




SPECIAL SECTION: MACHINE LEARNING IN AGRICULTURE

Artificial intelligence and satellite-based remote sensing can be used to predict soybean (*Glycine max*) yield

Deepak R. Joshi^{1,2} | Sharon A. Clay²  | Prakriti Sharma² | Hossein Moradi Rekabdarkolae³ | Tulsi Kharel⁴ | Donna M. Rizzo⁵ | Resham Thapa⁶  | David E. Clay² 

¹College of Agriculture, Arkansas State University, Jonesboro, Arkansas, USA

²Department of Agronomy, Horticulture and Plant Science, South Dakota State University, Brookings, South Dakota, USA

³Department of Mathematics and Statistics, South Dakota State University, Brookings, South Dakota, USA

⁴Crop Production Systems Research Unit, USDA–ARS, Stoneville, Mississippi, USA

⁵Department of Civil & Environmental Engineering, University of Vermont, Burlington, Vermont, USA

⁶Department of Agricultural and Environmental Sciences, Tennessee State University, Nashville, Tennessee, USA

Correspondence

David E. Clay, Department of Agronomy, Horticulture and Plant Science, South Dakota State University, Brookings, SD 57007, USA. Email:

David.clay@sdstate.edu

Assigned to Associate Editor Kathleen Yeater.

Funding information

National Science Foundation, Grant/Award Numbers: 2202706, 2026431; South Dakota Soybean Association

Abstract

Because the manual counting of soybean (*Glycine max*) plants, pods, and seeds/pods is unsuitable for soybean yield predictions, alternative methods are desired. Therefore, the objective was to determine if satellite remote sensing-based artificial intelligence (AI) models could be used to predict soybean yield. In the study, multiple remote sensing-based AI models were developed for soybean growth stage ranging from VE/VC (plant emergence) to R6/R7 (full seed to beginning maturity). The ability of the deep neural network (DNN), support vector machine (SVM), random forest (RF), least absolute shrinkage and selection operator (LASSO), and AdaBoost to predict soybean yield, based on blue, green, red, and near-infrared reflectance data collected by the PlanetScope satellite at six growth stages, was determined. Remote sensing and soybean yield monitor data from three different fields in 2 years (2019 and 2021) were aggregated into 24,282 grid cells that had the dimensions of 10 m by 10 m. A comparison across models showed that the DNN outperformed the other models. Moreover, as crops matured from VE/VC to R4/R5, the R^2 value of the models increased from 0.26 to over 0.70. These findings indicate that remote sensing data collected at different growth stages can be combined for soybean yield predictions. Moreover, additional work needs to be conducted to assess the model's ability to predict soybean yield with vegetation indices (VIs) data for fields not used to train the model.

Abbreviations: AI, artificial intelligence; DNN, deep neural network; LASSO, least absolute shrinkage and selection operator; NIR, near-infrared; RF, random forest; SVM, support vector machine; VI, vegetation index.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Agronomy Journal* published by Wiley Periodicals LLC on behalf of American Society of Agronomy.

1 | INTRODUCTION

Farmers and agronomists have estimated soybean (*Glycine max*) yield by counting the number of plants, pods per plant, and seeds per pod in relatively small areas and extrapolating to whole field areas. This information, while interesting, is labor-intensive and may not provide useful and accurate information when applied to the whole field scale. For example, de Souza et al. (2023) reported that to assess a plant's phenotypic characteristic, four traits from 21 soybean plants contained within a 2.7-m² area should be assessed. However, when this sampling protocol, designed for small plots, is extended across fields that may be greater than 650,000 m² (65 ha), the sampling requirement quickly becomes unmanageable. Therefore, an alternative approach to estimate soybean yield is needed for precision agriculture.

Previous efforts to predict soybean yield include using artificial intelligence (AI), specifically deep neural network (DNN) models, and regional data collected at the county level (Khaki et al., 2021; Sun et al., 2019). For example, Sun et al. (2019) predicted soybean yield at the county scale based on weather, moderate resolution imaging spectroradiometer (MODIS) surface reflectance, and historical yield data. Similarly, Khaki et al. (2021) created an AI model to predict corn (*Zea mays*) and soybean yield at the county scale. While these efforts provide meaningful data, it is unlikely that these models can provide useful information at the field scale. Therefore, the purpose of this study is to fill this gap and develop models that employ AI techniques using remotely sensed data to make soybean yield predictions. Our hypothesis was that AI techniques, when used with satellite-based plant reflectance indices, can provide reliable yield estimates. The specific objectives of this study were to (1) identify AI models that best predict field-scale soybean yield using remotely sensed plant reflectance indices data, (2) determine the optimal soybean growth stages at which the remote sensing data should be acquired, and (3) develop a pipeline for processing and feeding remotely sensed data into AI models for soybean yield estimation.

2 | MATERIALS AND METHODS

2.1 | Selection of study sites

The study sites were located within the South Dakota counties of Edmunds, Hamlin, and Miner (Figure 1a). These are located on the border between the Dfa (hot summer humid continental climate), Dfb (warm summer humid continental climate), and Bsk (cold semi-arid) Koppen climate regions. The crop rotation at all sites was corn (*Zea mays*), followed by soybeans, and the fields were chisel plowed, disked, and planted. Weather data were obtained from the NOAA website, where the respective weather stations for Edmunds,

Core Ideas

- Remote sensing-based artificial intelligence (AI) models can estimate soybean yield.
- Of the AI techniques tested, the deep neural network (DNN) generally explained the most yield variability (R^2).
- Yield predictions were improved by combining satellite images collected at multiple growth stages.

Hamlin, and Miner counties were located at (45.441975°, -98.751588°), (44.72683°, -97.02859°), and (44.004956°, -97.629446°), which are 17.5, 15, and 6.5 miles from the study sites, respectively (National Oceanic and Atmospheric Administration (NOAA), National Centers for Environmental Information, 2022).

At the Edmunds County (Figure 1b) field, the soil texture was silt loam (USDA-NRCS, 2023). Soybean was planted in this 41.1 ha field on May 20, 2019 and June 2, 2021, and harvested on October 26, 2019 and October 17, 2021, respectively.

At Hamlin (Figure 1c), the soil texture was a silty clay loam (USDA-NRCS, 2023). At this 53-ha site, soybean was planted on May 7, 2019 and harvested on October 18, 2019. In 2021, soybean was planted on April 20 and harvested on October 22.

At Miner (Figure 1d), the soil texture was loam (USDA-NRCS, 2023). The total field area was 39.4 ha. In 2019, the field was planted on April 24 and harvested on October 27, whereas in 2021, it was planted on April 20 and harvested on October 7.

2.2 | Selection of remote sensing platform and high-resolution imagery

Of the many space-based sensors to choose from, PlanetScope was selected because it has high spatial (3.12 m × 3.12 m) and temporal (daily) resolution and provides multispectral images (blue [455–515 nm], green [500–590 nm], red [590–670 nm], and near-infrared [780–860 nm]; Jain et al., 2016; Planet Team, 2017; Yang et al., 2012). The Ortho Scene-Analytics (Level 3B) surface reflectance imageries are orthorectified, radiometrically calibrated, and atmospherically corrected products that capture the reflectance characteristics of the lower atmosphere (Frazier & Hemingway, 2021; Houborg & McCabe, 2018; T. P. Kharel et al., 2023). Between planting and harvesting in 2019 and 2021, six cloud-free images at six growth stages (VE/VC [June 10]; V1/V3 [July 2], R1/R2 [July 24], and R2/R3 [August 10], R4/R5 [August 25], and

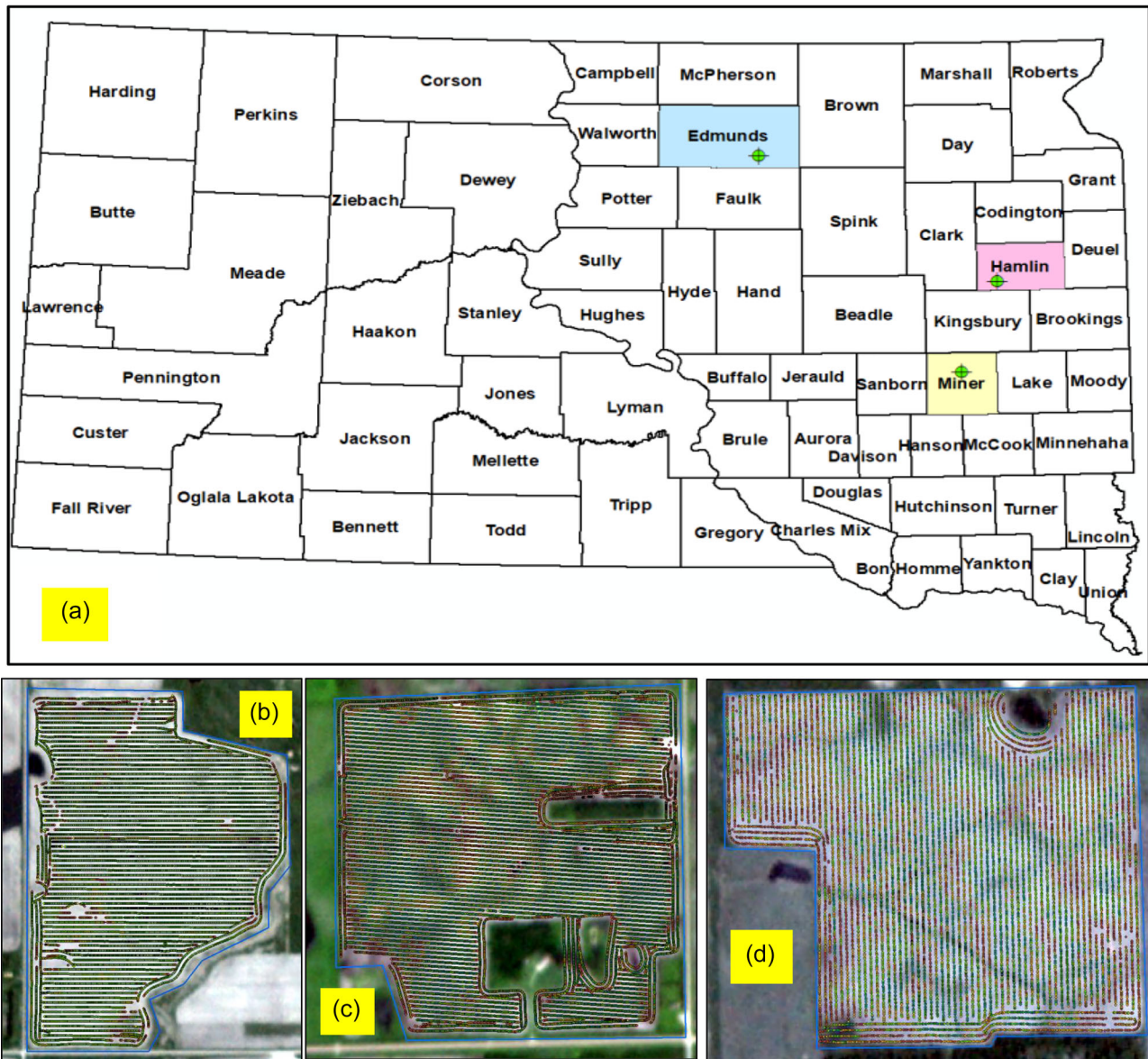


FIGURE 1 South Dakota County map (a) with the Edmunds (b), Hamlin (c), and Miner (d) study sites. Green dots on the map (a) represent the location of each study site.

R6/R7 [September 10]) were acquired. Because the images were chosen on the same dates each year, variations in geographic location, planting dates, and seasonal weather may have resulted in slight variations in soybean growth stages across the six site years.

2.3 | Yield data

Georeferenced yield data were collected with a calibrated yield monitor system at each site following recommended protocols (T. Kharel et al., 2018). At all study sites, the John Deere combine harvester equipped with a yield monitor system had a header width of 9.1 m. The distance between each yield data point across all three sites ranged from 1.5 to 1.8

m apart. Following data collection, the data were cleaned to remove issues associated with sensor delays, yield extremes, and start and end pass delays using SMS Advanced software (Ag Leader Technology; Cho et al., 2021; Dobermann & Ping, 2004; T. Kharel et al., 2018). All data were adjusted to 13% moisture.

2.4 | Data preprocessing and feature selection

The overall steps and processes utilized during data preprocessing and model development are provided in the model pipeline outlined in Figure 2. After collecting PlanetScope images and yield monitor data at all sites, a region of inter-

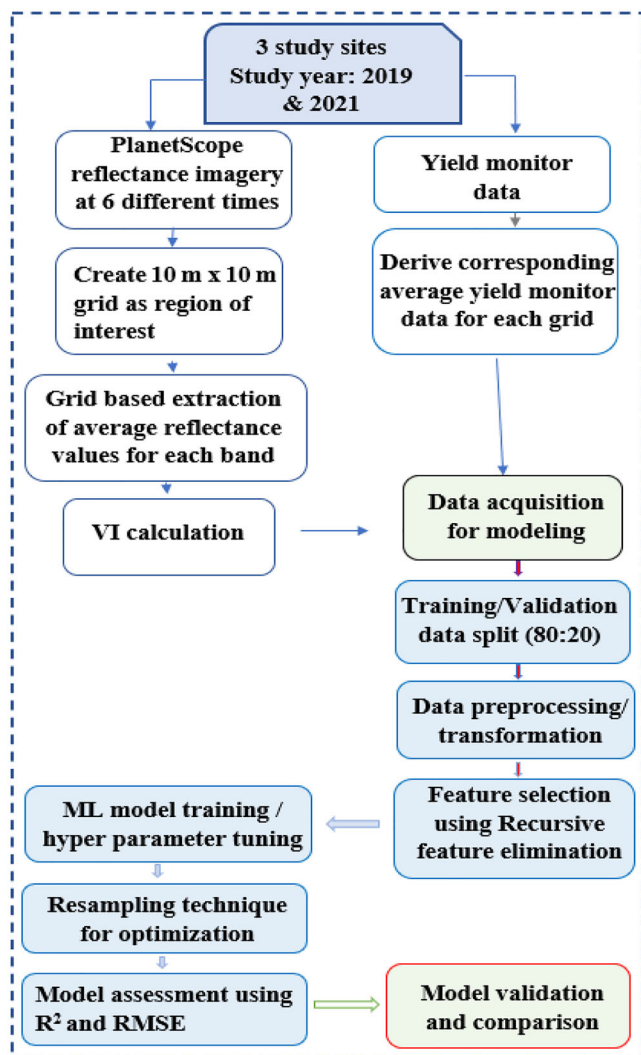


FIGURE 2 Model pipeline implemented during analysis. ML, machine learning; RMSE, root mean square error; VI, vegetation index.

est was created by dividing each field into 10-m by 10-m grid cells. Depending upon field size, total grid numbers were 3281, 5246, and 3614 at the Edmunds, Hamlin, and Miner sites, respectively. Over the 2 years and three sites, the total number of grid cells was 24,282.

After dividing fields into grid cells, the datasets were aligned by assigning the appropriate georeferenced remotely sensed reflectance data and yield values into 10-m by 10-m grid cells using the “Rasterio” library in Python (Bosch, 2019). On average, across all six site years, each field grid cell consisted of approximately five to six yield data points. Using the reflectance pixel values for each grid cell, 10 vegetation indices (VIs) were computed from each image (at each growth stage). These VIs that were used as predictor variables in the models were difference vegetation index (DVI), green chlorophyll vegetation index (GCI), green normalized differential vegetation index (GNDVI), normalized differential vegetation index (NDVI), normalized difference water index (NDWI), normalized red–green difference index (NGRDI),

renormalized difference vegetation index (RDVI), ratio vegetation index (RVI), soil adjusted vegetation index (SAVI), triangular vegetation index (TVI), and visual atmospheric resistance index (VARI) (Table 1).

After data acquisition, the soybean yield and VIs reflectance dataset for each field and year were randomly split into training (80%) and validation (20%) components. Following splitting and before developing the AI models, irrelevant VI’s (predictor) variables were removed using the recursive feature elimination with cross-validation (RFE-CV) method (Granitto et al., 2006). Beginning with every feature in the training dataset, RFE-CV deletes features one at a time until the best set of predictor variables is achieved.

2.5 | AI model development

Many AI techniques have been used in agricultural research to predict soil and plant characteristics (Acharya et al., 2021; D. R. Joshi et al., 2022; Sharma et al., 2022). Of these models, we selected DNN, Support Vector Machine (SVM), Random Forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO), and AdaBoost approaches for additional research. The DNN has a multilayer perceptron that is composed of multiple, non-linear layers and transforms raw data into a higher level representation (Joshi et al., 2023b; Khaki & Wang, 2019; LeCun et al., 2015). As the network gets deeper, it extracts more complex features that may improve prediction accuracy. The architecture of DNN is composed of hidden layers, neurons or nodes, activation functions, and input data that are passed from the input layer to the hidden layers, whose weights are used to store information (Muruganantham et al., 2022). An advantage of the DNN architecture is that it has been very successful at developing models to map complex datasets. In addition, DNNs such as those developed by Ball et al. (2017) and Schmidhuber (2015) have been used to create yield prediction algorithms based on crop reflectance. One major disadvantage is that it is difficult to explain how the results were obtained.

The SVM creates a line, or hyperplane, capable of separating the training data into a known number of output classes. The line or hyperplanes represent decision boundaries and are often used to classify continuous outputs (Brereton & Lloyd, 2010). However, the challenges associated with SVM are computational complexity and memory requirements, depending on the size of the dataset.

The RF is an ensemble learning method for regression based on the recursive partitioning principle. In recursive partitioning, populations are split into sub-populations (i.e., decision trees), each with unique characteristics. The strengths of the RF approach are that the accuracy and robustness of the model generally improves with the number of trees in the forest (Breiman, 2001). The RF approach has been used to predict corn, wheat (*Triticum* L.), and potato (*Solanum*

TABLE 1 Ten vegetation indices were calculated using remotely sensed plant reflectance data.

Index	Source	Formula
NDVI	Rouse (1974)	$\frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$
GNDVI	Moges et al. (2004)	$\frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$
TVI	Broge and Leblanc (2000)	$\frac{20(\text{NIR} - \text{Green}) - 200(\text{Red} - \text{Green})}{2}$
NGRDI	Tucker (1979)	$\frac{\text{Green} - \text{Red}}{\text{Green} + \text{Red}}$
VARI	Gitelson et al. (2002)	$\frac{\text{Green} + \text{Red} - \text{Blue}}{\text{Green} - \text{Red}}$
SAVI	Venancio et al. (2019)	$\frac{1.5(\text{NIR} - \text{Red})}{\text{NIR} + \text{Red} + 0.5}$
GCI	Gitelson et al. (2005)	$\frac{\text{NIR}}{\text{Green} - 1}$
NDWI	McFeeters (1996)	$\frac{\text{Green} - \text{NIR}}{\text{Green} + \text{NIR}}$
RDVI	Chen (1996)	$\frac{\text{NIR} - \text{Red}}{\sqrt{(\text{NIR} + \text{Red})}}$
RVI	Tucker (1979)	$\frac{\text{NIR}}{\text{Red}}$
DVI	Richardson and Wiegand (1977)	$\text{NIR} - \text{Red}$

Abbreviations: DVI, difference vegetation index; GCI, green chlorophyll vegetation index; GNDVI, green normalized differential vegetation index; NDVI, normalized differential vegetation index; NDWI, normalized difference water index; NGRDI, normalized red–green difference index; NIR, near-infrared; RDVI, renormalized difference vegetation index; RVI, ratio vegetation index; SAVI, soil adjusted vegetation index; TVI, triangular vegetation index; VARI, visual atmospheric resistance index.

tuberosum) yields at global and regional levels (Jeong et al., 2016). However, a weakness of the RF approach is that it can over smooth predictions when the training datasets are relatively small (Koparan et al., 2022).

The LASSO approach is well suited to the automation of some steps in the model specification process, whereas AdaBoost is an ensemble learning method that was developed to improve the performance of binary classifiers (Sai et al., 2022). AdaBoost employs an ensemble approach to improve the performance of classifiers by iteratively learning from mistakes (Sai et al., 2022). For each variable, a decision stump (decision tree with only one level) is created, and the decision tree's ability to assign samples to target classes is evaluated. The advantages of AdaBoost are that it is less prone to overfitting and exhibits reduced bias. The weaknesses are that AdaBoost is sensitive to outliers and noise.

In our analysis, after creating the training and validation datasets, the machine learning and deep learning modeling activities were performed six times for each year and field. This analysis developed a series of AI models that varied in complexity. As soybeans matured, more information was considered by each AI model. For example, for the VE growth stage, only data collected at the VE growth stage were used, whereas at the R6/7 growth stage, data collected at the previous five growth stages were used by the model. This means all AI models were first evaluated using VIs that were collected only during the VE/VC growth stage. After each analysis, VI information from the next growth stage (V1/V3) was added to VIs from the previous growth stage to test the effect of combining data from two growth stages for yield prediction. This process of adding VI information in the AI modeling from satellite passes at subsequent growth stages continued at the soybean R1/R2, R2/R3, R4/R5, and R6/R7 growth stages, that is, until the soybean reached full maturity. Thus, this modeling

approach resulted in training each AI model six times during the soybean growing season.

For the DNN, the activation function was the rectified linear unit (ReLU, or Rectifier). Different combinations of layers, neurons, and activation functions were analyzed to determine the optimal combination, given the input data (Figure 3). To improve prediction accuracy, various optimization techniques like batch normalization and dropping out were used. Dropout was used to reduce overfitting, whereas batch normalization was used to accelerate the learning process by reducing covariance shift. Various hidden layers were tested during the analysis, but in this case, five hidden layers gave the highest prediction accuracies. The optimal number of nodes to be assigned in each hidden layer, learning rate, batch size, and epochs were calculated using hyperparameter tuning.

The AI models (SVM, RF, LASSO, and AdaBoost) were implemented using the “Scikit-learn” (Pedregosa et al., 2011) Python library, whereas the DNN was implemented with the TensorFlow (Abadi et al., 2016) and Keras (Chollet, 2018) libraries. For all models, hyperparameter tuning was conducted to deploy the optimal set of parameters for each algorithm. This prevents both over- and under-fitting during model implementation. For the DNN model, different hyperparameters and their corresponding search grid values were defined and different ranges of learning rate values (0.1, 0.01, 0.001, 0.007, and 0.005) were tested. Similarly, each hidden layer's dropout, batch size, epoch, and neurons were also tested. A resampling method like 10-fold cross-validation was used to obtain validation evaluation matrices across each iteration to decide the hyperparameter setting for final training model selection. For SVM, the kernel function, cost, and gamma value were all considered to obtain the final robust training model. For the RF models, the number of trees and minimum and maximum number of data

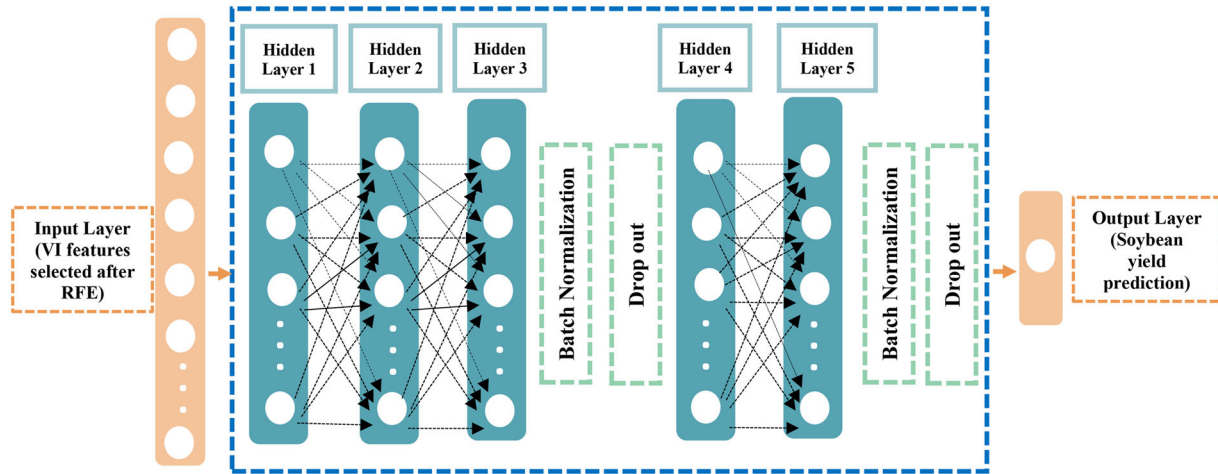


FIGURE 3 Modeling structure of the Deep Neural Network (DNN) model. RFE, recursive feature elimination; VI, vegetation index.

points in a node were determined, while for LASSO and AdaBoost, the alpha value and number of estimators with learning rate were considered during training model optimization. The optimal set of hyperparameters yielding the lowest error was then chosen as the final training model for that algorithm.

2.6 | Assessment of model performance

To evaluate the model performance, coefficients of determination (R^2) and root mean square errors (RMSEs) were calculated using Equations (1) and (2):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (1)$$

and

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

where y_i and \hat{y}_i were the measured and predicted soybean yield values, respectively, \bar{y} was the mean of all measured yield values, and n was the number of samples. The best performing models have high R^2 (closer to 1) and low RMSE values. To assess the transferability of the year and field specific models, the field specific model created in 2019 was used to predict the 2021 yield.

2.7 | Optimum time to acquire satellite images

The relationship between yield and surface reflectance (VIs) at the different growth stages was determined by calculating Pearson's correlation (r) using the "cor" function in Rstu-

dio (R core team, 2021). To determine the optimum time to acquire surface reflectance data, the model's ability to predict soybean yields at the six growth stages (VE/VC, V1/V3, R1/R2, R2/R3, R4/R5, and R6/R7) was compared using the DNN model. For this comparison, the DNN model was selected for its superior performance when compared to the other AI models. The DNN model was generated using VI information from a single image at a specific growth stage and did not include VI information from other images taken at other growth stages.

3 | RESULTS AND DISCUSSION

3.1 | Weather and climatic conditions

In 2019, rainfall was higher than the 30-year average for all sites (18.7%, 14.0%, and 41.1% greater in Edmunds, Hamlin, and Miner counties, respectively), and air temperature was lower than the 30-year average during the growing season (8.5%, 25.2%, and 5.3% lower in Edmunds, Hamlin, and Miner counties, respectively; Table 2). In 2021, growing season rainfall was lower than the 30-year average (26.3%, 32.7%, and 58.4% lower in Edmunds, Hamlin, and Miner counties, respectively), and air temperatures, relative to the 30-year average, were mixed.

At Edmunds in 2019, yields ranged from 1.0 to 3.9 Mg ha⁻¹ and averaged 3.0 Mg ha⁻¹; whereas in 2021, yields ranged from 0.3 to 4.6 Mg ha⁻¹ and averaged 2.9 Mg ha⁻¹ (Figure 4a). At Hamlin, the 2019 yields ranged from 0.3 to 4.4 Mg ha⁻¹ and averaged 2.4 Mg ha⁻¹; whereas in 2021, the yield ranged from 0.7 to 5.7 Mg ha⁻¹ and averaged 3.7 Mg ha⁻¹ (Figure 4b). At Miner, the 2019 yields ranged from 0.7 to 4.2 and averaged 2.9 Mg ha⁻¹, whereas in 2021, yields ranged from 0.1 to 1.3 Mg ha⁻¹ and averaged 0.6 Mg ha⁻¹ (Figure 4c). The low yields in 2021 were attributed to drought (58.4% lower rainfall during the

TABLE 2 Summary of temperature and rainfall at all three study counties during 2019 and 2021 (Data source: National Oceanic and Atmospheric Administration (NOAA), National Centers for Environmental Information, 2022).

Site	Year	Growing season		Annual		30-year average growing season	
		Temp. (°C)	Rainfall (mm)	Temp. (°C)	Rainfall (mm)	Temp. (°C)	Rainfall (mm)
Edmunds	2019	16.2	438	3.5	686	17.7	369
	2021	17.8	272	6.2	453		
Hamlin	2019	13.3	515	1.5	821	17.8	452
	2021	14.1	304	3.6	633		
Miner	2019	17.7	748	5.6	1098	18.7	529
	2021	19.5	220	8.4	449		

Abbreviation: Temp., temperature.

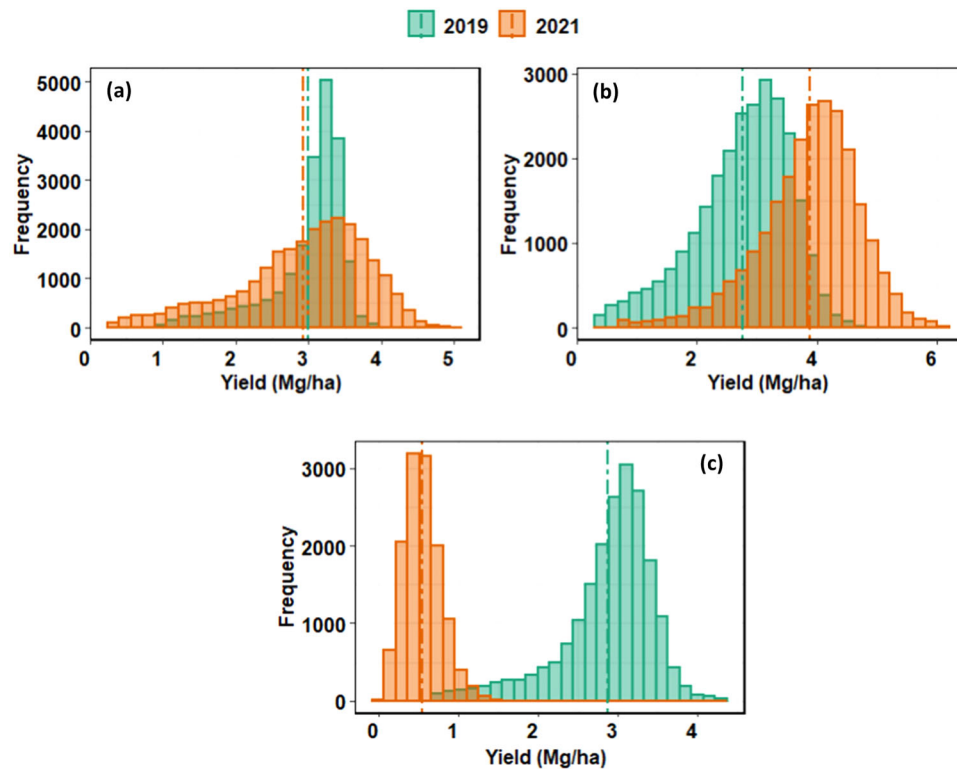


FIGURE 4 Histogram plot for sites in Edmunds (a), Hamlin (b), and Miner (c) showing yield distribution before aggregating to each grid in 2019 and 2021. Dashed vertical lines represent average yield.

growing season) when compared to the 30-year average (Table 2).

3.2 | Model performance

The amount of yield variation (R^2) explained by the remote sensing-based AI models at the VE/VC (seed emergence) growth stage was relatively low. This was attributed to the reflectance data providing more information about the soil color than plant health. Later-season model performance

improved with the greater amount of data collected. For example, at Edmunds in 2019, the R^2 values ranged from 0.26 to 0.32 when the AI models were based only on data collected at the VE/VC growth stage. Increasing the number of images (from one image at growth stage VE to two at VE + V3) increased the range of R^2 values (0.31–0.45). When all images that were collected prior to R5 were included, the percentage of yield variation explained by the models ranged from 58% to 67% (Figure 5). Including images collected after R5, that is, R6/R7, had a relatively low impact on model performance.

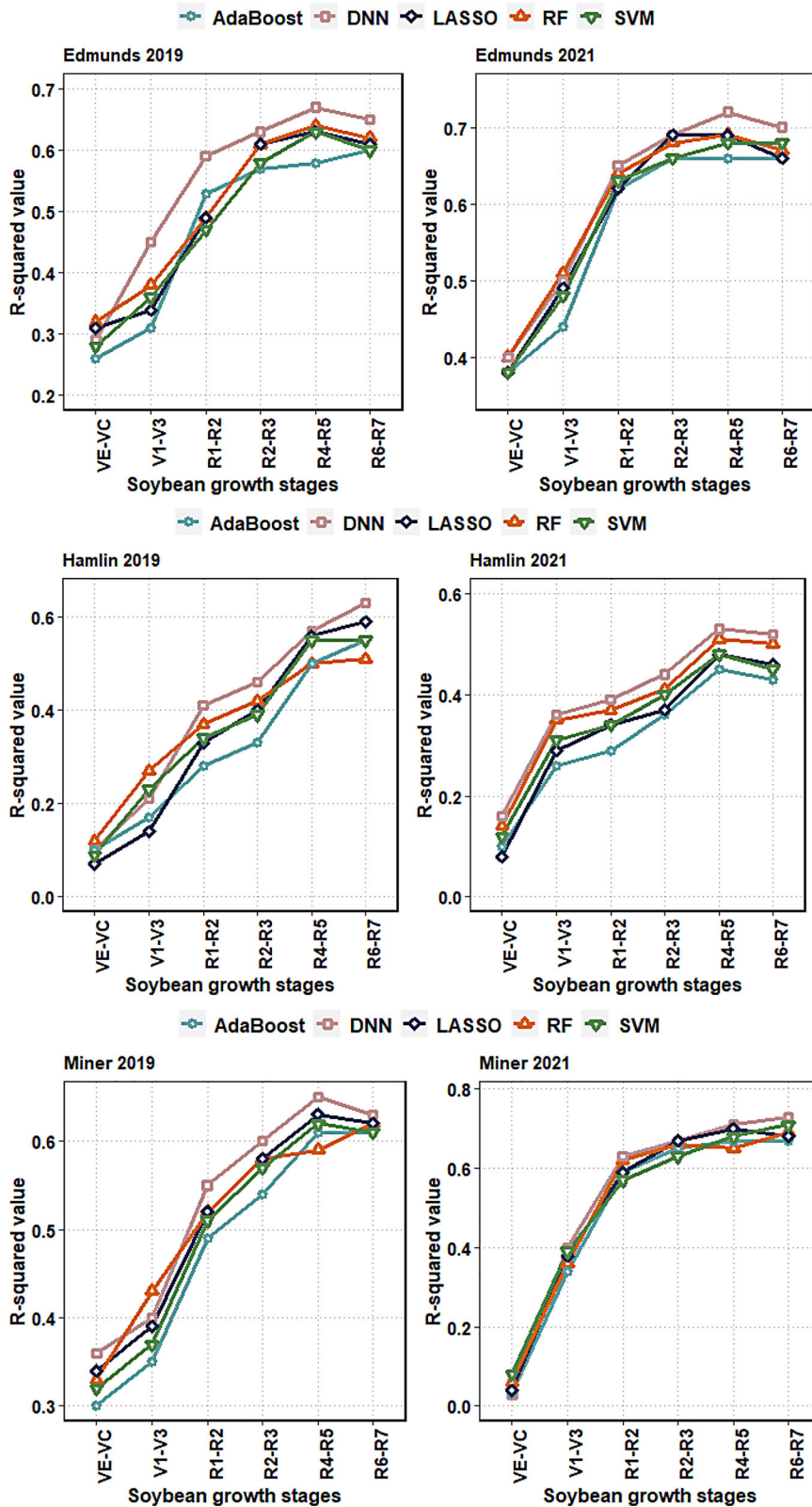
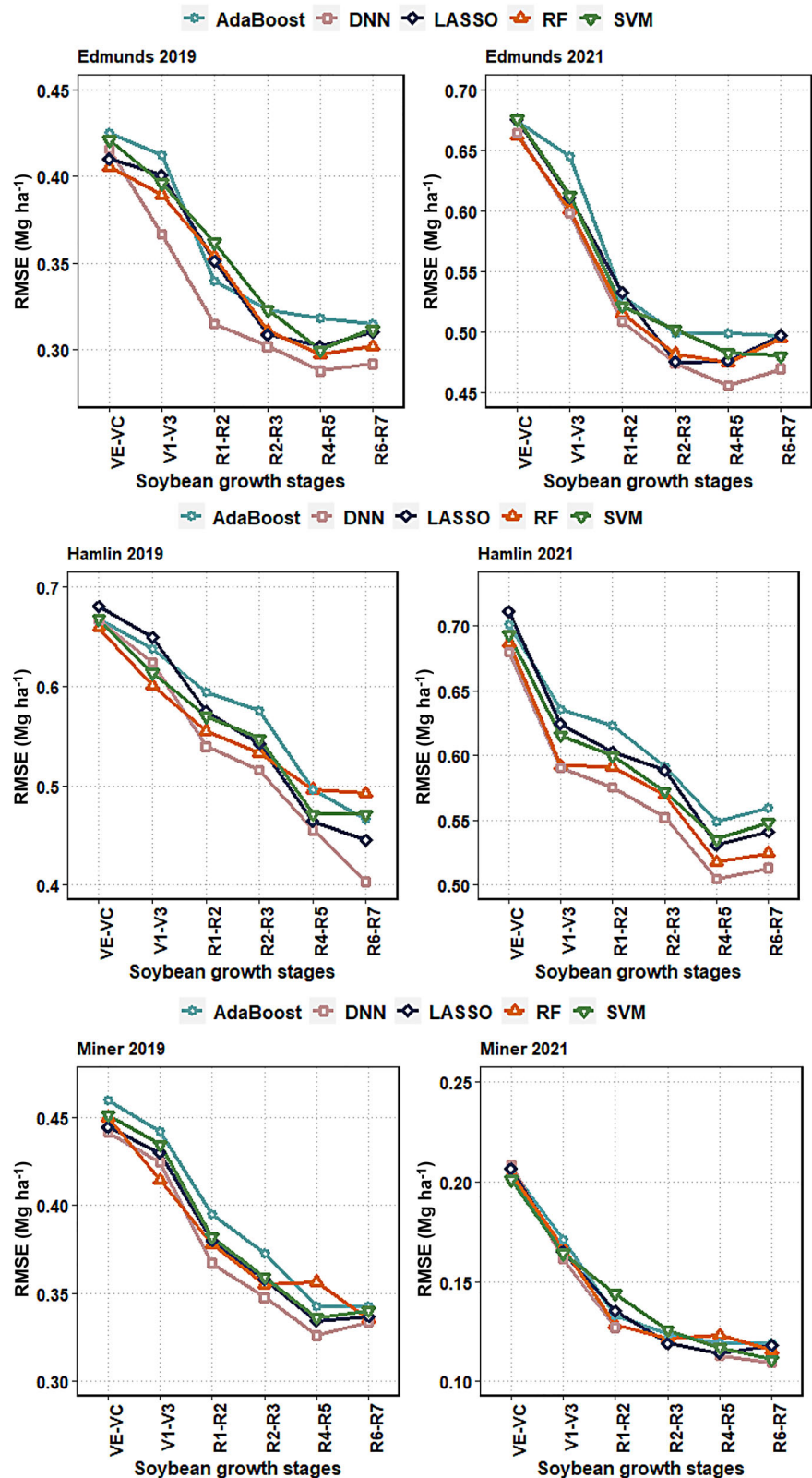


FIGURE 5 The proportion of soybean yield explained by our five artificial intelligence (AI) models during the testing phase at the Edmunds, Hamlin, and Miner fields in 2019 and 2021. In this analysis, the simplest AI model only utilized reflectance information from a single remote sensing image acquired at the soybean VE/VC growth stage. For each additional growth stage, reflectance information from an additional satellite image was acquired and included in the model. In other words, the AI models were built using reflectance data available up to and including the specific growth stage. DNN, deep neural network; LASSO, least absolute shrinkage and selection operator; RF, random forest; SVM, support vector machine.

At the early growth stages, model performance was somewhat mixed. For example, at the VE/VC stage, RF and AdaBoost outperformed DNN. These results may be attributed to the small number of input variables that were poorly correlated to yield (Kang & Kang, 2017; Schmidhu-

ber, 2015). As the season progressed and more input variable data were collected, DNN models generally outperformed the other models. For example, at Edmunds in 2019, the R6/R7 DNN model built with data from all six growth stages had a yield prediction R^2 value of 0.65 and a RMSE value of

FIGURE 6 The performance of various artificial intelligence (AI) models in predicting soybean yield, as indicated by root mean square errors (RMSEs), during the testing phase for the Edmunds, Hamlin and Miner fields in 2019 and 2021. For each additional growth stage, reflectance information from an additional satellite image was acquired and included in the model. In other words, the AI models were built using all reflectance data available up to and including the specific growth stage. DNN, deep neural network; LASSO, least absolute shrinkage and selection operator; RF, random forest; SVM, support vector machine.



0.29 Mg ha⁻¹. The RF ($R^2 = 0.62$ and RMSE = 0.3 Mg ha⁻¹), AdaBoost ($R^2 = 0.6$ and RMSE = 0.31 Mg ha⁻¹), LASSO ($R^2 = 0.61$ and RMSE = 0.31 Mg ha⁻¹), and SVM ($R^2 = 0.60$ and RMSE = 0.31 Mg ha⁻¹) models had slightly lower R^2 and RMSE values (Figures 5 and 6). These results

were attributed to batch normalization and drop-out methods used by the DNN model (Ioffe & Szegedy, 2015). Others have reported similar findings (Khaki & Wang, 2019; Maimaitijiang et al., 2020; Sun et al., 2019). For example, Maimaitijiang et al. (2020), in a small plot study, reported

that the DNN model based on unmanned aerial vehicle-based remote sensing explained 72% of yield prediction variability. In the modeling of county-level data, Sun et al. (2019) reported that a convolutional neural network based on multiple moderate resolution imaging spectroradiometer (MODIS) images performed better than the other models tested.

The model's transferability was assessed by comparing the yield predictions to measured values at a site where data from the site were not used in training. In this analysis, the DNN model based on data collected in 2019 was used to predict yields in 2021. In all cases, the R^2 values decreased from 2019 (data used in training) to 2021 (data not used in training). For example, the R^2 value of the DNN model for Edmunds decreased from 0.70 to 0.40. Similar results were observed at Hamlin, where the R^2 values decreased from 0.52 to 0.35, and at Miner, where the R^2 values decreased from 0.73 to 0.32. These decreases indicate that models created for one field and year have limited transferability to other fields. To assess if this limitation could be minimized, the datasets for all sites and years were combined. In this analysis, the R^2 values of models created with VI data up to R6/R7 across years and sites were 60%, 67%, 64%, 65%, and 54% for the AdaBoost, DNN, LASSO, RF, and SVM models, respectively. It was interesting that, when compared to the year- and field-specific models, combining the 2 years and three fields into a common dataset had a minimal impact on R^2 values. This analysis has several implications. First, additional work needs to be conducted to assess the model's ability to predict soybean yield with VI data for fields not used to train the model. Second, the DNN model performed better than the other models. Third, it may be possible to create models that provide accurate, precise, and realistic site-specific soybean yield predictions from VI data. Fourth, accurate soybean yield predictions might then be used for later precision crop management, harvest planning, and marketing.

3.3 | Selecting the best time to acquire satellite images

It is critical to determine the best time to acquire satellite images for the prediction of soybean yield. Just after soybean emergence (VE/VC), all VIs were poorly correlated to soybean yield at all three locations (Figure 7). As soybean plants matured, the correlations coefficients (r) improved. The correlation coefficients between individual VI and soybean yields were highest for satellite images acquired at either R2/R3, R4/R5, or R6/R7 growth stages.

To further determine the best time to acquire satellite images for soybean yield prediction, the DNN model was used because it generally outperformed the other approaches. For this purpose, separate DNN models were built to predict soybean yields using plant reflectance information acquired

from a single satellite image at a specific growth stage. Across all study fields and years, we found that the DNN model, using remote sensing data collected at VE/VC, had the lowest R^2 and largest RMSE (Figure 8). As the crop matured from VE/VC to R4/R5 (which occurred in the end of August), the R^2 values increased while the RMSE values decreased progressively. This indicates that the DNN model, based on satellite images acquired at later growth stages, particularly R4 and R5, explained more yield variability. Thus, surface reflectance information collected at the soybean R4/R5 growth stage was the most important, whereas images collected at the VE/VC growth stage were the least important.

At the Hamlin field, slightly different results were observed during the 2019 growing season, where the maximum R^2 and minimum RMSE occurred at the R6/R7 growth stage (Figure 8), which might be due to the indeterminate growth characteristic of the soybean crop. A study conducted by You et al. (2017) and Khaki et al. (2021) also found that August satellite images had the highest accuracy, with the lowest RMSE, for soybean yield prediction. Based on these results, we conclude that the best time to acquire satellite images is at later soybean reproductive stages. This is critical to fast and timely predictions of soybean yield over large geographic regions.

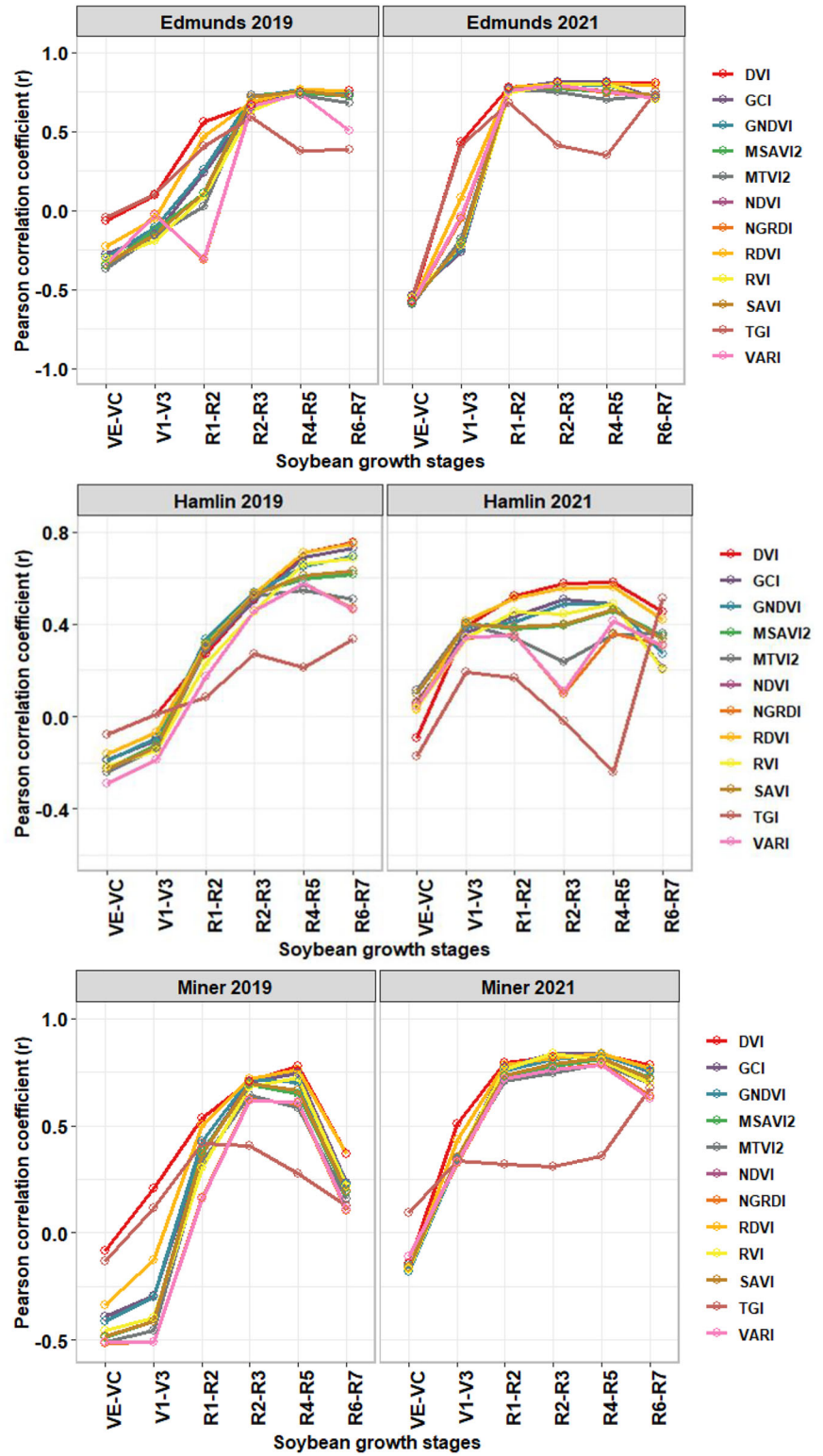
4 | CONCLUSIONS

An important component of implementing precision farming is the ability to quantify the potential benefits. This involves predicting yields. Historically, soybean yields are estimated by determining the number of plants per hectare, the number of pods per plant, the number of seeds or pods, and the weight of each seed. These estimates are often correlated to plant biomass, but this is not always the case. The primary problems with physical measurements are cost, accuracy, and low suitability for precision farming.

Estimating soybean yield is complicated by the soybean plant that can either have determinant or indeterminate growth characteristics and low harvest indexes. Indeterminant growth means that if the environment is favorable, the plants will continue growing (Schoving et al., 2022). This study showed that remote sensing could be used to predict soybean yield in large production fields planted with indeterminant cultivars.

Realistic yield estimates at earlier soybean growth stages are needed for numerous purposes, including pest management input decisions, crop marketing, price forecasting, insurance, and harvest planning. The analysis across fields and years suggests that AI models can be created to make relatively accurate within-field soybean yield predictions. In the past, most yield predictions were based on remote

FIGURE 7 Correlation coefficients (r) between yield and vegetation indices (VIs) calculated for the different soybean growth stages at the Edmunds, Hamlin, and Miner study sites. DVI, difference vegetation index; GCI, green chlorophyll vegetation index; GNDVI, green normalized differential vegetation index; MSAVI, modified soil-adjusted vegetation index; MTVI, modified triangular vegetation index; NDVI, normalized differential vegetation index; NGRDI, normalized red–green difference index; RDVI, renormalized difference vegetation index; RVI, ratio vegetation index; SAVI, soil adjusted vegetation index; TGI, triangular greenness index; VARI, visual atmospheric resistance index.



sensing data collected during the late vegetative and early reproductive stages.

Different models exhibited different predictive abilities. Generally, the DNN models outperformed the other models. We also discovered that the timing of surface reflectance collection is crucial for improved prediction accuracy. The

correlation analysis revealed that reflectance data collected at the R4/R5 or R6/R7 growth stages were highly correlated with yield and that models based on this information explained the most yield variability. Overall, field-scale soybean yield prediction can be accomplished by combining satellite-based remote sensing and AI algorithms. It is likely that the models

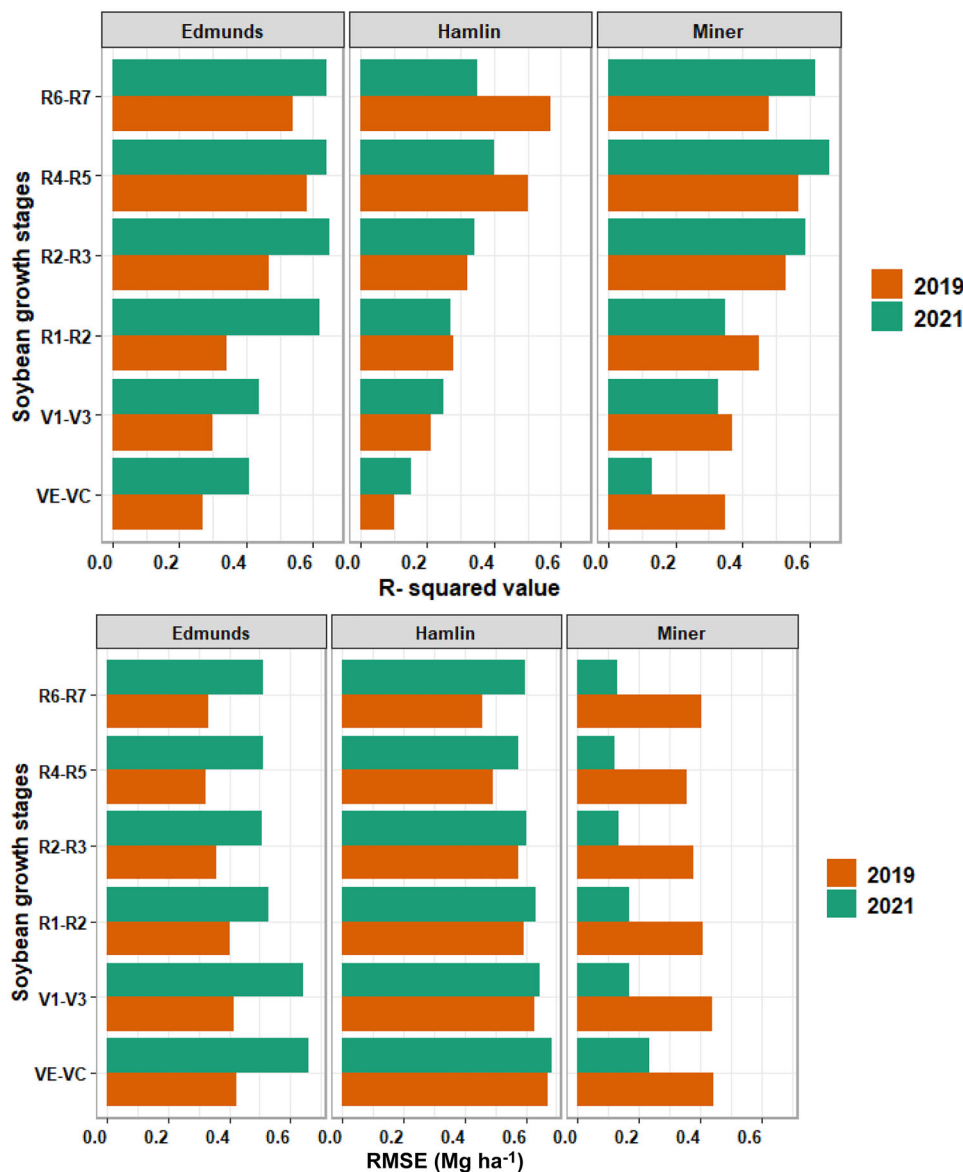


FIGURE 8 Change in deep neural network (DNN) model performance (R^2 and root mean square error [RMSE]) on the testing dataset due to the individual satellite image taken at six different growth stage for the Edmunds, Hamlin, and Miner study sites.

could be improved by including weather, chlorophyll fluorescence, and soil parameters (V. R. Joshi et al., 2021; Joshi et al., 2023a). Though DNN has been found to be the best prediction model, its limitations include the need for large datasets containing a large number of input variables and the need for high-speed computer processing.

AUTHOR CONTRIBUTIONS

Deepak R. Joshi: Conceptualization; data curation; formal analysis; investigation; methodology; software; writing—original draft; writing—review and editing. **Sharon A. Clay:** Conceptualization; investigation; methodology; validation; writing—review and editing. **Praakriti Sharma:** Data curation; formal analysis; investigation; software. **Hossein Moradi Rekabdarkolae:** Resources; validation; writing—

review and editing. **Donna M. Rizzo:** Resources; validation; writing—review and editing. **Resham Thapa:** Resources; validation; writing—review and editing. **Tulsi Kharel:** Supervision; validation; writing—review and editing. **David E. Clay:** Conceptualization; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; writing—review and editing.

ACKNOWLEDGMENTS

This research is based on work supported by the South Dakota Soybean Association and the National Science Foundation under grant numbers 2202706 and 2026431, respectively. Trade names or commercial products mentioned in this article are solely for the purpose of providing specific information and do not infer either recommendation or endorsement by

the US Department of Agriculture. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or US government determination or policy. USDA is an equal opportunity provider and employer.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Sharon A. Clay  <https://orcid.org/0000-0003-4166-6995>

Resham Thapa  <https://orcid.org/0000-0002-0059-764X>

David E. Clay  <https://orcid.org/0000-0002-5031-9744>

REFERENCES

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., ... Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX symposium on operating systems design and implementation (OSDI '16)* (pp. 265–283). Cornell University.
- Acharya, B. S., Bhandari, M., Bandini, F., Pizarro, A., Perks, M., Joshi, D. R., Wang, S., Dogwiler, T., Ray, R. L., Kharel, G., & Sharma, S. (2021). Unmanned aerial vehicles in hydrology and water management: Applications, challenges, and perspectives. *Water Resources Research*, 57(11), e2021WR029925.
- Ball, J. E., Anderson, D. T., & Chan, C. S. Jr. (2017). Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community. *Journal of Applied Remote Sensing*, 11, 042609.
- Bosch, M. (2019). PyLandStats: An open-source Pythonic library to compute landscape metrics. *PLoS ONE*, 14(12), e0225734. <https://doi.org/10.1371/journal.pone.0225734>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/a:1010933404324>
- Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst*, 135(2), 230–267. <https://doi.org/10.1039/b918972f>
- Broge, N. H., & Leblanc, E. (2000). Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. *Remote Sensing of Environment*, 76(2), 156–172.
- Chen, J. M. (1996). Evaluation of vegetation indices and a modified simple ratio for boreal applications. *Canadian Journal of Remote Sensing*, 22(3), 229–242.
- Cho, J. B., Guinness, J., Kharel, T. P., Sunoj, S., Kharel, D., Oware, E. K., van Aardt, J., & Ketterings, Q. M. (2021). Spatial estimation methods for mapping corn silage and grain yield monitor data. *Precision Agriculture*, 22, 1501–1520. <https://doi.org/10.1007/s11119-021-09793-z>
- Chollet, F. (2018). *Keras: The python deep learning library*. Astrophysics Source Code Library.
- de Souza, R., Toebe, M., Marchiorom, V. S., Filho, A. C., Bittencourt, K. C., Mello, A., & Paraginski, A. (2023). Sample size and modeling plant variability using precision statistics in soybean counting traits. *Field Crop Research*, 291, 108789.
- Dobermann, A., & Ping, J. (2004). Geostatistical integration of yield monitor data and remote sensing improves yield maps. *Agronomy Journal*, 96, 285–297. <https://doi.org/10.2134/agronj2004.2850>
- Frazier, A. E., & Hemingway, B. (2021). A technical review of planet smallsat data: Practical considerations for processing and using PlanetScope imagery. *Remote Sensing*, 13, 3930.
- Gitelson, A. A., Kaufman, Y. J., Stark, R., & Rundquist, D. (2002). Novel algorithms for remote estimation of vegetation fraction. *Remote Sensing of Environment*, 80(1), 76–87.
- Gitelson, A. A., Viña, A., Ciganda, V., Rundquist, D. C., & Arkebauer, T. J. (2005). Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters*, 32(8), 271–282. <https://doi.org/10.1029/2005GL022688>
- Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2), 83–90.
- Houborg, R., & McCabe, M. F. (2018). Daily retrieval of NDVI and LAI at 3 m resolution via the fusion of CubeSat, Landsat, and MODIS data. *Remote Sensing*, 10, 890.
- Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach, & D. Blei (Eds.), *International conference on machine learning, proceedings of machine learning research* (Vol. 37, pp. 448–456). Journal of Machine Learning Research.
- Jain, M., Srivastava, A. K., Singh, B., Joon, R. K., McDonald, A., Royal, K., Lisaius, M. C., & Lobell, D. B. (2016). Mapping smallholder wheat yields and sowing dates using micro-satellite data. *Remote Sensing*, 8, 860.
- Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., & Reddy, V. R. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE*, 11, e0156571.
- Joshi, A., Pradhan, B., Chakraborty, S., & Behera, M. D. (2023a). Winter wheat yield prediction in the conterminous United States using solar-induced chlorophyll fluorescence data and XGBoost and random forest algorithm. *Ecological Informatics*, 77, 102194.
- Joshi, A., Pradhan, B., Gite, S., & Chakraborty, S. (2023b). Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: A systematic review. *Remote Sensing*, 15, 2014. <https://doi.org/10.3390/rs15082014>
- Joshi, D. R., Clay, D. E., Clay, S. A., Moriles-Miller, J., Daigh, A. L. M., Reicks, G., & Westhoff, S. (2022). Quantification and machine learning based N₂O–N and CO₂–C emissions predictions from a decomposing rye cover crop. *Agronomy Journal*, In press. <https://doi.org/10.1002/agj2.21185>
- Joshi, V. R., Kazula, M. J., Coulter, J. A., Naeve, S. L., & Garcia y Garcia, A. (2021). In-season weather data provide reliable yield estimates of maize and soybean in the US central Corn Belt. *International Journal of Biometeorology*, 65, 489–502. <https://doi.org/10.1007/s00484-020-02039-z>
- Kang, H.-W., & Kang, H.-B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. *PLoS ONE*, 12, e0176244.
- Khaki, S., Pham, H., & Wang, L. (2021). Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. *Scientific Reports*, 11, 11132. <https://doi.org/10.1038/s41598-021-89779-z>

- Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, *10*, 621. <https://doi.org/10.3389/fpls.2019.00621>
- Kharel, T., Swink, S., Youngerman, C., Maresma, A., Czymbek, K., Ketterings, Q., Kyveryga, P., Lory, J., Musket, T. A., & Hubbard, V. (2018). *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. <https://hdl.handle.net/1813/56111>
- Kharel, T. P., Bhandari, A. B., Mubvumba, P., Tyler, H. L., Fletcher, R. S., & Reddy, K. N. (2023). Mixed-species cover crop biomass estimation using planet imagery. *Sensors*, *23*, 1541. <https://doi.org/10.3390/s23031541>
- Koparan, M. H., Rekabdarkolae, H. M., Sood, K., Westhoff, S. M., Reese, C. L., & Malo, D. D. (2022). Estimating soil organic carbon levels in cultivated soils from satellite image using parametric and data-driven methods. *International Journal of Remote Sensing*, *43*(9), 3429–3449.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.
- Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., & Fritsch, F. B. (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sensing of Environment*, *237*, 111599. <https://doi.org/10.1016/j.rse.2019.111599>
- McFeeters, S. K. (1996). The use of the normalized difference water index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, *17*(7), 1425–1432.
- Moges, S. M., Raun, W. R., Mullen, R. W., Freeman, K. W., Johnson, G. V., & Solie, J. B. (2004). Evaluation of green, red and near infrared bands for predicting winter wheat biomass, nitrogen uptake, and final grain yield. *Journal of Plant Nutrition*, *27*(8), 1431–1441.
- Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., & Islam, N. (2022). A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, *14*(9), 1990.
- National Oceanic and Atmospheric Administration (NOAA), National Centers for Environmental Information. (2022). *Climate dataonline: Dataset discovery*. NOAA. <https://www.ncdc.noaa.gov/cdo-web/datasets>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.
- Planet Team. (2017). Planet application program interface; Space for life on Earth. <https://api.planet.com>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Richardson, A. J., & Wiegand, C. L. (1977). Distinguishing vegetation from soil background information. *Photogrammetric Engineering and Remote Sensing*, *43*(12), 1541–1552.
- Rouse, J. W. Jr. (1974). *Monitoring the vernal advancement of retrogradation of natural vegetation* (Document ID 19740022555). NTRS–NASA Technical Reports Server. <https://ntrs.nasa.gov/citations/19740022555>
- Sai, K., Sood, N., & Saini, I. (2022). Abiotic stress classification through spectral analysis of enhanced electrophysiological signals of plants. *Biosystems Engineering*, *219*, 189–204.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.
- Schoving, C., Champolivier, L., Maury, P., & Debaeke, P. (2022). Combining multi-environmental trials and crop simulation to understand soybean response to early sowings under contrasting water conditions. *European Journal of Agronomy*, *133*, 126439.
- Sharma, P., Leigh, L., Chang, J., Maimaitijiang, M., & Caffé, M. (2022). Above-ground biomass estimation in oats using UAV remote sensing and machine learning. *Sensors*, *22*, 601. <https://doi.org/10.3390/s22020601>
- Sun, J., Di, L., Sun, Z., Shen, Y., & Lai, Z. (2019). County-level soybean yield prediction using deep CNN-LSTM model. *Sensors*, *19*(20), 4363.
- Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2), 127–150.
- USDA-NRCS. (2023). *Web soil survey*. <https://websoilsurvey.nrcs.usda.gov/app/>
- Venancio, L. P., Mantovani, E. C., do Amaral, C. H., Neale, C. M. U., Gonçalves, I. Z., Filgueiras, R., & Campos, I. (2019). Forecasting corn yield at the farm level in Brazil based on the FAO-66 approach and soil-adjusted vegetation index (SAVI). *Agricultural Water Management*, *225*, 105779.
- Yang, C., Everitt, J. H., Du, Q., Luo, B., & Chanussot, J. (2012). Using high-resolution airborne and satellite imagery to assess crop growth and yield variability for precision agriculture. *Proceedings of the IEEE*, *101*, 582–592. IEEE.
- You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep gaussian process for crop yield prediction based on remote sensing data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). <https://doi.org/10.1609/aaai.v31i1.11172>

How to cite this article: Joshi, D. R., Clay, S. A., Sharma, P., Rekabdarkolae, H. M., Kharel, T., Rizzo, D. M., Thapa, R., & Clay, D. E. (2023). Artificial intelligence and satellite-based remote sensing can be used to predict soybean (*Glycine max*) yield. *Agronomy Journal*, 1–14. <https://doi.org/10.1002/agj2.21473>