JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 07, ISSUE 09, 2020

Agricultural crop yield prediction using optimized artificial neural network approaches

Kamred Udham Singh,
Asst. Professor, SOC (School of computing),
GEHU-Dehradun Campus
DOI: 10.48047/jcr.07.09.590

Abstract: Big data in agriculture is an integration of computational methods and statistical analysis. The massive amounts of data generated in agriculture are no match for Big Data. When compared to more conventional approaches, it efficiently gathers and combines unique data for analysis. With the ability to recognize patterns in the data, Big Data may be of service in the agriculture sector. The sheer volume of data involved in processing such satellite photos might be daunting. The vast volumes of data generated during agricultural yield forecast are manageable with the help of Big Data Analysis. The goal of Machine Learning is to create efficient and quick learning algorithms that can anticipate outcomes based on data. In this study, we use two different methods for making predictions: the Multiple Linear Regression (MLR) model and an Artificial Neural Network (ANN). For the purpose of selecting the most effective characteristics of the yield components, multi-level regression (MLR) is a useful technique. The definition of an ANN is the structure of connections between several levels of neurons. The weights of connections are periodically adjusted via a learning process.

Keywords: Big Data, Artificial Neural Network (ANN), Multiple Linear Regression (MLR)

Introduction

The foundation of modern computing is the creation and integration of data at an unprecedented velocity. Content creation rates for the millions of online web apps that service people all over the globe are skyrocketing. There has been a significant data explosion because of the widespread use of the internet for things like socializing, buying, and doing research through Google. Big Data is shorthand for this deluge of information. Web apps are designed to provide services to consumers as quickly and easily as possible. However, as data volumes grow, it becomes more difficult to handle, store, and analyze. Big data analysis is growing in importance in the commercial world. By efficiently analyzing Petabytes of data, we are able to offer our consumers with first-rate services in the competitive business sector. Many factors influence crop yield in agricultural systems, and these factors are often hard to isolate using conventional methods of analysis. The field of machine learning is gaining prominence and has the potential to provide answers to the challenges posed by Big Data. To deal with the problems associated with Big Data, machine learning provides a data analysis strategy that is both modular and scalable. Crop protection refers to the study and practice of preventing agricultural losses due to weeds, insects, and diseases. In this study, we use Big Data and ML to the issue of estimating future harvests.

For controlling complicated interactions with substantial non-linearity among agricultural production parameters, the Artificial Neural Network (ANN) is suggested. Artificial neural networks (ANNs) are a kind of topological machine learning. The most popular Neural Networks (NN) for agricultural production prediction are the Back-Propagation (BP) method and the MultiLayered Perceptron (MLP). A workable topology for the data sample space cannot be determined automatically, though. Therefore, an empirical selection of topology is performed for practicable agricultural yield forecasting. An ANN is used when there are few input characteristics. In this paper, we present a Greedy Search and Tabu Search (TS) hybrid PBIL method for forecasting agricultural production using Big Data. Commercial crop production in agriculture is dependent on a wide range of environmental and economic factors, including but not limited to weeds, pesticides, harvesting, rainfall, temperature, fertilizers, irrigation, cultivation, climate, soil, and so on. Information on past crop output is also very important for the supply chain activities of businesses in the industrial sector. These sectors rely heavily on agricultural outputs such as paper, seed, pesticides, fertilizer, poultry, chemicals, animal feed, cattle, and raw materials. With the use of precise risk and crop production estimates, these businesses can manage supply chain decisions like production scheduling.

Companies in the agricultural equipment, agrochemical, fertilizer, and seed sectors rely on crop output projections as the foundation for marketing and production activity planning. By accounting for differences in crop type reflectance, crop identification and mapping benefits from the use of multi-temporal imaging, which permits classification. The audit of land usage and agricultural planning rely on the crop categorization data for efficient crop growing. Growing crops may be challenging since there are so many distinct ways to do it. The primary goal of agricultural production is to maximize crop output while minimizing input costs. Early identification and the application of agricultural yield indicators for managing difficulties may help boost crop yield and its profit. Changing the weather via the manipulation of meteorological phenomena on a global scale has a profound effect on agricultural output. Crop managers prepare forecasts to mitigate losses in output caused by unfavorable growth conditions. These forecasts are also used to their full potential when environmental conditions are optimal for plant growth. Studying how agriculture is practiced in a given region is made possible by satellite imaging, which shows crop development from planting through harvesting and reveals the erratic and unpredictable patterns of the seasons. A georeferenced, orthorectified picture may help identify trouble areas and estimate their magnitude. The use of satellite photography for monitoring farmland has resulted in a meteoric rise in the amount of data collected. For the recovery of all necessary specimen data sets, the available technologies for large data volume storage and prediction were insufficient. A structural step required to be defined in order to break down both and move towards real-time data. Big Data was proven to be more effective in terms of knowledge acquisition and decision-making. It was assumed that using Big Data will significantly alter any optimal model.

Machine Learning Approaches

computer learning (ML) is a method or idea that enables a computer to be trained without a pre-existing software by using its own history and normal processes. The use of many, potentially linked databases allows ML methods to handle large, non-linear problems without human intervention. Data collection, model development, and model deployment are only a few of the many steps involved in applying ML to real-world situations. With the incorporation of expert ideas into the program, ML is able to generate better decisions and meaningful responses in real-world settings with human intervention. Artificial intelligence methods are broken down into the supervised, unsupervised, and reinforcement learning subfields. The inputs and outputs of supervised learning are training examples, which are used to master a goal. In cases when sources seldom provide output values, unsupervised learning makes an effort to learn the patterns from the input data. The reinforcement learning system uses the current environment to reinforce a control pattern. Many different industries, but agriculture in particular, rely heavily on ML methods that use these algorithms because of their many benefits. According to Whelan and Taylor (2013), ML's primary goal for the cropping system is to offer data that will enable improved space-to-time management choices. In most cases, a large field area will be divided into smaller management zones that all adhere to the same set of treatment guidelines. Traditionally, yield maps and variation maps of agricultural fields have been used for this demarcation. Delineating management zones for ML applications is shown to benefit from the latest advancements in sensing technologies. Delineating management zones is now a feasible strategy in industrial farming because to advancements in data fusion, geostatistical study, and interpolation methods. Despite ML's widespread use and significant recent breakthroughs, it is not without its limitations if applied blindly in a data-driven manner. The quality of the dataset, the model representation, and the link between input and output variables in the gathered dataset all have a significant impact on the accuracy of the predictions and the inconsistencies caused by the ML algorithms. Prejudices in the data, the presence of an outlier, and very noisy data may all reduce the model's prediction power. Expert expertise in model selection, transfer learning, and outlier identification may help overcome these restrictions.

One approach to estimating future crop yields based on known variables is called Crop Yield Prediction (CYP). There are a number of elements, both trainable and untrainable, that influence yield prediction. Data mining and probability are the foundation of predictive modeling. There are four steps involved in data modeling for forecasting.

(i) historical data analysis,

(ii) data pre-processing,

(iii) modeling of data, and

(iv) performance estimationUsing the feature set to fine-tune the ML algorithm's parameters yields a reliable prediction. Scientists are hard at work creating effective means of gauging the precision of their forecasts. Using traditional statistical and ML techniques, data-driven models have become more popular and have been successfully applied to CYP. Parametric and nonparametric supervised ML methods, such as Artificial Neural Networks (ANNs),

Support Vector Regression (SVRs), k-Nearest Neighbors (k-NNs), and Random Forests (RFs), are now dominating the CYP across a variety of agricultural datasets.

Literature Review

Ronnie Concepcion et.al.,(2020) In the case of managed agriculture, crop water stress is entirely preventable. Measuring agricultural water usage is seldom done because of the high cost and complicated laboratory technique involved. Water management, however, necessitates the use of water stress indicators. This research used deep transfer image networks, feature-based machine learning, and computer vision to predict full moisture content (FMC) and equal water thickness (EWT) in lettuce canopies, two indicators of water stress. After 42 days of growth (plus an extra 7 days to account for drought-induced senescence), 330 heads of irrigated and non-irrigated aquaponic lettuces were harvested. The leaf wet and dry weights were obtained using an oven dehydrator. Pixels that weren't vegetation were removed using a graph-cut segmentation technology called lazysnapping. The R, G, B, a, and Y color components, morphological canopy area, and thermal crop water stress index (CWSI) were chosen as the most influential thermo-visible phenotypic characteristics using neighborhood component analysis (NCA). In terms of FMC and EWT predictions, the evolutionary strategy (ES)-optimized recurrent neural network (RNN) showed the greatest efficiency, with an average inference time of 7.5 seconds and R2 performance of 0.9233 and 0.8155, respectively. There is an increase in lettuce canopy surface FMC of 2.9845%°C _l and EWT of 0.4632 gcm2C1. When the canopy area grows by 1 square centimeter, the EWT increases by 3.1996%, and the FMC rises to 0.0212 gcm2. Greener lettuce canopies develop on plants with FMC more than 90% and EWT less than 0.4 gcm2 in the third week after germination. The computational evaluation of lettuce water stress using an ES-RNN model is shown to be accurate, sensitive, and real-time..

Tanhim Islam et.al.,(2019) The best method for crop selection and yield prediction with little time and money investment is presented in this study. One of the most reliable modeling and prediction techniques available is the artificial neural network. This research compares the accuracy and error rate of this algorithm against others, including the support vector machine, the Logistic Regression, and the random forest method, all of which strive to improve output and prediction. In addition, the purpose of using these algorithms is to evaluate their efficiency on a dataset of over 0.3 million examples. For this forecast, we compiled data on 46 variables, including high and low temperatures, average rainfall, humidity, climate, weather, and different kinds of land, chemical fertilizer, soil, soil structure, soil composition, soil moisture, consistency, reaction, and texture. In this research, we propose using a deep neural network to forecast agricultural crop yields.

A Suresh et.al.,(2018) K-means and Modified K-Nearest Neighbor (KNN) are proposed as a way of crop prediction in this research for the principal crops grown in Tamilnadu. Clustering is done in Matlab, while classification is done in WEKA. The numerical outcome proves that our strategy is superior than the conventional data mining technique.

For maximum annual production, farmers must keep a close eye on their crops. Unfortunately, many types of crop disease have a major negative impact on crop yield, product quality, and overall agricultural output. For a long time, sustainable agricultural growth was severely hampered by biological hazards including crop diseases and insect pests. Farmers may use this data to make educated decisions about which new crop varieties to adopt in order to maximize yields. Sometimes, new types don't need any adjustments to the farmer's current resource expenditures.

Therefore, the data must be reliable and useful to both farmers and seed companies. Data mining is the process of discovering actionable insights in unstructured data sets. Predicting future behavior helps in making well-informed, preventative choices. Results from a variety of data and analyses may be generated using the Linear Regression model.the model constructed using the innate mathematical connections in the existing dataset. Research in agriculture, crucial to the expansion of the economy, is bolstered by a number of technological advancements, instruments, and algorithms.

Data mining is one such method, and it is used for analyzing, evaluating, and deriving patterns and information from predictive results. Research built on the outcomes of the forecast is used to enhance manufacturing. The process of using data mining techniques to glean useful information from massive datasets is known as Knowledge Discovery in datasets (KDD). Machine learning, statistics, and pattern identification have all been used by scientists to agricultural challenges. The information may take the form of symbols, numbers, or both. Machine learning algorithms may infer causal connections in the data and use statistical analysis to corroborate such hypotheses.

Predictive and descriptive data mining methods exist, with predictive approaches further subdivided into classification, regression, and time series. The ability to foresee yield and crop illnesses is greatly aided by these methods. Crop and soil management research also benefits from this method. Support Vector Machine (SVM), Multiple Linear Regression (MLR), Neural Network (NN), and Bayesian Network (BNN) are the four machine learning techniques covered in this chapter. Preventing crop losses via disease management is done in response to forecasts and predictions. The symptoms are used to diagnose the condition. In order to better agriculture, experts are quite concerned about these autonomous disease prediction and control systems. The laws of Classification and prediction are explored in this chapter for their application in the agricultural arena. In order to classify sugarcane, paddy, and banana, this chapter proposes utilizing the Linear Discriminant Analysis technique.

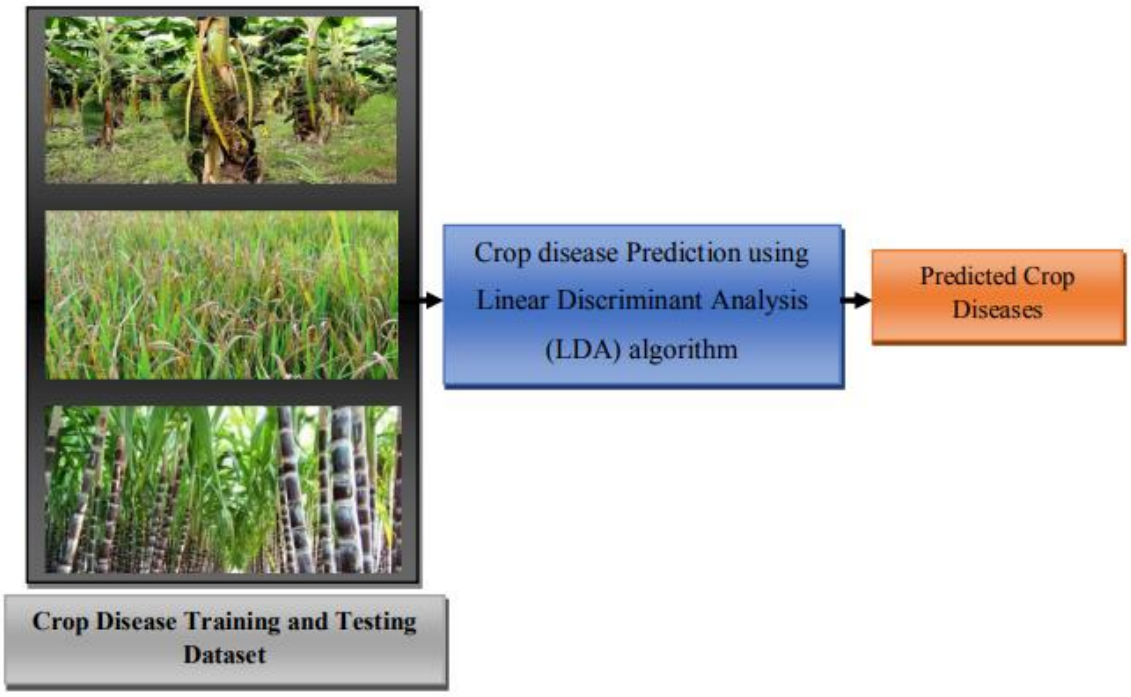

Figure 1. Shows the proposed crop disease prediction architecture

Proposed Linear Discriminant Analysis For Crop Prediction:

A decision tree induction approach was used for the prediction phase of categorizing the classes in this study. The nodes reflect the criteria for splitting, while the edges show the results of the test. Finally, the leaves represent labelled categorization results. The suggested approach classifies crop yields into low, medium, and high categories based on weather conditions. When there are two or more recognized groups or clusters that may be utilized to categorize the study data, a discriminant analysis can be performed. Applicable situations include a categorical dependent variable and an interval predictor or independent variable. Using discriminant analysis, the gaps between the categories may be maximized. For more complex classification issues, it was developed into Linear Discriminant Analysis (LDA). Similar to regression analysis, discriminant analysis establishes a connection between a predictor and a dependent variable from which a value for the predictor may be derived. Logistic regression, for instance, is a least-squares approach to a two-class classification issue.

The proposed linear discriminant analysis approach is evaluated against a C4.5 decision tree, a Support Vector Machine (SVM), and an Artificial Neural Network (ANN) using a dataset

consisting of training and testing data of 100 instances for a banana, paddy, and sugarcane.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
C4.5	0.54	0.065	0.592	0.54	0.511	0.931
SVM	0.78	0.029	0.792	0.78	0.777	0.967
ANN	0.56	0.055	0.631	0.56	0.551	0.875
LDA	1	0	1	1	1	1

Table 1: Comparative measures of each technique for banana

Table 1 shows that compared to other algorithms, LDA produces the best results in terms of sensitivity, specificity, recall, precision, F-measure, and area under the receiver operating characteristic curve (ROC). This includes C4.5, SVM, and ANN.

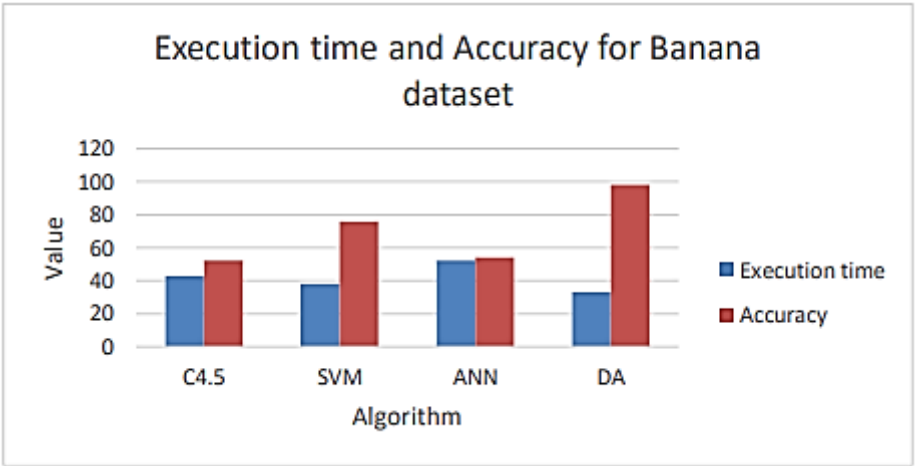

Figure 2 Execution Time and Accuracy comparison for Banana Dataset

Figure 2 displays the Banana dataset's execution time and accuracy graph. The C4.5 algorithm has a shorter runtime than ANN. However, when combined with SVM, performance lags significantly. The SVM method, on the other hand, has a relatively short running time. When compared to SVM, however, the suggested LDA technique has a faster runtime.

Table 2: Comparative measures of each technique for paddy

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
C4.5	0.51	0.068	0.536	0.51	0.467	0.908
SVM	0.98	0.003	0.983	0.98	0.98	0.998
ANN	0.72	0.032	0.751	0.72	0.713	0.901
LDA	1	0	1	1	1	1

Table 2 shows that when comparing the true positive rate, false positive rate, precision, recall, F-measure, and ROC area of several algorithms, LDA is the clear winner.

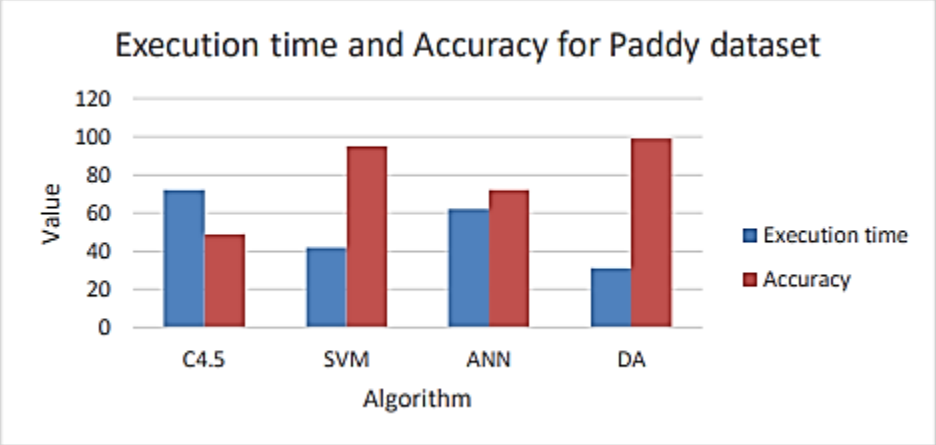

Figure 3: Execution Time and Accuracy comparison graph for paddy dataset

Conclusion

In this study, we use Linear Discriminant Analysis (LDA) to forecast crop diseases. To successfully forecast crop diseases, this step included comprehensive analysis of several data mining classifiers on various feature sets. The current study on banana, paddy, and sugarcane disease prediction demonstrates the utility of LDA models for future forecasting models due to its applicability in disease prediction. Better knowledge of system dynamics is possible via experimental techniques combined with dynamic models. Predictions of plant diseases are currently being made using just multiple regression and neural networks. Recent advances in machine learning, such as linear discriminant analysis (LDA), have been shown to improve forecasting, which will aid in the development of control systems that reduce yield losses.

References

1. R. Concepcion II, S. Lauguico, V. J. Almero, E. Dadios, A. Bandala and E. Sybingco, "Lettuce Leaf Water Stress Estimation Based on Thermo-Visible Signatures Using Recurrent Neural Network Optimized by Evolutionary Strategy," *2020 IEEE 8th R10*

Humanitarian Technology Conference (R10-HTC), Kuching, Malaysia, 2020, pp. 1-6, doi: 10.1109/R10-HTC49770.2020.9356963.

2. T. Islam, T. A. Chisty and A. Chakrabarty, "A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh," *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Malambe, Sri Lanka, 2018, pp. 1-6, doi: 10.1109/R10-HTC.2018.8629828.
3. A. Suresh, P. Ganesh Kumar and M. Ramalatha, "Prediction of major crop yields of Tamilnadu using K-means and Modified KNN," *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2018, pp. 88-93, doi: 10.1109/CESYS.2018.8723956.
4. X. Gao, J. Huang, H. Ma, W. Zhuo and D. Zhu, "Regional Winter Wheat Maturity Date Prediction Using Remote Sensing-Crop Model Data Assimilation and Numerical Weather Prediction," 2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics), Hangzhou, China, 2018, pp. 1-5, doi: 10.1109/Agro-Geoinformatics.2018.8476094.
5. H. Aghighi, M. Azadbakht, D. Ashourloo, H. S. Shahrabi and S. Radiom, "Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 12, pp. 4563-4577, Dec. 2018, doi: 10.1109/JSTARS.2018.2823361.
6. H. Qing and C. Zhongxin, "Predicting Winter Wheat Yield in 2030 and 2050 in North China Based on BioMa-Site and BioMa-Spatial," 2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics), Hangzhou, China, 2018, pp. 1-5, doi: 10.1109/Agro-Geoinformatics.2018.8475998.
7. R. L. F. Cunha, B. Silva and M. A. S. Netto, "A Scalable Machine Learning System for Pre-Season Agriculture Yield Forecast," *2018 IEEE 14th International Conference on e-Science (e-Science)*, Amsterdam, Netherlands, 2018, pp. 423-430, doi: 10.1109/eScience.2018.00131.
8. X. Huang, J. Liu, C. Atzberger and Q. Liu, "Research on the Optimal Thresholds for Crop Start and End of Season Retrieval from Remotely Sensed Time-Series Data Based on Ground Observations," *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Valencia, Spain, 2018, pp. 7727-7730, doi: 10.1109/IGARSS.2018.8519031.
9. M. R. S. Muthusinghe, P. S.T., W. A. N. D. Weerakkody, A. M. H. Saranga and W. H. Rankothge, "Towards Smart Farming: Accurate Prediction of Paddy Harvest and Rice Demand," *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Malambe, Sri Lanka, 2018, pp. 1-6, doi: 10.1109/R10-HTC.2018.8629843.
10. K. Teeda, N. Vallabhaneni and T. Sridevi, "Comparative Analysis of Data Mining Models for Crop Yield by Using Rainfall and Soil Attributes," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, India, 2018, pp. 1176-1180, doi: 10.1109/ICICCT.2018.8473074.
11. Y. Gandge and Sandhya, "A study on various data mining techniques for crop yield prediction," *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, Mysuru, India, 2017, pp. 420-423, doi: 10.1109/ICEECCOT.2017.8284541.
12. Y. M. Fernandez-Ordoñez and J. Soria-Ruiz, "Maize crop yield estimation with remote sensing and empirical models," *2017 IEEE International Geoscience and Remote Sensing*

Symposium (IGARSS), Fort Worth, TX, USA, 2017, pp. 3035-3038, doi: 10.1109/IGARSS.2017.8127638.

13. A. K. Mariappan and J. A. Ben Das, "A paradigm for rice yield prediction in Tamilnadu," *2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, Chennai, India, 2017, pp. 18-21, doi: 10.1109/TIAR.2017.8273679.
14. N. Gandhi, L. J. Armstrong and M. Nandawadekar, "Application of data mining techniques for predicting rice crop yield in semi-arid climatic zone of India," *2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, Chennai, India, 2017, pp. 116-120, doi: 10.1109/TIAR.2017.8273697.
15. M. A. Hossain, M. N. Uddin, M. A. Hossain and Y. M. Jang, "Predicting rice yield for Bangladesh by exploiting weather conditions," *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, Korea (South), 2017, pp. 589-594, doi: 10.1109/ICTC.2017.8191047.