Crop Yield Prediction at Gram Panchayat Scale using Deep Learning Framework

[1]**Lokeshwari M**
Ph.D. Scholar
The Graduate School
ICAR- Indian Agricultural Research Institute, New Delhi.

[2]**Girish Kumar Jha**
Head and Principle Scientist
ICAR- Indian Agricultural Statistics Research Institute, New Delhi.

[3]**Sunil Kumar Dubey**
Deputy Director
Mahalanobis National Crop Forecast Centre, New Delhi.

[4]**Rajeev Ranjan Kumar**
Scientist
ICAR- Indian Agricultural Statistics Research Institute, New Delhi.

[5]**P. Venkatesh**
Senior Scientist
ICAR- Indian Agricultural Research Institute, New Delhi.

Abstract- **Crop yield prediction is crucial for assurance of food security, implementation of policies and the evaluation of crop insurance losses from biotic and abiotic stress. This paper aims to explore the strength of spectral vegetation indices, specifically Normalized Difference Vegetation Index (NDVI) derived from Moderate Resolution Imaging Spectroradiometer (MODIS) data accessible through the google earth engine platform for predicting crop yields using deep learning framework. We proposed a long short-term memory neural network model, which captures the temporal dependencies within historically satellite-derived observations and weather patterns. The proposed model is developed for kharif paddy in the Krishna district of Andhra Pradesh state during 2013-2020. The result indicates that, in predicting paddy yield, the proposed model showed considerable superiority over other baseline models such as random forest regression and shallow neural network in terms of root mean square error (88.01 Kg/ha) and R-square value (91.76%). The findings also revealed that NDVI has significant impact on predicting crop yield compared to weather variables. Our study highlights that the proposed deep learning framework offers a simple, scalable, and cost-effective method for reliably predicting paddy yield based on NDVI before harvest. In addition, it is the first attempt to enhance the paddy yield prediction at gram panchayat level in India.**

Keywords: **paddy yield prediction, deep learning, LSTM, NDVI, crop insurance.**

1. Introduction

Crop yield prediction at a lower level helps in understanding farm variability, improving productivity, making informed decisions, and ensuring timely crop insurance settlement at village and gram panchayat levels (Tripathy et al., 2022). Therefore, in order to facilitate the settlement of insurance claims, estimating the crop yield at gram panchayat level before the harvest becomes imperative. This timely prediction enhances the credibility of settlements and ensures efficient processing.

Traditionally, yield prediction relied on historical yield data and farmer experience. However, the past few decades have seen a technological revolution in this field. Two primary approaches have been used for prediction of yield i.e., process-oriented crop growth models and empirical statistical models. Process-oriented models simulate crop growth processes based on environmental conditions and agronomic practices. They require extensive data on soil properties, weather patterns, and crop physiology (Battisti et al., 2017; Huang et al., 2015; Lobell, 2013). Empirical statistical models, on the other hand, use historical yield data and environmental variables to predict future yields, often employing machine learning algorithms for enhanced accuracy. The integration of satellite imagery and remote sensing technology has further transformed yield prediction. These technologies provide comprehensive and timely data on crop health, soil

moisture, and weather conditions. Integrating remote sensing data with machine learning algorithms presents a cost-effective and time-efficient method for predicting crop yield. This approach comes highly recommended for addressing interaction between complex factors in the process.

Cai et al., (2019) highlighted the superior performance of machine-learning methodologies over regression techniques for predicting crop yield. Neural networks, notably, have gained widespread usage within machine learning techniques due to their capacity for effectively capturing complex patterns embedded within data (Ferreira et al., 2019; Johnson et al., 2016; Pantazi et al., 2016). Recent years have witnessed the emergence of various neural network models aimed at yield estimation, including the shallow neural network (SNN), backpropagation neural network (BPNN), convolutional neural network (CNN), and long short-term memory (LSTM). Tian et al., (2021) employed an enhanced BPNN to predict winter wheat yield in the Guanzhong Plain, PR China, concentrating on the influence of normalized difference vegetation index (NDVI) and normalized difference water index (NDWI) at distinct growth stages on wheat yield. However, due to the BPNN's conventional and simplistic architecture among neural networks, its efficacy in resolving nonlinear and complex approximation problems remains relatively limited, consequently yielding lower accuracy in yield estimation ($R^2 = 0.34$).

However, to overcome the above issue, the recent researchers focus on yield prediction using CNN and LSTM, which are popular types of deep neural networks (DNNs). LSTM, a specialized form of recurrent neural network (RNN), stands out for its long memory capabilities, facilitating the retention of information over prolonged periods. Its gating mechanisms and recurrent structure regulate data flow into and out of the cell, allowing it to capture complex, nonlinear relationships. Notably, LSTM's capacity to handle sequential data, owing to its feedback connections, makes it particularly preferred for classifying, processing, and predicting based on time-series data. Its applications span various domains, such as predicting water table depth in agricultural areas (Zhang et al., 2018), sea surface temperatures (Xiao et al., 2019), runoff (Kratzert et al., 2018), and even crop yield (Jiang et al., 2020). Sun et al. (2019) proposed a CNN-LSTM hybrid model leveraging spatial-temporal features, exhibiting promise in predicting soybean and corn yield in the U.S. Corn Belt. Haider et al., (2019) demonstrated LSTM's superiority in wheat yield prediction in Pakistan, surpassing the accuracy of the machine learning model and RNN. The impressive performance of LSTM in crop yield prediction proved that it could capture not only the variation trend of data but also characterize the dependence relationship of time series data. Despite this, the utilization of LSTM in the domain of yield estimation for handling time series data remains relatively uncommon (Maimaitijiang et al., 2020; You et al., 2017).

Considering the limited application of LSTM in crop yield estimation, our study aims to bridge this research gap by developing an LSTM-based deep neural network framework to utilize multiple input features derived from remote sensing and meteorological data across various time steps. This approach aims to enhance the accuracy of yield estimation, specifically at the gram panchayat level in the Krishna district of Andhra Pradesh.

The subsequent sections of this paper are structured as follows: Section 2 delineates the data employed in this study. Section 3 furnishes an in-depth exposition of our LSTM model designed for yield prediction. Section 4 shows the outcomes yielded by our model. Finally, our conclusions are encapsulated in Section 5.

2. Study Area and Data

Study area

Our current study was conducted within the Krishna district of Andhra Pradesh in India, situated in the southern part of the state between the eastern longitudes 80º55' E and 81º56' E, and northern latitudes 16º71' N and 17º53' N. Covering approximately 8,727 square kilometres, this area encompasses 49 mandals housing 980 Gram Panchayats (Gumma, 2011). The climate predominantly tends toward semi-arid conditions, with some sub-humid areas in the eastern district regions. The average yearly precipitation stands at around 800mm (Milesi & Kukunuri, 2022). The cropping patterns revolve around two primary seasons known as kharif and rabi (Gumma, 2011). In this study, we focused solely on the kharif paddy yield within 609 Gram Panchayats due to data availability constraints.

Datasets and Description

This study combines satellite observations (NDVI and NDWI) at each paddy growth stage, reflecting water stress, photosynthesis, and dry matter accumulation, with meteorological data (Table 1).

Meteorological Data: The study used monthly weather variables, including air temperatures, precipitation, relative humidity, vapour pressure, and solar radiation from NASA power and climate research unit for the study period.

Satellite observations: The study uses MODIS surface reflectance data such as NDVI and NDWI from the Terra and Aqua satellites to analyze seasonal variations in vegetation vigour, surface water, and soil moisture. The data is downloaded using Google Earth Engine for every 16-day of growing period.

Paddy yield data: The Gram Panchayat (GP) level yearly crop cutting experiment (CCE) paddy yield data are obtained from Mahalanobis National Crop Forecast Centre (MNCFC), New Delhi. The yield performance dataset contained observed average yield for paddy between 2013 and 2020 across 609 GPs within the Krishna district of Andhra Pradesh state. This time range is determined by the availability of both satellite observation and paddy yield data for all GP.

Table 1: Data description

Category	Variables	Source
Satellite observations	NDVI, NDWI derived from MODIS Bands	MODIS MCD43A4V6, exported from GEE (16-day temporal resolution)
Weather variables/ metrological data	$T_{min}, T_{max},$ precipitation, relative humidity, vapour pressure, solar radiation	https://crudata.uea.ac.uk/cru/data/hrg/cru_ts_4.05/cruts.2103051243.v4.05/ https://power.larc.nasa.gov/data-access-viewer/
Paddy yield data (**CCE**)	Yield (2013-2020)	MNCFC, New Delhi

Note: The data are acquired using the google earth engine cloud computing platform

3. Methodology

An LSTM, a specified form of RNN (recurrent neural network), uses input as a sequential data (Hochreiter and Schmidhuber, 1997). This model follows a sequential structure that emulates the progression of time steps in crop growth modelling. It functions by incorporating memory cells and gates that regulate the flow of information. The memory cells, which are capable of retaining information for long periods, similar to our human memory of past occurrences. The gates, including the forget gate, input gate, and output gate, manage what information is remembered, discarded, or updated within the memory cells. During training, the LSTM learns patterns in data, excelling in sequences by selectively storing relevant information from previous steps and using it to influence predictions at each subsequent step. This method allows the model to understand and process sequences of information effectively. In this study, we constructed a deep neural network framework comprising five layers for predicting paddy yield, illustrated in Figure 1. The model architecture consists of total five layers which includes an input layer, two LSTM layers, a dense layer, and an output layer. The input data is structured as a time series encompassing satellite observations (NDVI and NDWI) recorded at 16-day intervals during the paddy growing season (June to November), along with monthly meteorological data (average precipitation, minimum and maximum temperatures, relative humidity, vapor pressure, and solar radiation). The network's input is defined by three parameters: n_samples, n_time_steps, and n_features. Here, n_samples denote the batch size during training, impacting computational speed, where after several trials, an optimal predictive performance of our model is achieved with a batch size of 48. The n_time_steps encompass three schemes 3, 4, and 5, respectively while n_features is set at 93 (11*10*72). The model's output corresponds to the predicted paddy yield. Prior to feeding the data into the model, all inputs ware normalized using the min-max method. To curb overfitting, a dropout mechanism was implemented with a dropout rate empirically set at 0.5 (Srivastava et al., 2022) for the dense layer inputs. Determining the ideal number of hidden nodes is not universally standardized and typically necessitates experimentation, hence, for this study, the LSTM model's performance is assessed using two LSTM layers comprising 60 and 40 hidden nodes, respectively. To optimize network parameters, we employed the Adam optimizer with a learning rate at 0.001. The dataset is spilted into 80% for training to develop the model and 20% for testing dataset to evaluate the model performance.

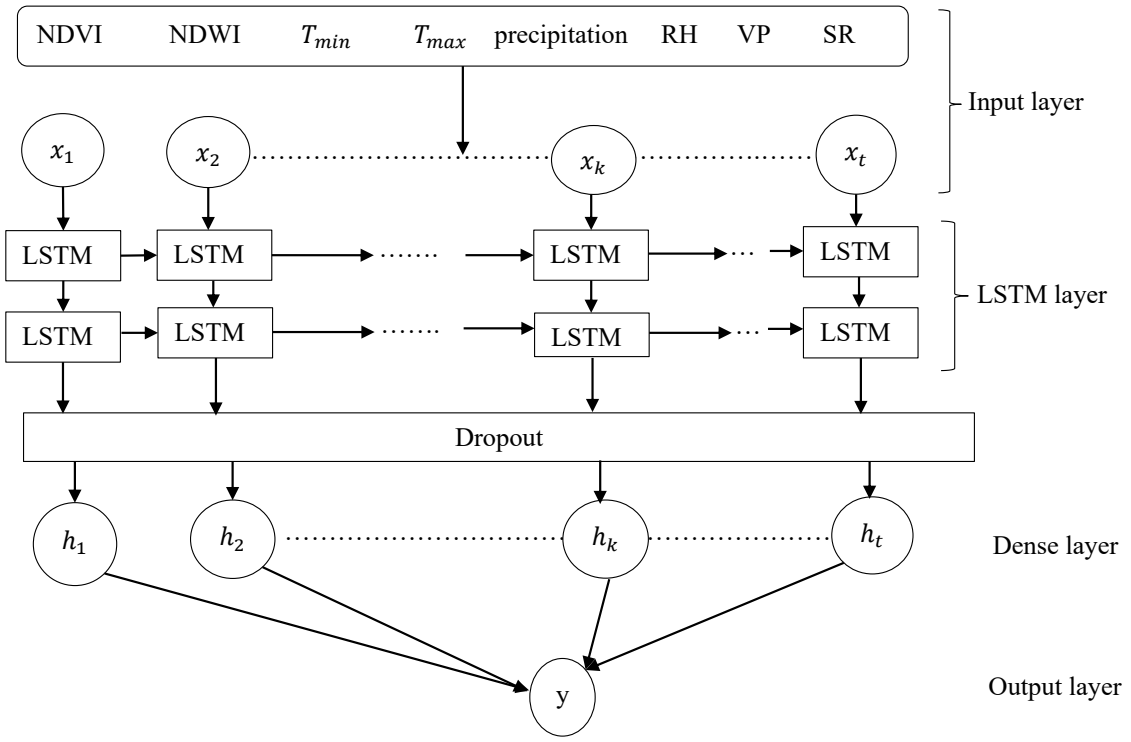

Figure:1 The proposed LSTM architecture for paddy yield prediction

Model Performance Evaluation

The model performance is evaluated using coefficient of determination (R^2), and root mean square error (RMSE). The equations are written as follows:

$$R^2 = \left[\frac{\sum (o_i - \bar{o})(y_i - \bar{y})}{\sqrt{\sum (o_i - \bar{o})^2 (y_i - \bar{y})^2}}\right]^2 * 100$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (o_i - y_i)^2}$$

where, n is the number of samples, and y_i and o_i are the measured and predicted values of paddy yield, respectively. $\bar{o}$ and $\bar{y}$ are the mean of measured and predicted values of paddy yield, respectively.

4. Result and discussion

The proposed model is implemented in Python using the TensorFlow open source library. Training the LSTM neural network model took around 30 minutes, performed on a Tesla K20m GPU. Additionally, for comparison, we implemented other baseline prediction models like random forest regression (RFR) and shallow neural network (SNN) (having a single hidden layer with 200 neurons). We also employed 10- fold cross validation to tune the hyperparameter of predictive models and also to ensure the generalisation power of the proposed model to predict the crop yield. The hyperparameters for the regression tree are configured as follows: a maximum tree depth of 10 to prevent overfitting and a minimum number of 2 samples required to split an internal node within the tree.

In Table 2, the predictive performance of three models for both training and testing datasets is outlined, considering metrics like RMSE and R^2. These results indicated that LSTM neural networks are found to be superior than other baseline models. While RFR showed comparable performance with SNN on the training dataset, it performed worse on the testing dataset. This disparity might be due to its vulnerability to overfitting, particularly when handling numerous features, which could limit its applicability to new data. SNN surpassed RFR across all performance measures as it possesses the capability to handle nonlinearities in data. However, it might lack interpretability and face challenges with sequential data. In sequential tasks, LSTMs tend to outperform shallow networks due to their adeptness in capturing temporal relationships and retaining long-term dependencies.

Table 2: Prediction performance of different models

Models	Training RMSE	Training R^2(%)	Testing RMSE	Testing R^2(%)
RFR	106.87	75.46	208.76	56.76
SNN	98.98	86.69	99.87	85.55
LSTM	88.12	91.76	89.01	88.21

We generated probability density functions for observed yield and the predicted yield through the LSTM model to evaluate whether the model preserves the distributional characteristics of the observed yield. Figure 2a shows that the LSTM model successfully approximated the distributional properties of the actual yield. Nonetheless, the predicted yield exhibited a smaller variance compared to the ground truth yield, indicating that the LSTM model's predictions tended to cluster more around the mean value.

Importance of input variables

To assess the significance of individual input variables in yield prediction, we utilized an LSTM model to capture the nonlinear effects of individual components. Figure 2b shows the yield prediction performance of LSTM model with three different input combinations i.e., NDVI, NDWI, and meteorological data. Our results revealed that yield estimates using NDVI (R^2 = 87.54 %, RMSE = 90.34 kg/ha) alone are more accurate than those using meteorological data (R^2 = 76.87 %, RMSE = 108.06 kg/ha) alone. The findings indicated that NDVI holds a significant influence over paddy yield prediction. This is attributed to its critical role in reflecting the potential for photosynthesis and dry matter accumulation. These aspects are pivotal in assessing crop growth conditions and estimating yield, emphasizing the significance of NDVI as a key variable in the yield prediction process.

5. Conclusion

The study has demonstrated a methodology for paddy yield prediction at gram panchayat scale using LSTM neural network model based on satellite observations and weather variables. The proposed model significantly outperformed other baseline models such as RFR and SNN. The study's findings highlighted the substantial utility of moderate-resolution remote sensing data for more precise yield estimation at the GP level. The proposed LSTM neural networks effectively grasp complex, nonlinear relationships among satellite observations, weather variables, and their interdependencies from historical data, resulting in reasonably precise yield predictions for paddy. These findings hold relevance for insurance settlements under the revised Pradhan Mantri Fasal Bima Yojana (PMFBY) policy, which now operates at the Gram Panchayat level rather than the sub-district, and for farm-level crop management. Yet, for practical implementation, it is crucial to validate the accuracy across greater number of years and different locations. Moreover, further examination is mandatory for exploring this methodology's applicability to other crops. A key area of focus will involve exploring advanced remote sensing techniques or incorporating additional indices to further elevate the accuracy of yield prediction models.

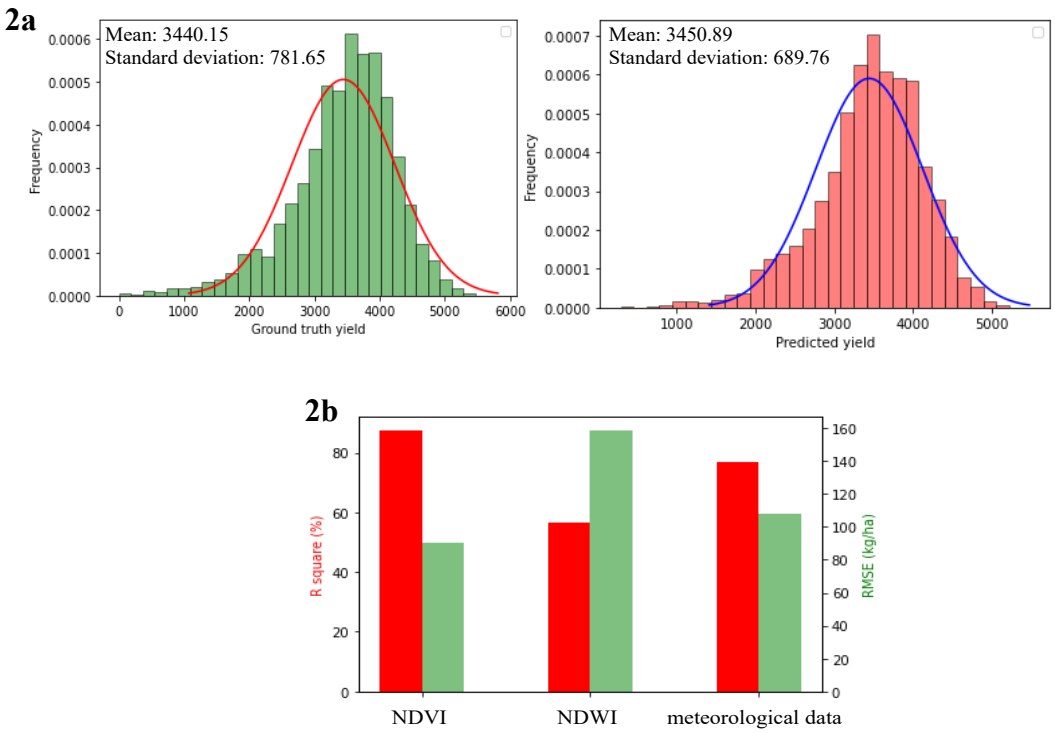

Figure 2aThe probability density functions for both the ground truth yield (observed yield) and the yield predicted by the LSTM model. 2b. Yield prediction performance of LSTM model for individual input variable.

Acknowledge
The first author acknowledges The Graduate School, ICAR-Indian Agricultural Research Institute and ICAR- Indian Agricultural Statistics Research Institute, for providing all the facility throughout her doctoral studies.

REFERENCES:

1. Cai Y, Guan K, Lobell D, Potgieter A B, Wang S, Peng J, Xu T, Asseng S, Zhang Y, You L, and Peng B. 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology* **274**: 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010
2. Ferreira L B, da Cunha F F, de Oliveira R A, and Fernandes Filho E I. 2019. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. *Journal of Hydrology* **572**: 556–570. https://doi.org/10.1016/j.jhydrol.2019.03.028
3. Guan K, Wu J, Kimball J S, Anderson M C, Frolking S, Li B, Hain C R, and Lobell D B. 2017. The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. *Remote Sensing of Environment* **199**: 333–349. https://doi.org/10.1016/j.rse.2017.06.043
4. Gumma M K. 2011. Mapping rice areas of South Asia using MODIS multitemporal data. *Journal of Applied Remote Sensing* **5**(1): 053547. https://doi.org/10.1117/1.3619838
5. Haider S, Naqvi S, Akram T, Umar G, Shahzad A, Sial M, Khaliq S, and Kamran M. 2019. LSTM Neural Network Based Forecasting Model for Wheat Production in Pakistan. *Agronomy* **9**(2): 72. https://doi.org/10.3390/agronomy9020072
6. Hochreiter S, and Schmidhuber J. 1997. Long Short-Term Memory. *Neural Computation* **9**(8): 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735
7. Jiang H, Hu H, Zhong R, Xu J, Xu J, Huang J, Wang S, Ying Y, and Lin T. 2020. A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US Corn Belt at the county level. *Global Change Biology* **26**(3): 1754–1766. https://doi.org/10.1111/gcb.14885
8. Johnson M D, Hsieh W, Cannon A J, Davidson A, and Bédard F. 2016. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology* **218**: 74–84. https://doi.org/10.1016/j.agrformet.2015.11.003

9. Kratzert F, Klotz D, Brenner C, Schulz K, and Herrnegger M. 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences* **22(11)**: 6005–6022. https://doi.org/10.5194/hess-22-6005-2018
10. Maimaitijiang M, Sagan V, Sidike P, Hartling S, Esposito F, and Fritschi 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
11. Milesi C, and Kukunuri M. 2022. Crop Yield Estimation at Gram Panchayat Scale by Integrating Field, Weather and Satellite Data with Crop Simulation Models. *Journal of the Indian Society of Remote Sensing* **50**(2): 239–255. https://doi.org/10.1007/s12524-021-01372-z
12. Pantazi X E, Moshou D, Alexandridis T, Whetton R L, and Mouazen A M. 2016. Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture* **121**: 57–65. https://doi.org/10.1016/j.compag.2015.11.018
13. Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., & Singh, R. (2016). Canopy Temperature and Vegetation Indices from High-Throughput Phenotyping Improve Accuracy of Pedigree and Genomic Selection for Grain Yield in Wheat. *G3 Genes|Genomes|Genetics 6*(9): 2799–2808. https://doi.org/10.1534/g3.116.032888
14. Srivastava A K, Safaei N, Khaki S, Lopez G, Zeng W, Ewert F, Gaiser T, and Rahimi J. 2022. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports* **12**(1). https://doi.org/10.1038/s41598-022-06249-w
15. Sultana S R, Ali A, Ahmad A, Mubeen M, Zia-Ul-Haq M, Ahmad S, Ercisli S, and Jaafar H. Z E. 2014. Normalized Difference Vegetation Index as a Tool for Wheat Yield Estimation: A Case Study from Faisalabad, Pakistan. *The Scientific World Journal* 1–8. https://doi.org/10.1155/2014/725326
16. Tian H, Wang P, Tansey K, Han D, Zhang J, Zhang S, and Li H. 2021. A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the Guanzhong Plain, PR China. *International Journal of Applied Earth Observation and Geoinformation* **102**. https://doi.org/10.1016/j.jag.2021.102375
17. 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. https://doi.org/10.1007/s12524-021-01361-2
18. Xiao C, Chen N, Hu C, Wang K, Gong J, and Chen Z. 2019. Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach. *Remote Sensing of Environment* **233**: 111358. https://doi.org/10.1016/j.rse.2019.111358
19. Yang Q, Shi L, Han J, Zha Y, and Zhu P. 2019. Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. *Field Crops Research* **235**: 142–153. https://doi.org/10.1016/j.fcr.2019.02.022
20. You J, Li X, Low M, Lobell D, and Ermon S. 2017. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:12045912
21. Zhang J, Zhu Y, Zhang X, Ye M, and Yang J. 2018. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *Journal of Hydrology* **561**: 918–929. https://doi.org/10.1016/j.jhydrol.2018.04.065