WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning Models for Accurate Crop Yield Prediction: A Comprehensive Study and Comparative Analysis

Mrs. Ramya B N[1], Sathvik Venkatesh K[2], Vikram Simha Reddy[3]

[1]*Department of AIML, Jyothy Institute of Technology, ramyabn@jyothyit.ac.in*
[2](1JT20AI037), *Department of AIML, Jyothy Institute of Technology, 1jt20ai037@jyothyit.ac.in*
[3](1JT20AI052), *Department of AIML, Jyothy Institute of Technology, 1jt20ai052@jyothyit.ac.in*

ABSTRACT –

Machine learning plays a pivotal role in aiding decisions related to crop yield prediction, guiding choices in crop selection and management throughout the growing season. Numerous machine learning algorithms have been utilized for this purpose. This study conducts a Systematic Literature Review (SLR) to systematically analyze and synthesize algorithms and features employed in crop yield prediction research. Our search, encompassing six electronic databases, yielded 567 relevant studies, from which 50 were selected based on predefined criteria. Thorough investigation and analysis of these studies revealed prevalent features such as temperature, rainfall, and soil type, with Artificial Neural Networks being the most utilized algorithm. Additionally, we conducted a supplementary search focusing on deep learning, identifying 30 relevant papers. The primary deep learning algorithm found was Convolutional Neural Networks (CNN), followed by Long-Short Term Memory (LSTM) and Deep Neural Networks (DNN). This comprehensive analysis provides insights and suggests avenues for future research in crop yield prediction using machine learning.

INTRODUCTION

Machine learning (ML) techniques have found diverse applications across numerous domains, from retail, where they analyze customer behavior (Ayodele, 2010), to telecommunications, predicting phone usage patterns (Witten et al., 2016). Agriculture has also embraced ML for several years (McQueen et al., 1995), particularly in the complex realm of precision agriculture, where crop yield prediction poses a significant challenge. Multiple models have been proposed and validated to tackle this intricate problem, as it involves diverse datasets encompassing factors like climate, weather, soil conditions, fertilizer usage, and seed variety (Xu et al., 2019). Evidently, crop yield prediction entails intricate processes and is far from being a straightforward task, with ongoing efforts aiming for enhanced predictive performance (Filippi et al., 2019a).

As a subset of Artificial Intelligence (AI) focusing on learning, ML offers a practical approach to achieving more accurate yield predictions based on various features. ML excels in recognizing patterns, establishing correlations, and extracting insights from datasets. The training of ML models involves using historical data to determine parameters, while the testing phase evaluates model performance on unseen data.

ML models can be either descriptive, aimed at understanding and explaining past events, or predictive, focusing on forecasting future outcomes (Alpaydin, 2010). Constructing high-performance predictive models in ML studies poses unique challenges, requiring careful algorithm selection tailored to the specific problem, and platforms capable of handling voluminous datasets.

To comprehensively explore the application of ML in crop yield prediction, a systematic literature review (SLR) was conducted. An SLR identifies research gaps and guides both practitioners and researchers in undertaking new studies. Adhering to a methodological framework, relevant studies were gathered from electronic databases, synthesized, and presented to address predefined research questions. Objectivity and transparency are crucial in SLR studies, ensuring replicability and a systematic coverage of all existing literature.

This paper proceeds as follows: Section 2 provides background information, Section 3 outlines the methodology, and Section 4 presents the SLR results. Section 5 delves into deep learning-based crop yield prediction research, followed by Section 6 for discussion. Finally, Section 7 concludes the paper, summarizing the empirical findings and responses to the research questions outlined in this review article.

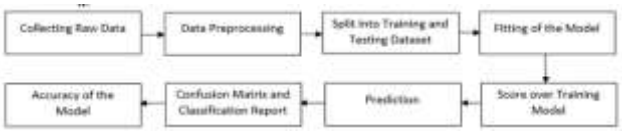
I. LITERARURE SURVEY

Crop yield prediction holds paramount importance for decision-makers at national and regional levels, facilitating rapid and informed decision-making, such as at the EU level. An accurate prediction model aids farmers in making crucial decisions regarding crop selection and optimal planting times. This literature review explores the utilization of machine learning in crop yield prediction, analyzing the existing body of work.

In our scrutiny of publications, one exclusion criterion was the omission of survey or traditional review papers, which are discussed in this section as related works. Chlingaryan and Sukkarieh conducted a review on nitrogen status estimation using machine learning, foreseeing cost-effective solutions in agriculture through advancements in sensing technologies and ML techniques. Elavarasan et al. surveyed machine learning models associated with crop yield prediction based on climatic parameters, emphasizing the need for comprehensive parameter consideration. Liakos et al. published a review on the broader application of machine learning in agriculture, covering crop, livestock, water, and soil management. Li, Lecourt, and Bishop focused on fruit ripeness determination and yield prediction. Mayuri and Priya addressed challenges and methodologies in image processing and machine learning for disease detection in agriculture. Somvanshi and Mishra discussed machine learning approaches in plant biology. Gandhi and Armstrong's review on data mining in agriculture highlighted the need for further research in implementing data mining into complex agricultural datasets. Beulah's survey explored various data mining techniques for crop yield prediction, concluding that data mining holds promise in solving this problem.

This study stands out as the first systematic literature review (SLR) concentrating on machine learning in crop yield prediction, distinct from existing surveys that often focus on specific aspects of the problem. Notably, we presented 30 deep learning-based studies in this article, shedding light on the specific deep learning algorithms employed in these studies.

II.METHODOLOGY



1. Collecting the Raw Data

The practice of cumulating and scrutinizing data from different sources is known as data collection. keep track of past occurrences so that one can utilize da patterns. The ‘Crop Recommendation’ dataset is collected from the Kaggle website. The dataset takes into account 22 different crops as class labels and 7 features- (i) Nitrogen content ratio ii) Phosphorus content ratio (P) (iii) Potassium content ratio (K) in the soil, v) Percentage of Relative Humidity (vi) ph value and different sources is known as data collection. keep track of past occurrences so that one can utilize data analysis to detect patterns. The ‘Crop Recommendation’ dataset is collected from the Kaggle website. The i) Nitrogen content ratio tio (K) in the soil, (iv) Temperature vi) ph value and (vii) Rainfall . The ‘Crop Recommendation’ dataset is collected from the Kaggle website. The dataset takes into account 22 different crops (N) (ii) Phosphorus content ratio (P) expressed in degree Celsius (v) Percentage of Relative Humidity measured in millimeters.

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice
...	...	...	...	...	...	...	...	...
2195	107	34	32	26.774637	66.413269	6.780064	177.774507	coffee
2196	99	15	27	27.417112	56.636362	6.086922	127.924610	coffee

2. Data Preprocessing

The process of modifying raw data into a form that analysts and data scientists can use in machine learning algorithms to find insights or forecast outcomes is called Data preprocessing. In this project the data processing method is to find missing values. Getting every data point for every record in the dataset is tough. Empty cells, values like null or a specific character, such as a question mark, might all indicate that data is missing. The dataset used in the project didn’t have any missing values.

3.Train and Test Split

It is a process of splitting the dataset into a training dataset and testing dataset using train_test_split() method of scikit learn module. 2200 data in the dataset has been divided as 80% of a dataset into training dataset-1760 and 20% of a dataset into testing dataset-440 data.

4. Fitting the model

Modifying the model’s parameters to increase accuracy is referred to as fitting. To construct a machine learning model, an algorithm is performed on data for which the target variable is known. The model’s accuracy is determined by comparing the model's outputs to the target variable's actual, observed values. Model fitting is the ability of a machine learning model to generalize data comparable to that with which it was trained. When given unknown inputs, a good model fit refers to a model that properly approximates the output.

5.Checking the score over a training dataset

Scoring, often known as prediction, is the act of creating values from new input data using a trained machine learning model. Using model.score() method calculating the score of each model over a training dataset shows how well the model has learned.

6.Predicting the model

When forecasting the likelihood of a specific result, "prediction" refers to the outcome of an algorithm after it has been trained on a previous dataset and applied to new data. Predicting the model using predict() method using test feature dataset. It has given the output as an array of predicted values.

7. Confusion Matrix and Classification Report

Confusion Matrix and Classification Report are the methods imported from the metrics module in the scikit learn library that are calculated using the actual labels of test datasets and predicted values. Confusion Matrix gives the matrix of frequency of true negatives, false negatives, true positives and false positives. Classification Report is a metric used for evaluating the performance of a classification algorithm's predictions. It gives three things: Precision, Recall and f1-score of the model. Precision refers to a classifier's ability to identify the number of positive predictions which are relatively correct. It is calculated as the ratio of true positives to the sum of true and false positives for each class.

$$Precision = \frac{TP}{TP + FP}$$

Where Precision-Positive Prediction Accuracy; TP-True Positive; FP-False Positive Recall is the capability of a classifier to discover all positive cases from the confusion matrix. It is calculated as the ratio of true positives to the sum of true positives and false negatives for each class.

$$Recall = \frac{TP}{TP + FN}$$

Where Recall- The percentage of positives that were correctly identified; FN-False Negative F1 score is a weighted harmonic mean of precision and recall, with 0.0 being the worst and 1.0 being the best. Since precision and recall are used in the computation, F1 scores are often lower than accuracy measurements.

$$F1\ score = \frac{2 * PR}{(P + R)}$$

The number of correct predictions divided by the total number of predictions accuracy. The accuracy of the mod metrics module.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

IV.RESULT

1. Predictive skill

We begin by comparing the accuracy of the various approaches in predicting yields in years that were not used to train the model; . The accuracy of the parametric model and the SNN was substantially improved by bagging, but the bagged SNN performed best. The fully-nonparametric neural net—which was trained identically to the SNN but lacked parametric terms—performed substantially worse then either the OLS regression or the SNN.

Model	Bagged	$\widehat{MSE}_{oob}$
Parametric	No	367.9
Semiparametric neural net	No	292.8
Parametric	Yes	334.4
Fully-nonparametric neural net	Yes	638.6
Semiparametric neural net	Yes	251.5

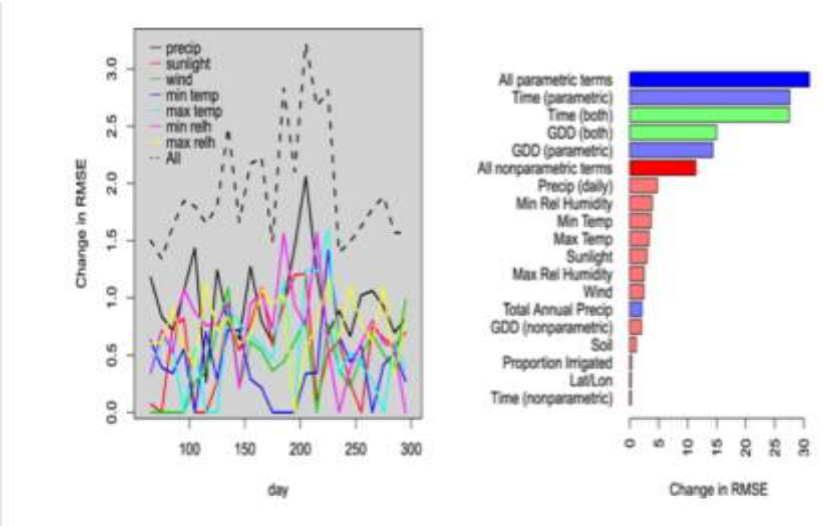
That bagging improves model fit—of both the OLS regression and the SNN—implies that certain years may have served as statistical leverage points, and as such that un-bagged yield models may overfit the data. This is because there are too few distinct years of data to determine whether the heat of an anomalously hot year is in fact the cause of that year's anomalously low yields. If bootstrap samples that omit such years estimate different relationships, then averaging such estimates will reduce the influence of such outliers.

That the SNN and the OLS regression both substantially out-perform the fully-nonparametric neural net is simply reflective of the general fact that parametric models are more efficient than nonparametric models, to the degree that they are correctly specified. That the SNN is more accurate than the OLS regression—but not wildly so—implies that model (1) is a useful but imperfect approximation of the true underlying data-generating process.

2.Variable importance

It can be desirable to determine which variables and groups of variables contribute most to predictive skill. Importance measures were developed in the context of random forests (Breiman 2001). Applied to bagged estimators, these statistics measure the decline in accuracy when a variable or set of variables in the out-of-bag sample is randomly permuted. Random permutation destroys their correlation with the outcome and with variables with which they interact, rendering them uninformative. We compute these measures for each set of variables as the average MSE difference across five random permutations.

These are plotted in the daily weather variables, those measured in mid-summer are most important by this measure, particularly daily precipitation and minimum relative humidity during the warmest part of the year. The later is consistent with findings of Anderson et al (2015) and Lobell et al (2013), who show that one of the main mechanisms by which high temperatures affect yield is through their influence on water demand.

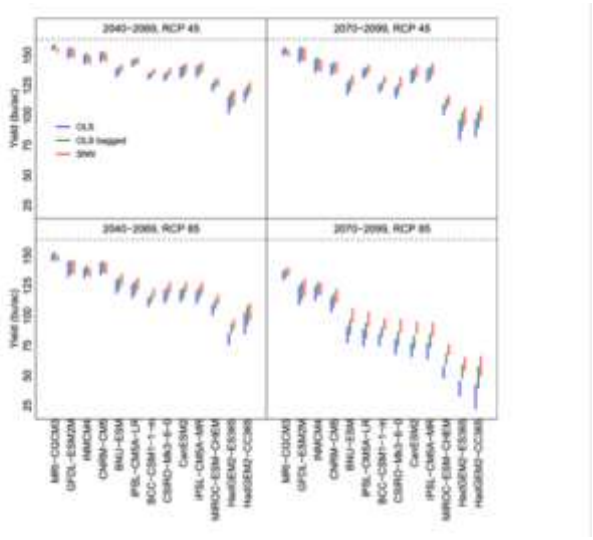


The parametric portion of the model is more important than the nonparametric part of the model (figure 2, right)—mostly through the incorporation of parametric time trends, but also through GDD. The nonparametric component of the model—taken as a whole—is less important than the parametric representation of GDD, though nonetheless responsible for the improvement in predictive skill over the baseline OLS regression. However, the predictive power of the time-varying temperature variables lends support to the findings of Ortiz-Bobea (2013), who notes that there is room for improvement in the additive separability approximation implicit in the baseline parametric specification. Soil variables, proportion of land irrigated, and geographic

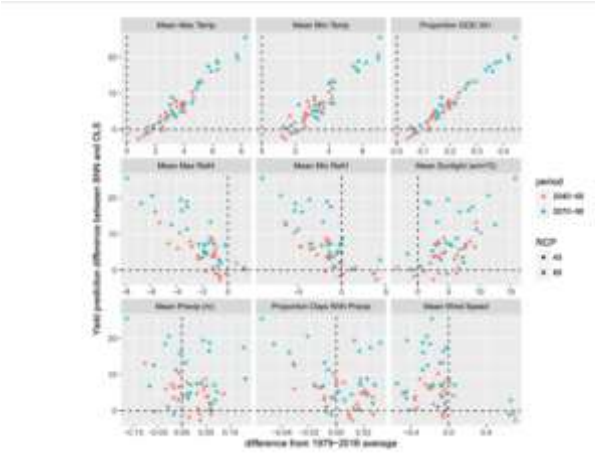
coordinates have low importance values. It is possible that these variables would be more important in a model trained over a larger, less homogeneous area—allowing them to moderate and localize the effects of daily weather variables. The centrality of time trends to predictive skill explains the poor performance of the fully-nonparametric neural net.

3. Projections

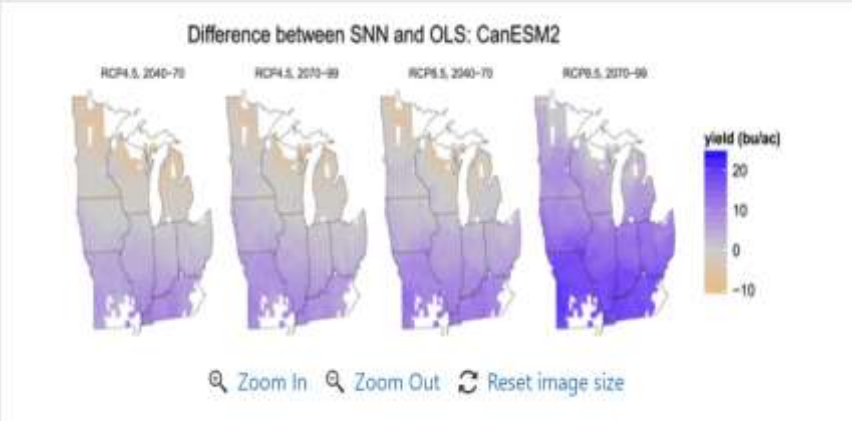
Projections for the periods 2040–2069 and 2070–2099, for both RCPs and all climate models, are reported in figure 3. Plotted confidence intervals are derived by averaging the pointwise standard errors of a smoothing spline applied to the underlying time series of projected yields, and as such reflect projected interannual variability. These projections make no assumptions about technological change, inasmuch as the value of the time trend is fixed at the year 2016 for the purpose of prediction. As such, projections only reflect change in response to weather, assuming 2016s technology and response to weather. Nor do we model adaptation or carbon fertilization. Omission of these factors is likely to bias statistical projections downwards, to the degree that these factors will increase expected yields, though we note that recent research raises uncertainty about the magnitude or existence of the CO2 fertilization effect (Obermeier et al 2018).



While there is little difference between the models in scenarios in which yields decline less, the SNN projects substantially less-severe impacts in scenarios where yields decline the most across all models, including most of the models in 2070–99 under RCP8.5. Notably, the bagged OLS specification is nearly always less pessimistic than the standard OLS model. It is likely that the pessimism of OLS relative to bagged OLS derives from a relatively small number of severe years in the historical period affecting model estimates by serving as outliers, in a manner which is diluted by the bootstrap aggregation process.



The spatial distribution of these projections for the Canadian Earth System Model—which is roughly in the middle of projected severity of yield impact, out of our suite of models—is presented in figure 5. While OLS specifications are generally more pessimistic, they project increases in the northernmost regions of our study area. SNN projections do not share this feature, though they are less pessimistic overall.



5. Discussion

On average, yield impacts projected by the neural net—in response to future weather scenarios simulated by global climate models—are somewhat less severe than those projected using parametric models. Nonetheless, these estimates are still among the more severe estimates for temperate-zone corn compared to the studies compiled for the IPCC 5th Assessment Report (Porter et al 2013). It is worth emphasizing that the difference in yield projections between the statistical approaches considered here is not as large as the difference in yield projections between climate models and emissions scenarios.We find that the timing of heat and moisture are important to predicting corn yields, along with the simple accumulation of heat. This is seen both in the less-pessimistic projections in the south of our study area, and the lack of positive responses to increasing warmth in the north of our study area. As such, we find that GDDs are a useful but imperfect proxy for the role of heat in predicting crop yield. Indeed, work has indicated important roles for VPD and soil moisture (Roberts et al 2012, Lobell et al 2013, Anderson et al 2015, Urban et al 2015) in explaining and building upon the baseline parametric specification. The complexity of these underlying response mechanisms is an argument for the explicit use of GCMs in assessing climate change impacts on agriculture, which explicitly capture the co-evolution of multiple climate variables.

While deep learning has led to substantial breakthroughs in predictive applications and artificial intelligence, classical statistical methods will remain central to scientific applications that seek to elucidate mechanisms governing cause and effect. We describe a semiparametric approach that fuses the two and works better than either alone in terms of predictive performance. This approach is suitable for any prediction problem in which there is some—potentially imperfect—prior knowledge about the functions mapping inputs to outcomes, and longitudinal or other structure in the data.

Ultimately, we find that combining ML with domain-area knowledge from empirical studies improves predictive skill, while altering conclusions about climate change impacts to agriculture. There is substantial scope to refine and extend this work, along four major avenues: (1) better representations of domain-area knowledge in the parameterization of the parametric component of the model, (2) extension to wider geographic areas, which will require more-explicit treatment of differences in seasonality of production over space, (3) bringing the nonparametric (neural network) part of the model closer to the research frontier in ML and artificial intelligence, and (4) finding ways to integrate elements from deterministic crop models that have heretofore been challenging to model statistically, such as CO2 fertilization.

V.CONCLUSION

A comparative analysis was conducted on three distinct supervised machine learning models (KNN, Decision Tree, and Random Forest) to determine the most suitable crop for specific land, aiding farmers in more efficient crop cultivation. Our findings revealed that the Random Forest Classifier exhibited the highest accuracy of 99.32% in predicting crop types within the crop prediction dataset, both in terms of Entropy and Gini Criterion. Conversely, the K-Nearest Neighbor model demonstrated the lowest accuracy at 97.04%, while the Decision Tree Classifier fell in between KNN and Random Forest Classifier in terms of accuracy. Notably, the Decision Tree Gini criterion outperformed the Decision Tree Entropy Criterion, achieving an accuracy of 98.86%. Future prospects involve gathering additional field data to enhance soil characterization and exploring the integration of other machine learning and deep learning algorithms such as ANN or CNN to broaden the classification spectrum for various crop varieties.

VI. ACKNOWLEDGEMENT

I would like to express my appreciation to my mentors, colleagues, and peers who have provided guidance, insights, and support throughout the exploration of this topic. Their feedback, discussions, and collaborations have been invaluable in shaping my understanding and enhancing the quality of the work.

The progress made was a collective effort, and it is through the contributions and collaboration of numerous individuals and institutions that we have been able to deepen our understanding and achieve advancements in this field.

VII. REFERENCES

[1] Rao, M.V., Sreeraman, Y., Mantena, S.V., (...), Roja, D., Vatambeti, R. (2024) Brinjal crop yield prediction using shuffled shepherd optimization algorithm based ACNN-OBDLSTM model in smart agriculture. *Journal of Integrated Science and Technology.*

[2] Chandar, A.G., Sivasankari, K., Lakshmi, S.L., (...), Kannadhasan, S., Balakumar, S. (2024) An innovative smart agriculture system utilizing a deep neural network and embedded system to enhance crop yield. *Multidisciplinary Science Journal.*

[3] Zhao, L., Qing, S., Wang, F., (...), Shi, Y., Cui, N. (2023) Prediction of Rice Yield Based on Multi-Source Data and Hybrid LSSVM Algorithms in China. *International Journal of Plant Production.*

[4] Li, C., Ren, X., Zhao, G. (2023) Machine-Learning-Based Imputation Method for Filling Missing Values in Ground Meteorological Observation Data. *Algorithms.*

[5] Garai, S., Paul, R.K., Rakshit, D., (...), Tashkandy, Y., Chesneau, C. (2023) Wavelets in Combination with Stochastic and Machine Learning Models to Predict Agricultural Prices. *Mathematics.*

[6] Guo, Y. (2023) Integrating genetic algorithm with ARIMA and reinforced random forest models to improve agriculture economy and yield forecasting. *Soft Computing.*

[7] Sharma, P., Dadheech, P., Aneja, N., Aneja, S. (2023) Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning. *IEEE Access.*

[8] Bansal, Y., Lillis, D., Kechadi, M.-T. (2023) A Deep Learning Model for Heterogeneous Dataset Analysis - Application to Winter Wheat Crop Yield Prediction. *Communications in Computer and Information Science.*

[9] Sikandar, S., Mahum, R., Aladhadh, S. (2023) Automatic Crop Expert System Using Improved LSTM with Attention Block. *Computer Systems Science and Engineering.*

[10] Sharma, P., Dadheech, P., Senthil, A.V.S.K. (2023) AI-Enabled Crop Recommendation System Based on Soil and Weather Patterns (Book Chapter). *Artificial Intelligence Tools and Technologies for Smart Farming and Agriculture Practices.*

[11] Kalinaki, K., Shafik, W., Gutu, T.J.L., Malik, O.A. (2023) Computer Vision and Machine Learning for Smart Farming and Agriculture Practices (Book Chapter). *Artificial Intelligence Tools and Technologies for Smart Farming and Agriculture Practices.*

[12] Qazi, U.K., Ahmad, I., Minallah, N., Zeeshan, M. (2023) Classification of tobacco using remote sensing and deep learning techniques. *Agronomy Journal.*

[13] Meeradevi, Mundada, M.R. (2023) Optimized Farming: Crop Recommendation System Using Predictive Analytics. *International Journal of Intelligent Engineering and Systems.*

[14] Wu, J., Zhao, F. (2023) Machine learning: An effective technical method for future use in assessing the effectiveness of phosphorus-dissolving microbial agroremediation. *Frontiers in Bioengineering and Biotechnology.*