Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 4 (2024)

A Systematic Review of Ml Techniques in Crop Yield Prediction

Dr. K. Chitra[1*], Leena. K. V.[2]

[1*]*Associate Professor, Department of Computer Science, Sri Krishna Adithya College of Arts and Science, Arivoli Nagar, Kovai Pudur, Coimbatore-641042, India*

[2]*Research scholar, Sri Krishna Adithya College of Arts and Science, Arivoli Nagar, Kovai Pudur, Coimbatore-641042, India*

Abstract

In India, there are several strategies to enhance crop productivity and boost economic growth in agriculture. One promising avenue involves leveraging recent technological advancements, such as Machine learning (ML), to predict crop outcomes based on atmospheric and soil parameters of agricultural land. ML serves as a crucial tool for aiding decisions related to crop yield prediction, offering valuable insights into determining optimal crops to cultivate and guiding actions throughout the crop's growing season. In our investigation, we systematically reviewed the literature to collect and merge data concerning the algorithms and characteristics employed in research centered on forecasting crop yields. Different ML algorithms have been employed for facilitating research in crop yield prediction. The study under consideration examines ML approaches, including Support Vector Machine (SVM) and Random Forest (RF), in the context of predicting crop yields.

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

1. Introduction

Agriculture holds a pivotal position within the Indian economy, being a linchpin of its sustenance. As the human population burges, ensuring food security becomes increasingly paramount, and at the core of this challenge lies the crucial task of predicting crop yields [1]. The yield of crops hinges on multifaceted variables, encompassing weather conditions like rainfall, humidity, and temperature, as well as insights into pesticide application. Moreover, having a comprehensive historical record of crop yield data is indispensable for accurate predictions and effective risk management in agriculture.

Crop yield predictions conducted at various geographic scales offer significant advantages to a diverse set of stakeholders, ranging from farmers to policymakers [2]. Traditionally, farmers relied heavily on their own experiences to forecast crop yields. However, in contemporary times, rapid changes in environmental conditions and agricultural practices demand a more informed approach. The current scenario necessitates farmers to cultivate a wider variety of crops, yet many lack the requisite knowledge about these newer crops and remain unaware of the environmental factors that influence crop production. The solution to these challenges lies in the field of crop prediction, which can provide valuable insights and assistance.

Numerous computational intelligence techniques have found application in the realm of agriculture. This paper explores various ML methods employed for predicting crop yields. ML 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. ML 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, ML 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. Moreover, the classifications are detailed in fig.1.

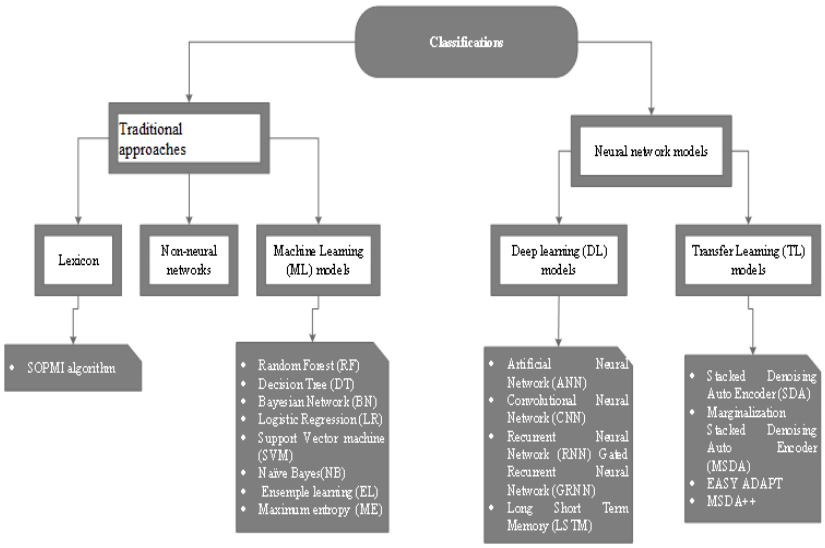

Fig.1 Classification Diagram

The main aim of this research is to explore machine-learning methods for predicting crop yields. This research aims to examine prevalent and well-known machine-learning methodologies that have played a crucial function in crop yield prediction. The subsequent organization of this paper is outlined as follows: section 2 presents the literature review of the proposed work and section 3 provides an overview of ML techniques in crop yield prediction. Section 4 explains the challenges and Section 5 serves as the conclusion of the paper.

2. Literature Review

The following sections provide a brief review of recent literature related to this topic

Prabakaran and colleagues [4] 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 (RBF), polynomial kernel function, and sigmoid function. Among these options, the RBF kernel demonstrated exceptional accuracy when compared to the others.

Gyamerah et.al. [5] introduced a probabilistic forecasting model, named QRF-E, to assess uncertainty in crop yield forecasts. This method integrates the Epanechnikov kernel function (E) and the Sheather and Jones (SJ) bandwidth selection technique with the quantile random forest (QRF) model to produce inclusive probability density curves for crop yield. Employing an appropriate bandwidth and kernel function in the proposed technique (QRF-E) enabled the comprehensive acquisition of conditional probability density for different time frames.

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. [6] carried out a research project centered on predicting potato yield using ML methods in conjunction with Sentinel 2 data. 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.

Li et al. [7] used the RReliefF feature selection algorithm to pinpoint the most influential features for crop yield prediction. RReliefF involves the utilization of probabilities, which are determined based on the relative distance between predicted values of two observations. In contrast to other feature selection methods that solely depend on statistical metrics, RReliefF takes into consideration the interrelationships among predictor variables, providing it with a unique advantage.

SVMs have hyper-parameters that need to be tuned for better performance. Shafiee and colleagues [8] employed a grid search technique to fine-tune the hyper-parameters of their model, specifically focusing on the "c" value and the kernel type. Grid search proved to be a valuable tool for enhancing model performance by systematically exploring various combinations of hyper-parameter values. The researchers create a pre-established list of values for different hyper-parameters, and the computer methodically evaluates the model's effectiveness for each set of these values. This process ultimately leads to the identification of the optimal hyperparameter values within the specified set.

Paidipati et.al. [9] investigated the prediction of Rice Cultivation in India using the Support Vector Regression (SVR) approach with different kernels to account for non-linear patterns. Obtaining an optimal configuration for the hyper-parameters involves the need for precise understanding and intuition, often achieved through an iterative trial-and-error process. Consequently, parameter tuning involves selecting values for the model's parameters to enhance its accuracy. The effectiveness of SVR is contingent upon the interrelated nature of the tuned parameters.

Iniyan and Jebakumar [10] developed a methodology called "Mutual Information Feature Selection (MIFS)" for predicting crop yield on both corn and soybean crops. Their approach involved employing a Multilayer Stacked Ensemble Regression 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.

Ramos et al. [11] developed an approach that employs random forest ranking to predict maize yield by utilizing UAV-based vegetation spectral indices. The study involved the individual ranking of vegetation indices (VIs) based on a merit metric, assessing the enhancement of Pearson's correlation coefficient through the application of the RF method compared to a baseline approach. Consequently, the RF model only took into account the most relevant VIs as input features.

Schwalbert et al. [12] proposed a method for predicting soybean yield based on satellite data, incorporating ML methods and weather information to improve crop yield prediction in southern Brazil. Random forests, known for their ease of training, insensitivity to outliers, computational efficiency, and resilience against overfitting, operate as an ensemble classifier. This method involves bootstrapping training samples and variables to generate multiple decision trees, followed by the aggregation of results from these individual trees for making predictions.

The feature selection step involves identifying pivotal attributes for forecasting crop yields. Random Forest can provide feature importances, which helps to select the most relevant variable. Gopal et al. [13] investigated different feature selection algorithms, such as 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.

Sakamoto and Toshihiro [14] showcased the enhancement of the VI-based crop yield finding process by integrating supplementary ecological factors (such as temperature, precipitation, Soil dampness, brief wave radiation, and statistical data associated with the proportion of farmland harvested through irrigation at the county level) using the random forest regression algorithm. The suggested method effectively minimized estimation errors and addressed concerns regarding fluctuations in the environment, leading to a notable increase in the method's estimation accuracy.

Obsie et.al. [15] elucidated a method for predicting the yield of natural blueberries using a combination of computer simulation and ML algorithms. The combination involved embedding the Random Forest (RF) with the Gradient Boosting method, known for its capability to reduce the variability of a statistical learning model. Additionally, the Random Forest model was utilized for the selection of significant predictors.

Cedric et.al. [16] 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. 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. [17] 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. Han et.al. [18] proposed a method for predicting winter wheat yield in China based on multi-source data and ML techniques. Over the past decade, Random Forest (RF) has demonstrated its effectiveness in managing datasets with numerous dimensions and mitigating overfitting. Furthermore, RF was capable of assessing the comparative significance of observed variables, making it a sound approach for the process of selecting variables. The advantages and disadvantages of different ML methods in crop yield prediction are outlined in Table 1.1,Table 1.2 and Table 1.3

Table 1.1 Advantages and Disadvantages of RF Techniques in crop yield prediction

Year	Author's Name	Techniques	Advantages	Disadvantages
2019	Gopal et.al. [13]	RF	The suggested method identifies the error minimum and enhances the accuracy of predictions	The computational duration of the suggested method was high
2020	Ramos et.al. [11]	RF	It delivers elevated precision in a cost-effective manner	In the field of precision agriculture, making decisions is a demanding task
2020	Schwalbert et.al. [12]	RF	This aids in predicting crop yield with a reasonable level of accuracy	Accurately estimating the financial specifics is difficult

2020	Sakamoto and Toshihiro [14]	RF	The information loss is minimal	The weather observation data is notably intricate, thereby introducing an additional layer of complexity to the analysis
2020	Obsie et.al. [15]	RF	The suggested method exhibited superior performance compared to alternative approaches	Acquiring extensive temporal and spatial datasets is both expensive and challenging
2020	Feng et.al. [17]	RF	The assignment is finished within a comparatively brief period	The prediction accuracy is not satisfactory
2020	Han et.al. [18]	RF	RF exhibited the most effective generalization capability	The prediction accuracy was influenced by varying agricultural zones and temporal training configurations
2022	Cedric et.al. [16]	RF	The execution requires a relatively short duration	Not well-suited for complex discoveries

Table 1.2 Advantages and Disadvantages of SVM Techniques in crop yield prediction

2020	Gyamerah et.al. [5]	SVM	It is capable of predicting non-parametric distribution	Acquiring authentic datasets poses a significant challenge
2020	Li et.al. [7]	SVM	The suggested method contributed to achieving a	Additional time is required to forecast the accuracy

			satisfactory level of prediction accuracy	
2021	Prabakaran and colleagues [4]	SVM	A reduced cost outlay is necessary	The proposed research focused solely on the design of the installed system, without considering the computer's speed and technical specifications
2021	Shafiee and colleagues [8]	SVM	The proposed approach exhibited a strong level of robustness	Insufficient analysis was conducted on the most significant vegetation indices for the prediction of grain yield
2021	Paidipati et.al. [9]	SVM	The suggested approach can analyze data from various dimensions, revealing diverse patterns	Fine-tuning the hyper-parameters requires careful attention due to their sensitivity
2022	Iniyan and Jebakumar [10]	SVM	The lowest error rate was observed during the execution process	Improving crop yield prediction poses a significant challenge

Table 1.3 Advantages and Disadvantages of Combination of RF and SVM Techniques in crop yield prediction

Year	Author's Name	Techniques	Advantages	Disadvantages
2019	Gómez et.al. [6]	SVM, RF	The SVM model demonstrated superior performance when no feature selection technique was applied, while the RF model achieved results that were deemed satisfactory.	A substantial quantity of samples is necessary in the initial dataset to attain a more resilient outcome.

When predicting the crop yields, SVM can be used to examine past crop data and forecast yields based on a range of input parameters, including crop types, weather, soil quality, and agricultural practices. Large datasets may provide challenges for SVM, as it may be sensitive to the selection of the kernel function and hyperparameters. Consequently, ML algorithms that are used to help computers recognise patterns in data and anticipate outcomes

without explicit programming. Also, ML techniques are used to estimate crop yields by training models on past crop data in order to find patterns and correlations between different elements that affect crop yields. However, the quality and quantity of training data greatly influence the performance of ML models, which may necessitate considerable feature engineering. In order to estimate crop yields, a variety of characteristics, including crop types, weather patterns, soil properties, and agricultural methods, can be analysed using RF and Transfer learning (TL) strategies. As a result of their ability to analyse historical data and find patterns and links between many elements influencing crop growth and production, SVM, ML, TL and RF algorithms are commonly employed to estimate agricultural yields.

3. Methodology

There are multiple ML methods are utilized to predict crop yields. But, this paper spotlighted two commonly used techniques SVM and Random forest. SVM are powerful ML techniques for crop yield prediction, especially when dealing with complex, non-linear relationships in the data. Similarly, 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 both ML techniques are successfully reviewed in the above section. In this review, the radial bias function (RBF) was performed as the best kernel selection method to avoid kernel selection problem. Similarly, rqlasso and LeapBack techniques performed as the best feature selection methods. In addition the tuning of hyper parameter problems was cleared by using Grid search algorithm. Also, MIFS-based Multi-layer Stacked Ensemble Regression model helped to achieve optimum crop yield prediction accuracy. The both ML techniques provide lot of benefits while applied in the field of crop yield prediction; however, it presents several obstacles that must be tackled to ensure precise and dependable forecasts. By applying ML techniques, the obtaining of high-quality and comprehensive data on crop yields, weather, soil, and other relevant factors can be challenging. Also, in some regions, historical data may be limited or unreliable, posing difficulties in training models with high accuracy. Similarly, the ML techniques creates more complexity in the process of identifying the most relevant features and transforming them into suitable representations. Also, selecting the right ML algorithm and optimizing its hyper-parameters can be time-consuming and may require expertise. In addition, the models trained on data from one region or period may not generalize well to different regions or years with varying conditions. Ensuring model generalization is a significant challenge. Similarly, crop yield predictions may need to account for local variations, such as micro climates or soil types within a region, which can have a significant impact on outcomes.

Transfer Learning (TL) can potentially improve the performance of crop yield prediction compared to Random Forest (RF) and Support Vector Machine (SVM) in several ways:

Data Efficiency: Crop yield prediction models often suffer from limited labeled data, especially in specific regions or for certain crops. TL can leverage knowledge gained from related tasks or domains, allowing the model to be more data-efficient. Pre-trained models on larger datasets related to agriculture or environmental conditions can capture useful features and patterns that benefit crop yield prediction.

Domain Adaptation: Transfer Learning is particularly effective in handling domain shifts. Agriculture and environmental conditions can vary across regions and seasons, leading to changes in data distribution. TL can adapt the knowledge learned from one set of conditions to another, improving the model's ability to generalize across diverse environments.

Feature Extraction: TL often involves feature extraction from pre-trained models. In the context of crop yield prediction, this means that the model can automatically learn relevant features from a source domain (e.g., general agricultural data) and transfer this knowledge to the target domain (specific crop yield prediction). This can be advantageous when extracting meaningful features manually is challenging.

Improved Initialization: TL often involves using pre-trained models as initializations for the target task. This can accelerate the convergence of the learning process, especially when compared to starting with randomly initialized

models. Faster convergence can be crucial in applications like crop yield prediction, where timely insights are valuable.

Adaptation to New Crops or Regions: Crop yield prediction models often need to adapt to different crops or regions with varying characteristics. TL allows the model to transfer knowledge from one crop or region to another, providing a foundation for adaptation and potentially enhancing predictive performance.

Handling Seasonal Variability: Crop yield prediction is influenced by seasonal changes and environmental conditions. TL can capture seasonal patterns and variations in the source domain, contributing to better generalization and prediction in the target domain.

4. Discussion

TL techniques for predicting crop yields involve leveraging knowledge from pre-trained models, adapting features, fine-tuning, and sharing information learned from related tasks or domains. These approaches can help improve model performance, especially when dealing with limited labeled data in the specific crop yield prediction domain.

Table.2 Performance metrics

Sl.no	Parameters / Techniques	RMSE	MAE
1	SVM, RF	0.93	8.64
2	RF	0.99	0.041
3	SVM	0.90	0.16

5. Conclusion

This article offers a comprehensive review of the literature on crop yield prediction using ML approaches. This paper successfully reviewed two ML techniques commonly employed in crop yield prediction, including methodologies like Support Vector Machine and Random Forest. In this survey, the proposed ML techniques provided a maximum accuracy with a correlation coefficient and MAE (Mean Absolute Error) of 0.77 and 852.13 kg ha−1. As a review result, the proposed ML techniques provide optimum prediction accuracy and robustness performance when compared to other techniques. Moreover, TL models have offer advantages such as improved generalization, reduced training time, effective feature extraction, robustness to data scarcity, adaptability to domain shifts, integration of domain-specific knowledge, and interpretability. These merits make TL an attractive approach for crop yield prediction and other agricultural applications.

Compliance with Ethical Standards

Conflict of interest

The authors declare that they have no conflict of interest.

Human and Animal Rights

This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed Consent

Informed consent does not apply as this was a retrospective review with no identifying patient information.

Funding: Not applicable

Conflicts of interest Statement: Not applicable

Consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and material:

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Code availability: Not applicable

Competing Interests: Not applicable

References

1. Pant, Janmejay, 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. "MLfor regional crop yield forecasting in Europe." *Field Crops Research* 276 (2022): 108377.
3. Paudel, Dilli, et al. "MLfor large-scale crop yield forecasting." *Agricultural Systems* 187 (2021): 103016.
4. 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.
5. Gyamerah, Samuel Asante, Philip Ngare, and Dennis Ikpe. "Probabilistic forecasting of crop yields via quantile random forest and Epanechnikov Kernel function." *Agricultural and Forest Meteorology* 280 (2020): 107808.
6. Gómez, Diego, et al. "Potato yield prediction using MLtechniques and sentinel 2 data." *Remote Sensing* 11.15 (2019): 1745.
7. Li, Bo, et al. "Above-ground biomass estimation and yield prediction in potato by using UAV-based RGB and hyperspectral imaging." *ISPRS Journal of Photogrammetry and Remote Sensing* 162 (2020): 161-172.
8. 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.
9. Paidipati, Kiran Kumar, et al. "Prediction of rice cultivation in india—support vector regression approach with various kernels for non-linear patterns." *AgriEngineering* 3.2 (2021): 182-198.
10. Iniyan, 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.
11. Ramos, Ana Paula Marques, et al. "A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices." *Computers and Electronics in Agriculture* 178 (2020): 105791.
12. Schwalbert, Raí A., et al. "Satellite-based soybean yield forecast: Integrating MLand weather data for improving crop yield prediction in southern Brazil." *Agricultural and Forest Meteorology* 284 (2020): 107886.
13. Gopal, PS Maya, and R. Bhargavi. "A novel approach for efficient crop yield prediction." *Computers and Electronics in Agriculture* 165 (2019): 104968.
14. Sakamoto, Toshihiro. "Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm." *ISPRS Journal of Photogrammetry and Remote Sensing* 160 (2020): 208-228.
15. Obsie, Efrem Yohannes, Hongchun Qu, and Francis Drummond. "Wild blueberry yield prediction using a combination of computer simulation and MLalgorithms." *Computers and Electronics in Agriculture* 178 (2020): 105778.
16. Cedric, Lontsi Saadio, et al. "Crops yield prediction based on MLmodels: Case of West African countries." *Smart Agricultural Technology* 2 (2022): 100049.
17. Feng, Puyu, et al. "Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and MLtechnique." *Agricultural and Forest Meteorology* 285 (2020): 107922.
18. Han, Jichong, et al. "Prediction of winter wheat yield based on multi-source data and MLin China." *Remote Sensing* 12.2 (2020): 236.