IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AGRICULTURAL KEY ENVIRONMENTS AND SOIL PARAMETERS THAT INFLUENCE THE CROP YIELD PREDICTION AND ITS SUSTAINABILITY

[1]A. Anbarasi, [2]Dr. A. Kangaiammal,

[1]Ph.D Research scholar, [2]Assistant Professor, Department of Computer Applications, Government Arts College (A), Salem-7,

[1]Department of Computer Science,

[1]Periyar University, Salem, India

Abstract: Accurate crop yield prediction is extremely challenging for sustainable agricultural planning and food security in Indian's day-to-day life style. The traditional statistical models and the machine learning techniques are used to achieve the accuracy values, but often they focus on only predicting the single crop, individual models, and prediction accuracy values. To overcome these issues, this study proposes a multi-model machine learning framework with multi crops to identify the agricultural key environments and soil parameters that are influencing the crop yield prediction and its sustainability using the historical crop yield data consisting of soil and weather attributes. Four ensemble-based machine learning models such as Random Forest (RF), Gradient Boosting Regressor (GBR), Extra Tree Regressor (ETR), and XGBoost are used and evaluated. Performance metrics observed are Mean Square Error (MSE), Root Mean Square Error (RMSE), and R^2. A Consensus Feature Importance (CFI) methodology is implemented to identify the stable factors that influences the model's outcome in terms of crop yield. Additionally, the Sustainability-Aware Crop Sensitivity Analysis is also implemented to rank the crops based on their response to soil and environmental conditions. These two approaches enhance the agriculture yield sustainability. The obtained result shows that the Random Forest model achieved the highest predictive performance accuracy value $R^2 = 0.67$. The most influencing soil parameters are potassium(K), area cultivated, PH level of the soil and the pesticides used. This study identifies the most sensitive crops that are influenced by the soil and environmental conditions based on the sustainability ranking of the crops. As a result, Coconut, Rice, Sugarcane and Wheat are the top four crops identified. This study would not only enhance the approach of agricultural practices but also it enhances data driven decision-making and sustainable agricultural methodologies.

Index Terms **-** Crop Yield Prediction, Machine Learning, Soil Parameters, Environmental Factors, Sustainability Index, Consensus Feature Importance.

1. INTRODUCTION

Agriculture plays an important role in the Indian economy. Nowadays the farmer's contribution towards the food supply, employment in rural development is more challenging. Due to the increasing demand for food and the challenges faced by the farmers such as climate variability, sustainable and accurate prediction of crop yield has become a national priority. Also, the farmers face a dynamic change in environment for the agricultural production. Yield Prediction helps the policymakers and farmers to make an informed decisions about resource allocation, market planning, and food security. Even though there are so many traditional statistical approaches the crop yield prediction is influenced by the soil and environmental factors [1, 2, 4]. Yield prediction research is essential in current agricultural systems since it serves as a key point of reference for farm management during planning, agronomical technologies investment intervention, and preharvest procedures [2].

In recent years many studies using the machine learning algorithms mainly focusing on the improving accuracy prediction of specific crops or region that often overlooking on the identification of stable key parameters across multiple crops and models [2]. Also, the sustainability implication of yield influencing factors. In agricultural crop yield prediction, they mainly deal on how the soil properties interact with plant growth and understanding which soil parameters is essential for analyzing and predicting the crop yield. Specifically, the soil characteristics such as Nitrogen, Phosphorus and Potassium, PH values and moisture level are directly influencing the plants health and overall productivity of the crop. The accurate crop yield evaluation of these soil factors is now enhanced by applying machine learning approaches, that allows for the crop yield prediction. Different variations occur in these parameters can lead to significant spatial and temporal changes in productivity, even under similar climatic and management conditions and understanding how soil parameters affect crop yield. The soil key parameters are the most essential parameter for analyzing the fertilizer application, improving the soils health condition, and that promotes the sustainable agricultural practices. By analyzing these soil attributes and their relationships with crop yield, researchers and policymakers can develop data-driven strategies to enhance their productivity, and support long-term agricultural sustainability, especially in regions facing soil degradation and climate variability.

2. LITERATURE REVIEW

The traditional statistical techniques combined with advanced machine learning methodologies are used to enhance the predictions accuracy value and data-driven insights regarding agricultural productivity [7]. Many studies use the machine learning models and give the high prediction accuracy values for individual crops such as rice, wheat, paddy, maize, etc and they used weather and soil data, or satellite images [2, 4]. Also, many researches have highlighted the importance of of soil properties, including nitrogen (N), phosphorus (P), potassium (K), and pH, as major determinants of crop productivity [6, 7]. Environmental factors such as temperature, rainfall, and humidity have also been shown to significantly affect yield variability [6,7]. Ensemble-based models, particularly Random Forest and Gradient Boosting, are frequently reported to outperform linear and single-tree models due to their robustness and ability to model nonlinear interactions [17]. Most studies focus on single-crop scenarios, applying a single ML model, or lack interpretability in identifying stable influencing parameters [18]. The reviewed literature work collectively highlights the importance of combining the multiple ML models, identifying the consensus influencing parameters, and integrating a sustainability crop ranking across multiple crops.

S. No	Title	Author	Techniques and Dataset Used	Problem Addressed	Problem Unaddressed	Future Enhancement
1.	Exploring the potential role of environmental and multi-source satellite data in crop yield prediction	Zhenwang Li et al. (2022).	ML models (RF, SVM, ensemble), climate + soil + satellite datasets have been used.	Accurate large-scale crop yield prediction is done by using the multi-source data	It is high in computational cost. Very limited generalization to other regions.	Real-time forecasting, region-transferable models, lightweight ML frameworks
2.	Evaluating ML models and identifying key factors influencing spatial maize yield predictions	S. Maseko et al. (2024).	Random Forest, Decision Tree, MLP are used for prediction On-farm soil, NDVI, fertilizer dataset are used here.	Identifying the yield. Only limited factors are used for precision agriculture.	Limited Seasonal variability is handled.	Multi-season adaptive models can be applied. Sustainability-focused optimization can also applied in future
3.	Agriculture Yield Prediction: AI-Driven Optimization for Sustainable Farming	A. D. Chokhat et al. (2025).	For yield prediction the RF, SVM, DL are used. soil + weather + historical crop datasets used.	AI-based crop yield prediction also the recommendation system is also applied.	Interpretability of DL models but they are not addressed properly.	Can apply the Explainable AI and sustainability indices integration.
4.	Artificial Intelligence in Agriculture: A Systematic Review of Crop Yield Prediction	C. R. Screpnik et al. (2025).	Systematic review of SVM, RF, XGBoost, KNN.	Overview of AI methods for yield prediction	No experimental validation or any unified framework applied.	Development of benchmark datasets and applying any hybrid models for predictions.
5.	Crop prediction based on soil and environmental characteristics using feature selection	A. Suruliandi et al. (2021).	Wrapper feature selection, RFE + Adaptive Bagging. soil & climate datasets are used.	Identifying the optimal crops by selecting the features.	Yield wise quantity predictions are not considered here.	Extension to the yield estimation and applying any sustainability metrics.
6.	Enhancing prediction of crop yield and soil health assessment using ML	K. N. Vhatkar et al. (2025).	BPNN, IP-EF feature selection methods applied here. GIS + multi-source datasets are used.	Joint prediction of crop yield and identifying the soils health.	High dependency on DL. It is a complex architecture that have been used.	Can integrate with the Lightweight ML models. IoT integration for real-time use.

7.	Predictive Analysis of Crop Yield Based on Environmental and Soil Conditions	Vineet Goyal (2024).	Linear Regression, Decision Trees are applied for the experimental farm data.	Comparison of simple ML models for yield prediction.	Scalability and sustainability aspects are missing	Can apply Ensemble models and long-term sustainability for analysis.
8.	Enhanced wheat yield prediction through integrated climate and satellite data	M. Ashfaq et al. (2025).	RF, SVM, CNN, RNN models are implmented. satellite + soil + climate datasets are used.	It gives the High-accuracy wheat yield prediction	It has very high computational and data dependency	Generalized crop-independent frameworks can be applied in future.
9.	Enhancing Crop Yield Prediction Using IoT-Based Soil Sensors	Zaid Alsalami et al. (2025).	IoT sensors + RF are applied for yield prediction. NPK, soil moisture, weather dataset has been used here.	Real-time yield prediction is obtained and the resource optimization is also obtained.	They are limited in scalability and in regional testing.	Large-scale deployment and AI-driven sustainability planning.

3. PROPOSED METHODOLOGY

3.1 DATASET AND PREPROCESSING

The study uses three datasets such as 1) Historical Crop Yield Dataset that includes state-wise crops, year-wise production and area of land for cultivation multiple crops. 2) Soil Dataset consists of soils average nutrition properties such as nitrogen (N), phosphorous (P), potassium (K) and PH level values. 3) Weather Dataset that consists of temperature and humidity values with aggregated annual climatic reports. These datasets have been sourced from the publicly accessible data repository, Kaggle's "public agricultural data". The dataset covers the crop yields across multiple Indian states between the years 1997 to 2020. There are 10 features across three datasets that are used in this study such as crop area, fertilizers, pesticide, N, P, K, PH values, average temperature in Celsius, average humidity in percent, and the temporal variable year with crop yield is the target variable.

Data preprocessing is a crucial and important step in preparing the dataset for analysis. It involves several key steps for processing the data. All the three datasets have been cleaned and standardized by harmonizing column names. The Aggregation is performed earlier to merging and reduce the dimensionality and memory usage. All the missing values are removed and only the numeric features are focused for the model processing. Here the crop yield is taken as a target variable. The cultivated area, fertilizers and pesticides used, soil nutrients such as N, P, K, PH values and the climatic factors average temperature, average humidity are the input features considered in this study.

3.2 MACHINE LEARNING MODELS

In this study, Four ensemble machine learning models have been implemented such as: i) Random Forest Regressor is implemented for robust to noise and capable of capturing the nonlinear relationships ii) Gradient Boost Regressor is a sequential ensemble learning that used for improving the accuracy value iii) Extra Tree Regressor is used to reduce the variance of high randomness and iv) XGBoost Regressor is implemented for boosting the model with correct regularizations. These models use 80% training and 20% testing for the process.

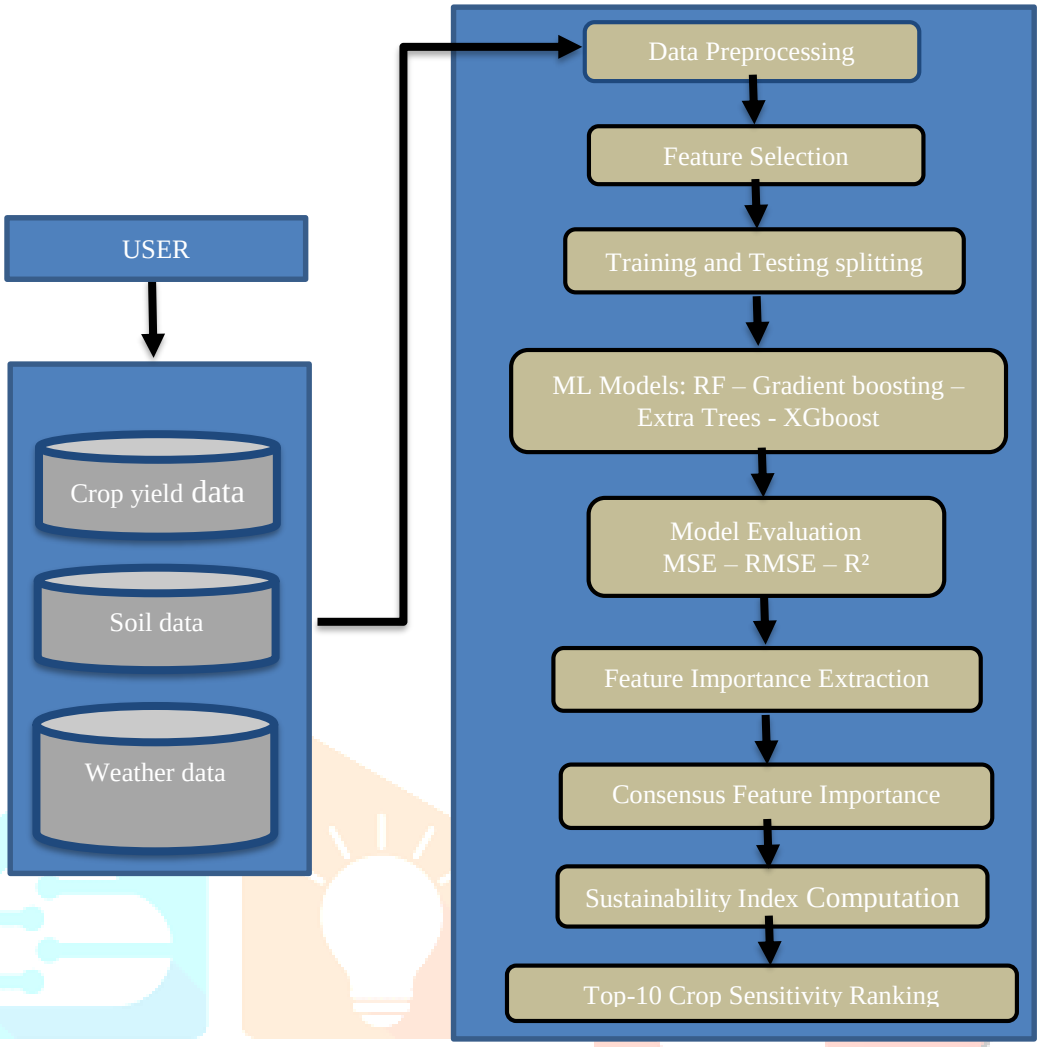

Fig 1: Architecture of the Proposed Model

3.3 EVALUATION METRICS

Since yield prediction is a regression-based model, the performance was calculated using the three-standard metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE**),** Coefficient of Determination (R^2) [18,19]. These are the three standard metrics that are used to evaluate the prediction capability and that has generalized the ensemble machine learning models.

3.4 CONSENSUS FEATURE IMPORTANCE

To reduce a model bias and improve the robustness, the feature importance value from all the ensemble machine learning models is normalized and averaged to compute the Mean Importance Score. This approach reduces the models specific bias and provides a reliable identification of the parameters that consistently influence the yield. The consensus feature ranking model is more accurate than the single model interpretations. Fig 2, represent the Architecture diagram of Consensus Feature Importance model.

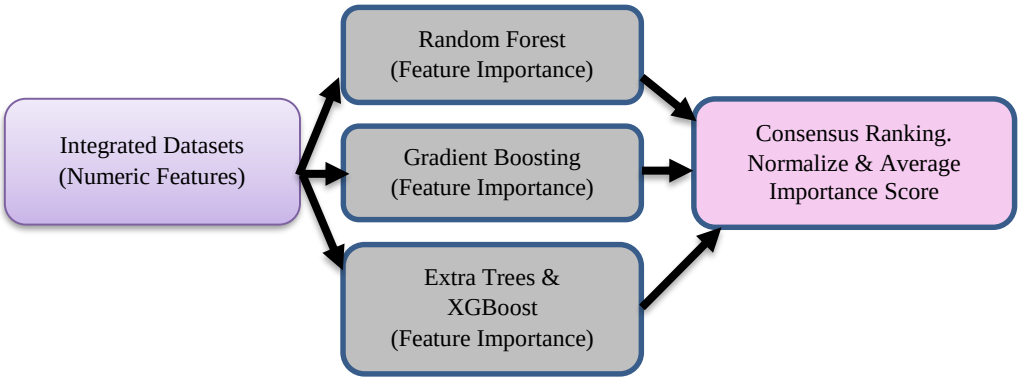

Fig 2: Architecture of Consensus Feature Importance

3.5 SUSTAINABILITY INDEX

A sustainability index is defined as a relative measure that indicates the crop yield and their key input sensitivity. This index evaluates the captured yield efficiently that related to environmental and soil resource usages. When the rainfall data is not available, then the yield alone is utilized as a reliable alternative to ensure the robustness. Fig 3 represents the Architecture diagram of the sustainability index.

Also, the sustainability index is defined as the

$$SI = \frac{Yield}{Rainfall+Nitrogen+1}$$

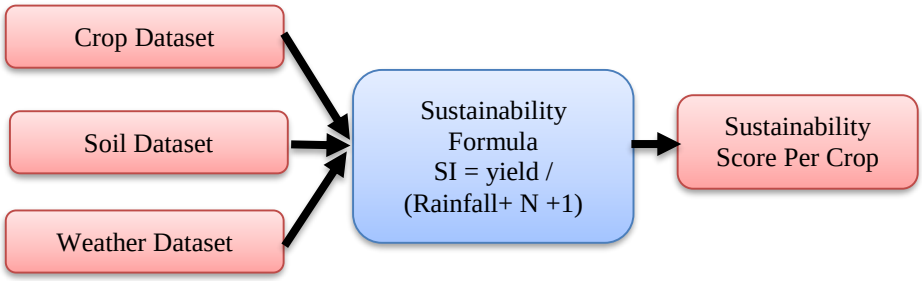

Fig 3: Architecture of Sustainability Index.

3.6 CROP LEVEL SENSITIVITY ANALYSIS

Crop level sensitivity analysis in this study is evaluated by top 10 crops that are ranked based on their average sustainability index and their crop yields variability. Also, the crops are identified by most influencing soil and environmental conditions. Fig 4 represents the Architecture diagram of the sensitivity analysis in crop level.

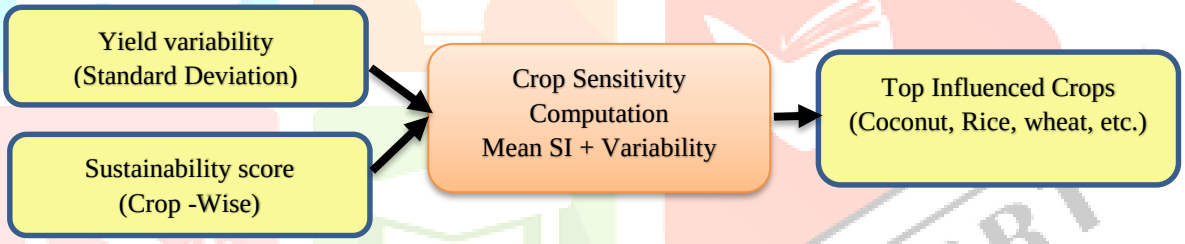

Fig 4: Architecture of Crop Level Sensitivity Analysis

4. RESULTS AND DISCUSSION

Many research papers and studies show that the yield prediction is done by just applying traditional machine learning models and error metrics. In this study, the result is obtained by applying a machine learning framework that uses the ensemble models for crop yield prediction. The framework mainly focuses on the performance of the model, key influential parameters, sustainability assessment and the crop level sensitivity analysis.

4.1 PERFORMANCE ANALYSIS OF MODELS

There are four ensemble machine learning models that have been used. They are Random Forest (RF), Gradient Boosting Regressor (GBR), Extra Tree Regressor (ETR), and XGBoost. These models are evaluated by using the regression metrics Mean Squared Error (MSE), Root Mean Squared Error (RMSE)**, Coefficient of determination (R^2). It is observed from the evaluation results that the Random Forest model has achieved the high accuracy R^2 value of 0.67. While both** Gradient Boosting Regressor (GBR) and XGBoost performed to be very effective in predicting the agricultural yields by using the ensemble machine learning models. The Extra Tree exhibits comparatively lower generalization and that indicates the sensitivity of a data variability. The overall result shows that Ensemble machine learning models effectively captures the nonlinear relationships between the multi-crop agricultural data (i.e., by using the soil and environmental parameters) over the traditional single model approaches. Fig. 5 shows the performance of four ensemble machine learning models in terms of **R^2** values**.**

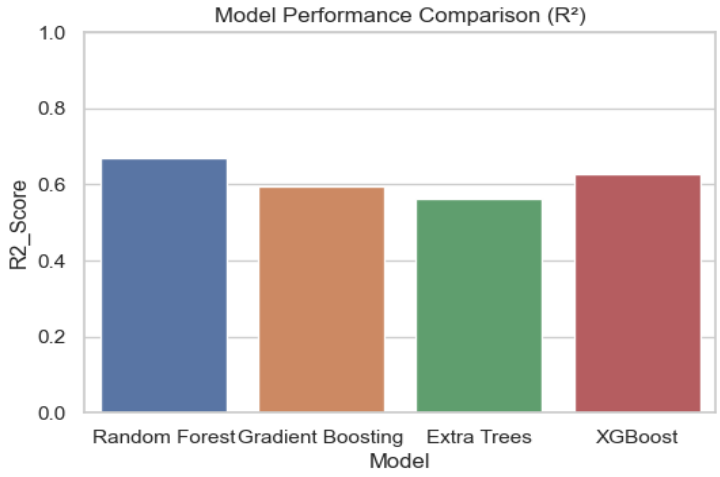

Fig 5: Model performance comparison.

4.2 IDENTIFICATION OF SOIL AND KEY ENVIRONMENTAL PARAMETERS:

To identify the soil and key environmental parameters the Consensus feature importance method was employed by normalizing and averaging the feature importance score from all the four ensemble machine learning models**.** The outcome of this method gives the most influential parameters that affects the crop yield are soil PH, potassium (K), cultivated area, fertilizer and pesticides used. The dominance of the soil nutrients such as potassium (K) and PH levels, along with the various management practices validates the agronomic principles. This highlights the performance of the proposed model of Consensus-based approach. Fig. 6 represents the obtained mean importance value of key environmental soil parameters.

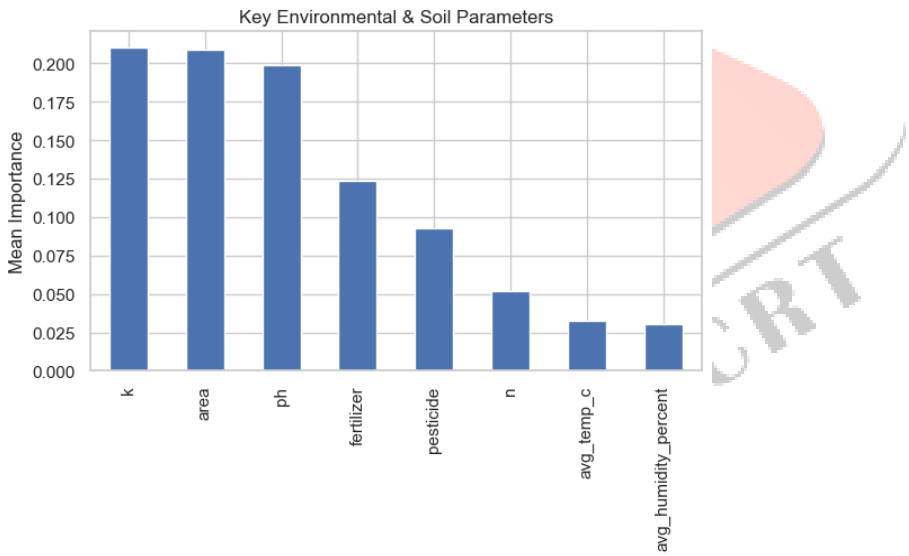

Fig 6: Key Environmental & Soil Parameters

4.3 SUSTAINABILITY INDEX AND CROP LEVEL SESITIVITY ANALYSIS

The Sustainability Index (SI) is calculated by integrating the crop yield with the nitrogen input and the rainfall values when it is available. The outcome of the SI is, the crop that gives high yield when it is having lower resource dependence, then it achieves the higher sustainability score. This proves the robustness of Sustainability Index (SI) methods over maintaining the consistent assessments through the crops. Also, it proves the quantitative bridge between the yield prediction and sustainability agricultural practices that is rarely addressed in many conventional yield prediction studies.

The crop level sensitivity index was evaluated using the sustainability index that combined with the yield variability. Also, this study gives the top 10 crops that are influenced by the soil and environmental parameters. The resulted top 10 influenced parameters are Coconut, Sugarcane, Wheat, Rice, Cotton, Jute, Potato, Banana, Tapioca, and maize. The high valued and water intensive crops have been demonstrated the greater sensitivity that emphasize the need of crop's specific soil and resource management plans. Fig 7: represents the top 10 crops that are influenced by the soil and environmental parameters.

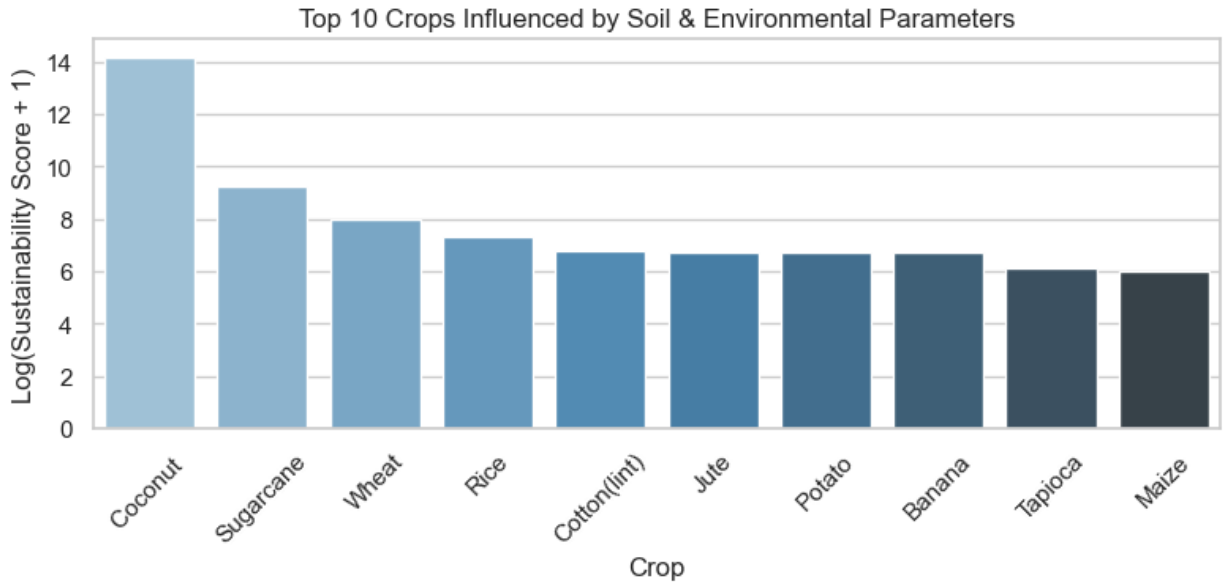

Fig 7: Top 10 crops influenced by soil and environmental parameters.

It gives more stable and powerful approach for the farmer-friendly predictions. This study demonstrates that the Ensemble machine learning model improves the crop yield prediction accuracy score and its robustness. The consensus feature importance method reduces the model's overall bias and enhances the interpretabilities. The sustainability aware method evaluates the deeper insights beyond the accuracy predictions and the final crop level sensitivity analysis reinforces the targeted agricultural decision making.

5. CONCLUSION

This study presents an innovative, interpretable and proposed ensemble machine learning framework to identify the agricultural key environmental and soil parameters that influences the crop yield prediction and its sustainability. By combining the heterogeneous crop, soil, and weather datasets through an aggregation method for the purpose of memory efficient, and this proposed approach effectively addresses the scalability and the computational challenges that are often faced with the agricultural datasets. The experimental result shows that, from the four ensemble machine learning models the Random forest model achieved the high predictive performance accuracy value with $R^2 = 0.67$, confirming the model's effectiveness. Also, it captures the nonlinear interactions between the agronomic factors. The implemented Consensus Feature Importance method provides the robustness and the reliability identification of dominant parameters. It reveals that the potassium content, cultivated area, soil PH, fertilizer and pesticides usage are the most influential parameters that affects the crop yield.

To enhance the crop yield prediction towards the sustainability assessment, a Sustainability Index has been implemented to evaluate the crop yields efficiency that corresponds to the environment and soil resource usages. The sustainability index enabled a quantitative understanding of how the resource inputs impact the agricultural productivity. The Crop level Sensitivity Analysis results the crop such as Coconut, Sugarcane, Wheat, and Rice are highly sensitive to the soil and key environmental parameters. This method highlights and enhances the necessity for the specific crop managements. Overall, the proposed framework effectively predicts the high accuracy, interpretabilities, and sustainability's. It also insights for the researchers, farmers, and policymakers. This study represents its robustness, transparency, and practical applications. The future work may incorporate with decision support system and developed using deep learning models for the precision agricultural practices.

REFERENCES:

[1] Li, Z., Ding, L., & Xu, D. (2022). Exploring the potential role of environmental and multi-source satellite data in crop yield prediction across Northeast China. Science of the Total Environment, 815, 152880. https://doi.org/10.1016/j.scitotenv.2021.152880.

[2] Maseko, S., van der Laan, M., Tesfamariam, E. H., Delport, M., & Otterman, H. (2024). Evaluating machine learning models and identifying key factors influencing spatial maize yield predictions in data-intensive farm management. European Journal of Agronomy, 157, 127193. https://doi.org/10.1016/j.eja.2024.127193.

[3] Chokhat, A. D., Pawar, S. P., Gadpayle, A. R., Ghate, O. V., & Kshirsagar, B. M. (2025). Agriculture yield prediction: AI-driven optimization for sustainable farming. International Journal for Multidisciplinary Research, 7(2), 1–10.

[4] Screpnik, C. R., Zamudio, E., & Gimenez, L. I. (2025). Artificial intelligence in agriculture: A systematic review of crop yield prediction and optimization. IEEE Access, 13, 70691–70705. https://doi.org/10.1109/ACCESS.2025.3560631.

[5] Suruliandi, A., Mariammal, G., & Raja, S. P. (2021). Crop prediction based on soil and environmental characteristics using feature selection techniques. Mathematical and Computer Modelling of Dynamical Systems, 27(1), 117–140. https://doi.org/10.1080/13873954.2021.1882505.

[6] Vhatkar, K. N., Koparde, S. A., Kothari, S., Sarwade, J., & Sakur, K. (2025). Enhancing prediction of crop yield and soil health assessment for sustainable agriculture using a machine learning approach. MethodsX, 14, 103418. https://doi.org/10.1016/j.mex.2025.103418.

[7] Goyal, V. (2024). Predictive analysis of crop yield based on environmental and soil conditions. International Journal on Computational Modelling Applications, 1(2), 50–63.

[8] Ashfaq, M., Khan, I., Afzal, R. F., Shah, D., Ali, S., & Tahir, M. (2025). Enhanced wheat yield prediction through integrated climate and satellite data using advanced AI techniques. Scientific Reports, 15, 18093. https://doi.org/10.1038/s41598-025-02700-w.

[9] Alsalami, Z., Mohammed, G., & Srinivas, T. (2025). Enhancing crop yield prediction using IoT-based soil moisture and nutrient sensors. SHS Web of Conferences, 216, 01029. https://doi.org/10.1051/shsconf/202521601029.

[10] van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709. https://doi.org/10.1016/j.compag.2020.105709.

[11] Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy, 10(7), 1046. https://doi.org/10.3390/agronomy10071046

[12] Nyéki, A., & Neményi, M. (2022). Crop yield prediction in precision agriculture. Agronomy, 12(10), 2460. https://doi.org/10.3390/agronomy12102460

[13] Al-Adhaileh, M. H., & Aldhyani, T. H. H. (2022). Artificial intelligence framework for modeling and predicting crop yield to enhance food security in Saudi Arabia. PeerJ Computer Science, 8, e1104. https://doi.org/10.7717/peerj-cs.1104

[14] Talaat, F. M. (2023). Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. Neural Computing and Applications, 35, 17281–17292. https://doi.org/10.1007/s00521-023-08619-5

[15] Sbai, Z. (2025). Deep learning models and their ensembles for robust agricultural yield prediction in Saudi Arabia. Sustainability, 17(13), 5807. https://doi.org/10.3390/su17135807

[16] Elavarasan, D., & Durairaj Vincent, P. M. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access, 8, 86886–86901. https://doi.org/10.1109/ACCESS.2020.2992480.

[17] Li, X., Tan, J., Li, H., Wang, L., Niu, G., & Wang, X. (2023). Sensitivity Analysis of the WOFOST Crop Model Parameters Using the EFAST Method and Verification of Its Adaptability in the Yellow River Irrigation Area, Northwest China. Agronomy, 13(9), 2294. https://doi.org/10.3390/agronomy13092294.

[18] Nikhil, U. V., Pandiyan, A. M., Raja, S. P., & Stamenkovic, Z. (2024). Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models. Computers, 13(6), 137. https://doi.org/10.3390/computers13060137.

[19] Hove, K., Nyamugure, P., Mdlongwa, P. *et al.* Advancements in maize yield estimation: a comprehensive review of methods and models. *Discov Sustain* **6**, 1417 (2025). https://doi.org/10.1007/s43621-025-02180-y.

[20] Hamim, A.M & Mohaimin, Abdul & Ishmum, Al & Ifty, Rashedul & Patwary, M.. (2025). Advances in Machine Learning for Crop Yield Prediction: A Comprehensive Review of Techniques, Trends, and Challenges. 1-6. 10.1109/ECCE64574.2025.11013031.