



A SURVEY OF MACHINE LEARNING TECHNIQUES FOR CROP YIELD PREDICTION

Leena K. V.¹, Dr. K. Chitra²

¹Research scholar, Sri Krishna Adithya College of Arts and Science,
Arivoli Nagar, Kovai Pudur, Coimbatore, India

²Associate Professor, Department of Computer Science,
Sri Krishna Adithya College of Arts and Science, Arivoli Nagar, Kovai Pudur, Coimbatore, India

ABSTRACT

Machine Learning has emerged as a pivotal tool across various domains, including agriculture. In this context, crop prediction relies on the amalgamation of historical and current data for a given month, thus validating the accuracy of climatic information. The proposed system revolves around the application of supervised Machine Learning techniques to assign the most probable class, which, in this case, signifies the predicted crop based on input parameters. Once the crop prediction is made, it serves as a valuable resource for farmers in selecting the most suitable crop for their specific plot of land. Furthermore, a mobile application assists farmers in comprehending which type of seeds to sow for optimal yield. Traditionally, crop prediction was heavily reliant on farmers' past experiences with climate conditions. However, predicting yield remains a challenging issue given the available data. Machine Learning methods present a more effective alternative for this purpose. This paper introduces and implements a system designed to forecast crop yield using historical data. The approach involves applying Machine Learning algorithms such as Support Vector Machine (SVM) and Random Forest (RF) to agricultural data.

Keywords: Support Vector Machine, Crop yield prediction, Machine learning, Agriculture Random Forest.

1. INTRODUCTION

This paper explores various machine learning methods employed for predicting crop yields. Machine learning approaches adopt an empirical and data-driven perspective, aiming to discern valuable patterns and connections within input data. This avenue holds substantial promise for enhancing crop yield predictions. Machine learning algorithms, in essence, construct approximations of functions that establish relationships between input features or predictors and outcomes, such as crop yield. Much like statistical models, machine learning algorithms can incorporate outputs from other methodologies as features. Furthermore, they boast several unique advantages [3]. They excel at modeling non-linear associations across multiple data sources, tend to perform better with larger training datasets and exhibit robustness against noisy data through the implementation of regularization techniques that mitigate variability as well as enhance summarization. Consequently, ML can amalgamate the strengths of different methods, including models for agricultural crop growth as well as the utilization of remote sensing technology, to furnish dependable predictions of crop yields.



2. APPLICATION OF SUPERVISED MACHINE LEARNING APPROACHES IN CROP YIELD PREDICTION

Supervised machine learning approaches have been mostly applied in crop yield prediction to improve agricultural productivity and decision-making. These approaches involve training models on historical data where both input features (such as weather data, soil quality, and crop management practices) and the corresponding crop yields are known. Once trained, these models can make predictions about forthcoming crop harvests using fresh input data. Below are some typical applications of supervised machine learning in forecasting crop yields:

2.1. Crop Type Classification:

Supervised ML may be utilized for classifying different crop types based on satellite imagery or other data sources. Knowing the crop type is important for yield prediction because different crops have different growth patterns and requirements. Momm et al. [4] introduced a methodology for generating historical crop-type information, which was subsequently assessed. The study delved into the use of the Crop Data Layers (CDL) algorithm for classifying crop types in historical datasets. This endeavor posed challenges mainly related to the temporal resolution of the sensors and the presence of datasets without cloud cover. To address these difficulties, robust classification algorithms, particularly machine learning algorithms of the non-parametric kind, were employed. Among them, the Random Forest algorithm garnered attention. This algorithm is defined as an ensemble data mining and classification technique due to it operates by evolving and concurrently evaluating a collection of autonomous classifiers based on tree structures. Throughout the process of generating potential answers, the procedure employs bootstrapping methods to enhance convergence by ensuring variety in the model. It's essential to highlight that this approach comes with specific constraints. One notable drawback is that the separate decision trees generated by the Random Forest algorithm are not directly accessible to the user. Additionally, the computational demands associated with this method can be quite intensive.

2.2. Optimizing Irrigation:

Supervised learning can help optimize irrigation schedules by predicting soil moisture levels and crop water requirements. This helps conserve water resources and maintain optimal crop health. Chen et al. [5] introduced an irrigation decision-making strategy founded on Deep Q-Learning (DQN) and hinged upon short-term weather forecasts. This approach was designed to identify the most advantageous irrigation decisions. Two distinct irrigation decision-making strategies were assessed in the study: conventional irrigation, which represents the commonly practiced flooded irrigation approach among local farmers, and the DQN-based irrigation approach. The study aimed to evaluate their effectiveness in terms of water conservation. The DQN irrigation strategy demonstrated remarkable proficiency in generalization following the training process, enabling it to effectively make irrigation decisions based on weather forecasts.

2.3. Fertilizer Recommendation:

ML models can analyze soil nutrient levels and recommend appropriate fertilizer application rates, reducing waste and ensuring crops receive the nutrients they need for optimal growth. Ransom et al. [6] introduced a study that explored the integration of statistical as well as ML approaches for enhancing the accuracy of nitrogen (N) recommendations for corn crops, utilizing both soil and weather information. Various modeling

scenarios were examined, considering factors such as addressing multicollinearity through adjustments and incorporating interaction terms between variables. The objective was to find soil as well as weather changes that could enhance three distinct N recommendation tools. Among the ML techniques assessed, the random forest approach emerged as the most effective for refining N recommendation tools. **2.4. Harvest Timing:**

Predicting the optimal time for harvesting is crucial for maximizing crop yield and quality. ML models can consider factors like crop maturity, weather forecasts, and market conditions to make informed harvesting recommendations. Feng et al. [7] employed the random forest technique to predict the ideal time for harvesting. The system that depended on Random Forest consistently demonstrated superior performance compared to the forecasting system based on multiple linear regression in all forecasting scenarios. This approach enabled the timely harvesting of crops using well-maintained equipment, ensuring that harvesting occurred as soon as the plots reached physiological maturity. This proactive approach helped minimize potential grain losses due to factors like wind, rain, or pest damage.

2.5. Climate Change Adaptation:

As climate change continues to impact agriculture, supervised learning is applicable for assessing the effects of changing climate conditions on crop yields and developing adaptation strategies. Feng et al. [8] introduced a novel hybrid model, combining APSIM (Agricultural Production Systems Simulator) with Random Forest (RF), to assess the influence of impending climate change on wheat yield. This research involved a comparative analysis between the APSIM + RF hybrid model and the standalone APSIM model for estimating the consequences of climate variation. The study yielded intriguing results, revealing that the future yield projections generated by the single APSIM model could potentially overestimate the outcomes by 1-10% when contrasted with the APSIM + RF hybrid model. This innovative approach effectively and accurately estimated the influence of climate change on wheat yield.

3. OVERVIEW OF MACHINE LEARNING TECHNIQUES IN CROP YIELD PREDICTION

ML methods have been widely employed in crop yield prediction because of their capability to analyze difficult data and make accurate forecasts. Crop yield prediction using machine learning involves using data-driven models to forecast the expected crop output based on various factors. Figure 1 illustrates the architecture of crop yield prediction using machine learning.

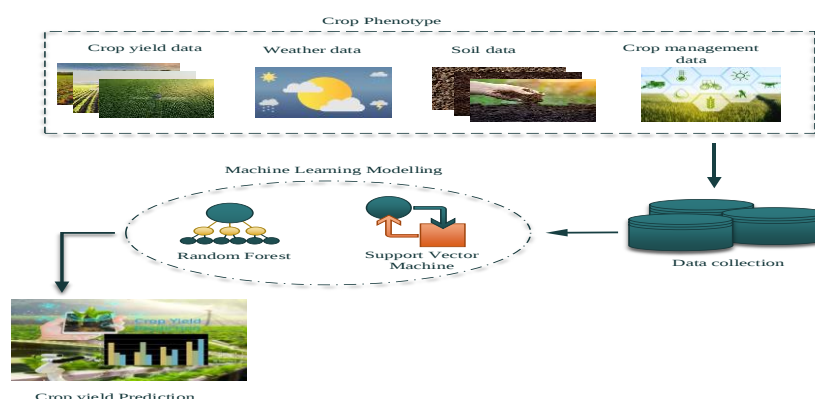


Figure 1. Architecture of Crop Yield Prediction Using Machine Learning



Initially, historical data on crop yields are gathered for the specific region and crop of interest. This data should include information about past yields, as well as factors that can affect yields, such as weather data, soil characteristics, crop yield data, and crop management data. Then these data are pre-processed from which the most important features that have the greatest influence on agricultural yields are extracted. Finally, crop yields will be predicted by using machine learning techniques. There are multiple ML methods are utilized to predict crop yields. But, this section will spotlight two commonly used techniques SVM and Random forest.

3.1. Support Vector Machines:

SVM are powerful ML techniques for crop yield prediction, especially when dealing with complex, non-linear relationships in the data. Here are some techniques and considerations specific to using SVMs for crop yield prediction:

A. Kernel Selection:

SVM can use various kernel functions to transform the data into an elevated-dimensional space where it becomes linearly separable. The choice of the kernel depends on the nature of our data. Prabakaran and colleagues [9] designed an FPGA-based system aimed at predicting agricultural productivity effectively. They employed a Fuzzy Support Vector Machine approach for this purpose, which involved several crucial steps. One of these steps involved the selection of appropriate kernels, as well as the final dataset and parameter configurations. These choices were contingent on the intricacy of the variables being considered. It became evident that relying solely on a linear kernel was inadequate when dealing with real-world agricultural problems. Consequently, the team explored different kernel functions, including the radial bias function, polynomial kernel function, and sigmoid function. Among these options, the RBF kernel demonstrated exceptional accuracy when compared to the others.

B. Feature Selection:

Carefully select and engineer features that are relevant to crop yield prediction. Consider factors like weather data, soil quality, crop type, and agricultural practices. Feature scaling, transformation, and encoding categorical variables may be necessary for optimal SVM performance. Gómez et.al. [10] carried out a research project centered on predicting potato yield using ML methods in conjunction with Sentinel 2 data. Their research findings emphasized the critical role of feature selection during the data preprocessing stage, particularly in mitigating multicollinearity issues among predictors. In their investigation, Gómez and colleagues explored two noteworthy algorithms, namely rqlasso and LeapBack, which possess built-in capabilities for feature selection. These algorithms aim at finding the optimal model that incorporates a predefined quantity of predictors. Additionally, they evaluated models that didn't undergo any feature selection process. It's worth noting that while rqlasso and LeapBack inherently include mechanisms for automatic variable reduction, their performance suffered when confronted with a larger quantity of predictors as well as when the selection of features had been undertaken.

C. Hyper-parameter Tuning:

SVMs have hyper-parameters that need to be tuned for better performance. Shafiee and colleagues [11] employed a grid search technique to fine-tune the hyper-parameters of their model, specifically focusing on the "c" value and the kernel type. The goal was to identify the most effective machine learning regression model for predicting yield. Grid search proved to be a valuable tool for enhancing model performance by systematically



exploring various combinations of hyper-parameter values. This approach operates on a straightforward, exhaustive search principle. Researchers compile a predefined list of values for various hyper-parameters, and the computer systematically assesses the model's performance for each combination of these values. This process ultimately leads to the identification of the optimal hyperparameter values within the specified set.

D. Ensemble Techniques:

Consider combining SVMs with other machine learning models in an ensemble to improve prediction accuracy. Iniyan et.al. [12] developed a methodology termed "Mutual Information Feature Selection (MIFS)" for predicting crop yield on both corn and soybean crops. Their approach involved employing a Multilayer Stacked Ensemble Regression (MSER) technique. The core of this research centered on predicting crop yield accurately, with a specific focus on phenotype factors. To carry out the yield prediction, the team utilized Gradient Boosting Regression within a sequence of learning models. Among the various ensemble methods, the MIFS-based MSER model exhibited superior performance compared to other bagging and boosting methods.

3.2. Random Forest

RF is a commonly chosen ML technique utilized for various prediction activities, including crop yield prediction. It's an ensemble learning technique that amalgamates numerous decision trees to enhance prediction accuracy. When applied to crop yield prediction, RF can provide valuable insights for farmers, agronomists, and policymakers. The common Random Forest methods for crop yield prediction are described below,

A. Data Collection:

Collect historical data about crop yields and pertinent attributes. These features can include weather data (temperature, precipitation, and humidity), soil characteristics (pH, nutrient levels), crop type, planting dates, and more.

B. Pre-processing:

Prep-processing involves cleaning the data by handling missing values, outliers, and inconsistencies and encoding the categorical variables into numerical values using optimal encoding techniques. Pant et.al. [13] Employed the one-hot encoding method to transform categorical variables into numerical values. Their analysis revealed that this particular technique yielded the most favorable results when compared to alternative encoding methods.

C. Feature Selection:

The feature selection step involves identifying pivotal attributes for forecasting crop yields. Random Forest can provide feature importances, which help you select the most relevant variable. Gopal et.al. [14] Explored various feature selection algorithms, including sequential forward feature selection, correlation-based feature selection, variance inflation factor analysis, and random forest, to identify various subsets of features. These features were then integrated into the Multiple Linear Regression model to determine the most optimal feature subset. The inclusion of these specific features ultimately led to improved prediction accuracy.

D. Hyper-parameter Tuning:

This process involves the tuning of hyper-parameters of the Random Forest algorithm. Typical parameters to fine-tune encompass the number of trees within the forest, the maximum depth of individual trees, and the minimum sample size needed for node splitting. Cedric et.al. [15] Employed the GridsearchCV library to facilitate hyper-parameter tuning through the implementation of cross-validation. This approach proved

instrumental in identifying the most suitable model that effectively matched the dataset, all while avoiding the issue of overfitting.

E. Predictions:

Once we are satisfied with the model's performance, we can use it to make crop yield predictions for future seasons. Input the relevant features (e.g., weather forecasts, and soil conditions) to obtain yield prediction. Feng et.al. [16] utilized a nonparametric approach known as Random Forest (RF). This method involves the construction of multiple independent decision trees, which are then combined to achieve a more precise and robust prediction. The advantages and disadvantages of machine learning techniques in crop yield prediction are shown in Table 2

Table 1. Advantages and Disadvantages of Machine Learning Techniques in crop yield prediction

Year	Author's Name	Techniques	Advantages	Disadvantages
2021	Prabakaran and colleagues [9]	Kernel selection	Kernels provide machine learning models with the capability to capture intricate, non-linear connections within the data. In the domain of crop yield prediction, the outcomes are frequently shaped by intricate, non-linear interplays among diverse factors such as weather patterns, soil conditions, and agricultural techniques.	Kernel methods can impose a significant computational burden, especially when working with extensive datasets.
2019	Gómez et.al. [10]	Feature selection	By opting for the most pertinent features, we can diminish the presence of irrelevant information and duplicated data in the dataset.	Eliminating features during the feature selection process can result in the omission of potentially valuable information.
2021	Shafiee and colleagues [11]	Hyper-parameter tuning	Fine-tuning hyper-parameters contributes to the development of more resilient models that exhibit reduced sensitivity to fluctuations in the data.	Efforts to find the optimal hyper-parameters frequently demand a substantial investment of time and a series of experimental iterations



2022	Iniyan et.al. [12]	Ensemble methods	Ensemble methods tend to yield predictions that are more precise compared to those generated by individual base models.	Developing and training multiple models within an ensemble can impose a significant computational load.
2021	Pant et.al. [13]	Pre-processing	Utilizing pre-processing techniques can contribute to the provision of more dependable data for the modeling process.	Pre-processing may add extra intricacy to the modeling pipeline, which can result in increased challenges in terms of comprehension, debugging, and maintenance.
2020	Feng et.al. [16]	Prediction	Predictive models assist in enhancing resource allocation, which ultimately results in increased agricultural efficiency.	The process of gathering, handling, and upholding high-quality agricultural data can incur significant costs and demand substantial resources.

5. CONCLUSION

This paper presents a systematic literature review of machine learning-based crop yield prediction techniques and methodologies. This paper emphasized two machine learning techniques commonly employed in the field of crop yield prediction. The crop yield prediction was done by using ML algorithms like SVM and RF. Also, this paper successfully analyzed the application of supervised machine learning approaches in the field of crop yield prediction.

REFERENCES

1. Pant, Janmejey, et al. "Analysis of agricultural crop yield prediction using statistical techniques of machine learning." *Materials Today: Proceedings* 46 (2021): 10922-10926.
2. Paudel, Dilli, et al. "Machine learning for regional crop yield forecasting in Europe." *Field Crops Research* 276 (2022): 108377.
3. Paudel, Dilli, et al. "Machine learning for large-scale crop yield forecasting." *Agricultural Systems* 187 (2021): 103016.
4. Momm, Henrique G., Racha ElKadiri, and Wesley Porter. "Crop-type classification for long-term modeling: An integrated remote sensing and machine learning approach." *Remote Sensing* 12.3 (2020): 449.



5. Chen, Mengting, et al. "A reinforcement learning approach to irrigation decision-making for rice using weather forecasts." *Agricultural Water Management* 250 (2021): 106838.
6. Ransom, Curtis J., et al. "Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen recommendations." *Computers and Electronics in Agriculture* 164 (2019): 104872.
7. Feng, Puyu, et al. "Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique." *Agricultural and Forest Meteorology* 285 (2020): 107922.
8. Feng, Puyu, et al. "Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia." *Agricultural and Forest Meteorology* 275 (2019): 100-113.
9. Prabakaran, G., D. Vaithiyanathan, and Madhavi Ganesan. "FPGA based effective agriculture productivity prediction system using fuzzy support vector machine." *Mathematics and Computers in Simulation* 185 (2021): 1-16.
10. Gómez, Diego, et al. "Potato yield prediction using machine learning techniques and sentinel 2 data." *Remote Sensing* 11.15 (2019): 1745.
11. Shafiee, Sahameh, et al. "Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery." *Computers and Electronics in Agriculture* 183 (2021): 106036.
12. Iniyar, S., and R. Jebakumar. "Mutual information feature selection (MIFS) based crop yield prediction on corn and soybean crops using multilayer stacked ensemble regression (MSER)." *Wireless Personal Communications* 126.3 (2022): 1935-1964.
13. Pant, Janmejy, et al. "Analysis of agricultural crop yield prediction using statistical techniques of machine learning." *Materials Today: Proceedings* 46 (2021): 10922-10926.
14. Gopal, PS Maya, and R. Bhargavi. "A novel approach for efficient crop yield prediction." *Computers and Electronics in Agriculture* 165 (2019): 104968.
15. Cedric, Lontsi Saadio, et al. "Crops yield prediction based on machine learning models: Case of West African countries." *Smart Agricultural Technology* 2 (2022): 100049.
16. Feng, Puyu, et al. "Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique." *Agricultural and Forest Meteorology* 285 (2020): 107922.