Review of Machine learning models for Crop Yield Prediction

Aravind T [1], Dr.D.Indumathi[2]
{20mz01@psgtech.ac.in [1], ind.cse@psgtech.ac.in[2]}

PG Scholar, Dept. of Computer Science, and Engineering,
PSG College of Technology, Coimbatore, India [1], Assistant Professor (SG), Dept. of Computer Science and Engineering, PSG College of Technology, Coimbatore, India[2]

Abstract. Agriculture is not only a necessary component of human life, but it is also one of India's most important sources of employment. Agriculture employs more than half of our country's population. It is the economic backbone of our country. Farmers can benefit from early diagnosis and control of issues in order to increase crop productivity. Crop yield prediction is an essential scientific area that aids in food security. Based on many parameters, a machine learning model will grasp the pattern of the crop and yield and estimate the yield of the region in which the farmer would crop. The machine learning algorithm can then be distributed after implementation with a web-based visual application that is convenient to use. The results obtained will be granted access to the farmers. This study investigates various machine learning models that are employed in predicting the crop yield.

Keywords: Agriculture, Machine Learning, Parameters.

1 Introduction

India is a densely populated country with erratic weather circumstances that need securing the country's food supplies. Drought conditions pose significant problems to farmers. Figure 1 shows that drought has seriously affected the yield of a farmer. Climate change is also quite common nowadays. Agriculture is a partner trade industry that has benefited greatly in recent years as information science and machine learning approaches have progressed. These advances return to our society's round-faced environmental and demographic challenges, with research indicating a need for a significant rise in international agricultural production to feed an expanding population on a warming globe. For the most part, the metric capacity unit takes use of some rather remote sensing information over the farm, leaving the world of production prediction behind. Agriculture strives to enhance and improve agricultural yields, as well as crop quality, in order to support human existence. As a result, knowing meteorological conditions makes it difficult to cultivate crops. There is a need to employ technology to locate or comprehend crop specifics so that farmers can be advised accordingly. Crop yield forecasting will undoubtedly benefit farmers. Improving crop quantity is the primary objective of precision agriculture, which entails gaining a better knowledge of the crop via the use of information technology tools. Profitability and sustainability are the primary goals of precision agriculture. The farmer can make crop selection decisions and contribute more to the farm's earnings. There are several crop yield prediction models available, some of which use weather

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2314568

actual factors while others use static parameters. Machine learning has been identified as a particularly interesting subject that can contribute to agriculture. Machine learning models can accept a variety of precise inputs and provide real results. The goal is to develop the most efficient model to forecast the crop's production, therefore experiment is done with several algorithms and comparison is made to see which one has the least error and loss, then choose that model and estimate the crop's yield. In this article, the performance of various algorithms are compared .

Figure 1: Drought affecting farmers crop yield

2 Literature Survey

Dilli Paudel et al presented a machine learning methodology for crop production prediction using MARS - Crop Yield Forecasting System data [1]. The process is assessed by forecasting agricultural yield for five crops and three nations at NUTS2 or NUTS3 levels. Experiments were conducted for each crop and nation to forecast early season (30 days after planting) and end of season crop production with and without the use of previous year's projected yield trend. For each experiment, regional predictions were compared to prior MCYFS forecasts using a basic technique with no prediction skill. The predictions were also aggregated to national (NUTS0) level and compared to past MCYFS forecasts. Machine learning algorithms beat the conventional model, according to the findings.

During the growth season of the crops, A.Sharma et al suggested assessing soil moisture over winter wheat fields using machine-learning algorithms [2]. RADARSAT-2 data was gathered using quad polarizations and 240 sample plots. To broaden the SAR feature space, polar metric decomposition parameters were recovered in addition to the four linear polarization channels. Support vector regression, random forests (RF), and gradient boosting regression tree were three sophisticated machine-learning models presented. The results reveal that the RF outperformed the SVM in terms of estimated accuracy.

A study of crop image classification in order to assess crop quality and production using machine learning and computer vision was presented [3]. ML models based on data gathered by collar sensors improved livestock productivity by forecasting reproductive patterns, detecting eating problems, and predicting cow behavior. It illustrates how knowledge-based agriculture may boost long-term production and product quality. For weed prediction, Convolution Neural Networks is the best, Random Forest is better for crop production prediction, and the regression method is superior for weather forecast.

A detailed examination of the benefits and drawbacks of machine learning-based crop production prediction, as well as appropriate identification of present and future agricultural sector issues was presented [4]. It is primarily concerned with the yield of palm oil. Soil characteristics are the most important element determining palm oil yield. Several algorithms were employed, and their results were compared. Linear Regression, Random Forest, and Neural Networks have greater accuracy than Deep Neural Networks, Convolution Neural Network, and Long Short Term Memory, according to the comparison results.

For the efficient use of fertilizers for the crop, a random forest algorithm was introduced [5]. The back propagation algorithm was compared with the random forest algorithm. The back propagation method produced a 94 % average accuracy, whereas the random forest approach earned a 94.25 overall accuracy. The accuracy of the results is related to the amount of the dataset collected.

Crop yield analysis using Support Vector Regression (SVR) and Linear Regression (LR), was presented by Haque et al [6], these algorithms were highly suited for verifying the variable parameters in the forecasting the continuous variable estimate using 140 data points that were obtained. The mean square error (MSE) and coefficient of determination were used to calculate the error rate (R2). The findings revealed that the support vector machine was on par with linear regression in terms of accuracy, with both methods achieving comparable outcomes.

A system that analyses data collected from real-time sensors in agricultural areas was presented [7]. The data was pre-trained, and the model was built using the naïve bays classifier. The technology would assist farmers in predicting the most cost-effective crop for their specific area. The farmers are then directed via an app on their phones to help them understand what kind of seeds they should plant in their fields to increase yields. The information is gathered using real-time sensors in farm fields. A boosting approach is used to increase the accuracy of the Naive Bayes model. The model's accuracy was determined to be 97 %.

Jrip Classifier is a novel method proposed by Jyoti et al [8]. This study aims to examine the soil and its fertility in India on a broad scale, as well as their influence on crop forecast. The study discovered that soil has a significant impact on crop choices. The JRip classifier and the naïve bayes classifier were compared for accuracy. The JRip Classifier is utilized, and it's an effective approach for mistake pruning. The coaching material is divided into two sections throughout this algorithm, and the error is minimized with the help of pruning machinists.

The application of several data mining methods to improve accuracy rate was proposed[9]. Variables such as ph, humidity, rainfall, temperature, and so on are included in the dataset. During training, a huge number of decision trees are formed, and the outcome or output is split into classes based on the number of classes. A Decision Tree Classifier is also used in this study to compare the two and select the right choice. Supervised Learning techniques were used to predict the outcome. For training the model, Random forest has been compared with Decision tree.

Crop production using one of the most used boosting methods was introduced [10]. They evaluated two boosting algorithms: AdaBoost and Gradient Boost. The goal of employing a boosting algorithm is to increase a poor learner's performance so that a better outcome may be produced. The results indicated that the AdaBoost Regressor with Decision Tree has 94.67 percent accuracy compared to 94.9 percent for the AdaBoost Regressor with Random Forest.

A study of the literature on machine learning models for predicting agricultural productivity using meteorological data was presented[11]. According to the report, extension of the search to include additional crop yield-related parameters. Rainfall, temperature, and soil fertility were among the issues addressed. When comparing the experimental values and outcomes for the crop paddy dataset, the deep reinforcement learning model is shown to predict the data with a 93.7 percent higher accuracy and precision than the other methods tested.

A classifier-based crop recommendation system was introduced[12]. The "Decision Tree" and "KNN" ML classification algorithms were compared in this study. Soil parameters, climatic parameters, and production parameters make up the data set. This study uses machine learning methods to calculate. This paper compared both methods separately, but did not combine them. This is a model for a recommendation system that uses classifiers.

Yield of wheat crop higher resolution images was collected, the weather data was acquired with the help of NDVI time-series data, and the amount of yield, harvesting equipment's m crop type and geo location were discussed [13]. Nine machine learning algorithms are used to create ensemble-based learners. The support vector regression with radial basis function has the lowest RMSE of 0.59, according to the results.

Traditional regression methods such as boosted regression tree, random forest regression, support vector regression, and Gaussian process regression to assess the effectiveness of sophisticated Machine Learning approaches were introduced [14] .The NDVI values of all silage maize fields were averaged and integrated for this purpose, resulting in a two-dimensional dataset for each year. The machine learning approaches were applied and their assessment criteria were used to assess their effectiveness and stability. Finally, the yields of silage maize were calculated by averaging the findings of each ML method.

The ability of an Extreme Learning Machine (ELM) model to evaluate soil fertility features and deliver an accurate estimate of Robusta coffee yield was presented[15]. The efficiency of 18 distinct ELM-based models based on soil organic matter (SOM), accessible potassium, boron, sulphate, zinc, phosphoric, ammonia, exchangeable calcium, magnesium, and pH was assessed using single and multiple combinations of predictor variables. The performance of the ELM model was compared to that of current prediction techniques such as MLR and Random Forest (RF). Table 1 summarizes about the existing models

3 Research Questions

This study seeks to learn more about the research that have been published in the field of machine learning and crop production prediction. Studies have been evaluated from a variety of perspectives in order to gain understanding. The following four research questions (Qs) have been established for this study.

• **Q1**: Which machine learning methods have been used to forecast agricultural yields in the literature?

• **Q2**- In the literature, what characteristics have been utilized to forecast agricultural yields using machine learning?

• **Q3**- Which evaluation factors and techniques for crop yield prediction have been utilized in the literature?

• **Q4**- What are the obstacles in applying machine learning?

4 Discussion

• **Q1-related (algorithms) discussion :** Random Forest and Decision Tree are the most used algorithm, according to table 1. In most situations, the Decision tree is used as a benchmarking method to see if the proposed algorithm is superior than Decision tree.

• **Q2-related (features) discussion**: To display the major features and algorithms, groups for features and algorithms are formed. Although specific information has been lost as a result of this decision, clarity has been retained. Soil type, rainfall, temperature, land area, and season are the most commonly utilized characteristics.

• **Q3-related (evaluation parameters and approaches) discussion:** There aren't many assessment factors mentioned. RMSE was almost commonly used as a metric for assessing model quality. MSE, R2, and MAE are some of the other assessment metrics that were used.

• **Q4-related (challenges) discussion:** The difficulties are mostly in the area of improving a functional model. Much more may be stated about the model's accuracy as more data is needed for training and testing. The integration of the models into farm management systems is also a difficulty. When specific tools are developed to utilize these models, farmers can effectively make use of it to make crop choices.

I. Machine learning algorithms

Machine learning is a field of computer science that allows computers to learn without needing to be programmed directly. In 1950, Alan Turing proposed the concept of learning machines and released "The Turing Test for Machine Intelligence," a research paper. The machine was put through a series of tests to determine whether it could exhibit intelligent behavior similar to that of humans. A machine or intelligent computer algorithm learns from data and extracts the information from it, allowing it to make predictions or intelligent judgments. As a result, the machine learning process is separated into three parts: data collection, model development, and generalizations.

The process of anticipating the outcome for inputs with which the algorithm has never been taught is known as generalization. Prediction of weather, spam filtering, detection of plant disease, and pattern recognition are all examples of issues where machine learning algorithms are utilized to tackle difficult problems where human knowledge fails. Companies and academic organizations are increasingly turning to machine learning algorithms to tackle a wide range of challenges, thanks to the abundance of innovative algorithms and large data sets available through internet resources. Deep Learning is a branch of machine learning techniques that is learned from huge datasets and makes intelligent choices using an artificial neural network .

Machine learning algorithms comprise of supervised learning, unsupervised learning, and reinforcement learning., Supervised Learning as the name implies, is learning with the supervision or guidance of a supervisor. This group of algorithms works with labeled data-sets, which means that there are outputs for each input. With this labeled data set, the algorithm constructs an input-output connection and then generalizes or predicts outcomes for unknown inputs. Supervised learning techniques that predict category values are known as classification algorithms, whereas methods that predict numerical values are known as regression algorithms.

Unsupervised learning algorithms operate with unlabeled data and group like things together to discover unknown objects. Unsupervised learning algorithms are more difficult to construct than supervised learning algorithms since their objective is to extract hidden knowledge from the training data set. Reinforcement learning is a machine learning training approach that rewards desirable actions while penalizing undesirable ones.

Supervised machine learning algorithms such as Random Forest, Ada Boost, Gradient Boosting using bagging and boosting techniques are highly effective in predicting the crop yield.

5 Random Forest

Random Forest is an ensemble strategy that solves regression and classification problems by combining several decision trees using a technique called as Bootstrap and Aggregation, often known as bagging. Instead of relying on individual decision trees to decide the ultimate outcome, the main idea is to combine many decision trees.

Random Forest employs a large number of decision trees as a foundational learning paradigm. Row and feature selection from the dataset are done at random, resulting in sample datasets for each model. This section is called Bootstrap. Figure 2 summarizes the working of random forest algorithm.

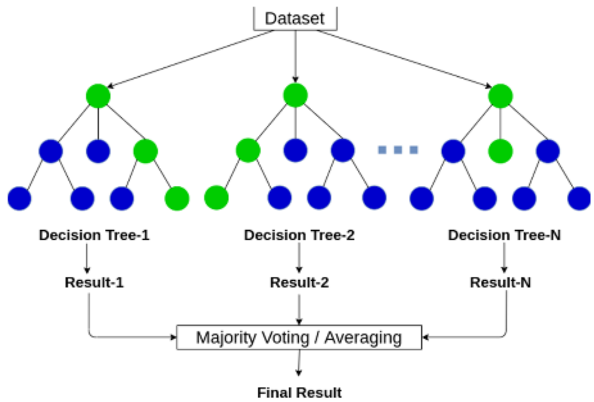

Figure 2: Random Forest Algorithm

Step 1: In Random forest, n random records are chosen at random from a data collection of k records.

Step 2: For each sample, an individual decision tree is built.

Step 3: Each decision tree will provide a result.

Step 4: For Classification and Regression, the final outcome is based on Majority Voting.

Random Forest is a particularly useful method for crop yield prediction. Understanding the hyper parameters of crop are simpler because of the default hyper parameters it uses. It typically provides a decent prediction result for the yield of crops.

6 AdaBoost Algorithm

The AdaBoost algorithm is a Boosting approach used in Machine Learning as an Ensemble Method. The weights are re-allocated to each instance, with larger weights applied to erroneously categorized instances. During the data training phase, it creates a certain number of decision trees. The wrongly categorized record in the first model is given precedence when the first decision tree/model is constructed. Only these records are given to the second model as input. The procedure continues until we have decided on a number of base learners to construct. Figure 3 summarizes the working of AdaBoost algorithm.

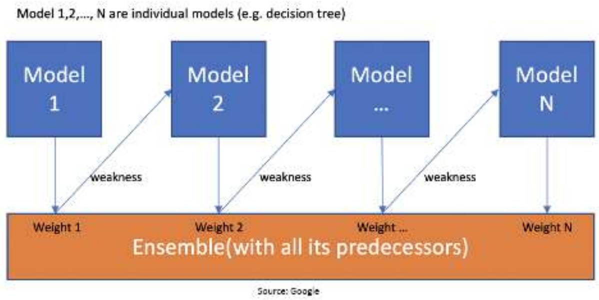

Figure 3: Ada Boost Algorithm

Step 1 :Adaboost begins by selecting a training subset at random.

Step 2 : It continually trains the model by selecting the training set based on the accuracy of the previous training.

Step 3: It gives incorrectly categorized observations a larger weight so that they have a better likelihood of categorization in the following iteration.

Step 4 : It also provides weight to the trained classifier in each iteration based on the classifier's accuracy. The classifier with the highest accuracy will be given the most weight.

Step 5 : This method is repeated until the entire training data fits without error or until the stated maximum number of estimators is achieved.

Since Adaboost is an ensemble approach that combines a lot of weak classifiers to produce a powerful classifier. When used in combination with other machine learning algorithms, it learns from the weak classifier and improves accuracy in forecasting the crop yield.

7 Gradient Boosting

Gradient Boost is a robust machine learning method that combines gradient descent with boosting. The methodology provides a straightforward interpretation of boosting methods from the standpoint of numerical optimization in a function space, and it generalizes them by enabling the optimization of any loss function. Gradient Boosting is made up of three major parts: an additive model, a loss function, and a weak learner.

Gradient Boosting is based on the idea that combining the best feasible future model with the prior models reduces overall prediction errors. To minimize errors, the fundamental notion is to carry over the goal outcomes from prior models to the next model. Figure 4 summarizes the working of Gradient Boosting algorithm.

Gradient boosting is a boosting technique that can be used to improve the yield of the crop. It improves the performance of the algorithm by reducing overfitting of the crop yield data set.

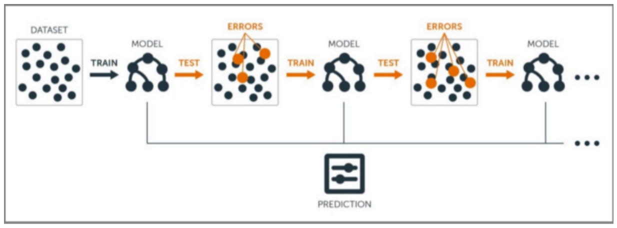

Figure 4 :Gradient Boosting Algorithm

Table 1 : Summary of Existing Models

S.NO	Paper Title , Author , Year of Publication	Methodology	Pros	Cons
1	Machine learning for large-scale crop yield forecasting Dilli Paudel et al (2021)	• Random Forest, • K-nearest Neighbors • Support Vector Regression • Decision Tree • Gradient Boosting	In comparison with MYCFS forecasts, Machine Learning baseline has achieved higher accuracy.	The baseline was not used for extremely large data analysis.
2	Estimating Soil Moisture Over Winter Wheat Fields During Growing Season Using Machine-Learning Methods L. Chen et al (2021)	• Support Vector Regression • Random Forest Regression	Random Forest produced the most accurate estimation results for both the training and validation sets.	The sampling area was limited to one wheat field and the number of samples was minimal.
3	Machine Learning Applications for Precision Agriculture: A Comprehensive Review A.Sharma et al (2021)	• Convolution Neural Networks • Random Forest • Support Vector Machine • Decision Tree	Random Forest algorithm performed better than all other algorithms .	Unsupervised learning algorithms was not used.
4	A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction M. Rashid et al (2021)	• Random Forest, • Support Vector Regression • Artificial Neural Networks • Linear Regression • Decision Tree	Linear Regression, Random Forest, and Neural Networks have greater accuracy than Deep Neural Networks, Convolution Neural Networks and Long Short Term Memory Neural Networks.	Complex data set.

5	Crop Yield Prediction and Efficient use of Fertilizers S. Bhanumathi et al (2019)	● Random Forest , Back Tracking algorithm	Accuracy of the Random forest was higher than the back propagation algorithm.	More Algorithms could have been used for the comparison of crop yield.
6	Crop Yield Analysis Using Machine Learning Algorithms F. F . Haque et al (2020)	● Support Vector Regression ● Linear Regression	Linear Regression achieved a lesser error rate .	Classification techniques were not used.
7	Crop Prediction using Machine Learning. M.Kalimuthu et al (2020)	● Naive Bayes	Naive Bayes model achieved 97 % accuracy.	Naive Bayes is an outdated algorithm.
8	Prediction Of Soil Accuracy Using Data Mining Techniques . D. Jyoti Bhanudas et al (2019)	● JRip Classifier ● Naive Bayes	JRip classifier result is superior, at 0.982, than Naive Bayes classifier result, which is 0.86.	JRip Classifier could have been compared with ensembling learning algorithms.
9	Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector Y.J.N Kumar et al (2020)	● Decision Tree, ● Random Forest	Random Forest achieved higher accuracy than decision tree.	More innovative ensemble learning algorithms could have been used.
10	An Ensemble Algorithm for Crop Yield Prediction M.Keerthana et al (2021)	● Ada Boost ● Gradient Boost ● Random Forest ● SVM, ● LR ● Decision Tree	AdaBoost Regressor with Decision Tree Achieved 95.7% accuracy.	Advanced sophisticated algorithms with greater prediction can also be used to improve accuracy.
11	Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications Elavarasan et al (2020)	● Deep reinforcement learning	Precision of this model is higher than the other models.	Time Complexity was higher.

12	Crop Prediction System using Machine Learning Algorithms Patil et al (2020)	● Decision Tree ● K Nearest Neighbor	Precision of K Nearest Neighbor was higher .	The model did not combine two algorithms.
13	Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods Elisa kamir et al (2020)	● Random Forest ● K Nearest Neighbor ● Support Vector Regression	Support vector regression with radial basis functions predicted yield with lesser error rate.	The data set was complex.
14	Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI Aghighi et al (2018)	● Boosted regression tree ● Random forest regression ● Support vector regression ● Gaussian process regression	Performance of machine learning models was higher than the conventional regression methods.	Inconsistent and inaccurate images.
15	Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties Kouadio et al (2018)	● Extreme learning machine	Extreme learning machine achieved higher accuracy than multiple linear regression.	Hybrid approaches were not used.

8 Conclusion

Several crop yield prediction techniques are discussed in depth utilizing a synthetic and comparative analysis of existing systems.Performance of Machine Learning algorithms such as Random Forest , Adaboost and Gradient Boosting show that they are extremely beneficial to the agricultural community in overcoming the problems faced by the farmers. It provides feedback to the farmers regarding the effective grow of the crop and the maximum yield. The efficiency of the crop yield is achieved by using such methodologies and sowing the right crops to obtain the maximum yield at any given time.

References

[1] Dilli Paudel et al., Machine learning for large-scale crop yield forecasting,Agricultural Systems, Volume 187, 2021, 103016, ISSN 0308-521X, https://doi.org/10.1016/j.agsy.2020.10306

[2] L. Chen et al., "Estimating Soil Moisture Over Winter Wheat Fields During Growing Season Using Machine-Learning Methods," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 3706-3718, 2021, doi: 10.1109/JSTARS.2021.3067890.

[3] A. Sharma, A. Jain, P. Gupta and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," in IEEE Access, vol. 9, pp. 4843-4873, 2021, doi: 10.1109/ACCESS.2020.3048415

[4] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin and N. Khan, "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction," in IEEE Access, vol. 9, pp. 63406-63439, 2021, doi: 10.1109/ACCESS.2021.3075159.

[5] S. Bhanumathi, M. Vineeth and N. Rohit, "Crop Yield Prediction and Efficient use of Fertilizers," 2019 International Conference on Communication and Signal Processing (ICCSP), 2019, pp. 0769-0773, doi: 10.1109/ICCSP.2019.8698087.

[6] F. F. Haque, A. Abdelgawad, V. P. Yanambaka and K. Yelamarthi, "Crop Yield Analysis Using Machine Learning Algorithms," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1-2, doi: 10.1109/WF-IoT48130.2020.9221459.

[7] M. Kalimuthu, P. Vaishnavi and M. Kishore, "Crop Prediction using Machine Learning," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 926-932, doi: 10.1109/ICSSIT48917.2020.9214190.

[8] D. Jyoti Bhanudas and K. Rahat Afreen, "Prediction of Soil Accuracy Using Data Mining Techniques," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), 2019, pp. 1-5, doi: 10.1109/ICCUBEA47591.2019.912957

[9] Y. J. N. Kumar, V. Spandana, V. S. Vaishnavi, K. Neha and V. G. R. R. Devi, "Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 736-741,doi: 10.1109/ICCES48766.2020.9137868.

[10] M. Keerthana, K. J. M. Meghana, S. Pravallika and M. Kavitha, "An Ensemble Algorithm for Crop Yield Prediction," 2021 Third International Conference on IntelligentCommunication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 963-970

[11] Elavarasan, Dhivya, and PM Durairaj Vincent. "Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications." IEEE Access 8 (2020): 86886-86901.

[12] Patil, Ajinkya, et al. "Crop Prediction using Machine Learning Algorithms." Kapila Journal of Research 1.1 (2020): 1-8.

[13] Elisa Kamir, François Waldner, Zvi Hochman, estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods,ISPRS Journal of Photogrammetry and Remote Sensing,Volume 160,2020,Pages 124-135,ISSN, 0924-2716,https://doi.org/10.1016/j.isprsjprs.2019.11.008.

[14] Aghighi, Hossein Azadbakht, MohsenAshourloo, DavoudShahrabi, Hamid Radiom, Soheil. (2018). Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. PP. 1-15. 10.1109/JSTARS.2018.2823361.

[15] Kouadio, Louis, et al. "Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties." Computers and electronics in agriculture 155 (2018): 324-338.