DOI: 10.15201/hungeobull.72.4.4 Hungarian Geographical Bulletin **72** 2023 (4) 383–398.

Effectiveness of machine learning and deep learning models at county-level soybean yield forecasting

Nizom FARMONOV[1], Khilola AMANKULOVA[1], Shahid Nawaz KHAN[2], Mokhigul ABDURAKHIMOVA[3], József SZATMÁRI[1], Tukhtaeva KHABIBA[4], Radjabova MAKHLIYO[4], Meiliyeva KHODICHA[5] and László MUCSI[1]

Abstract

Crop yield forecasting is critical in modern agriculture to ensure food security, economic stability, and effective resource management. The main goal of this study was to combine historical multisource satellite and environmental datasets with a deep learning (DL) model for soybean yield forecasting in the United States' Corn Belt. The following Moderate Resolution Imaging Spectroradiometer (MODIS) products were aggregated at the county level. The crop data layer (CDL) in Google Earth Engine (GEE) was used to mask the data so that only soybean pixels were selected. Several machine learning (ML) models were trained by using 5 years of data from 2012 to 2016: random forest (RF), least absolute shrinkable and selection operator (LASSO) regression, extreme gradient boosting (XGBoost), and decision tree regression (DTR) as well as DL-based one-dimensional convolutional neural network (1D-CNN). The best model was determined by comparing their performances at forecasting the soybean yield in 2017–2021 at the county scale. The RF model outperformed all other ML models with the lowest RMSE of 0.342 t/ha, followed by XGBoost (0.373 t/ha), DTR (0.437 t/ha), and LASSO (0.452 t/ha) regression. However, the 1D-CNN model showed the highest forecasting accuracy for the 2018 growing season with RMSE of 0.280 t/ha. The developed 1D-CNN model has great potential for crop yield forecasting because it effectively captures temporal dependencies and extracts meaningful input features from sequential data.

Keywords: agriculture, remote sensing, farmers, random forest, soybean, machine learning

Received September 2023, accepted October 2023.

Introduction

Crop yield forecasting is crucial in agricultural decision-making for food security, crop insurance, and improving overall food production (Tantalaki, N. *et al.* 2019). Traditionally, farmers relied on their personal experience and incorporated weather and other relevant data to forecast their individual crop yields and make informed decisions (Liakos, K. *et al.* 2018). However, this traditional approach is associated with inherent uncertainties, particularly when extrapolated to large-scale scenarios, because of influencing factors that differ depending on the region and crop type (Shahhosseini, M. *et al.* 2020). Crop yield forecasting and prediction are distinct approaches that differ in methodology and

[1] Department of Geoinformatics, Physical and Environmental Geography, University of Szeged, Egyetem utca 2, 6722 Szeged, Hungary. Corresponding author's e-mail: farmonov.nizom@stud.u-szeged.hu

[2] Geospatial Sciences Center of Excellence, Department of Geography and Geospatial Sciences, South Dakota State University, Brookings, SD 57007, USA.

[3] Department of State Cadastre, Tashkent Institute of Irrigation and Agricultural Mechanization Engineers, National Research University, Tashkent, Uzbekistan

[4] Department of Hydrology and Ecology, "TIIAME" NRU Bukhara Institute of Natural Resources Management, Gazli Avenue 32, Bukhara, Uzbekistan.

[5] Department of Land Resources, Cadastre and Geoinformatics, Karshi Institute of Irrigation and Agrotechnology, "TIIAME" National Research University, Karshi, Uzbekistan.

data utilization (Shahhosseini, M. *et al.* 2020). Forecasting entails using historical observations to train a model and subsequently generating forecasts by employing input features specific to the future (Paudel, D. *et al.* 2021). The forecasting process acknowledges the influence of various biotic and abiotic factors on crop yield and necessitates a comprehensive understanding and management of these elements for a full grasp of the yield dynamics. A critical issue is the scarcity of extensive data encompassing all pertinent factors on a large scale.

Crop yield prediction differs from crop yield forecasting in that the former can utilize the target variable of the current year in the training phase (Shahhosseini, M. *et al.* 2020). Current crop yield prediction approaches can be divided into two main categories: physical models and statistical models (Tripathy, R. *et al.* 2022). Physical models simulate crop-growing conditions in combination with parameters that affect crop yield. Representative physical models include the Agricultural Production Simulator (Keating, B.A. *et al.* 2003) and Decision Support System for Agrotechnology Transfer (Jones, J.W. *et al.* 2003). Physical models are widely used, but they require extensive data for calibration, which limits their capabilities for large-scale crop monitoring. In contrast, statistical models are simpler with fewer input requirements, which make them very convenient for large-scale studies. Machine learning (ML) models use historical data to characterize the relationship between input variables and crop yield (Ma, Y. *et al.* 2021). A major advantage of ML models is that they can be used even when some specific crop parameters are not available. With the increased availability of data and computing power, robust ML methods have been developed and applied to crop yield prediction.

The increased availability of extensive remote sensing data has greatly facilitated their utilization in various agricultural applications, including crop classification, identification, and drought characterization (Sun, J. *et al.* 2019; Khan, S.N. *et al.* 2023), which has opened new avenues for extracting meaningful input features from remote sensing data for crop yield prediction. Vegetation indices (VIs) include factors such as greenness, vegetation health, and stress, and they have received substantial interest for research in vegetation dynamics (Panda, S.S. *et al.* 2010; Khan, K. *et al.* 2020). The normalized difference vegetation index (NDVI) is widely used for crop health characterization and yield prediction (Fernandes, J.L. *et al.* 2017). The NDVI measures the difference in reflectance of near-infrared (Nir) and red light to capture variations in plant biomass and photosynthetic activity, which makes it very useful for assessing crop productivity. Other VIs that have been employed for crop yield prediction include the enhanced vegetation index (EVI), soil-adjusted vegetation index, and normalized difference water index. In addition to remote sensing data, some studies have utilized climatic variables and soil data for crop yield prediction.

Recent studies have combined remote sensing data with ML models for crop yield prediction (Shahhosseini, M. *et al.* 2020). For example, Piekutowska, M. *et al.* (2021) used multiple linear regression on phenological and meteorological data to predict the potato yield and obtained a mean absolute percentage error of < 15 percent. Zeng, W. *et al.* (2018) used partial least-squares regression on weather data to predict the sunflower yield with an accuracy of $R^2 = 0.69$. Other ML models, including random forest (RF) (Khan, S.N. *et al.* 2022), support vector regression (Khosla, E. *et al.* 2020), and decision tree regression (Khan, S.N. *et al.* 2022) have been used for crop yield prediction at different scales and with different variables. In addition to conventional ML models, several deep learning (DL) models have recently been applied to crop yield prediction (Kang, Y. *et al.* 2020). Sun, J. *et al.* (2019) used a convolutional neural network and long short-term memory (CNN-LSTM) model with Moderate Resolution Imaging Spectroradiometer (MODIS) data to predict the soybean yield at the county level.

They classified data into different stages of the crop-growing season and evaluated the model performance in the stages. The end-of-season models outperformed the in-season models. (Ma, Y. *et al.* 2021) used a Bayesian neural network (BNN) with VI, climate, and soil data to predict the corn yield at the county level. They compared the performances of several models and found that the BNN outperformed other ML models such as RF, support vector regression, and LSTM. Although Earth Observation Systems (EOS) play a crucial role in monitoring crop yield through satellite data, there exists a significant research gap in enhancing the integration of multiplatform data, including data processing techniques, and the effective application of this technology in precision agricultural management. Novel research efforts are needed to bridge this gap and further optimize the utilization of EOS for informed decision-making in agriculture.

In this study, we applied several ML and DL models to forecast the soybean yield of the United States (US) at the county level and evaluated their performances. The spatial patterns between the observed and forecasted yields were also analysed. To address the above tasks, the following research questions were selected:

1. What are the most significant input features to predict the crop yield at the county scale?

2. How does the deep learning-based 1D-CNN model compare to traditional ML models for forecasting soybean yield in the US Corn Belt using historical satellite and environmental data?

Data

Study area

The study area comprised fourteen soybean-producing states: North Dakota (ND), South Dakota (SD), Nebraska (NE), Kansas (KS), Minnesota (MN), Iowa (IA), Missouri (MO), Arkansas (AR), Wisconsin (WI), Illinois (IL), Mississippi (MS), Michigan (MI), Indiana (IN), and Ohio (OH) *(Figure 1).* Most of the study area is in the Midwest, also known as the Corn Belt (Green, T.R. *et al.* 2018), and it is responsible for producing most of the corn

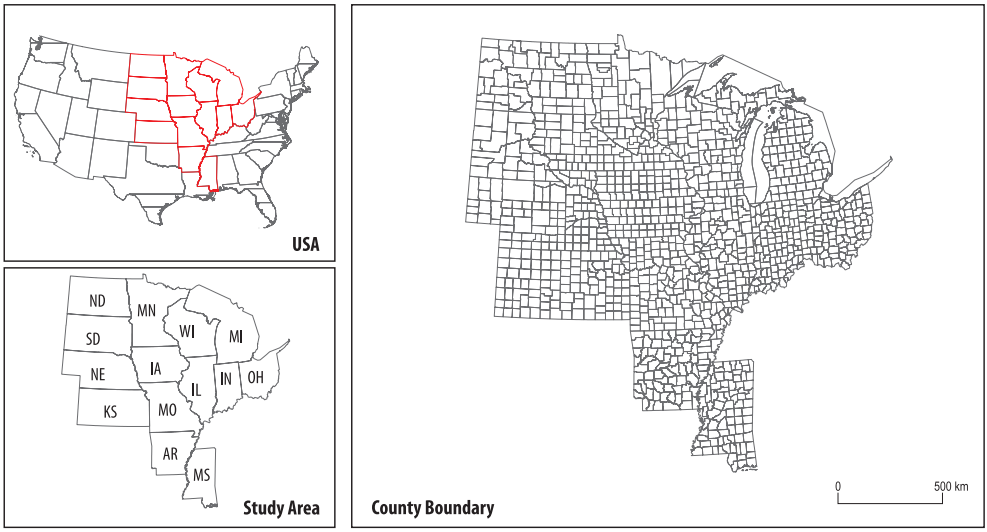

Fig. 1. Map of the study area. For state codes see the text.

and soybean in the US. Corn and soybean are often grown in rotation. The study area is primarily rainfed, and only a very small part is irrigated.

Crop yield data

County-level soybean yield data were obtained from the United States Department of Agriculture (USDA) for 2012–2021 (USDA/NASS 2021). The data were originally in bushels per acre, but we converted the values to tons per hectare (t/ha). The crop yield data were used for training, testing, and validating the models considered in this study. Annual soybean yields at a county level were different in the years from 2012 to 2021 *(Figure 2).*

Cropland data layer

The cropland data layer (CDL) is a crop-specific land-cover data layer created for the continental US that is generated each year by using MODIS and ground truth data at a resolution of 30 m (Boryan, C. *et al.* 2011). The CDL was developed by the Geospatial Information Branch, Spatial Analysis Research Section of the Research and Development Division at the National Agricultural Statistics Service (NASS), which is part of the USDA. In this study, we utilized the CDL to create a mask that identifies soybean pixels while excluding other data. We also used the CDL to derive statistics that helped identify counties with zero soybean pixels, which aided in the selection and exclusion of counties for further analysis.

County boundaries

The geographic boundaries of the counties in the study area were derived from the US Census Topologically Integrated Geographic Encoding and Referencing project, which provides accurate and comprehensive spatial data on county boundaries in the US (https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html). The geographic boundaries were used to collect and represent county-level data each year within the specified period.

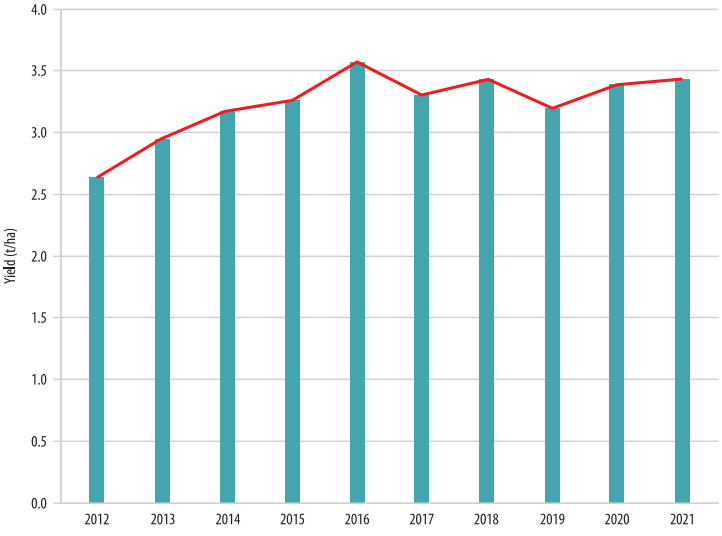

Fig. 2. Average soybean yield at the county level for 2012–2021.

Remote sensing and weather data

We employed MODIS data from the NASA Earth Observing System Data and Information System (https://search.earthdata.nasa.gov/), which are collected at regular intervals of 16 days and are available at spatial resolutions of 250, 500, and 1,000 m. We utilized MODIS NDVI and EVI products from the MOD13Q1.061 Terra Vegetation Indices 16-Day Global 250 m dataset (Didan, K. 2021). The MOD09A1 product

provides a surface reflectance product for Terra MODIS bands 1–7 at a spatial resolution of 500 m and considers atmospheric conditions such as gases, aerosols, and Rayleigh scattering. The reflectance of each pixel is selected from the best value of multiple acquisitions within an 8-day composite based on factors such as high observation coverage, low view angle, no clouds or cloud shadows, and aerosol loading (Vermote, E. 2021). The MOD11A2 product offers the average land surface temperature (LST) over 8 days at a spatial resolution of 1 km. The average LST for each pixel is calculated by taking the simple average of all corresponding LST values collected from the MOD11A1 product within that specific 8-day period (Wan, Z. *et al.* 2021). The MOD11A2 product incorporates both daytime and night time surface temperature bands to account for long-term soil factors (Saravanan, V. and Tamburi, V.N. 2022). Meteorological data were obtained from Daymet (Thornton, M.M. *et al.* 2022) and included daily measurements of six parameters, of which four were selected as climatic factors: precipitation, vapour pressure, minimum temperature, and maximum temperature. These parameters are generated on a gridded surface with a resolution of 1 km × 1 km. *Table 1* presents an overview of the satellite and climatic data utilized in this study.

Data preprocessing

The above datasets were preprocessed and downloaded by using Google Earth Engine (GEE) *(Figure 3).* All datasets were aggregated to a temporal resolution of 16 days corresponding to the period of 14–31 July for consistency. This period was selected because it coincides with the peak vegetation phase of soybeans. We previously investigated the contributions of dynamic input features to the crop yield, including the growth stage and phenology. We found that VI data from mid-July are more significant to forecasting the soybean yield than from other periods. This period corresponds to the crucial pod-setting stage for soybeans. The significance of the dynamic input features implies that the presence of pods and leaves predominantly influences the soybean crop yield (Li, Y. *et al.* 2023). Thus, all data within this period should have a significant relationship with the crop yield.

First, a dataset (e.g., MODIS NDVI) and the CDL were loaded into the GEE environment. The region of interest was selected, and both datasets were clipped to that region. We used the eq() function to create a binary mask based on the CDL values. Pixels with CDL values that corresponded to cropland were assigned a value of 1, while all other pixels were assigned a value of 0. We used the updateMask() function to apply the CDL mask to the MODIS NDVI dataset, which masked out all NDVI values that corresponded to non-cropland areas and left only soybean pixels. We then calculated the mean values of input features for each dataset at the county level. The soybean yield and input features were merged for each year using the county and year as common columns.

Method

Machine learning models

We compared various ML models in an effort to find a relationship between geospatial data and the soybean yield. We selected four different ML models: RF, least absolute shrinkable and selection operator (LASSO) regression, extreme gradient boosting (XGBoost), and decision tree regression (DTR). In addition, we selected a one-dimensional convolutional neural network (1D-CNN) as a DL model. All models were trained, tested, and validated. Models were trained and tested on data for 2012–2016. The models were then validated by forecasting the soybean yield for 2017–2021. The ML models were implemented in Python 3.11.3 with the scikit-learn package (Pedregosa, F. *et al.* 2012). The 1D-CNN model was implemented in Keras

Table 1. Overview of the satellite and climatic data used in this study

Data source	Spatial resolution	Temporal resolution	Feature	Products description
MOD13Q1	250 m	16-day	NDVI	Vegetation index value at a per-pixel basis.
			EVI	
MOD09A1	500 m	8-day	Red	The surface spectral reflectance of Terra MODIS bands 1–7, accounting for atmospheric factors. It includes seven reflectance bands, a quality layer, and four observation bands. Pixel values are chosen based on criteria like extensive coverage, optimal view angle, clear sky conditions, and minimal aerosol.
			Nir	
			Blue	
			Green	
			Nir 1	
			Swir 1	
			Swir 2	
MOD11A2	1,000 m	8-day	LST	The 8-day period aligns with the ground track repeat period of the Terra and Aqua satellites. This product includes day- and night-time LST bands, quality indicator layers, additional MODIS bands, and observation layers.
Daymet	1,000 m	Daily	Precipitation	Daymet V4 is an updated version that offers gridded estimates of daily weather parameters for Continental North America, Hawaii, and Puerto Rico. It addresses known issues by reducing timing bias, enhancing regression models, and introducing a novel approach to handle high elevation temperature measurement biases.
			Water vapour	
			Temperature min.	
			Temperature max.	

(Ketkar, N. 2017), which is an open-source DL framework written in Python.

RF estimates the crop yield by combining multiple regression trees. Each regression tree captures relationships between the input features and the target variable. Subsamples are randomly selected from the training set, which comprises 70 percent recorded yield samples and 30 percent test data. Each subsample is fitted to a regression tree. The final forecast is obtained by averaging the forecasts of all trees. RF has been demonstrated to be effective at mitigating overfitting (Wang, H. *et al.* 2016). We employed the Gaussian kernel function to investigate the non-linear association between input features (i.e., climatic and remote sensing data) and the target variable (i.e., crop yield). We applied RF to identify the most important input features utilizing Python 3.11.3 and the scikit-learn library, which offers Random Forest Regressor classes.

LASSO regression uses linear regression to minimize the residual sum of squares while constraining the absolute values of coefficients below a specified threshold (Tibshirani, R. 1996). LASSO regression addresses overfitting by automatically selecting relevant input features, which leads to a more concise regression model.

XGBoost is designed for tree boosting, which involves constructing multiple weak learners and combining their results to improve the regression or classification performance. It incorporates regularization techniques to prevent overfitting, and the weak learners can be regression trees or linear models (Chen, T. and Guestrin, C. 2016;

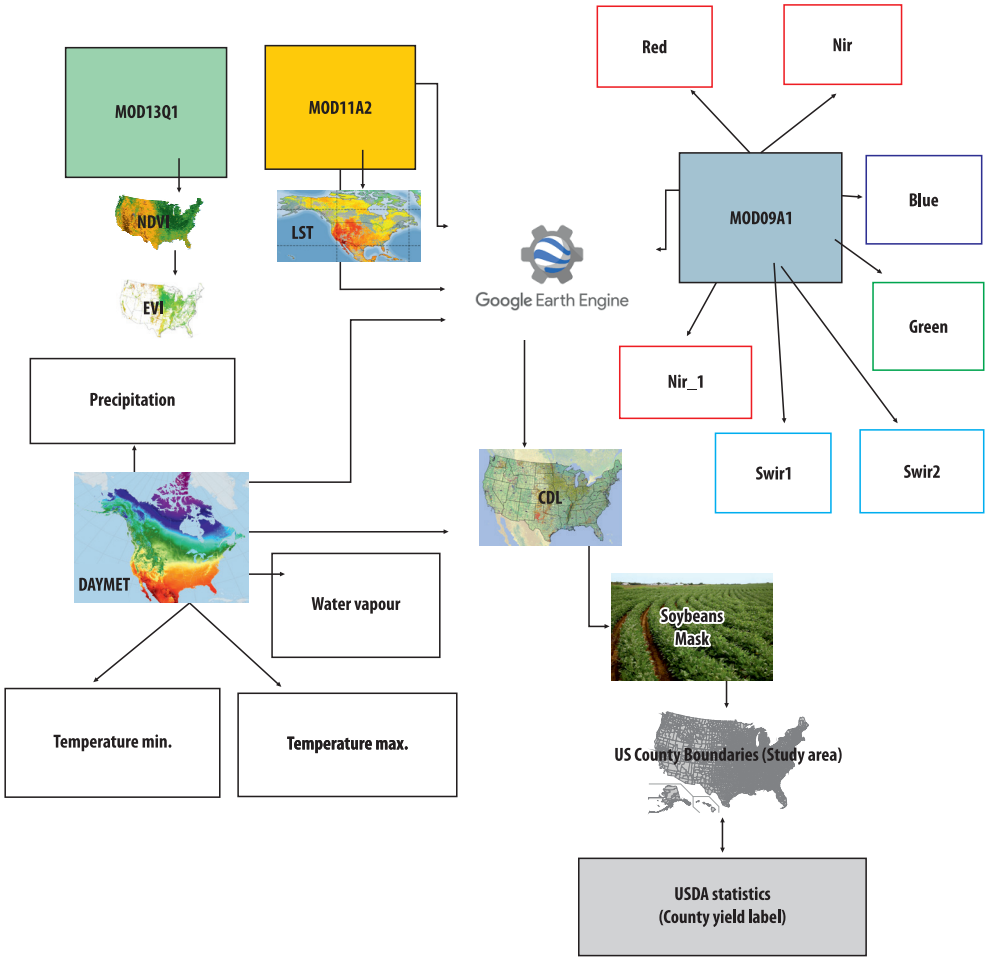

Fig. 3. GEE-based data preprocessing workflow

Song, Y. *et al.* 2019). We opted to use XGBoost based on decision trees. XGBoost makes forecasts by summing the weights of the leaves in all decision trees. We used the GridSearchCV package (Molinaro, A.M. *et al.* 2005) to determine the optimal parameters for XGBoost.

DTR constructs a tree structure based on input features in the training data to forecast the target variable. It is suitable for both classification and regression tasks and offers the benefit of interpretable results in the form of a tree. DTR utilizes binary splits to divide data into two groups and minimizes the sum of squared deviations from the mean within each group. This process is continued until a minimum node size specified by the user is achieved (San Millan-Castillo, R. *et al.* 2020).

Finally, 1D-CNN is a type of neural network commonly used for sequential data analysis (Kiranyaz, S. *et al.* 2021). Different filters are applied to the input data to extract meaningful input features, which allows the model to learn the representation more efficiently *(Figure 4).* The convolutional layers are followed by pooling layers that reduce the dimensions of data and only keep the

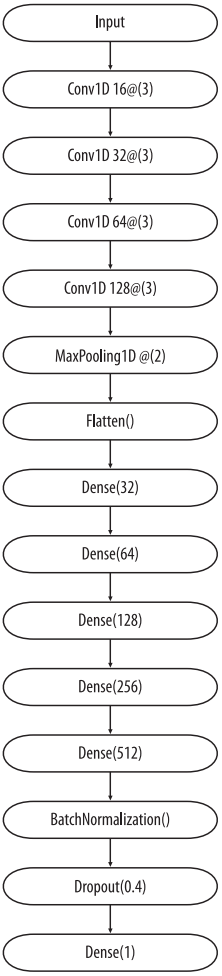

Fig. 4. 1D-CNN model architecture

useful information. This information is then passed to fully connected (FC) layers for the final forecast. We used a 1D-CNN comprising four convolutional layers, one max-pooling layer, and five FC layers followed by a dropout layer and then the final FC layer for crop yield forecasting. We used three convolutional layers with 16–128 filters. The FC layers had 32–512 neurons. The dropout rate was set to 0.4. In Keras, training a DL model requires tuning several hyper-parameters to find the optimal model, which include the learning rate, batch size, and optimization function. We considered learning rates of 0.00001–0.01, batch sizes of 16–128, and the Adam optimizer, which uses stochastic gradient descent to optimize the model performance and reduce the loss.

Model evaluation

We evaluated the model performances by using several metrics. We used the R^2 value to measure the proportion of variance explained by each model. We used the normalized root mean squared error (NRMSE%) and root mean squared error (NRMSE) to measure the forecasting error as a percentage area. We used the mean absolute error (MAE), which represents the average absolute difference between the forecasted and actual values, and the mean squared error (MSE), which measures the average squared difference between the forecasted and actual values. These metrics are calculated as follows:

$$R^2 = 1 - \frac{\sum_i \gamma_i^2}{\sum_i (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(yp^i - y^i)^2} \quad (2)$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|yp^i - y^i\right| \quad (3)$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(yp^i - y^i)^2 \quad (4)$$

$$NRMSE\% = \frac{RMSE}{y_{i,max} - y_{i,min}} \quad (5)$$

These metrics can be used to quantify the probability that a model correctly forecasts new samples from the underlying data distribution . The residuals (r_i) are the differences between the predicted values (yp^i) and actual labels (y^i) as well as the average of the labels ($\bar{y}$).

Results

Input feature importance

We identified surface reflectance (SR) band 2, EVI, and NDVI as the most influential input features with importance scores of 0.30, 0.29, and 0.13, respectively *(Figure 5).* These input features exhibited strong predictive power, which indicates their significance to the crop yield. The LST was found to have moderate importance with a score of 0.08. Conversely, climatic data and other SR bands had relatively low importance scores of 0.01–0.05. These findings emphasize the potential of incorporating MODIS LST and SR data with climatic data to enhance the accuracy of crop yield forecasting models.

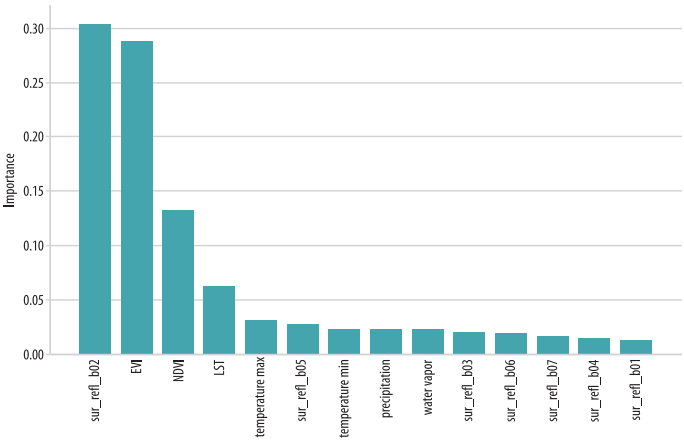

Fig. 5. Importance of input features according to the random forest model

1D-CNN performance

The 1D-CNN model was also applied to forecasting the soybean yield in the study area. *Table 2* presents the forecasting results for 2017–2021. Overall, the 1D-CNN model performed better than classical ML models with MSE values of 0.076–0.108 t/ha and R^2 values of 0.76–0.86. The highest R^2 value was in 2021 while the lowest R^2 value was in 2019. The scatterplots of the actual and forecasted yields were created for 2017–2021 *(Figure 8).*

Performance comparison of machine learning models

The RF model consistently demonstrated the best performance across all metrics *(Figure 6).* The RF model had the highest R^2 value of 0.75, which surpassed the R^2 values of XGBoost (0.70), DTR (0.59), and LASSO regression (0.60). The RF model also had the lowest RMSE of 0.342 t/ha, MAE of 0.264 t/ha, MSE of 0.117 t/ha, and NRMSE of 8.4 percent. XGBoost had the second-best performance with an RMSE of 0.373 t/ha, MAE of 0.287 t/ha, MSE of 0.139 t/ha, and NRMSE of 9.15 percent. Therefore, the RF model was selected for forecasting the soybean yield.

Crop yield forecasting

The year-to-year results indicated that the RF model achieved MAE values of 0.112–0.119 t/ha, RMSE values of 0.334–0.345 t/ha, R^2 values of 0.75–0.77, and NRMSE% values of 7.85–8.46 percent *(Figure 7).* The best results were obtained in 2021, for which the RF model achieved a MAE of 0.112 t/ha, RMSE of 0.334 t/ha, R^2 of 0.77, and NRMSE% of 7.85 percent.

Spatial patterns of crop yield forecasts

Spatial distribution of the forecasted soybean yields was generated over the period of 2017–2021 with the RF model *(Figure 9).* The soybean yield was forecasted as exceptionally high in the central and northeast (i.e., Iowa, Indiana, and Nebraska) at 1–5 t/ha while the observed yield was 1.0–4.5 t/ha. The error between the observed and forecasted yields was minimal, which indicated a high level of accuracy. The south (i.e., Arkansas and Mississippi) was forecasted with an average soybean yield of 2.0–3.5 t/ha, while Ohio in the west was forecasted with a soybean yield of 2.5–3.5 t/ha. The north (i.e., Minnesota and North Dakota) were forecasted with lower yields of 1.5–2.5 t/ha. The difference between the observed and forecast yields was not substantial. However, these results do suggest that the growing conditions in the north were less favourable for soybean cultivation than elsewhere in the study area. The variations in yield across different regions can be attributed to a multitude of factors, including climate, soil quality, agricultural practices, and other local conditions. The central and

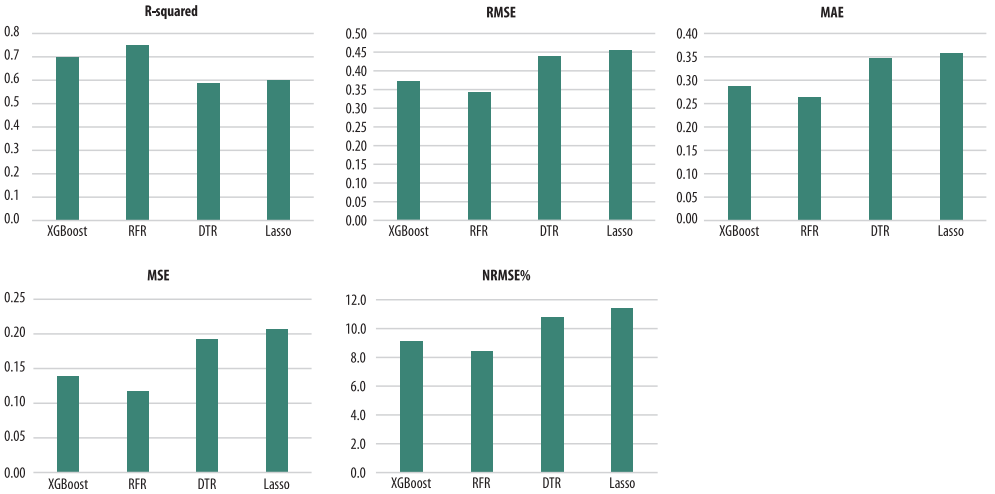

Fig. 6. Performance metrics of several ML models

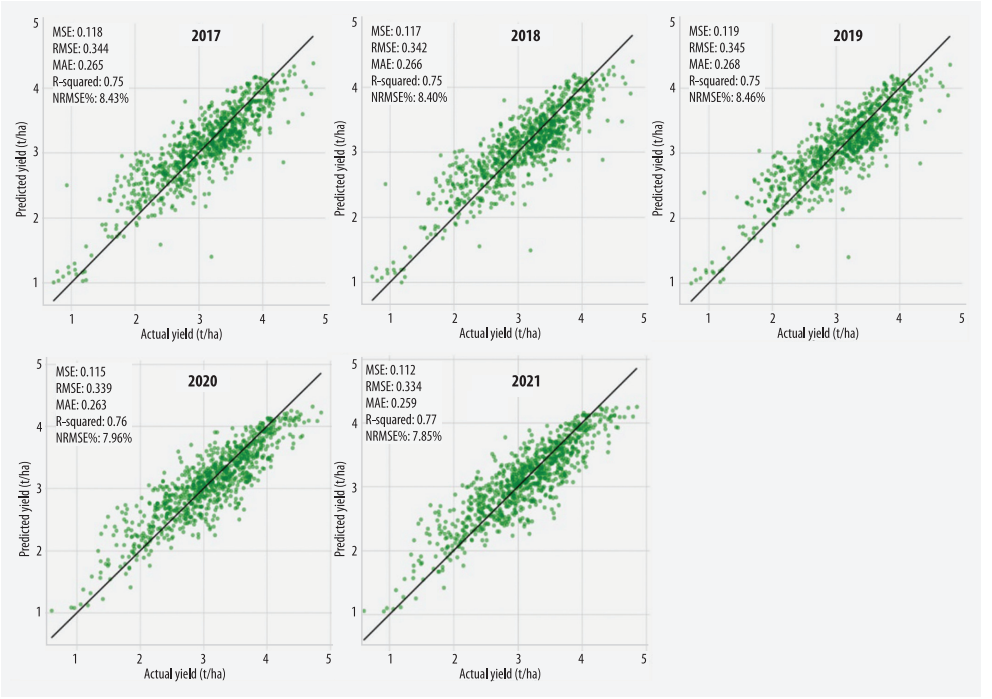

Fig. 7. Scatterplots between actual and forecasted soybean yields for 2017–2021 with the RF model

Table 2. Performance of the 1D-CNN soybean yield forecast model for 2017–2021.

Year	MSE	RMSE	MAE	R^2	NRMSE
	t/ha				%
2017	0.076	0.276	0.202	0.81	6.99
2018	0.079	0.280	0.208	0.83	6.96
2019	0.080	0.283	0.206	0.76	9.26
2020	0.090	0.300	0.222	0.80	6.74
2021	0.108	0.329	0.240	0.86	7.24

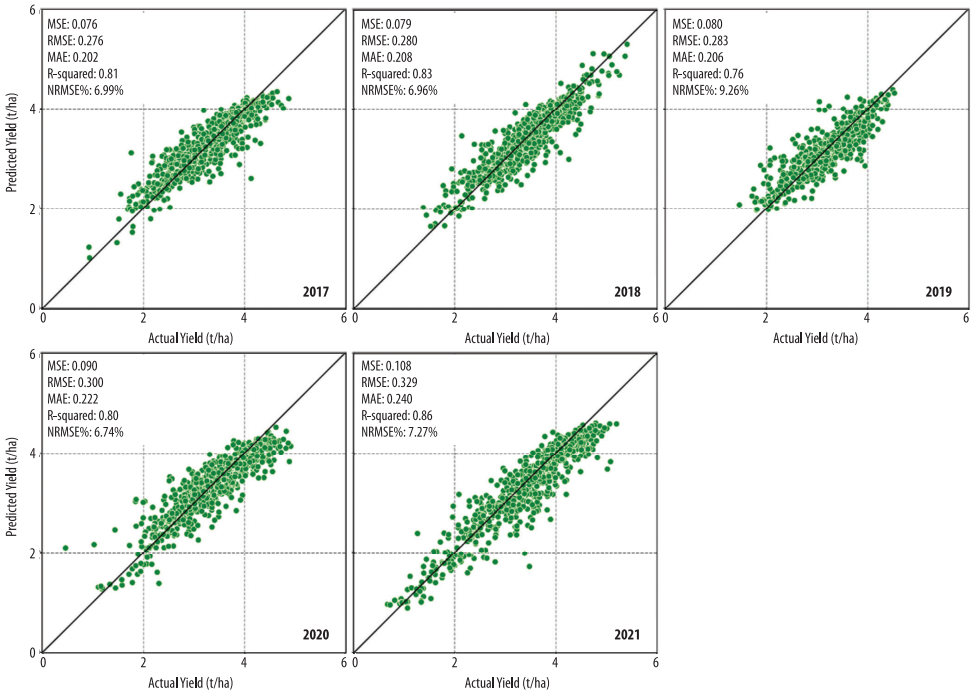

Fig. 8. Scatterplots between the actual and forecasted soybean yields for 2017–2021 with the 1D-CNN model

north-eastern states benefit from a highly conducive agricultural environment, which led to the forecast of higher yields. Conversely, the southern and northern states have less optimal conditions, which lowered the forecasted yields. Note that these forecasts were based on the RF model's analysis of historical data and other relevant factors. However, local factors, unforeseen events, and changes in agricultural practices can potentially affect the actual yield. Therefore, continuous monitoring and adjustment of the model's output to consider real-time data are essential for accurate and up-to-date forecasting.

Discussion

Input feature importance

The RF model found the SR band 2, EVI, and NDVI as the most influential input features

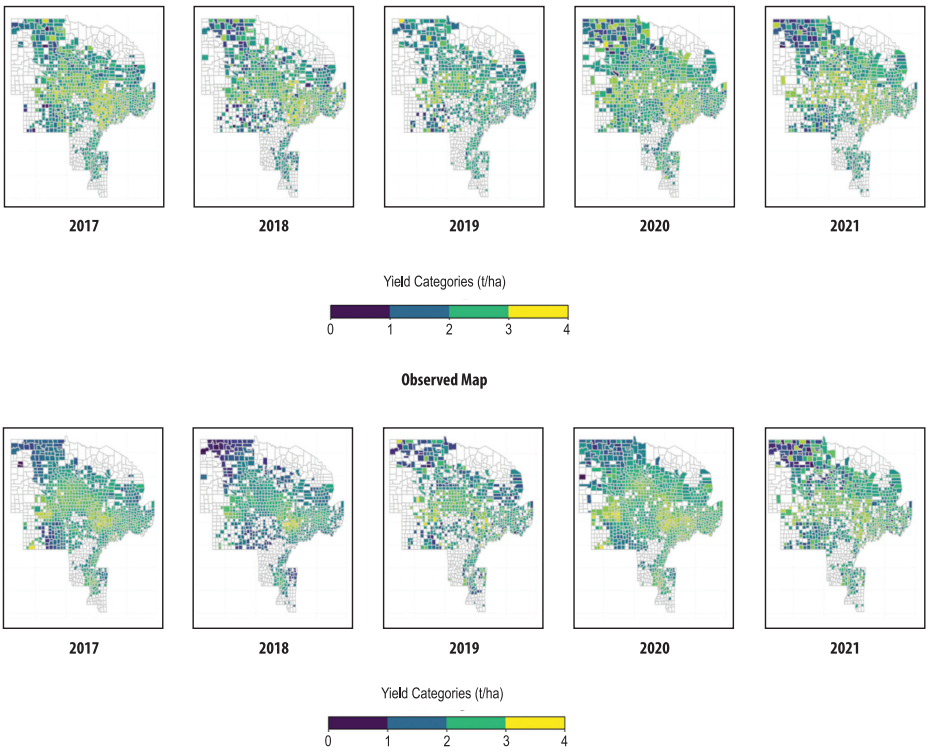

Fig. 9. Spatial distributions of the forecasted soybean yield with the RF model and the observed soybean yield for 2017–2021.

for forecasting the soybean yield in the study area, which is consistent with previous studies that have highlighted the importance of VI data on crop yield forecasting. Sun, J. *et al.* (2019) demonstrated a strong correlation between MODIS SR data and soybean yield. Kuwata, K. and Shibasaki, R. (2016) found that EVI exhibited a significant relationship with corn yield. Our study reinforces the existing literature by highlighting the significance of these VIs in different crop yield forecasting models. The LST was found to have moderate importance for forecasting the soybean yield. This aligns with previous studies that have emphasized the effect of temperature on crop development and productivity (Pede, T. *et al*. 2019; Mirhoseini, N. *et al.* 2022). Pede, T. *et al*. (2019) demonstrated the importance of MODIS LST to corn yield forecasting before the harvesting stage, especially during periods of extreme drought and water stress. Thus, our study reaffirmed the importance of considering temperature in crop yield forecasting. The climatic data and other SR bands were assigned a lower importance for soybean yield forecasting. This is consistent with previous studies that have shown mixed results regarding the contribution of climatic variables to crop yield forecasting (Cai, Y. *et al.* 2019; Hunt, M.L. *et al.* 2019; Farmonov, N. *et al.* 2022). Hunt, M.L. *et al.* (2019) found that climatic variables had limited predictive power of the wheat yield, but Cai, Y. *et al.* (2019) reported that climatic factors had a stronger influence on the wheat yield in Australia. These discrepancies may arise due to

variations in the crop type, geographic location, and specific weather variables considered. Nevertheless, our study highlights the potential of combining MODIS LST and SR data with climatic data to improve the accuracy of crop yiefigld forecasting models.

Model performance

We compared the performances of four ML models for soybean yield forecasting and found that the RF model consistently outperformed the others across all metrics with an RMSE value of 0.334 t/ha and R^2 value of 0.77. These results are consistent with those of Ji, Z. *et al.* (2022) who used an RF model to predict the corn and soybean yields in the Corn Belt with R^2 values of 0.55–0.75 and RMSE values of 1,000–1,500 kg/ha. Our findings align with previous studies that have reported the effectiveness of the RF model at crop yield prediction (Barbosa dos Santos, V. *et al.* 2022; Dhillon, M.S. *et al.* 2023). Barbosa dos Santos, V. *et al.* (2022) found that the RF model provided the most accurate soybean yield prediction in the Brazilian Cerrado with an R^2 value of 0.81 and RMSE value of 176.93 kg/ha, whereas Dhillon, M.S. *et al.* (2023) successfully applied the RF model to predicting the winter wheat and rapeseed yields in Germany. The robustness and flexibility of RF, combined with its ability to handle complex interactions and nonlinear relationships, make it a reliable choice for crop yield forecasting.

However, the comparison between the RF and 1D-CNN models revealed that the latter generally outperformed the former with an R^2 value of 0.86 in 2021 and RMSE value of 0.276 t/ha for 2017. This finding is consistent with other studies that have highlighted the effectiveness of DL models at crop yield prediction (Khaki, S. and Wang, L. 2019; Khaki, S. *et al.* 2020). Khaki, S. and Wang, L. (2019) employed a deep neural network for crop yield prediction and achieved better accuracy than with traditional ML models. Khaki, S. *et al.* (2020) developed a DL-based model to predict corn and soybean yields across the Corn Belt and demonstrated its superior performance. Our study adds to the growing body of literature supporting the potential of DL models for improving the accuracy of crop yield forecasting.

Conclusions

We explored the various contributions of dynamic input features to soybean yield forecasting and compared the performances of different ML models. The findings emphasized the importance of VI data from mid-July, particularly SR band 2, EVI, and NDVI, to the soybean yield, which align with previous studies. The RF model consistently outperformed the other ML models in forecasting the soybean yield. However, the 1D-CNN model performed even better, which highlights the potential of DL models for crop yield forecasting. The spatial patterns of the forecasted yields indicated higher yields in the central and north-eastern states and lower yields in the southern and northern states. These variations can be attributed to multiple factors, including the climate, soil quality, and local agricultural practices.

Overall, this study provides insights into the importance of different input features to crop yield forecasting and the performances of ML and DL models. These findings can help researchers, practitioners, and policymakers in making informed decisions to enhance crop productivity and ensure food security. Future research can focus on integrating additional variables and exploring advanced DL techniques to further improve the accuracy of crop yield forecasting.

Disclosure statement: The authors report there are no competing interests to declare.

Authors' ORCID:
Farmonov, N. – https://orcid.org/0000-0002-2491-9340
Amankulova, K. – https://orcid.org/0000-0001-6562-5616
Khan, S.N. – https://orcid.org/0000-0003-2185-7276
Szatmári, J. – https://orcid.org/0000-0002-7896-3363
Mucsi, L. – https://orcid.org/0000-0002-5807-3742

REFERENCES

Barbosa dos Santos, V., Moreno Ferreira dos Santos, A., da Silva Cabral de Moraes, J.R., de Oliveira Vieira, I.C. and de Souza Rolim, G. 2022. Machine learning algorithms for soybean yield forecasting in the Brazilian Cerrado. *Journal of the Science of Food and Agriculture* 102. (9): 3665–3672. Available at https://doi.org/10.1002/jsfa.11713

Boryan, C., Yang, Z., Mueller, R. and Craig, M. 2011. Monitoring US agriculture: The US Department of Agriculture, National Agricultural Statistics Service. Cropland data layer program. *Geocarto International* 26. (5): 341–358. Available at https://doi.org/10.1080/10106049.2011.562309

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J. Xu, T. *et al.* 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology* 274. (August): 144–159. Available at https://doi.org/10.1016/j.agrformet.2019.03.010

Chen, T. and Guestrin, C. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco CA, USA, ACM. 785–794. Available at https://doi.org/10.1145/2939672.2939785

Dhillon, M.S., Dahms, T., Kuebert-Flock, C., Rummler, T., Arnault, J., Steffan-Dewenter, I. and Ullmann, T. 2023. Integrating random forest and crop modelling improves the crop yield prediction of winter wheat and oil seed rape. *Frontiers in Remote Sensing* 3. (January): 1010978. Available at https://doi.org/10.3389/frsen.2022.1010978

Didan, K. 2021. *MODIS/terra vegetation indices 16-day L3 global 250 m SIN grid V061.* NASA EOSDIS Land Processes DAAC. Available at https://doi.org/10.5067/MODIS/MOD13Q1.061

Farmonov, N., Amankulova, K., Szatmári, J., Urinov, J., Narmanov, Z., Nosirov, J. and Mucsi, L. 2023. Combining planet scope and Sentinel-2 images with environmental data for improved wheat yield estimation. *International Journal of Digital Earth* 16. (1): 847–867.Available at https://doi.org/10.1080/17538947.2023.2186505

Fernandes, J.L., Ebecken, N.F. and dalla Mora Esquerdo, J.C. 2017. Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. *International Journal of Remote Sensing* 38. (16): 4631–4644. Available at https://doi.org/10.1080/01431161.2017.1325531

Green, T.R., Kipka, H., David, O. and McMaster, G.S. 2018. Where is the USA Corn Belt, and how is it changing? *Science of The Total Environment* 618. (March): 1613–1618. Available at https://doi.org/10.1016/j.scitotenv.2017.09.325

Hunt, M.L., Blackburn, G.A., Carrasco, L., Redhead, J.W. and Rowland, C.S. 2019. High resolution wheat yield mapping using Sentinel-2. *Remote Sensing of Environment* 233. (November): 111410. Available at https://doi.org/10.1016/j.rse.2019.111410

Ji, Z., Pan, Y., Zhu, X., Zhang, D. and Wang, J. 2022. A generalized model to predict large-scale crop yields integrating satellite-based vegetation index time series and phenology metrics. *Ecological Indicators* 137. (April): 108759. Available at https://doi.org/10.1016/j.ecolind.2022.108759

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J. and Ritchie, J.T. 2003. The DSSAT Cropping System Model. *European Journal of Agronomy* 18. (3): 235–265. Available at https://doi.org/10.1016/S1161-0301(02)00107-7

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C. and Anderson, M. 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environmental Research Letters* 15. (6): 064005. Available at https://doi.org/10.1088/1748-9326/ab7df9

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D.N., Huth, I. *et al.* 2003. An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy* 18. (3): 267–288. Available at https://doi.org/10.1016/S1161-0301(02)00108-9

Ketkar, N. 2017. Introduction to Keras. In *Deep Learning with Python*. Berkeley, CA, USA. Apress, 97–111. Available at https://doi.org/10.1007/978-1-4842-2766-4_7

Khaki, S. and Wang, L. 2019. Crop yield prediction using deep neural networks. *Frontiers in Plant Science* 10. (May): 621. Available at https://doi.org/10.3389/fpls.2019.00621

Khaki, S., Wang, L. and Archontoulis, S.V. 2020. A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science* 10. (January): 1750. Available at https://doi.org/10.3389/fpls.2019.01750

Khan, K., Iqbal, J., Ali, A. and Khan, S.N. 2020. Assessment of Sentinel-2-derived vegetation indices for the estimation of above-ground biomass/carbon stock, temporal deforestation and carbon emission estimation in the moist temperate forests of Pakistan. *Applied Ecology and Environmental Research* 18. (1): 783–815. Available at https://doi.org/10.15666/aeer/1801_783815

Khan, S.N., Li, D. and Maimaitijiang, M. 2022. A geographically weighted random forest approach to predict corn yield in the US Corn Belt. *Remote Sensing* 14. (12): 2843. Available at https://doi.org/10.3390/rs14122843

Khan, S.N., Khan, A.N., Tariq, A., Lu, L., Malik, N.A., Umair, M., Hatamleh, W.A. and Zawaideh, F.H. 2023. County-level corn yield prediction using supervised machine learning. *European Journal of Remote Sensing* 56. (1): 2253985. Available at https://doi.org/10.1080/22797254.2023.2253985

Khosla, E., Dharavath, R. and Priya, R. 2020. Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. *Environment, Development and Sustainability* 22. (6): 5687–5708. Available at https://doi.org/10.1007/s10668-019-00445-x

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M. and Inman, D.J. 2021. 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing* 151. (April): 107398. Available at https://doi.org/10.1016/j.ymssp.2020.107398

Kuwata, K. and Shibasaki, R. 2016. Estimating cord yield in the United States with MODIS EVI and machine learning methods. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* III–8. (June): 131–136. Available at https://doi.org/10.5194/isprsannals-III-8-131-2016

Li, Y., Zeng, H., Zhang, M., Wu, B., Zhao, Y., Yao, X., Cheng, T., Qin, X. and Wu, F. 2023. A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering. *International Journal of Applied Earth Observation and Geoinformation* 118. (April): 103269. Available at https://doi.org/10.1016/j.jag.2023.103269

Liakos, K., Busato, P., Moshou, D., Pearson, S. and Bochtis, D. 2018. Machine learning in agriculture: A review. *Sensors* 18. (8): 2674. Available at https://doi.org/10.3390/s18082674

Ma, Y., Zhang, Z., Kang, Y. and Özdoğan, M. 2021. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sensing of Environment* 259. (June): 112408. Available at https://doi.org/10.1016/j.rse.2021.112408

Mirhoseini, N., Mahdi, S., Abbasi-Moghadam, D., Sharifi, A., Farmonov, N., Amankulova, K. and Mucsi, L. 2022. Multi-spectral crop yield prediction using 3D-convolutional neural networks and attention convolutional LSTM approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 1–14. Available at https://doi.org/10.1109/JSTARS.2022.3223423

Molinaro, A.M., Simon, R. and Pfeiffer, R.M. 2005. Prediction error estimation: A comparison of resampling methods. *Bioinformatics* 21. (15): 3301–3307. Available at https://doi.org/10.1093/bioinformatics/bti499

Panda, S.S., Ames, D.P. and Panigrahi, S. 2010. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing* 2. (3): 673–696. Available at https://doi.org/10.3390/rs2030673

Paudel, D., Boogaard, H. de Wit, A., Janssen, S., Osinga, S., Pylianidis, C. and Athanasiadis, I.N. 2021. Machine learning for large-scale crop yield forecasting. *Agricultural Systems* 187. 103016. Available at https://doi.org/10.1016/j.agsy.2020.103016

Pede, T., Mountrakis, G. and Shaw, S.B. 2019. Improving corn yield prediction across the US Corn Belt by replacing air temperature with daily MODIS land surface temperature. *Agricultural and Forest Meteorology* 276–277. (October): 107615. Available at https://doi.org/10.1016/j.agrformet.2019.107615

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M. *et al.* 2012. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830. Available at https://doi.org/10.48550/ARXIV.1201.0490

Piekutowska, M., Niedbała, G., Piskier, T., Lenartowicz, T., Pilarski, K., Wojciechowski, T., Pilarska, A.A. and Czechowska-Kosacka, A. 2021. The application of multiple linear regression and artificial neural network models for yield prediction of very early potato cultivars before harvest. *Agronomy* 11. (5): 10.3390. Available at https://doi.org/10.3390/agronomy11050885

San Millan-Castillo, R., Morgado, E. and Goya-Esteban, R. 2020. On the use of decision tree regression for predicting vibration frequency response of handheld probes. *IEEE Sensors Journal* 20. (8): 4120–4130. Available at https://doi.org/10.1109/JSEN.2019.2962497

Saravanan, V. and Tamburi, V.N. 2022. Assessment of land surface temperature (LST) using MODIS MOD11A2 thermal satellite images using zero to null pixel averaging method for the Bengaluru urban district. Preprint. In Review. Available at https://doi.org/10.21203/rs.3.rs-1932983/v1

Shahhosseini, M., Hu, G. and Archontoulis, S.V. 2020. Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science* 11. 1120. Available at https://doi.org/10.3389/fpls.2020.01120

Song, Y., Jiao, X., Qiao, Y., Liu, X., Qiang, Y., Liu, Z. and Zhang, L. 2019. Prediction of double-high biochemical indicators based on LightGBM and XGBoost. In *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science*. Wuhan, Hubei Province, China, ACM, 189–193. Available at https://doi.org/10.1145/3349341.3349400

Sun, J., Di, L., Sun, Z., Shen, Y. and Lai, Z. 2019. County-level soybean yield prediction using deep CNN-LSTM model. *Sensors* 19. (20): 4363. Available at https://doi.org/10.3390/s19204363

Tantalaki, N., Souravlas, S. and Roumeliotis, M. 2019. Data-driven decision making in precision agriculture: The rise of Big Data in agricultural systems. *Journal of Agricultural & Food Information* 20. (4): 344–380. Available at https://doi.org/10.1080/10496505.2019.1638264

Thornton, M.M., Shrestha, R., Wei, Y., Thornton, P.E., Kao, S-C. and Wilson, B.E. 2022. *Daymet: Daily surface weather data on a 1-km grid for North America. Version 4 R1.* NetCDF, November, 0 MB. Available at https://doi.org/10.3334/ORNLDAAC/2129

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58. (1): 267–288. Available at https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tripathy, R., Chaudhari, K.N., Bairagi, G.D., Pal, O., Das, R. and Bhattacharya, B.K. 2022. Towards fine-scale yield prediction of three major crops of India using data from multiple satellite. *Journal of the Indian Society of Remote Sensing* 50. (2): 271–284. Available at https://doi.org/10.1007/s12524-021-01361-2

Vermote, E. 2021. *MODIS/terra surface reflectance 8-day L3 global 500 m SIN grid V061*. NASA EOSDIS Land Processes DAAC. Available at https://doi.org/10.5067/MODIS/MOD09A1.061

Wan, Z., Hook, S. and Hulley, G. 2021. *MODIS/terra land surface temperature/emissivity 8-day L3 global 1 km SIN grid V061*. NASA EOSDIS Land Processes DAAC. Available at https://doi.org/10.5067/MODIS/MOD11A2.061

Wang, H., Yang, F. and Luo, Z. 2016. Once measures. *BMC Bioinformatics* 17. 60. Available at https://doi.org/10.1186/s12859-016-0900-5

Zeng, W., Xu, C., Gang, Z., Wu, J. and Huang, J. 2018. Estimation of sunflower seed yield using partial least squares regression and artificial neural network models. *Pedosphere* 28. (5): 764–774. Available at https://doi.org/10.1016/S1002-0160(17)60336-9