YUGATO

ISSN: 0387-5695, eISSN: 0387-5695
Vol. 76 No. 1 (2024)

Received: January 13, 2023 / Revised: January 29, 2023 / Accepted: February 27, 2024 / Published: March 12, 2024

FORECASTING SUGARCANE PRODUCTIVITY IN COIMBATORE USING AGRO-METEOROLOGICAL INDICES: A COMPARATIVE STUDY BETWEEN XG BOOST & RECURRENT NEURAL NETWORK APPROACH

S R Krishna Priya[1*] and N Kausalya[2]
Assistant Professor[1*], Research Scholar[2]
Department of Statistics, PSG College of Arts & Science, Coimbatore.

Abstract: At each stage of development, the crop is affected by weather conditions in a different way. Thus, the amount of weather-related factors that affect crop output can be ascertained by considering both the fluctuation of weather during the crop season and their magnitude. Weather-related crop productivity forecasting techniques offer a more accurate productivity prediction by taking into consideration the relative contributions of each weather component. This study used machine learning techniques to examine and forecast sugarcane Productivity in Coimbatore district from 1960 to 2021. Extreme Gradient Boosting (XG Boost), and Recurrent Neural Network models were used to estimate the annual Productivity data of Sugarcane. The models were tested on a validation set of 2017 to 2021 after being trained on data from 1960 to 2016. Accuracy has led to the consolidation of error measures such as Mean Absolute Error, Mean Absolute Percentage Error, Mean Square Error, and Root Mean Square error. With respect to the training set, XG Boost performed the best. Nonetheless, Recurrent Neural Network considerably outperformed XG Boost on the training set, suggesting over fitting problems. Applying cutting-edge machine learning algorithms in a rigorous comparative manner delivers insightful data about future sugarcane productivity supplies. Planning for food security and developing agricultural policies in the area can benefit from the forecasts.

Keywords: Machine Learning. Sugarcane Productivity. XG Boost. Recurrent Neural Network. Long Short Term Memory

1. Introduction

Sugarcane Productivity in Coimbatore, a major agricultural region in Tamil Nadu, India, is influenced by various factors. Coimbatore's climate, characterized by moderate temperatures and adequate rainfall, provides favourable conditions for sugarcane cultivation. Additionally, the region's fertile soils and well-established agricultural practices contribute to its potential for high yields. Farmers in Coimbatore typically grow sugarcane using modern farming techniques, including efficient irrigation systems and improved varieties of sugarcane. However, fluctuations in weather patterns, pests, diseases, and market conditions can affect yield outcomes from season to season. Coimbatore's sugarcane productivity plays a crucial role in the region's economy, providing income for farmers and supporting local sugar mills and related industries. Continuous research and development efforts aim to enhance Sugarcane Productivity in Coimbatore and ensure sustainable agricultural practices for the future. The Sugarcane Productivity in Coimbatore, like in any agricultural area, can vary depending on several factors such as climate conditions, soil quality, agricultural practices, and water availability. Typically, Coimbatore, located in the state of Tamil Nadu in India, has favourable conditions for sugarcane cultivation.

However, specific yield figures can fluctuate annually and may be influenced by various factors, including weather patterns, disease outbreaks, and changes in farming techniques.

2. Literature Review

A variety of machine learning techniques, including statistical methods, can be used to predict crops. Here are some of the techniques that have already been examined are listed below,

Abdikan et al., 2023 represented a paper for estimating the sunflower crop height using various machine learning algorithms such as SLR, MLR, ANN, XG Boost and CNN. Gopi et al., 2023 used Red fox optimization with ensemble recurrent neural network models for the crop recommendation and yield prediction model. Krishna et al., 2023 constructed and optimised Recurrent Neural Network models for prediction of Fruit Rot Disease incidence in Areca Nut Crop Based on Weather Parameters. Li et al., 2023 proposed a framework coupled with XG Boost and multidimensional feature engineering for the county-level soybean yield prediction. Ge et al., 2022 used XG Boost machine learning model for the Prediction of greenhouse tomato crop evapotranspiration. Oikonomidis et al., 2022 used Hybrid Deep Learning-based models for crop yield prediction. Bali et al., 2021 used Deep learning method on wheat crop yield to predict yield in Punjab region of North India. Patel et al., 2021 employed LSTM-RNN Combined Approach for Crop Yield Prediction On various weather constraints. Prashant et al., 2021 used Deep Learning methodologies for the crop yield prediction in Indian Districts. Shahhosseini et al., 2021 utilized coupling machine learning and crop modelling to improve the crop yield prediction in the US Corn Belt. Choudhary et al., 2020 used machine algorithms for the yield prediction for the smart farms. Khaki et al., 2020 used the Convolutional Neural Network-Recurrent Neural Network framework for the prediction of crop yield. Ravi et al., 2020 employed XG Boost algorithm for Crop yield Prediction. Sharma et al., 2020 represented a paper using deep LSTM model for wheat crop yield prediction. Sivanandhini et al., 2020 used feed forward and recurrent neural network for Crop yield prediction analysis. Zhong et al., 2019 used Deep learning based multi-temporal crop classification.

3. Materials & Methods

3.1 Data Description

The Sugarcane Productivity data and Weather data for this study was obtained from the Season and Crop Report, Directorate of Statistics and Economics, Government of Tamilnadu, India and Agro Climate Research Centre, Tamilnadu Agricultural University, Coimbatore, which provides annual statistics on sugarcane Productivity for Coimbatore. The data covers the period from 1960 to 2021. Data from 1960 to 2016 for a period of 57 years is taken for training set and the remaining data of 5 years from 2017 to 2021 is taken for testing set.

3.2 Architecture of Planned model

The optimal crop output is predicted using machine learning and deep learning approaches in the framework that has been suggested. The suggested model runs an experiment on a crop dataset. The crop is selected taking into account the meteorological factors, and the Crop Productivity. When there are multiple possibilities accessible, deep learning is utilized to determine which crop is the most appropriate, leading to a multitude of successful computations. This method provides reliable crop predictions. As seen in figure 1, the XG Boost algorithm is applied under machine learning, whereas RNN are applied under deep learning techniques.

CROP DATASET
DATA PREPROCESSING
TRAINING SET

XG BOOST ALGORITHM	RECURRENT NEURAL NETWORK
TESTING SET	
XG BOOST ALGORITHM (CHECKING THE MODEL)	RECURRENT NEURAL NETWORK (CHECKING THE MODEL)
COMPARING THE ACCURACY	

Figure 1. Architecture of Planned Model

3.3 Procedures for Execution of the Planned Model

Step 1: Load the crop dataset with several parameters included.
Step 2: Load the necessary packages and libraries.
Step 3: Pre-processing the data is done.
Step 4: The data is split into training and testing sets in order to prepare the dataset.
Step 5: Following this, a model is built using deep learning (Recurrent Neural Network- Long Short Term Memory) and machine learning (Extreme Gradient Boosting – "XG Boost") techniques, which in turn forecasts the ideal crop that should be planted.
Step 6: The test set evaluates the model's performance. The model returns an error with the message "value mismatch or wrong prediction" if any trash values are entered.

3.4 Algorithms

Recurrent Neural Network

One type of artificial neural network specifically made for sequential data processing is called a recurrent neural network (RNN). RNNs have connections that create directed cycles, which allows them to display dynamic temporal behaviour, in contrast to standard feed forward neural networks, which analyse input data independently of each other. The main steps for utilizing RNNs are as follows:

Preparing Data: Sort your data in a progressive fashion into relevant input-output pairs. Text may need to be converted into word or character sequences, or time series data may need to be converted into a sequence of previous observations and future goals.

Architecture Model: Create the RNN's architecture. This includes identifying the kind of recurrent units (e.g., simple RNN, LSTM, GRU) the quantity of recurrent layers, and any other layers (e.g., dense layers) that are incorporated into the network.

Model Gathering: Assemble the RNN model, being sure to include the evaluation metrics, optimization technique, and loss function that will be utilized for training.

Training Models: Use the fit method to train the RNN model on your training set of data. Through back propagation through time (BPTT), the model learns to capture temporal dependencies in the data during training by varying the weights of its connections.

Assessment of the Model: Assess the trained RNN model's performance using suitable evaluation metrics (e.g., accuracy, mean squared error) on a different validation dataset. This stage aids in evaluating the model's ability to generalize to new data.

Adjusting Hyper Parameters: To maximize performance, adjust the model's hyper parameters (learning rate, number of recurrent units, dropout rate, etc.). Techniques like grid search and randomized search can be used for this.

Model Distribution: After you are happy with the model's performance, use it to make inferences on fresh data. This could entail integrating the trained model into a production environment and storing it on disk.
Monitoring and Maintenance: Keep an eye on the performance of the deployed model at all times, and retrain it as fresh data become available. This guarantees that the model will always be current and correct.
Because RNNs can collect and represent temporal connections in data, they are frequently employed in tasks including speech recognition, natural language processing (NLP), sequence prediction, and time series forecasting, among others.

Extreme Gradient Boosting - XG Boost

XG Boost stands for Extreme Gradient Boosting, and it's a popular and powerful machine learning algorithm known for its efficiency and effectiveness in handling structured data. It belongs to the ensemble learning methods, specifically the gradient boosting family.
Boosting Method: It constructs several decision trees one after the other, fixing the mistakes of the first tree. The overall prediction error is reduced by this iterative procedure.
Gradient Boosting: To reduce the loss function, XG Boost applies gradient boosting techniques. By minimizing the mistakes, gradient descent algorithms are used to optimize the performance of the model.
Regularization: To reduce over fitting and enhance generalization, XG Boost integrates L1 (Lasso regression) and L2 (Ridge regression) regularization approaches.
Parallel processing: It has been speed and efficiency optimized. Through parallelizing tree construction across all available CPU cores during training, XG Boost may take use of the computational capabilities of contemporary technology.
Tree Pruning: It uses a technique known as "pruning" to cut out splits that don't do much to lessen the loss. This aids in avoiding over fitting.
Handling Missing Values: You don't need to pre-process missing values before training because XG Boost has built-in capability to manage missing data. All things considered, XG Boost's great performance, adaptability, and simplicity of use make it a popular choice for a variety of machine learning contests and practical uses.

4. Results & Discussion

The study has chosen the following weather variables:

X_1 - Maximum Temperature; X_2 - Minimum Temperature; X_3 - Relative Humidity 7 Hours; X_4 - Relative Humidity 14 Hours; X_5 – Rainfall

4.1 Recurrent Neural Network

For LSTM training set, only the time series data have been taken. Here to predict the current time prediction, previous 5 years productivity has used. Based on the last 5 years productivity the model will predict the current year productivity. The data is pre-processed in window based style for the model to train on previous 5 year data to predict the current year value. Pre-processing has done to the dataset as the initial step and the Parameter Estimates of the model is given in Table 1.

Table 1. Parameter Estimates of Long Short Term Memory

Layer (Type)	Output Shape	Parameters
LSTM	None, 8	320
Dense	None,1	9
Total Parameters: 329		

Trainable Parameters: 329
Non-Trainable Parameters: 0

The above table shows the number of parameters of LSTM model. The model is trained for 200 epochs with Mean Squared Error as loss function, adam with a learning rate of 0.001 as optimizer. At the end of 200 epochs, the train and test loss were 0.4083 and 0.5132. Below are the results for the Long Short Term Memory.

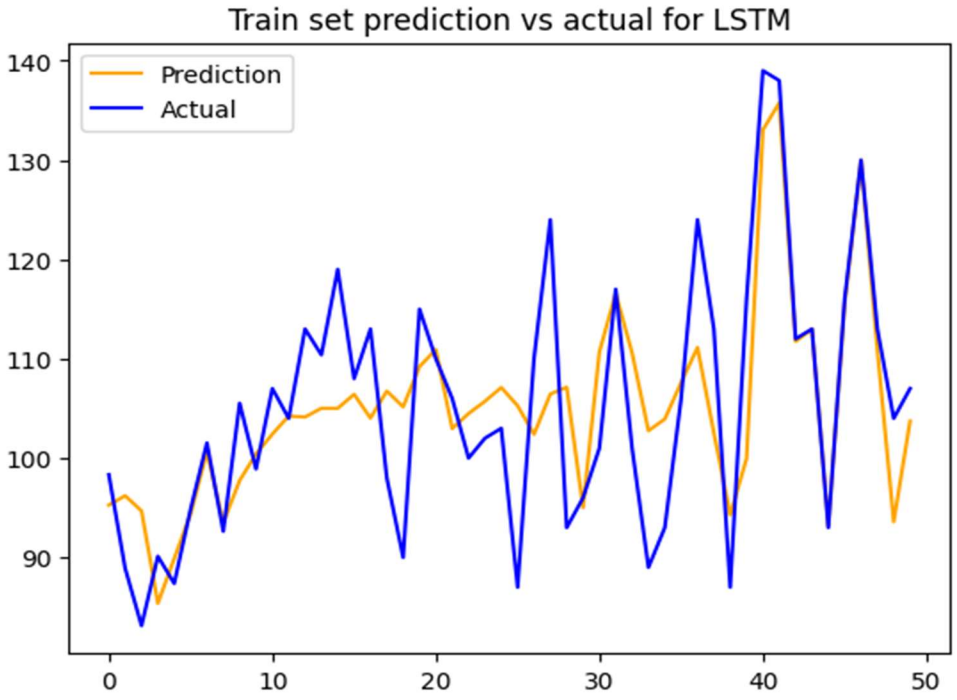

Figure 2. Training Set of Long Short Term Memory

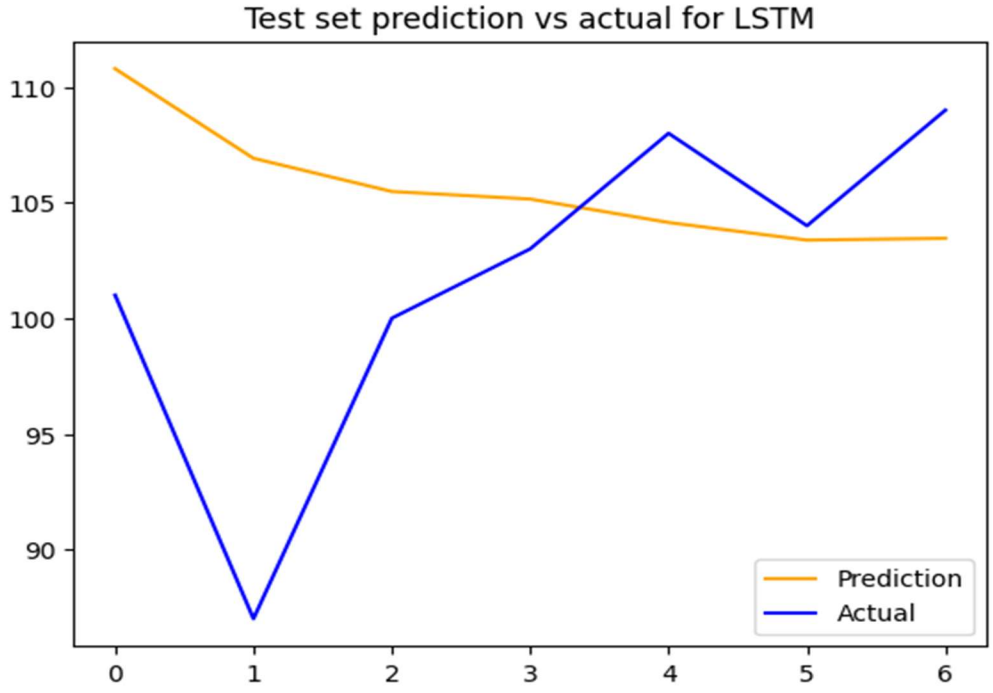

Figure 3. Testing Set of Long Short Term Memory

From the figure 2 & 3 and the metric score, it is observed that the model has performed better, despite a minor over-fit.

Table 2. Error Measures of Long Short Term Memory

Error Metrics	Training Set	Testing Set
R^2 Value	0.600	-0.794
MSE	64.891	81.909
RMSE	8.055	9.050
MAE	6.125	6.767
MAPE	0.059	0.070

4.2 XG Boost

For the XG Boost model, the number of estimators, maximum depth, sub-samples, and column samples by tree are the hyper parameters that are adjusted. Below are the optimal hyper parameter settings and their outcomes. In this case, the train R^2 score is positive despite the negative test R^2 score, and other error levels are significantly lower than in earlier models. This illustrates the model's performance in Figure 4.

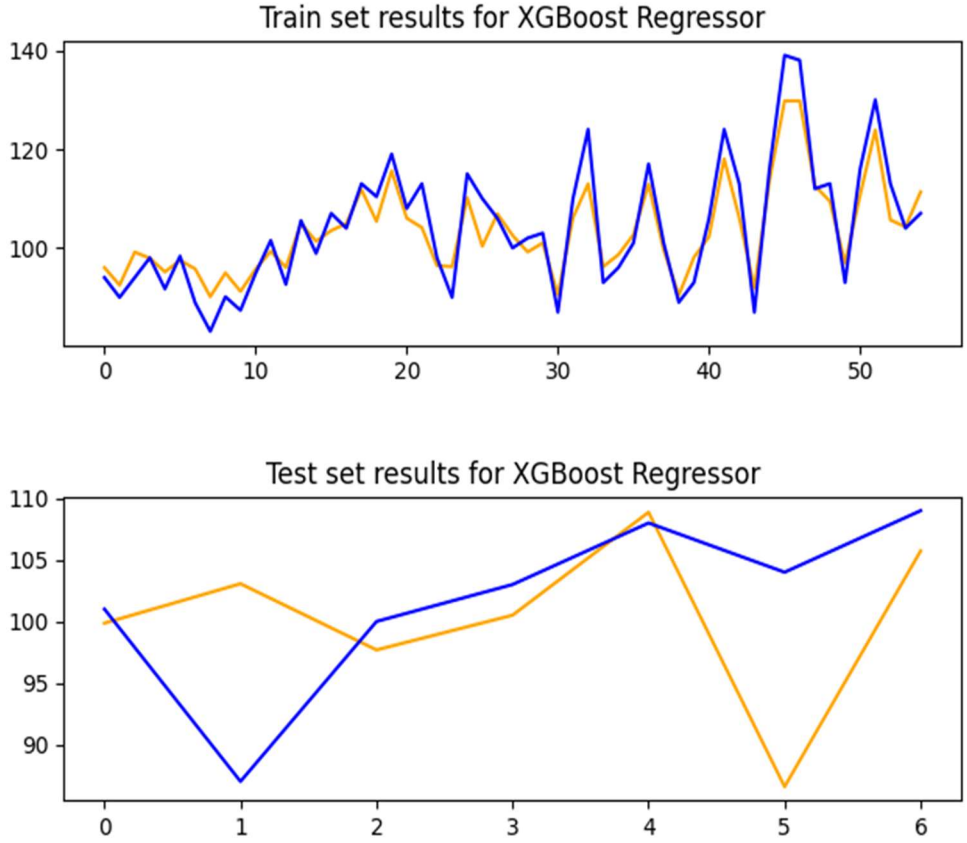

Figure 4. Training Set and Testing Set of XG Booster

From the figure 5, it is shown that Maximum Temperature, Minimum Temperature, Relative Humidity 7 Hours, Relative Humidity 14 Hours, and Rainfall have given more or less equal significant contribution to forecasting.

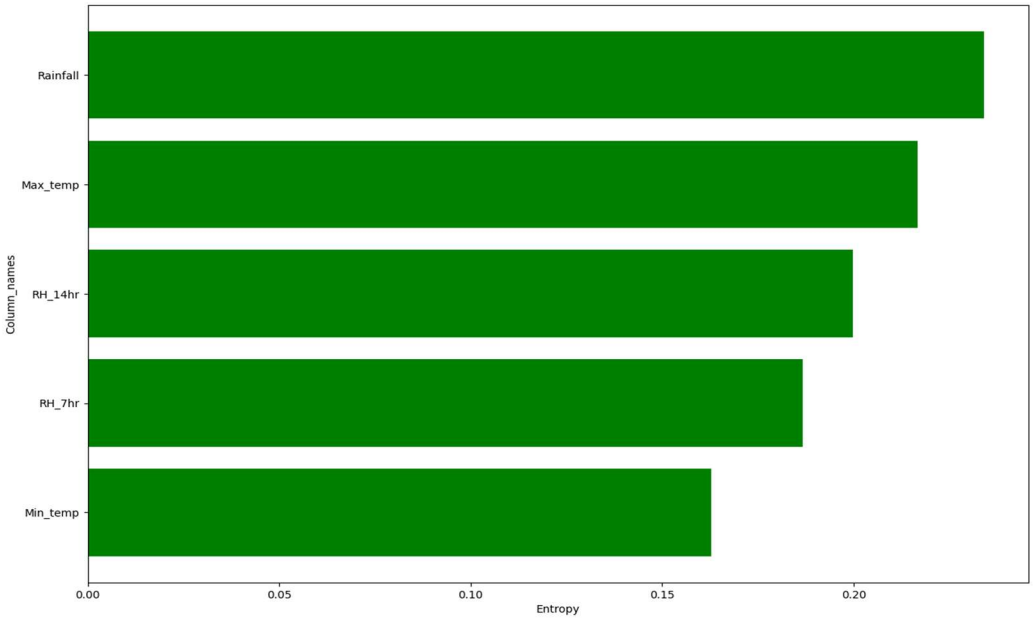

Figure 5. Significance of Weather Parameters in XG Booster

Relative Importance of the predicted Weather Parameters

Table 3. Relative Importance of XG Booster

Parameters	Rank
X_1	2
X_2	5
X_3	4
X_4	3
X_5	1

Rainfall, a weather metric, increased significantly in entropy, as seen in the graph above, improving forecast accuracy.

Error Measures of the XG Booster

Table 4. Error Measures of XG Booster

Error Metrics	Training Set	Testing Set
Depth of Tree	3	3
Estimator	15	15
Sub-Samples	0.3	0.3
Col.Samples Leaf	0.8	0.8
R^2 Value	0.865	-0.834
MSE	21.447	83.731
RMSE	4.631	9.150
MAE	3.858	6.226
MAPE	0.036	0.064

4.3 Comparison of Model Accuracy

Table 5. Comparing the Error Values of both the models

Error Metrics	LSTM	XG Booster
R^2 Value	0.600	0.865
MSE	64.891	21.447
RMSE	8.055	4.631
MAE	6.125	3.858
MAPE	0.059	0.036

For both models, the R^2values showed the least difference. In spite of this, R^2 values of XG Booster is still much lower than Long Short-Term Memory. Consequently, it can be concluded that for the Sugarcane Productivity of Coimbatore district, the XG Booster produces better results than the Long Short Term Memory.

5. Conclusion

The study used machine learning techniques to provide a thorough examination of Coimbatore sugarcane productivity. Data on Coimbatore annual Sugarcane Productivity from 1960 to 2021 was modelled using XG Boost and Recurrent Neural Network models. The appropriate model was found using the coefficient of determination (R^2) and error measures. From the training of models, it is concluded that XG Boost Regressor has done better compared to Recurrent Neural Network model. XG Boost Regressor can be used to make future predictions. The forecasts for Sugarcane Productivity give decision-makers, planners, stakeholders, agriculturists, and law makers essential information. The techniques and results could direct future research on food crop modelling and improve Coimbatore food security. In summary, this research highlights the advantages of employing scientific modelling methods to extract meaningful information from agricultural data.

6. References

Abdikan, S., Sekertekin, A., Narin, O. G., Delen, A., & Sanli, F. B. (2023). A comparative analysis of SLR, MLR, ANN, XGBoost and CNN for crop height estimation of sunflower using Sentinel-1 and Sentinel-2. *Advances in Space Research*, *71*(7), 3045-3059.

Bali, N., & Singla, A. (2021). Deep learning based wheat crop yield prediction model in Punjab region of North India. *Applied Artificial Intelligence*, *35*(15), 1304-1328.

Choudhary, N. K., Chukkapalli, S. S. L., Mittal, S., Gupta, M., Abdelsalam, M., & Joshi, A. (2020, December). Yieldpredict: A crop yield prediction framework for smart farms. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 2340-2349). IEEE.

Ge, J., Zhao, L., Yu, Z., Liu, H., Zhang, L., Gong, X., & Sun, H. (2022). Prediction of greenhouse tomato crop evapotranspiration using XGBoost machine learning model. *Plants*, *11*(15), 1923.

Gopi, P. S. S., & Karthikeyan, M. (2023). Red fox optimization with ensemble recurrent neural network for crop recommendation and yield prediction model. *Multimedia Tools and Applications*, 1-21.

Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, *10*, 1750.

Krishna, R., & Prema, K. V. (2023). Constructing and Optimising RNN Models to Predict Fruit Rot Disease incidence in Areca Nut Crop Based on Weather Parameters. *IEEE Access*.

Li, Y., Zeng, H., Zhang, M., Wu, B., Zhao, Y., Yao, X., & 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*, 103269.

Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Hybrid deep learning-based models for crop yield prediction. *Applied artificial intelligence*, *36*(1), 2031822.

Patel, J., Vala, B., & Saiyad, M. (2021, April). LSTM-RNN Combined Approach for Crop Yield Prediction On Climatic Constraints. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1477-1483). IEEE.

Prashant, P., Ponkshe, K., Garg, C., Pendse, I., & Muley, P. (2021, November). Crop Yield Prediction of Indian Districts Using Deep Learning. In *2021 Sixth International Conference on Image Information Processing (ICIIP)* (Vol. 6, pp. 250-255). IEEE.

Ravi, R., & Baranidharan, B. (2020). Crop yield Prediction using XG Boost algorithm. *Int. J. Recent Technol. Eng*, *8*(5), 3516-3520.

Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific reports*, *11*(1), 1606.

Sharma, S., Rai, S., & Krishnan, N. C. (2020). Wheat crop yield prediction using deep LSTM model. *arXiv preprint arXiv:2011.01498*.

Sivanandhini, P., & Prakash, J. (2020). Crop yield prediction analysis using feed forward and recurrent neural network. *International Journal of Innovative Science and Research Technology*, *5*(5), 1092-1096.

Zhong, L., Hu, L., & Zhou, H. (2019). Deep learning based multi-temporal crop classification. *Remote sensing of environment*, *221*, 430-443.