International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23s, 2025
https://theaspd.com/index.php

Prediction Of Suitable Crop For Cultivation Using Ensemble Machine Learning

Vaishali Kadwey[1], Anil Kumar Gupta[2]
[1]Research Scholar, Department of Computer Science & Applications, Barkatullah University Bhopal
vaishali2498@gmail.com
[2]Department of Computer Science & Applications, Barkatullah University Bhopal
akgupta_bu@yahoo.co.in

Abstract
In India Agriculture is influenced by the diverse climatic conditions and varied land types, which leads to diversity in cropping patterns and farming methods. In agriculture, cropping patterns depend on several factors like seasons, soil type, climate, and water availability. Due to variations in environmental and agronomic parameters; the selection of best crop for cultivation is a major challenge for farmers. The farmer often faces difficulties in selecting the right crop for sowing, and needs expert's advice for achieving maximum yield. The integration of agricultural practices with Machine Learning (ML) offers a transformative approach to tackle this challenge. The prediction of best suitable crop plays a significant role in optimizing agricultural output, ensures food security, and manages resources efficiently. This research involves developing hybrid models (HVM1, HSM2, HVM3, HSM4) using diverse classification and regression ML models such as RF, DT, KNN, SVM, LogRes and Stacking and Voting ensemble techniques. The trained models used for predicting best suitable crop to be grown for particular environmental and soil conditions. The potential of Machine Learning made revolution in crop prediction by providing robust, accurate, and actionable forecasts. The HSM4 model developed using diverse classification models and ensemble stacking technique shows the excellent results with 99.4% accuracy. By leveraging HSM4 model, this research help farmers in optimizing resource use, reducing input costs, and improve agricultural productivity and sustainability.
Keywords: *Machine Learning, Regression, Classification, crop recommendation, agricultural sustainability.*

1. INTRODUCTION

The convergence of agriculture with technology has provides the inveterate way for advancements in crop yield prediction. The emergence of machine learning algorithms in agriculture presents a transformative approach to addressing the multifaceted challenges like climate variability, soil health, and efficient resource utilization. Achieving the maximum crop is a vital component of agricultural management and planning. An accurate crop prediction enables farmers to make informed decisions about planting schedules, crop selection and resource allocation. Furthermore, crop yield forecasts are instrumental in mitigating risks associated with extreme weather and market fluctuations, thereby enhancing the resilience as well as sustainability in agricultural. By leverage the strength of machine learning and adopting data-driven techniques, the traditional methods used for farming can be upgraded. In agriculture, ML algorithms can analyze diverse data sources, including soil characteristics, weather conditions, crop health indicators, and historical yield records, to build predictive models with high accuracy and reliability. The application of ML in crop recommendation not only improves forecast precision but also facilitates the identification of key factors influencing crop performance, thereby guiding targeted interventions to enhance productivity.

1.1. Application of Machine Learning in Agriculture

Machine learning (ML) techniques have a significant role in the agriculture sector, it transform traditional farming into smart agriculture. It can be analyze large amount of agricultural data collected from various sources using drones, satellite imagery and sensors and identify meaningful patterns and relationships between various factors influencing growth of the crop. The machine learning has number of application in agriculture sector, such as monitor Soil health, detects crop diseases and pest infestations, suggest irrigation needs, predict crop yield and help farmers in schedule planting and harvesting etc. This study focused on the development of hybrid model. The trained hybrid models developed using regression and classifications methods (logistic regression, RF, KNN, DT, SVC) are able to predict best suitable crop for cultivation to achieve optimal crop yield with significant accuracy. The integration of machine learning in agriculture thus

enhances productivity, sustainability, and resilience against climatic variations, ultimately this research is contributing to food security and economic stability.

2 REVIEW OF LITERATURE

The Researchers have been exploring the use of various data sources, related agriculture. The machine learning has been used to draw meaningful insights, develop models for predicting crop yield. Several studies investigated have been given promising results.

[1], the researchers employed machine learning techniques Polynomial Regression, Random forest and Support Vector Regressor to predict potatoes and maize crop in Irish. The Random Forest model has been observed as the most effective model, with exhibiting RMSE 510.8, 129.9 and the R2 values 87.5 and 81.7 percent accuracy for potato and maize crop respectively. [2] Proposed study forecast the yields of six crops maize, rice, cassava, yams, bananas, seed cotton across West African countries. This research they integrated climatic information, chemical data, agricultural yields to facilitate predictions for farmers and decision-makers. The researchers employed multivariate logistic regression, decision tree, and k-nearest neighbor models, implementing hyper-parameter tuning during cross-validation to mitigate over fitting. The decision tree model perform well with R2 of 95.3%, while the logistic regression and K-Nearest Neighbor models exhibited R2 values of 89.78% and 93.15% , respectively. The primary focus of [3], on predicting crop yield , dataset contain information about rainfall, temperature, fertilizer and past crop yield. The Ensemble techniques such as XGBoost Random Forest, XGBoost-RF and Gradient Boosting have been used for yield prediction. The Ensemble XGBoost-RF achieves highest accuracy 97% and 0.00216 MSE.

[4] Undertook research to estimates the crop yields in the Rajasthan for identified five crops only using various machine learning algorithms, including SVM, long short-term memory, Random Forest, Gradient Descent, and Lasso regression techniques. The findings revealed that the Random Forest algorithm outperformed others, achieving a notable MAE (.0251) RMSE (0.035) and R2 score 0.963. The cross-validation techniques have been used for results validation. The research [5] focuses on predicting crop yield through various machine learning techniques, comparing algorithms such as, Decision Tree, K-Nearest Neighbor (KNN), Random Forest Classifier using Gini and Entropy criteria. The most accurate, result for selecting suitable crops based on climatic conditions and soil nutrients have been given by Random Forest. The study aims to investigate the optimal model for crop prediction. The study, [6]predicted regional crop yields for five crops in three European countries. Performance comparisons with a baseline method highlighted comparable early-season predictions but revealed room for improvement, suggesting the incorporation of new data sources, enhanced features, and diverse machine learning algorithms to refine large-scale crop yield forecasting. Both studies underscore the pivotal role of machine learning in revolutionizing agricultural decision-making and forecasting for optimal crop management and environmental sustainability.

[7] Discussed the influence of climate change on agriculture in India over the last two decades and highlights the need for predicting crop yield in advance to aid policymakers and farmers. The proposed solution involves creating an interactive web based prediction system by employing the Random Forest algorithm. This system aims to help farmers make informed decisions by providing crop yield predictions based on various climatic and soil factors. [8] The study used machine linearing regression techniques such as linear regression for of the agriculture data and achieving optimal parameters to maximize the crop production. [9] This study explored the application of statistical machine learning techniques for crop yield prediction and has aim to predict crop yield using methods such as decision trees, linear regression, and random forests. The data used in this study are taken from weather sensors, satellite imagery, soil moisture sensors, and other sources. The random forest model has exhibiting superior performance with 92.5% accuracy. The research highlighted the potential of combining sensing technologies and machine learning for cost-effective and comprehensive solutions in crop yield prediction, emphasizing the ongoing advancements in these areas. In[10], the growth period of barley in one of Iran's major production areas was divided into three parts. The research evaluated the performance of a model integrating field, meteorological and remote sensing data, and, employing machine learning methods. The Gaussian regression observed the most effective model, estimating barley yield with 0.84 R2 score. In [11] the authors introduced a model incorporating machine learning algorithms such as Artificial Neural Network, Decision Tree, and Random Forest to determine the best crop, incorporating deep learning techniques for improved accuracy. The proposed model not only predicts the

optimal crop but also provides detailed information about the required amounts of soil ingredients along with their respective expenses. Analyzing climatic and soil conditions, the model aids farmers in predicting profitable crops, contributing to increased profits. Employing SVM as a machine learning algorithm RNN and LSTM as deep learning algorithms, the study demonstrates the integration of smart strategies for anticipating the most productive crop under specific conditions with minimal expenses.

This study [12] investigated the potential improvement in corn yield predictions in the Corn Belt of US by coupling crop modeling and machine learning (ML). The study aimed to explore the efficacy of a hybrid approach (crop modeling + ML), obtained the most perfect combinations of hybrid models, and determines the influential features to integrate with ML for corn yield prediction. Utilizing five ML models and six ensemble models, the authors found that incorporating simulation crop model variables (APSIM) as input features to ML models reduced root mean squared error (RMSE) by 7 to 20% and improves the result.

3. Objectives

The objectives of research are understanding the robust power of machine learning techniques and how can be used to enhanced crop yield by predicting most suitable crop for given agronomic and environmental conditions and helps to advance the field of agriculture sustainability. The objectives are:-

- Develop hybrid models using diverse classification, regression methods and stacking and voting ensemble technique.
- Investigate the effectiveness of machine learning techniques for finding the best crop for cultivation according to soil condition and environmental parameters.
- To evaluate performance matrices and find efficient machine learning models to obtained maximum crop yields.

4. METHODOLOGY

The primary aim of the research is to develop a machine learning model that would aid more accurate crop recommendation for crop to be grown to produce more yields. The dataset considered for this research has been taken from publically available secondary source kaggle and government site https://www.soilhealth.dac.gov.in. This dataset has the detailed information of 22 types of crops and about agronomical and environmental factors such as soil nutrients (nitrogen, phosphorus, potassium), environmental conditions (temperature, humidity, rainfall), and crop-specific parameters pH that significantly influence on the growth of the crops.

4.1 The Work Flow of Proposed Study

The figure1 shows the work flow of the research. the preprocessed data has been split into training and test datasets The various classifications and Regression techniques like logistic, KNN, RF, DT, SVC are used to developed and trained models using training data. The best combination of the parameters of each model have been found using GridSearchCV for enhancing the performance of the base models. The outputs of trained base models are used as an input for the Meta model and hybrid models are developed. The hybrid models HVM1 and HSM2 are developed using logistic, KNN, RF, DT, SVC regressions models, voting and stacking ensemble techniques. The classifications models such as LogRes, KNN, RF, DT and SVC, ensemble stacking and voting are used to developed hybrid model, HVM3, HSM4. The performance is evaluated by calculating performance metrics for both training and test dataset for each model. The results obtained are compared and find best model for predicting best suitable crop for cultivation in given environmental and agronomic conditions.

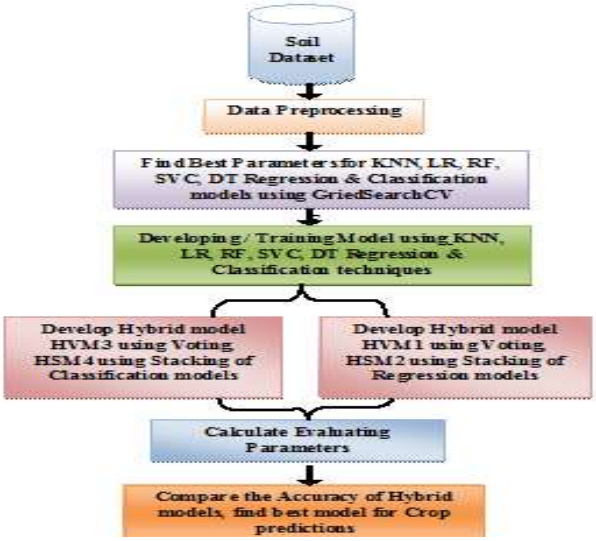

Figure 1 Work Flow

4.2. The Development of Hybrid Models

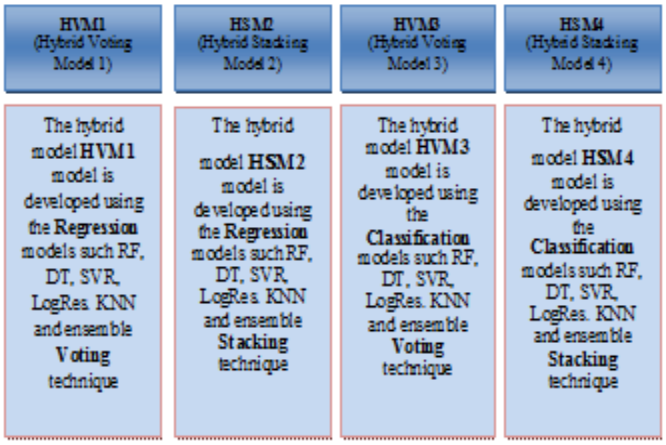

Figure 2 The Development of Hybrid Models

The hybrid model HVM1(Hybrid Voting Model 3) is developed using the Regression models such as Random Forest, Decision Tree, Support Vector Regressor, Logistic Regressor and ensemble Voting technique. The hybrid model HSM2 (Hybrid Stacking Model 2) is developed using Regression models such as Random Forest, Decision Tree, Support Vector Classifier, Logistic Regressor and ensemble Stacking technique. The hybrid model HVM3 (Hybrid Voting Model 3) is developed using Classification models such as Random Forest, Decision Tree, Support Vector Classifier, Logistic Regressor ensemble Voting technique. The hybrid model HSM4 (Hybrid Stacking Model 4) is developed using Classification models such as Random Forest, Decision Tree, Support Vector Regressor, Logistic Regressor and ensemble Stacking technique.

Random Forest Regression: Random Forest Regression is an ensemble learning method. It constructs multiple decision trees during training. Every tree is built by selecting random subset of features as well as data. It integrates the prediction of multiple decision trees to improve the accuracy and reduces over fitting. They also provide feature importance metrics, helping identify the most influential variables in the prediction process.

K-Nearest Neighbors: KNN (k-Nearest Neighbors) is a supervised machine learning algorithm applicable for both classification and regression. It makes predictions by finding k number of neighbors (closest data points). KNN used uses the average value for regression and majority class for classification of the nearest neighbors. It gives good results for small datasets.

SVC: The Support Vector Classifier, a type of Support Vector Machine (SVM) used for supervised classification tasks**.** It is useful for Binary classification (e.g., spam vs. not spam) as well as Multiclass classification (e.g., digit recognition). The Support Vector Classifier works with both linear and non-linear data.

Decision Tree: A Decision Tree is used for classification and regression tasks. It can handle both categorical as well as numerical data. It starts with root node, split data using metrics like Gini impurity, mean squared error or entropy.

4.3. Find Best Parameters for Improvement of the Model Performance

The hyper parameters turning used to set best combination of parameters to improve the performance of the model. The different combinations of hyper parameters make affect on the performance of the model and help to developed more accurate as well as robust model. The parameters turning can be done by adopting Grid Search, Random Search and Bayesian Optimizations (Automated ML) machine learning approach. In this study the GridSearchCV is used for finding the best combination of hyperparameter of the model so that the specific model can achieve optimal output with maximum accuracy.

5. RESULTS AND DISCUSSION

The various models like LR, KNN, RF, DT, SCV of Classifications and Regressions techniques of machine Learning has been developed and trained using 80% training data, the output (the predictions) made by these models are used as input to developed the hybrid models such as HVM1, HVM3 using ensemble Voting and HSM2 and HSM4 by ensemble Stacking. The accuracy of models has been tested using test data. The performance of the regression models HVM1, HSM2 has been evaluated by calculating RMSE, MSE, MAE and accuracy. The performance parameter F1 Score, Precision, Recall and accuracy has been calculated for HVM3 and HSM4 Classification model.

5.1 Evaluating the Performance of the Regression Models

Table 1 Accuracy of Regressions Model

Model	Training Accuracy	Testing Accuracy
LogReg	95.62	90.24
KNN	100	95.41
RF	99.55	94.16
DT	100	95.41
SVR	95.53	88.22
HSM2	99.19	96.16
HVM1	92.62	89.59

The table 1 shows the accuracy of various regression models as well as hybrid models. The models shown in table such as KNN, DT has 100% accuracy. The LogReg, SVR model has 95% accuracy, RF and HSM2 model shows 99% accuracy for
the training data. The RF, DT, KNN achieved 95% accuracy, LogRes, SVR, HVM1 shows approximate 90% accuracy and hybrid model HSM2 has shows excellent performance with 96.16% accuracy for test
data.

5.2 Evaluating the Performance of the Classification Models

Table 2 Accuracy of Classification Model

Model	Training Accuracy	Test Accuracy
LogReg	97.79	96.37
KNN	99.1	97.28
RF	100	99.32
DT	99.83	98.19
SVC	99.32	97.96
HSM4	99.99	99.4
HVM3	99.99	98.64

Table 2 shows the accuracy of various classification models; hybrid models HSM4 and HVM3 developed and trained by training data. It has been observed that the accuracy performance of RF is 100%, the models such as KNN, DT, SVC, HVM3 and HSM4 shows good performance with 99% average accuracy, LogReg shows 97.79% accuracy for the training datasets. The LogReg, KNN and SVC have

approximate 97% accuracy, RF, DT and HVM3 has 99% accuracy. The HSM4 excellent performance shows Hybrid model performed well among all the models ith 99.4% testing accuracy

5.3 The Analysis of Performance Accuracy of Classification and Regression Models

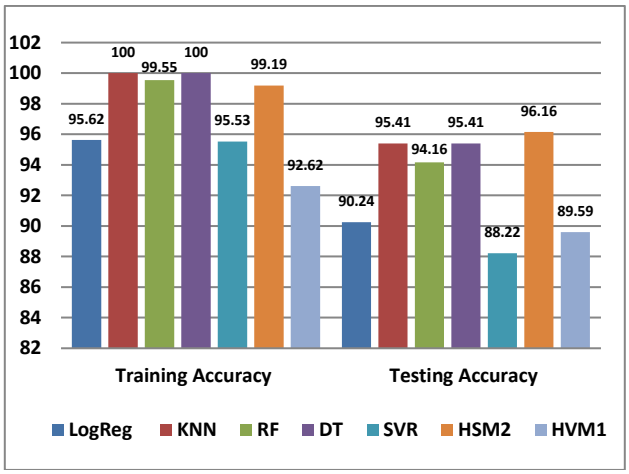

Figure 3 Compression of Accuracy of Regression Models

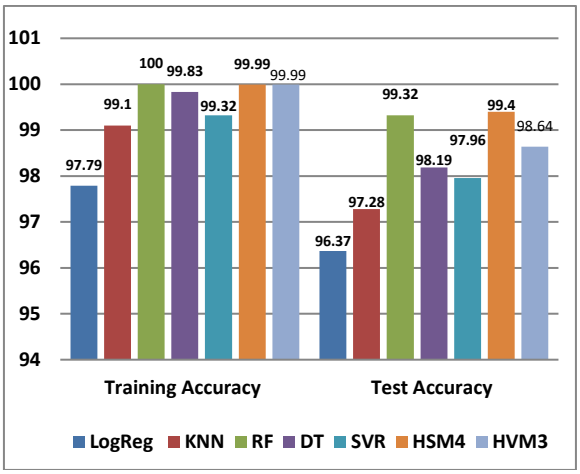

Figure 4 Compression of Accuracy of Classification Models

The figure 2 shows the comparison of accuracy of various regression models, HVM3 and HSM2 hybrid models. The figure 3 shows the compression of various classification models, HVM3 and HSM4 hybrid models. It has been observed that the hybrid models HSM2 and HSM4 developed using classification method and ensemble technique has achieved highest performance.

5.4 The Comparison of Training and Test Accuracy of Various Classification and Regression Models

Figure 5 shows the performance accuracy of various regression and classification models i.e. LogRes, KNN, SVC, RF and DT for training dataset. It has been observed that KNN DT regression models and classification model RF performed excellent with 100% accuracy. The classification models KNN, DT and SVR achieve more than 99% accuracy for training dataset.

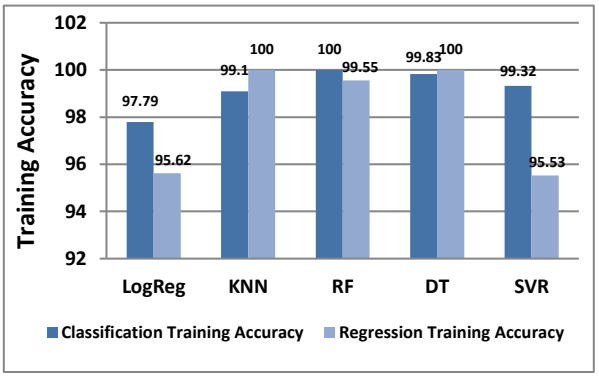

Figure 5 Comparison of Training Accuracy of Classification and Regression Model

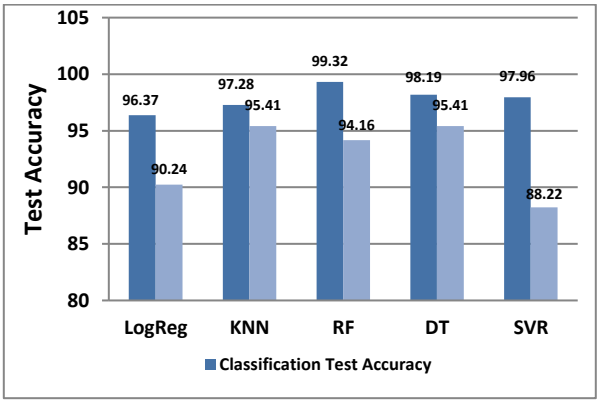

Figure 6 Comparison of Test Accuracy of Classification and Regression Model

Figure 6 shows the performance accuracy of various regression and classification models. The classification model LogRes, KNN, SVC and DT perform well with approximate 97% accuracy. It has been observed that

all regression models except SVR achieved with 100% accuracy for training dataset but the performance of classification models excellent than regression models for test data. The RF model developed using classification method shows robust performance with 99.32% accuracy for test data.

5.5 The Comparison of Performance Accuracy of Hybrid Models

Table 3 Comparison of Accuracy of Hybrid Models

Model	Training Accuracy	Testing Accuracy
HVM1	92.62	89.59
HSM2	99.19	96.16
HVM3	99.99	98.64
HSM4	99.99	99.4

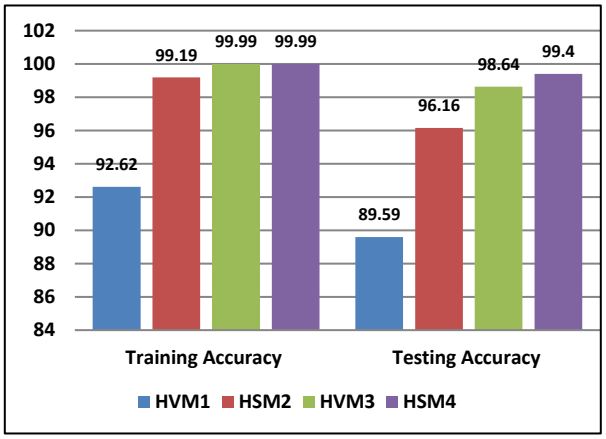

Figure 7 Compression of Accuracy of Hybrid Models

The evaluated performance accuracy of hybrid models HVM1, HSM2, HVM3 and HSM4 developed using Voting and Stacking techniques; ensemble by various regression and classification models has been shown in Table 3, the same result represented by Figure7 visually. it has been found that HVM1 model has 92.62 % accuracy and rest of all models HSM2, HVM3 and HSM4 shown approximately 99% accuracy for training data. The accuracy achieved by
HVM1 89.59%, HSM2 96.16% and HVM3 98.64% for test data. The hybrid classification model developed using ensemble stacking method; HSM4 has achieved maximum accuracy 99.4% for test data.

6. CONCLUSION

The study shows the significant potential of ensemble machine learning techniques in enhancing crop prediction for achieving maximum yield. The crop-specific analysis revealed that different crops have unique nutrient and climatic requirements, which can be precisely identified and optimized using machine learning models. The performance accuracy of regression and classification models such as KNN, LogReg, DT, RF, SVC, SVM and ensemble hybrid models HVM1, HVM3, HSM2 and HSM4 has been evaluated, compared, analyzed. All the models performed excellent with approximate more than 95% accuracy for test data. The

HSM4 model shows robust performance among all the models for selecting best crop for cultivation according to given soil and environmental parameters for specific area with 99.4 % accuracy. This precision agriculture approach enables farmers to take strategic decisions, ensuring optimal utilization of resource and maximizing yield.

7. REFERENCES

[1] M. Kuradusenge *et al.*, "Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize," *Agric.*, vol. 13, no. 1, 2023, doi: 10.3390/agriculture13010225.

[2] L. S. Cedric *et al.*, "Crops yield prediction based on machine learning models: Case of West African countries," *Smart Agric. Technol.*, vol. 2, no. December 2021, 2022, doi: 10.1016/j.atech.2022.100049.

[3] B. Rama Devi, P. Ragam, S. P. Godishala, V. S. K. N. Gandham, G. Panuganti, and S. S. Annavajjula, "Crop Yield Prediction Using Machine Learning Algorithms," *Lect. Notes Networks Syst.*, vol. 606, pp. 397–405, 2023, doi: 10.1007/978-981-19-8563-8_38.

[4] K. Jhajharia, P. Mathur, S. Jain, and S. Nijhawan, "Crop Yield Prediction using Machine Learning and Deep Learning Techniques," *Procedia Comput. Sci.*, vol. 218, pp. 406–417, 2022, doi: 10.1016/j.procs.2023.01.023.

[5] M. S. Rao, A. Singh, N. V. S. Reddy, and D. U. Acharya, "Crop prediction using machine learning," *J. Phys. Conf. Ser.*, vol. 2161, no. 1, 2022, doi: 10.1088/1742-6596/2161/1/012033.

[6] D. Paudel *et al.*, "Machine learning for large-scale crop yield forecasting," *Agric. Syst.*, vol. 187, no. December 2020, p. 103016, 2021, doi: 10.1016/j.agsy.2020.103016.

[7] D. J. Reddy and M. R. Kumar, "Crop yield prediction using machine learning algorithm," *Proc. - 5th Int. Conf. Intell. Comput. Control Syst. ICICCS 2021*, vol. 9, no. 4, pp. 1466–1470, 2021, doi: 10.1109/ICICCS51141.2021.9432236.

[8] S. S. and M. Sujithra, "Agricultural Data Analysis," *Int. J. Adv. Res.*, vol. 9, no. 08, pp. 807–815, 2021, doi: 10.21474/ijar01/13330.

[9] J. Pant, R. P. Pant, M. Kumar Singh, D. Pratap Singh, and H. Pant, "Analysis of agricultural crop yield prediction using statistical techniques of machine learning," *Mater. Today Proc.*, vol. 46, no. xxxx, pp. 10922–10926, 2021, doi: 10.1016/j.matpr.2021.01.948.

[10] A. Sharifi, "Yield prediction with machine learning algorithms and satellite images," *J. Sci. Food Agric.*, vol. 101, no. 3, pp. 891–896, 2021, doi: 10.1002/jsfa.10696.

[11] S. Agarwal and S. Tarar, "A hybrid approach for crop yield prediction using machine learning and deep learning algorithms," *J. Phys. Conf. Ser.*, vol. 1714, no. 1, 2021, doi: 10.1088/1742-6596/1714/1/012012.

[12] M. Shahhosseini, G. Hu, I. Huber, and S. V. Archontoulis, "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt," *Sci. Rep.*, vol. 11, no. 1, pp. 1–15, 2021, doi: 10.1038/s41598-020-80820-1.