International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

A Machine Learning Approach for Sustainable Crop Yield Prediction Using Climatic and Soil Attributes

[1]. Khushbu Rajput, [2]. Bhavesh Jain
[1] Student of Computer Studies and Emerging Technology, TransStadia University, Ahmedabad
[2] Assistant Professor of Computer Studies and Emerging Technology, TransStadia University, Ahmedabad.

Abstract- Agriculture is an important sector in terms of food security and economic development, especially in developing nations. Precise crop yield estimation is required for efficient agricultural planning and management in the context of the increasing effects of climate change. Crop yield is affected by various factors, including climate variability, soil type, and availability of nutrients. Conventional crop yield estimation techniques, which rely on average values and traditional knowledge, are not reliable due to the complexities involved in crop yield estimation. Proposed in this paper is a framework for crop yield prediction using machine learning, incorporating climatic and soil variables. The climatic variables of rainfall, temperature, and humidity, and soil variables of soil pH and necessary nutrients (nitrogen, phosphorus, and potassium) are used as input variables. Three supervised machine learning algorithms—Linear Regression, Random Forest, and Gradient Boosting—are applied and compared to assess their predictive capability. Linear Regression is applied as a baseline algorithm, while ensemble methods are applied to deal with non-linearities in agricultural data. The performance of the models is measured using typical regression evaluation criteria, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2). The experimental outcomes show that the models based on ensemble methods perform better than the baseline model in terms of prediction accuracy and generalization ability. The results confirm that the combination of climatic and soil properties helps to improve crop yield prediction.

Keywords: Crop Yield Prediction, Machine Learning Models, Climate Data, Soil Properties, Sustainable Farming.

I. INTRODUCTION

The challenges involved when attempting to predict yields from crops has dramatically increased due to the changes that have occurred in climate related and environmental factors making the traditional methods of predicting yield now obsolete. Therefore, a new method for yield prediction has been developed utilizing machine-learning techniques that will assist in measuring climate and soil in order to accurately predict yield [1].

1. High Sensitivity to Change in Climate: The accuracy of predictions has been affected by even minimal variations in precipitation or temperature thus rendering any alternative techniques unfeasible. [1]

2. Poor Scalability: As most statistical and machine learning models require scaling up, this has been made more difficult because the majority are specific to particular regions or crops, therefore, applying them to different farms is difficult [2].

3. Static Models: The majority of traditional crop yield prediction models are static models; therefore, they will not change once developed and will require updating manually when there is a change in the environment. To alleviate this problem, it is suggested that a "machine-learning-based yield prediction model" be established.

4. Data Integrity and Reliability: The agricultural and climatic data must be adequately processed to ensure its integrity before it can be used to analyze yield predictions [1].

5. Yield Prediction Using Machine Learning: Machine-learning models operate in a continuously evolving environment, where they are able to assess the climatic and environmental factors affecting yield based on the changing patterns in the ecology of the area.

II. LITERATURE REVIEW

I.Introduction to Yield Predictions in Crops

Predicting harvest production has emerged as a significant domain of research within agriculture; its significance relates directly to food security and ongoing sustainable agricultural practices. The accuracy of yield estimates gives farmers and policymakers the necessary tools for accurate planning and effective risk management. The traditional method used to produce crop yield estimations relies heavily on data from previous seasons and is mostly based on personal experience and judgement; this method of estimating crop yields has clear limitations because of current climate change. To address these limitations, most recent studies have focused on developing data-oriented models or forecasting procedures that utilize the data gathered from multiple sources to create accurate crop yield predictions.

II. Climate Information as Data for Agricultural Forecasting and Data Trustworthiness

Precipitation, atmospheric temperature, relative humidity and other variables are some of the most significant climatic data variables that can affect both the development of a crop and the total amount harvested. Researchers have repeatedly documented how variations in temperature, ten-year weather conditions and precipitation impact crop yields. The creation of statistically based models to predict yield estimates has shown that data gathered from the climatic variables can serve as one of the best possible methods of estimating yields. Generally speaking, all of the climatic variables mentioned above are considered uncertain data when conducting an analysis of climatic-related data. Researchers have also conducted research that has indicated how vital it is to manage climatic-related data in order to achieve accurate and reliable crop yield estimates; any attempt to manage climate-related data must take into consideration the growing need for ethical and legally correct ways of managing the data collected within the agricultural environment.

III. Machine Learning in Crop Yield Prediction

Several machine learning (ML) approaches have been seen as effective tools for managing the complexity of agricultural data. Linear regression, decision trees, random forest, and gradient boosting are some quantitative approaches applied to forecast crop yield. Furthermore, ML model approaches can be trained using climate and agricultural data to increase the accuracy of the crop yield forecasting models than traditional statistical approaches. Some studies show that some sophisticated machine learning models can be continuously trained to enable them to adapt to agricultural changes. Neural networks and Gaussian process are two deep learning approaches widely used to improve the accuracy of crop yield estimation, especially when dealing with large datasets and remote sensing inputs [7], [8], [9].

IV. Adaptive and Automated Agricultural Prediction Systems

The contemporary agricultural prediction system requires a high degree of adaptation and automation, as the environment in which it is set up keeps changing all the time. The adaptive approach involves updating the prediction models on its own whenever any new input becomes available or there is a drastic change in climate or soil conditions. The process occurs automatically and is not influenced by humans at all. Studies have shown that the adaptive approach was found to be more effective than the static one. It was most useful in areas where there were daily climate variations.

V. Knowledge Sharing and Decision Support in Agriculture

A system for sharing knowledge in collaboration has been found to be a key component in improving agricultural productivity. There have been decision support systems developed in agriculture by integrating methods of yield predictions with insights that can help in decision-making. A recent study highlights the importance of anonymously sharing agricultural insights in agriculture. This is considered to be an essential framework for sharing valuable insights in various agricultural groups.

VI. Evaluation through Simulation and Involvement of the Farmers

Simulation-based approaches are gradually becoming common in agriculture research whereby the techniques aim at evaluating farming approaches and the predictions. These methods give an opportunity for farmers, agricultural researchers, and professionals to participate in the simulation process while analyzing the effects of different climatic and soil conditions on crop productivity. From research conducted, simulation-based approaches are effective in involving farming communities in developing strategies that will help maximize crop production and understanding the agricultural process better.

VII. The Challenges Facing Current Crop Yield Prediction Models

Although there have been numerous developments and advancement in technologies associated with the prediction of the crop yields through the application of machine learning techniques, numerous challenges that limit these models from achieving their maximum potentials continue to emerge. One of the key challenges is the quality of data. The agricultural data used in developing the models could be of poor quality in that they would contain gaps or incomplete data as a result of the use of several different types of information from both the weather, soil and manual information. Additionally, there has been another significant challenge that is facing crop prediction models, which is the development and testing of these models in controlled environments.

A further challenge that limits these systems is that they are not able to incorporate current information within their frameworks. This challenge arises in view of the nature of crop development that involves dynamic elements such as climatic and soil changes. However, the challenge is that some of the crop prediction systems are based on historical trends rather than the changes in the environments. Finally, lack of transparency in the model makes these systems ineffective.

VIII. Conclusion and Research Gaps:

While considerable efforts have been made for agricultural sciences with respect to the application of machine learning techniques, climatic studies, and soil-related prediction techniques, still there lacks significant effort being made on the development of comprehensive and automated crop yield prediction systems. The second point highlighted by this literature review is that adequate attention has not yet been paid to decision support systems related to combined climatic and soil parameters for crop yield predictions.

This literature review hence reflects the need for development of an integrated and adaptive crop yield prediction system, which can potentially be used in sustainable agriculture applications.

III. Problem Statement and Research Objectives

A. Problem Statement

There are several limitations within existing systems used for projecting crop yield which include:

1. The Risk Resulting from Climate Change: Conventional models of prediction rely on variables such as temperature and precipitation with great precision; therefore any subtle shifts to be anticipated will make projections unreliable. [1]

2. Limited Applicability: Because most statistical/machine learning models have been developed for particular regions and specific types of crops, they cannot be applied appropriately to predict crop yields in other areas. [2]

3. Models Currently Existing Will not Provide an Effective Comprehensive Solution - Typical models that address climate-related issues examine these variables with no regard for the relationships that exist between all of the numerous variables affecting crop yield and/or the variability associated with the unique climate where the crops are grown. [3]

4. Inflexible models: Traditional models are unable to change on their own after going through training and therefore require constant retraining to remain effective (4).

5. Cannot help with decision-making—modern prediction systems are limited to providing estimates of expected production.

With this in mind, the goal of this research is to develop a self-adapting model to predict crop yields.

B. Research Objectives

This research demonstrates how machine learning can be useful in creating a robust and scientific way of predicting the yield of crops based upon climate and soil conditions. The goal is to develop a predictive model that provides farmers with better information with greater accuracy as well as flexibility to use land for the benefit of their business.

III. PROPOSED SOLUTION

The ml-cyps (Machine Learning - Crop Yield Prediction System) is an advanced, tech-driven, and machine-learning-based system used in agriculture that uses climatic and soil factors to accurately predict crop yield based on past yields.

By using machine learning to estimate crop yield, ml-cyps helps farmers increase their productivity and develop sustainable practices for farming through improved data preparation, model adaptation, and decision making.

1. Architecture that is based on a combination of Data

ML-CYPS (Machine Learning Crop Yield Prediction System) utilizes climate and soil datasets as a combined source to solve challenges of present crop yield estimate techniques that use distinct data sets and place limitations on features while being unable to perform in a constantly shifting climate ([4] R), by using the combined datasets to create a more accurate crop yield prediction based on an accurate prediction of rainfall, temperature, humidity, level of soil acidity, and available amounts of nitrogen, phosphorus and potassium in the soil.

Assuring the datasets are reliable and consistent: The preprocessing must be performed accurately to eliminate noise and ensure the integrity of the data.
Increase the size of the datasets: This method can be utilized to process very large datasets that come from various locations.

1.1. Machine Learning Models to Predict Yield

Using machine learning allows one to make predictions on crop yields on the basis of the ability of the machine learning (ML) model to discern patterns from climate and soil data (also called agrometeorological and agrolimnological) in relation to past crop yields. The set of ML models will feature typical methodologies by researchers, including linear regression, a random forest, and gradient boosting, as shown below.

1. 1.Finding Patterns – Based on the past data of these two variables, the ML models decide the relation between climate conditions and soil conditions in terms of crop yield.

2. 2.Non-Linear Predictions- The ML models use ensemble methods for the non-linear nature of the relationship between these variables.

3. 3.Adaptive Learning – The ML model will get new data/information over time that will be used by the model to enhance its precision of predicting crop yield.

1.2. Automated Model Adjustment for Changing Environmental Factors

Classic most models are to be changed manually to keep with changes in the surrounding environment. However, the ML-CYPS has certain built-in capabilities concerning automatic adjustment based on these sorts of changes.

Automatic Refinement of Models: Automatically, models are refined when there is a decline in performance due to environmental shifts.

Dynamic Adjustment of Models: Models automatically revise

and adjust the estimated parameters to suit the changed climatic environment and soil alterations.
Improved Predictive Performance: Prediction accuracy is the result of continual automatic fine-tuning.

1.3. Knowledge Sharing and Support for Decision-Making

The cooperative sharing of knowledge and making decisions together increases the efficiency of farming.

Anonymity of Shared Data - Data about the crop yield is shared anonymously with both farmers and researchers.

Support for Decision-Making is offered by the Decision Support System. This system can provide farmers with recommendations for planting crops, irrigating those crops, and applying fertiliser.

Lessens Amounts of Resources used - Additionally, farmers may be able to reduce the amount of water, as well as soil nutrients used for plant growth.

1.4. Simulation-Based Assessment

Simulations methods are employed for assessing agriculture-related strategies and prediction impacts.

Scenario Analysis: Farmers are able to analyze various weather and soil factors and their influence on production.
Strategy Creation: Provides means for developing effective farming strategies relying on predictions.
Interaction With Farmer: Increases awareness regarding agriculture-related processes using simulations.

2. Technical Implementation

The algorithm that predicts the yield of crops is a combination of machine learning algorithms, data processing procedures, and scalable deployment methodologies that help ensure accuracy and efficiency. Here are some technical considerations about the system [6].

2.1. Tech Stack

Machine Learning Models: Linear Regression, Random Forest, Gradient Boosting
Data Processing: Python libraries such as Pandas, NumPy for data preprocessing and feature engineering
Visualization Techniques: Data analysis using Matplotlib and Seaborn
Deployment Techniques: Deployment tools can be cloud-based or local.

2.2. Essential ML Capabilities

Feature Selection: Determines crucial climatic and soil factors that influence crop productivity.
Predictive Algorithms: Predictive algorithms create forecasts of crop yields depending on learning.
Evaluation Techniques: Evaluates algorithms based on metrics such as MAE, RMSE, and R^2.
Algorithm Optimization: Refines algorithms through hyperparameter tuning.

2.3. Deployment & Management

The system guarantees scalability, reliability, and uninterrupted operation of prediction models.

Deployment of Prediction Models: The models are deployed using Python-based frameworks for real-time predictions.
Scalability: Accommodates large data sets in various agricultural regions.
Maintenance of the System: Ensures timely updates and performance management.

Stage 1: Data Acquisition from Climatic and Soils Databases

The system obtains agricultural information from different sources like weather stations, soils sensors, or government databases.
Parameters related to climatic factors such as rainfall, temperature, humidity, as well as soil conditions such as pH and nutrients are obtained.
The database is set up to mimic practical farm scenarios for analysis purposes.
After obtaining the data, it undergoes continuous monitoring and preparation for processing [9].

Step 2: Machine Learning Modeling and Pattern Recognition

The machine learning model examines weather and soil patterns in order to determine their impact on the crop yield.
The model receives information about rain, temperature fluctuations, and soil fertility among others.
Pattern recognition techniques are then utilized to establish the relationship between these factors and the yield outcomes.

Stage 3: Data Analysis and Model Assessment

All data and predictions after processing are stored in an organized manner for assessment purposes.
The data is always kept consistent and reliable within the whole process.
Measures of performance like MAE, RMSE, and R^2 are used to assess prediction accuracy.
This helps in enhancing the overall performance of models and determining major influencing factors.

Stage 4: Automatic Updating of Prediction Models

If there is any change in the environment, the prediction model automatically updates itself.
In order to ensure the accuracy of the prediction model, it will retrain on the latest climatic and soil data.
Thus, in each update process, the model improves itself based on past experiences [6].

Stage 5: Decision Support and Knowledge Exchange

Predictive information regarding crop yields is conveyed to farmers, scientists, and other organizations involved in agriculture.
Advice is given on the cultivation of crops, irrigation practices, and fertilizer utilization.
Various approaches to farming can be considered to enhance efficiency and sustainability.
This cooperative process facilitates better decision-making in agriculture.

In closing:

In summary, the application of the ML-CYPS system uses machine learning algorithms along with data processing and model adaptation in order to build the intelligent framework for prediction of agricultural output. The use of Python models and systems allows developing a viable solution to this task.

3. Evaluation & Potential Effects

Machine Learning-Enabled Crop Yield Prediction System (ML-CYPS) uses highly developed methods for making predictions within agriculture. With the use of machine learning algorithms, adaptive approaches, data handling processes, and decision-making aids, the system maintains high levels of accuracy, adaptability, and sustainability in predicting crop yields. Below is the potential effects analysis:

3.1. Prediction Accuracy and Reliability

Predicting Crop Yield: The use of climate and soil parameters helps in predicting crop yield effectively.
Data Preprocessing: Effective preprocessing guarantees reliable data processing and mitigates issues related to noise and missing data.
Performance Evaluation Metrics: The performance of the model is evaluated using metrics such as MAE, RMSE, and R^2.
Increased Trust in Predictions: Accurate predictions boost the confidence level of farmers and other stakeholders.

3.2. Flexibility

Self-learning Models: The machine learning algorithms learn continuously from the latest weather and soil information to enhance predictive efficiency.
Automated Model Updating: The model updates itself automatically when its performance declines due to the changing environment.
Prediction Flexibility: The system efficiently accommodates changes in seasons and the environment.
Flexibility in Making Decisions: Suggestions are made accordingly based on new predictions.

3.3. Collaboration & Agricultural Knowledge Sharing

Data Sharing: The model shares agricultural knowledge anonymously with the farmer community and institutes.
Collaborative Learning: Through collaboration, it becomes easier to find improved farm management techniques.
Regional Adaptation: Knowledge sharing aids in understanding crop behavior in different regions.
Decision-Making: Assists decision-makers in formulating agriculture policies.

3.4. Sustainability and Resource Optimization

Optimized Resource Utilization: The system aids in optimizing water, fertilizer, and soil nutrient utilization through predictions.
Environmental Protection: Prevents excessive resource use and encourages environmentally friendly agricultural methods.
Sustainability: Enables sustainable agricultural development in response to climatic changes.
Increased Productivity: Increases crop yields without compromising soil sustainability.

IV. FUTURE WORK

Although ML-CYPS is efficient in crop yield prediction, there are many challenges associated with its application that need to be addressed. Some of them are presented below:

4.1. Problems in Data Quality and Availability

Insufficient Datasets: The agricultural data may have missing entries, making accurate predictions challenging [1].
Possible Solution: Clean and pre-process data using data processing techniques to enhance its quality.

Variability in Data: The differences in climatic data and soil composition from different locations limit the generality of the model.
Possible Solution: Employ large-scale datasets that increase generality.

Lack of Real-time Data: The inability to access real-time data limits the functionality of the system.
Possible Solution: Employ IoT based sensor technology [3].

4.2. Machine Learning Model Drawbacks

Inaccurate Predictions: The model may generate predictions that are incorrect because of intricate patterns in agriculture [2].
Suggested Solution: Employ ensemble machine learning approaches [6].

Training Data Bias: Inadequate data or bias in training data could lead to poor results.
Suggested Solution: Train the models with varied datasets in agriculture [2].

Predictive Uncertainty: Agricultural farmers may be unable to comprehend prediction outcomes.
Suggested Solution: Implement explainable artificial intelligence approaches [1].

The deep learning algorithms used in advanced neural networks are data-intensive and consume significant computing power, making them difficult to implement in small agricultural farms [9].

4.3. Real-World Application & Adoption

System Integration Problem: Challenging integration with current agricultural systems.
Possible Solution: Create application programming interfaces and intuitive interfaces for seamless integration [5].

Privacy Issues: Exchange of agricultural information could pose privacy threats .
Possible Solution: Adopt methods of secure information exchange and anonymization.

User Knowledge: There might be insufficient awareness among farmers about utilizing sophisticated systems.
Possible Solution: Conduct training sessions and awareness campaigns for effective system utilization [4].

Concluding Thoughts & Future Actions

There can be several obstacles when developing an ML-based crop yield prediction system. Still, the ML-CYPS method is a unique technique for crop prediction and decision-making. The future study should concentrate on:
Improving the efficiency of prediction through ML models and data sets [6].

V. CONCLUSION

Agriculture is becoming increasingly challenging with respect to variability in climate conditions and changes in the environment, hence the need for an intelligent predictive system. The proposed Machine Learning–Integrated Crop Yield Prediction System (ML-CYPS) can be considered the future of agriculture because it utilizes various techniques including machine learning algorithms, climate information, and soil characteristics [3].

Key Contributions of ML-CYPS:

1. Data Integration: Considers climate and soil information to enhance the prediction efficiency and accuracy.

2. Prediction via Machine Learning: Applies sophisticated models to analyze the complexities of agricultural production and make precise yields forecasts.

3. Self-adjusting Prediction Model: Adjusts itself according to environmental variations, thereby maintaining high levels of performance [5].

4. Decision-Making Tool: Assists farmers and other decision-makers in optimizing the cultivation process and adopting sustainable farming techniques.

5. Just like an assisting mechanism, constant improvement of the machine learning algorithm is required to lower the number of erroneous predictions and accommodate for real-time environmental fluctuations.

6. Integration with the current agricultural systems becomes a prerequisite for implementation.

7. Effective management and policies for the use of agricultural data need to be put into place.

8. Knowledge Sharing for Sustainable Agriculture: The agricultural communities can exchange knowledge and ideas that could help boost their efficiency.

As problems keep changing in agriculture, ML-CYPS has brought forth an efficient, intelligent, and systematic way that is more effective than existing predictive systems. In future, developments in machine learning models and large-scale application would make a greater impact on sustainable agriculture [2,3,4].

REFERENCES

[1] T. van Klompenburg, A. Kassahun, and C. Catal. Crop yield prediction using machine learning: A systematic literature review (2020)
https://doi.org/10.1016/j.compag.2020.105709
[2] S. M. Shawon et al. Crop yield prediction using machine learning: An extensive and systematic literature review (2025)
https://doi.org/10.1016/j.atech.2024.100718
[3] U. V. Nikhil et al. Machine learning-based crop yield prediction in South India (2024)
https://doi.org/10.3390/computers13060137
[4] M. M. Islam et al. Crop yield prediction for sustainable agriculture (2024)
https://doi.org/10.3934/agrfood.2024053

[5] H. Afzal et al. Incorporating soil information with machine learning (2025)
https://doi.org/10.1038/s41598-025-88676-z
[6] R. Prabavathi and B. J. Chelliah. Machine learning approaches using soil nutrients (2022)
https://doi.org/10.13189/ujar.2022.100302
[7] S. Khaki and L. Wang.
Crop yield prediction using deep neural networks (2019)
https://doi.org/10.1016/j.compag.2019.05.005
[8] J. You et al.
Deep Gaussian Process for crop yield prediction using remote sensing (2017)
https://doi.org/10.1609/aaai.v31i1.11084
[9] K. Kamilaris and F. X. Prenafeta-Boldú.
Deep learning in agriculture: A survey (2018)
https://doi.org/10.1016/j.compag.2018.02.016