield prediction, especially when considering long-term applications. Variability observed in feature importance ranking particularly in Farm B, shows complex interactions between environmental variables and crop productivity highlighting the importance of a tailored approach that is farm-specific.

The consistent appearance of both NIR and elevation across all years in both farms underscores their importance as robust predictors of corn silage yield across diverse agricultural contexts. While climate variables play a crucial role in monitoring crop health, their diminished importance could be their limited relevance for farm-scale applications due to insufficient spatial variability, resulting in their absence as a prominent feature. These findings emphasize the importance of integrating farm-specific knowledge and temporally sensitive analyses to enhance the effectiveness of actual silage yield prediction for agricultural management practices.

4.2. Selection and Comparison of Models

The ensemble-based models such as XGBoost, Random Forest, and the Extra Tree regressor outperformed the other regression and ML algorithms for predicting corn silage yield. Earlier studies [41–43] have also reported good performance of the ensemble-based models for yield predictions and other agricultural applications. Lower-performing models such as Bayesian Ridge, Least Angle, and Passive-Aggressive tend to rely on linear or simplistic algorithms. These models might struggle with capturing complex relationships and patterns in the data, leading to higher prediction errors.

When the developed models were tested using the same field or year the prediction using XGBoost was the better performing among the three. However, when the same trained XGBoost model was validated with an unseen field or year it tended to perform poorly. This could reflect overtraining of the model, where XGBoost exhibits high variance, leading to strong performance on the training dataset but failing to generalize well to unseen data. From the results, it is evident that the Extra Tree regressor was a good tradeoff between bias and variance that led to better performance during the validation step. Validation results tested using the unseen data (year and field) showed that Extra Trees, closely followed by Random Forest, emerged as better models. Earlier studies have shown that Extra Tree regressors as competitive models for validation [44,45].

4.3. Yield Prediction Potential for Missing Field

The notable decrease in prediction performance when applying models developed using one field to unseen nearby fields, highlights the limitations of generalizing models across different field environments. Farm B outperformed Farm A by predicting missing fields across all years highlighting the necessity for farm-based prediction approaches due to varying conditions and agricultural management practices. The best performance showed an RMSE of 6.46 Mg/ha, which exceeds the 4.94 Mg/ha currently used to decide if an increase in nitrogen fertilizer is warranted under state adaptive management policy in New York [5]. This emphasizes that the models cannot be generalized and highlights the critical need for farm-specific prediction approaches tailored to the unique conditions and management practices of each farm. By focusing on farm-based models, farmers can achieve more accurate yield predictions, leading to better-informed decisions regarding resource allocation, such as nitrogen fertilization.

4.4. Yield Prediction Potential for the Missing Year

Models trained and tested within the same year consistently exhibited better predictive accuracy, aligning with farm-wise prediction results. In this study, models generally perform better in validating prior years compared to future years, highlighting the temporal dynamics influencing prediction performance. The best validation showed an RMSE of 5.51 Mg/ha, indicating that missing data from previous years can be estimated, allowing for a more comprehensive analysis of historical yield patterns. However, as this study used only three years of data, expanding the dataset to include more years would provide

a deeper understanding of these patterns. Additionally, incorporating yield data from upcoming years into the training process can enhance the accuracy of predictive models. Similar to farm-wise, year-wise underscored Farm B's superior predictive performance over Farm A, highlighting the need for farm-specific approaches. These insights can then be leveraged to optimize future management practices.

5. Conclusions

This study underscores the importance of leveraging advanced machine-learning techniques and farm-specific data for accurate corn silage yield prediction and evaluating model performance across fields and years. The consistent emergence of EVI2 as a reliable indicator of crop health and productivity in Farm A suggests its potential for effective monitoring of crop performance across diverse seasons. Additionally, Farm B's reliance on elevation as the most influential feature highlights the significance of topographical factors in yield estimation. These findings indicate the necessity for tailored, farm-specific modeling approaches. The robust appearance of NIR and elevation across both farms and all years further emphasizes their importance as consistent predictors of corn silage yield, despite variations in environmental conditions. Ensemble-based models such as XGBoost, Random Forest, and Extra Tree regressors demonstrated superior predictive performance compared to other regression and machine learning algorithms, particularly when tested within the same field or year. However, when applied to unseen fields or years, XGBoost performed poorly, suggesting potential overfitting issues. In contrast, the Extra Tree regressor and Random Forest emerge as more reliable models for predicting missing data, underscoring the importance of considering model robustness and generalizability. While satellite data are effective for large-scale agricultural applications, our findings indicate limitations in capturing within-field and field to field variability, even with PlanetScope's highest resolution (3 m) imagery available for research. This highlights the need for finer resolution data, such as drone imagery, which offers greater granularity for more accurate yield estimation and better captures in-field variability in farm-scale applications and on-farm research.

Overall, our study emphasizes the importance of integrating farm-specific knowledge and temporally sensitive analyses to enhance the effectiveness of corn silage yield prediction models for informed agricultural management practices. Furthermore, this research draws on interdisciplinary knowledge from fields such as remote sensing technology, machine learning, and topography, highlighting the importance of collaborative efforts across these domains to achieve sustainable agricultural productivity.

Author Contributions: Conceptualization, Q.M.K.; Methodology, S.N.S.; Software, S.N.S.; Formal analysis, S.N.S.; Investigation, S.N.S.; Resources, S.S. and Q.M.K.; Data curation, S.N.S., M.M.III and S.S.; Writing—original draft, S.N.S.; Writing—review & editing, M.M.III, S.S., D.R.K., L.J.T. and Q.M.K.; Visualization, S.N.S.; Supervision, Q.M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Northern New York Agricultural Development Program, New York Farm Viability Institute, Atkinson Center for Sustainability, New York State Department of Agriculture, New York State Department of Environmental Conservation.

Data Availability Statement: The datasets presented in this article are not readily available because farmer yield data are confidential.

Acknowledgments: We extend our gratitude to Mike Contessa and Eric Bever from Champlain Valley Agronomics, along with the participating farmers, for generously sharing their yield data. We also gratefully acknowledge Planet's Education and Research Program, which provided satellite images at no cost through our university affiliation (https://www.planet.com/industries/education-and-research/ (accessed on 29 August 2024)).

Conflicts of Interest: Author Daniel R. Kindred was employed by the company Anglo American. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. McFadden, J.; Njuki, E.; Griffin, T. Precision Agriculture in the Digital Era: Recent Adoption on U.S. Farms. Available online: http://www.ers.usda.gov/publications/pub-details/?pubid=105893 (accessed on 22 August 2024).
2. Maestrini, B.; Basso, B. Drivers of Within-Field Spatial and Temporal Variability of Crop Yield across the US Midwest. *Sci. Rep.* **2018**, *8*, 14833. [CrossRef] [PubMed]
3. Kharel, T.P.; Maresma, A.; Czymmek, K.J.; Oware, E.K.; Ketterings, Q.M. Combining Spatial and Temporal Corn Silage Yield Variability for Management Zone Development. *Agron. J.* **2019**, *111*, 2703–2711. [CrossRef]
4. Kayad, A.; Sozzi, M.; Gatto, S.; Marinello, F.; Pirotti, F. Monitoring Within-Field Variability of Corn Yield Using Sentinel-2 and Machine Learning Techniques. *Remote Sens.* **2019**, *11*, 2873. [CrossRef]
5. Ketterings, Q.M.; Workman, K.; Gates, D.; Hornesky, J.; Latessa, S.; Bush, R.; Jordan, B.; Albercht, G. *Adaptive Nitrogen Management for Field Crops in New York*; Cornell University: Ithaca, NY, USA, 2023. Available online: http://nmsp.cals.cornell.edu/publications/extension/AdaptiveNitrogenManagement2023.pdf (accessed on 22 August 2024).
6. Yield Monitor Calibration: Garbage In, Garbage Out (Purdue University). Available online: https://www.agry.purdue.edu/ext/corn/news/timeless/yldmoncalibr.html (accessed on 4 April 2024).
7. Kharel, T.P.; Swink, S.; Youngerman, C.; Maresma, Á.; Czymmek, K.J.; Ketterings, Q.M.; Kyveryga, P.; Lory, J.; Musket, T.A.; Hubbard, V. *Processing/Cleaning Corn Silage and Grain Yield Monitor Data for Standardized Yield Maps across Farms, Fields, and Years*; Nutrient Management Spear Program, Department of Animal Science, Cornell University: Ithaca, NY, USA, 2018.
8. Killeen, P.; Kiringa, I.; Yeap, T.; Branco, P. Corn Grain Yield Prediction Using UAV-Based High Spatiotemporal Resolution Imagery, Machine Learning, and Spatial Cross-Validation. *Remote Sens.* **2024**, *16*, 683. [CrossRef]
9. de Oliveira, M.F.; Ortiz, B.V.; Morata, G.T.; Jiménez, A.-F.; de Rolim, G.S.; da Silva, R.P. Training Machine Learning Algorithms Using Remote Sensing and Topographic Indices for Corn Yield Prediction. *Remote Sens.* **2022**, *14*, 6171. [CrossRef]
10. Khan, S.N.; Li, D.; Maimaitijiang, M. Using Gross Primary Production Data and Deep Transfer Learning for Crop Yield Prediction in the US Corn Belt. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *131*, 103965. [CrossRef]
11. Khanal, S.; Fulton, J.; Klopfenstein, A.; Douridas, N.; Shearer, S. Integration of High Resolution Remotely Sensed Data and Machine Learning Techniques for Spatial Prediction of Soil Properties and Corn Yield. *Comput. Electron. Agric.* **2018**, *153*, 213–225. [CrossRef]
12. Sakamoto, T. Incorporating Environmental Variables into a MODIS-Based Crop Yield Estimation Method for United States Corn and Soybeans through the Use of a Random Forest Regression Algorithm. *ISPRS J. Photogramm. Remote Sens.* **2020**, *160*, 208–228. [CrossRef]
13. Ma, Y.; Zhang, Z.; Kang, Y.; Özdoğan, M. Corn Yield Prediction and Uncertainty Analysis Based on Remotely Sensed Variables Using a Bayesian Neural Network Approach. *Remote Sens. Environ.* **2021**, *259*, 112408. [CrossRef]
14. Skakun, S.; Kalecinski, N.I.; Brown, M.G.L.; Johnson, D.M.; Vermote, E.F.; Roger, J.-C.; Franch, B. Assessing Within-Field Corn and Soybean Yield Variability from WorldView-3, Planet, Sentinel-2, and Landsat 8 Satellite Imagery. *Remote Sens.* **2021**, *13*, 872. [CrossRef]
15. Zhang, L.; Zhang, Z.; Luo, Y.; Cao, J.; Xie, R.; Li, S. Integrating Satellite-Derived Climatic and Vegetation Indices to Predict Smallholder Maize Yield Using Deep Learning. *Agric. For. Meteorol.* **2021**, *311*, 108666. [CrossRef]
16. Li, F.; Miao, Y.; Chen, X.; Sun, Z.; Stueve, K.; Yuan, F. In-Season Prediction of Corn Grain Yield through PlanetScope and Sentinel-2 Images. *Agronomy* **2022**, *12*, 3176. [CrossRef]
17. Deines, J.M.; Patel, R.; Liang, S.-Z.; Dado, W.; Lobell, D.B. A Million Kernels of Truth: Insights into Scalable Satellite Maize Yield Mapping and Yield Gap Analysis from an Extensive Ground Dataset in the US Corn Belt. *Remote Sens. Environ.* **2021**, *253*, 112174. [CrossRef]
18. Pinto, A.A.; Zerbato, C.; Rolim, G.D.S.; Barbosa Júnior, M.R.; Silva, L.F.V.D.; Oliveira, R.P.D. Corn Grain Yield Forecasting by Satellite Remote Sensing and Machine-Learning Models. *Agron. J.* **2022**, *114*, 2956–2968. [CrossRef]
19. Sunoj, S.; Cho, J.; Guinness, J.; van Aardt, J.; Czymmek, K.J.; Ketterings, Q.M. Corn Grain Yield Prediction and Mapping from Unmanned Aerial System (UAS) Multispectral Imagery. *Remote Sens.* **2021**, *13*, 3948. [CrossRef]
20. Kayad, A.; Sozzi, M.; Gatto, S.; Whelan, B.; Sartori, L.; Marinello, F. Ten Years of Corn Yield Dynamics at Field Scale under Digital Agriculture Solutions: A Case Study from North Italy. *Comput. Electron. Agric.* **2021**, *185*, 106126. [CrossRef]
21. Marques Ramos, A.P.; Prado Osco, L.; Elis Garcia Furuya, D.; Nunes Gonçalves, W.; Cordeiro Santana, D.; Pereira Ribeiro Teodoro, L.; Antonio da Silva Junior, C.; Fernando Capristo-Silva, G.; Li, J.; Henrique Rojo Baio, F.; et al. A Random Forest Ranking Approach to Predict Yield in Maize with Uav-Based Vegetation Spectral Indices. *Comput. Electron. Agric.* **2020**, *178*, 105791. [CrossRef]
22. Burdett, H.; Wellen, C. Statistical and Machine Learning Methods for Crop Yield Prediction in the Context of Precision Agriculture. *Precis. Agric.* **2022**, *23*, 1553–1574. [CrossRef]
23. Cho, J.B.; Guinness, J.; Kharel, T.P.; Sunoj, S.; Kharel, D.; Oware, E.K.; van Aardt, J.; Ketterings, Q.M. Spatial Estimation Methods for Mapping Corn Silage and Grain Yield Monitor Data. *Precis. Agric.* **2021**, *22*, 1501–1520. [CrossRef]
24. Planet Labs PBC. Planet Application Program Interface: In Space for Life on Earth; Planet, 2017–2023. Available online: https://www.planet.com (accessed on 22 August 2024).
25. De Reu, J.; Bourgeois, J.; Bats, M.; Zwertvaegher, A.; Gelorini, V.; De Smedt, P.; Chu, W.; Antrop, M.; De Maeyer, P.; Finke, P.; et al. Application of the Topographic Position Index to Heterogeneous Landscapes. *Geomorphology* **2013**, *186*, 39–49. [CrossRef]

26. Weiss, A. Topographic Position and Landforms Analysis. In Proceedings of the Poster Presentation, ESRI User Conference, San Diego, CA, USA, 9–13 July 2001; Volume 200.
27. Mathieu, R.; Pouget, M.; Cervelle, B.; Escadafal, R. Relationships between Satellite-Based Radiometric Indices Simulated Using Laboratory Reflectance Data and Typic Soil Color of an Arid Environment. *Remote Sens. Environ.* **1998**, *66*, 17–28. [CrossRef]
28. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring Vegetation Systems in the Great Plains with ERTS. *NASA Spec. Publ.* **1974**, *351*, 309.
29. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a Green Channel in Remote Sensing of Global Vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. [CrossRef]
30. Jiang, Z.; Huete, A.R.; Didan, K.; Miura, T. Development of a Two-Band Enhanced Vegetation Index without a Blue Band. *Remote Sens. Environ.* **2008**, *112*, 3833–3845. [CrossRef]
31. Gitelson, A.A.; Viña, A.; Arkebauer, T.J.; Rundquist, D.C.; Keydan, G.; Leavitt, B. Remote Estimation of Leaf Area Index and Green Leaf Biomass in Maize Canopies. *Geophys. Res. Lett.* **2003**, *30*, 1248. [CrossRef]
32. Birth, G.S.; McVey, G.R. Measuring the Color of Growing Turf with a Reflectance Spectrophotometer1. *Agron. J.* **1968**, *60*, 640–643. [CrossRef]
33. Woebbecke, D.M.; Meyer, G.E.; Von Bargen, K.; Mortensen, D.A. Color Indices for Weed Identification Under Various Soil, Residue, and Lighting Conditions. *Trans. ASAE* **1994**, *38*, 259–269. [CrossRef]
34. Hunt, E.R., Jr.; Daughtry, C.S.T.; Eitel, J.U.H.; Long, D.S. Remote Sensing Leaf Chlorophyll Content Using a Visible Band Index. *Agron. J.* **2011**, *103*, 1090–1099. [CrossRef]
35. Pargent, F.; Pfisterer, F.; Thomas, J.; Bischl, B. Regularized Target Encoding Outperforms Traditional Methods in Supervised Machine Learning with High Cardinality Features. *Comput. Stat.* **2022**, *37*, 2671–2692. [CrossRef]
36. Moez, A. PyCaret: An Open Source, Low-Code Machine Learning Library in Python Version 1.0. Available online: https://pycaret.org/ (accessed on 11 July 2024).
37. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
38. Ji, Z.; Pan, Y.; Zhu, X.; Zhang, D.; Dai, J. Prediction of Corn Yield in the USA Corn Belt Using Satellite Data and Machine Learning: From an Evapotranspiration Perspective. *Agriculture* **2022**, *12*, 1263. [CrossRef]
39. Medina, H.; Tian, D.; Abebe, A. On Optimizing a MODIS-Based Framework for in-Season Corn Yield Forecast. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *95*, 102258. [CrossRef]
40. Kumhálová, J.; Kumhála, F.; Kroulík, M.; Matějková, Š. The Impact of Topography on Soil Properties and Yield and the Effects of Weather Conditions. *Precis. Agric.* **2011**, *12*, 813–830. [CrossRef]
41. Awasthi, P.; Mishra, S.; Gupta, N. Performance Assessment of Machine Learning Techniques for Corn Yield Prediction. In *Proceedings of the Advanced Network Technologies and Intelligent Computing*; Woungang, I., Dhurandher, S.K., Pattanaik, K.K., Verma, A., Verma, P., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2023; pp. 320–335.
42. Shahhosseini, M.; Hu, G.; Archontoulis, S.V. Forecasting Corn Yield With Machine Learning Ensembles. *Front. Plant Sci.* **2020**, *11*, 1120. [CrossRef]
43. Khan, S.N.; Li, D.; Maimaitijiang, M. A Geographically Weighted Random Forest Approach to Predict Corn Yield in the US Corn Belt. *Remote Sens.* **2022**, *14*, 2843. [CrossRef]
44. Patel, N.; Patel, D.; Patel, S.; Patel, V. Crop Yield Estimation Using Machine Learning. In *Proceedings of the Soft Computing and its Engineering Applications*; Patel, K.K., Garg, D., Patel, A., Lingras, P., Eds.; Springer: Singapore, 2021; pp. 328–342.
45. Nikhil, U.V.; Pandiyan, A.M.; Raja, S.P.; Stamenkovic, Z. Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models. *Computers* **2024**, *13*, 137. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.