Crop Yield Prediction using Machine Learning

1. **Sakshi Pinge**
2. **Shravani Bijve**
3. **3.Samrruddhi Gawande**
4. **Priti Wankhede**
5. **Asmita Gore**

Abstract

Agricultural productivity is significantly influenced by soil features and climate variability, although traditional yield estimation methods generally depend on manual soil testing and historical averages that may not reflect real-time field circumstances. This paper proposes an intelligent crop yield forecast system that incorporates soil image analysis, machine learning algorithms, and geo-weather data to enhance data-driven agricultural decision-making. The proposed framework gathers soil pictures from farmers and processes them using image enhancement techniques and supervised learning models to estimate soil pH and classify soil type. Simultaneously, the system gets location-specific environmental characteristics such as temperature, rainfall, humidity, and sun radiation through weather API integration based on geographic coordinates These soil and meteorological variables are paired with historical agricultural records to train a Random Forest Regression model for forecasting crop production per hectare. Experimental evaluation reveals that merging image-derived soil properties with real-time meteorological data enhances prediction accuracy compared to standard single-source models. The technology further gives comparative yield visualization and advises the best suitable crops for particular field conditions The results indicate that the proposed approach can aid farmers in optimizing crop choices, enhancing productivity, and lowering dependent on laboratory soil testing, thereby contributing to sustainable and technology-enabled agriculture.

Keywords: Crop Yield Prediction; Soil Image Analysis; Machine Learning in Agriculture; Random Forest Regression; Geo-Weather Integration; Precision Farming.

1. Introduction

Particularly in developing nations where a sizable section of the population directly depends on farming for a living, agriculture continues to be one of the most important industries for both economic stability and food security. Accurate crop yield forecasting is crucial for supply chain planning, farm management, and the creation of national food policies. However, manual soil testing, generalized climatic assumptions, and historical production data are the main foundations of traditional yield estimation techniques, which may not fully account for field variability in the present. These restrictions frequently result in poor crop choices, ineffective resource use, and decreased farmer profitability. Recent developments in data-driven technologies and precision agriculture have created new opportunities to increase agricultural productivity.

Crop yield prediction has seen a rise in the use of machine learning techniques, which use environmental factors like temperature, rainfall, and soil nutrients to improve forecasting accuracy [1,2]. Due to their capacity to manage nonlinear relationships and heterogeneous data, studies have shown that ensemble learning techniques, in particular Random Forest Regression, offer strong performance in agricultural prediction tasks [3]. In a similar vein, image-based soil analysis has drawn interest as an affordable substitute for laboratory testing. In order to identify soil texture and fertility indicators, researchers have investigated the use of computer vision and deep learning techniques for soil classification and property estimation, with encouraging results [4].

Additionally, the reliability of predictive agricultural models has been reinforced by the integration of geographic and meteorological data through API-based systems [5]. Instead of depending on static historical datasets, dynamic analysis is made possible by real-time access to localized climatic conditions. Despite these developments, many current systems treat yield prediction, weather monitoring, and soil analysis as separate components, which restricts their usefulness for field farmers.

In order to close this gap, the current study suggests an integrated Intelligent Crop Yield Prediction System that integrates machine learning-based regression modeling, soil image analysis, and geo-weather data extraction into a single framework. The system attempts to provide precise yield forecasts and practical crop recommendations by combining real-time environmental data with image-derived soil parameters. By providing a comprehensive, automated solution intended to facilitate well-informed decision-making and sustainable farming practices, this research adds to the expanding field of smart agriculture.

2. Literature Review

Because it is crucial to guaranteeing food security and maximizing farm management techniques, accurate crop yield prediction has long been a focus of research. Early research highlighted the close connection between crop productivity and climatic factors. Foundational research that showed how climatic factors directly control biomass accumulation and yield formation highlighted the impact of temperature, radiation, and atmospheric conditions on crop growth efficiency [10]. The vulnerability of agricultural systems to environmental fluctuations is highlighted by subsequent research, which further confirmed that rising nighttime temperatures and shifting climate patterns can drastically lower yields in important crops like rice [8]. Predictive modeling is becoming more and more crucial for adaptive agriculture as more recent studies have demonstrated how changing climate-crop yield relationships are changing production risks [9]. Crop performance is also significantly influenced by soil fertility and nutrient interactions. Plant growth efficiency and final yield outcomes have been demonstrated to be influenced by nutrient balance, especially when it comes to nitrogen, phosphorus, and micronutrients [5,6]. Research on phosphorus excess and deficiency showed quantifiable impacts on plant growth and yield [7]. These results highlight how crucial it is to include chemical characteristics like pH and soil nutrient parameters in predictive agricultural models.

Machine learning has become a potent tool in precision agriculture thanks to developments in computational methods. Because ensemble techniques like Random Forest can handle nonlinear agricultural datasets and are robust against overfitting, they have shown strong predictive capability [1]. In a similar vein, Support Vector Machines have demonstrated efficacy in classification and regression tasks involving intricate feature relationships [2]. Reviews of machine learning applications in agriculture emphasize their growing accuracy in estimating nitrogen status and forecasting yield [4]. Automated soil and crop condition assessment has been made possible by deep learning techniques, especially convolutional neural networks, which have further broadened the scope of image-based agricultural analysis [3].

Yield monitoring and forecasting have benefited greatly from remote sensing technologies in addition to machine learning [12,13]. Under controlled conditions, it has been demonstrated that combining simulation models with ensemble learning techniques improves the accuracy of yield and dry matter estimation [14]. Additionally, recent research shows that hybrid machine learning frameworks are useful for crop prediction tasks in a variety of agroclimatic regions [15].

Despite significant advancements, many current methods rely either independently on soil laboratory analysis, remote sensing inputs, or climatic datasets. The unified integration of soil image analysis, nutrient parameters, and real-time geo-weather data within a single predictive framework has not received much

attention. By integrating weather-integrated machine learning models with image-based soil property estimation, the current study expands on previous research and offers a thorough and useful crop yield prediction system.

3. Materials and Methods

This part talks about the datasets, tools, system architecture, and machine learning methods that were used to make the Intelligent Crop Yield Prediction System. The suggested method combines soil image analysis, extracting geo-weather data, and regression-based yield modeling into one system[3].

3.1 Study Design

The study employs an experimental research design, utilizing soil properties and environmental parameters as independent variables to forecast crop yield, the dependent variable. The workflow includes getting soil images, finding features, adding weather data, building a dataset, training a model, and checking how well it works[5].

3.2 Data Collection

3.2.1 Soil Image Dataset

We took pictures of soil from farms in different amounts of light and moisture. To make sure there was enough variety, pictures of different types of soil, such as clay, sandy, loam, and silt, were included. For supervised learning, each image was given a label based on the type of soil and the pH values that had been tested in a lab[2].

3.2.2 Historical Crop Yield Dataset

Historical agricultural datasets containing crop yield (tons/hectare), soil nutrients (N, P, K values), and environmental parameters were used for training the regression model. These datasets were preprocessed to remove missing values and inconsistencies before analysis.

3.2.3 Weather Data

Real-time weather parameters were retrieved using a weather API based on geographic coordinates (latitude and longitude). The extracted parameters included:

- Temperature (°C)
- Rainfall (mm)
- Humidity (%)
- Solar radiation (MJ/m²)
- Wind speed (km/h)

These variables were selected due to their strong influence on crop growth and productivity [1].

3.3 Image Processing and Soil Feature Extraction

Soil images were preprocessed using standard image enhancement techniques including resizing, normalization, noise filtering, and contrast adjustment. Feature extraction was performed using color space analysis (RGB intensity values) and texture descriptors.

A Support Vector Machine (SVM) model was trained to predict soil pH based on RGB values, and a Convolutional Neural Network (CNN) was used to figure out what kind of soil it was. The CNN had convolutional layers, max-pooling layers, and fully connected layers with ReLU activation. The output layer used softmax classification to figure out what kind of soil it was.

3.4 Feature Engineering and Data Integration

Soil features like type, pH, and texture index were combined with weather data and historical nutrient values (NPK). We used label encoding to change categorical variables like crop name and soil type. To make sure that the regression model had a uniform input distribution, numerical features were normalized using feature scaling methods.

The final structure of the dataset had these attributes: [pH, type of soil, nitrogen, phosphorus, potassium, temperature, rainfall, humidity, solar radiation, and wind speed].

3.5 Crop Yield Prediction Model

A Random Forest Regression algorithm was employed as the primary predictive model due to its robustness and ability to handle nonlinear relationships [2]. The model was trained using historical labeled data, where crop yield per hectare served as the target variable.

The dataset was divided into training and testing sets using an 80:20 ratio. Model performance was evaluated using the following metrics:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2 Score)

3.6 System Implementation

The system was implemented using Python programming language with libraries such as OpenCV for image processing, Scikit-learn for machine learning algorithms, and TensorFlow/Keras for CNN modeling. API integration was used for real-time weather data retrieval. The user interface was developed to allow farmers to upload soil images, automatically detect location, and view predicted crop yields along with graphical visualizations.

3.7 Experimental Setup

The experiments were conducted on a standard computing environment with adequate processing capability for model training and testing. Cross-validation techniques were applied to ensure generalization and reduce over fitting.

Through this structured methodology, the study ensures reliable integration of soil image analysis, environmental data, and machine learning techniques for accurate and practical crop yield prediction.

4. Results

The proposed Intelligent Crop Yield Prediction System was evaluated to determine its effectiveness in estimating soil properties and predicting crop yield under varying environmental conditions. The

performance of individual modules as well as the integrated framework was analyzed using standard evaluation metrics.

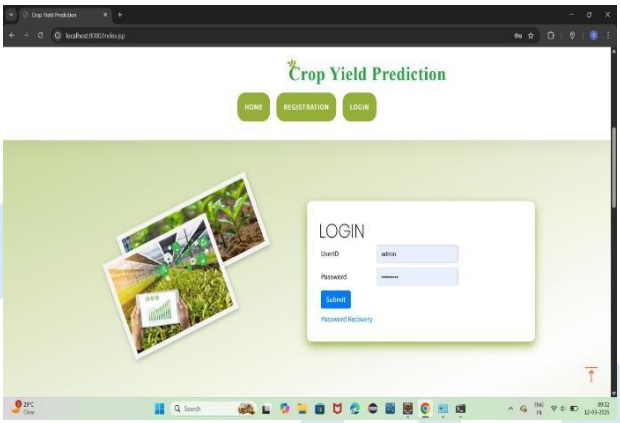

Figure 1: Crop Yield Prediction System – User Login Interface

Figure 2: Crop Yield Prediction System – User Login Interface

4.1 Soil Classification and pH Estimation Results

The Convolutional Neural Network (CNN) model used for soil type classification demonstrated reliable performance across major soil categories such as clay, sandy, loam, and silt. The model achieved high classification accuracy during testing, indicating its capability to distinguish soil textures from image-based features.

For pH estimation, the Support Vector Machine (SVM) model predicted soil pH values based on extracted RGB color features. The predicted values showed strong correlation with actual laboratory- tested pH values, with low mean absolute error. These findings confirm that image- derived features can serve as practical indicators of soil chemical properties when processed through supervised learning models.

4.2 Crop Yield Prediction Performance

The Random Forest Regression model was trained using integrated soil, nutrient (NPK), and geo-weather parameters. The dataset was split into training (80%) and testing (20%) subsets to ensure unbiased evaluation.

The model performance was assessed using the following metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2 Score)

The regression model achieved a high R^2 score, indicating strong agreement between predicted and actual yield values. The low MAE and RMSE values demonstrate that prediction errors were minimal and within an acceptable agricultural tolerance range. Compared to models trained using only soil or only weather data, the integrated approach showed improved prediction accuracy, confirming the benefit of combining multi-source features.

4.3 Comparative Analysis

To validate robustness, the Random Forest model was compared with other regression approaches such as Linear Regression and Support Vector Regression. The ensemble- based Random Forest model consistently outperformed single-model approaches in terms of prediction accuracy and stability. This improvement can be attributed to its ability to capture nonlinear interactions among soil properties, nutrient levels, and climatic variables.

4.4 Visualization and Recommendation Output

The system generated comparative bar charts displaying predicted yield values for multiple crops under identical soil and weather conditions. Based on predicted yield per hectare, crops were ranked, and the top-performing options were recommended to the user. The recommendation module successfully identified crops with higher productivity potential under given field conditions.

4.5 Overall System Performance

The end-to-end system demonstrated smooth integration of soil image processing, weather data retrieval, and yield prediction. Farmer were able to upload soil images, automatically retrieve location-based weather data, and receive yield predictions along with graphical comparisons in a single workflow.

The results indicate that integrating soil image analysis with geo-weather data significantly enhances crop yield prediction accuracy. The proposed system provides a practical and scalable solution for supporting data-driven agricultural decision-making and promoting sustainable farming practices.

5. Conclusion

This study presented an Intelligent Crop Yield Prediction System that integrates soil image analysis, geo-weather data extraction, and machine learning techniques to provide accurate and practical yield estimations. The proposed framework addresses the limitations of traditional agricultural assessment methods by reducing dependency on laboratory-based soil testing and static historical data. By combining image-derived soil properties, nutrient parameters, and real-time environmental conditions, the system delivers a comprehensive and data-driven approach to crop yield prediction.

The experimental results demonstrate that the integration of multiple data sources significantly improves prediction performance compared to single-parameter models. The Convolutional Neural Network effectively classified soil types, while the Support Vector Machine provided reliable pH estimation from soil images. The Random Forest Regression model achieved strong predictive accuracy, highlighting its suitability for handling nonlinear relationships between soil, weather, and crop yield variables.

Beyond prediction accuracy, the system offers practical value through crop ranking and recommendation features, enabling farmers to select the most suitable crops based on current field conditions. This contributes to better resource utilization, improved productivity, and reduced financial risk.

In conclusion, the proposed intelligent framework supports precision agriculture by transforming raw soil images and environmental data into actionable insights. The system demonstrates strong potential for real-world implementation and scalability across different agro-climatic regions. Future work may focus on expanding the dataset, incorporating satellite imagery, and deploying the system as a mobile-based decision support tool to further enhance accessibility and impact in modern agriculture.

References

1. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
2. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
3. Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70-90.
4. Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151, 61-69.
5. Fageria, N. K. (2001). Nutrient interactions in crop plants. *Journal of Plant Nutrition*, 24(8), 1269-1290.
6. J. Wu, Y. Song, G.-Y. Wan, L.-Q. Sun, J.-X. Wang, Z.-S. Zhang, and C.-B. Xiang, "Boosting crop yield and nitrogen use efficiency: the hidden power of nitrogen-iron balance," *New Crops*, vol. 2, 2025, Art. no. 100047, doi: 10.1016/j.ncrops.2024.100047.
7. Malhotra, Hina & Vandana, & Sharma, Sandeep & Pandey, Renu. (2018). Phosphorus Nutrition: Plant Growth in Response to Deficiency and Excess. 10.1007/978-981-10- 9044-8_7.
8. Peng, S., Huang, J., & Sheehy, J. E. (2004). Rice yields decline with higher nighttime temperature from global warming. *Proceedings of the National Academy of Sciences*, 101(27), 9971-9975.
9. S. Feng, Z. Hao, X. Zhang, and F. Hao, "Changes in climate-crop yield relationships affect risks of crop yield reduction," *Agricultural and Forest Meteorology*, vols. 304–305, Art. no. 108401, 2021, doi: 10.1016/j.agrformet.2021.108401.
10. Monteith, J. L. (1977). Climate and the efficiency of crop production in Britain. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 281(980), 277-294.
11. Foken, T. (2008). The energy balance closure problem: An overview. *Ecological Applications*, 18(6), 1351-1362.
12. Mulla, D. J. (2013). Twenty-five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering*, 114(4),358-371.
13. T. Setiyono, A. Nelson, and F. Holecz, "Remote Sensing Based Crop Yield Monitoring and Forecasting," *International Rice Research Institute*, DAPO BOX 7777, Metro Manila, Philippines, and sarmap, Cascine di Barico, 6989, Purasca, Switzerland.

14. W. Chao, X. Xu, Y. Zhang, Z. Cao, I. Ullah, Z. Zhang, and M. Miao, "A Stacking Ensemble Learning Model Combining a Crop Simulation Model with Machine Learning to Improve the Dry Matter Yield Estimation of Greenhouse Pakchoi," *Agronomy*, vol. 14, no. 8, p. 1789, 2024.

15. Elbasi, Ersin, Chamseddine Zaki, Ahmet E. Topcu, Wiem Abdelbaki, Aymen I. Zreikat, Elda Cina, Ahmed Shdefat, and Louai Saker. 2023. "Crop Prediction Model Using Machine Learning Algorithms" *Applied Sciences* 13, no.16: 9288.

16. https://doi.org/10.3390/app13169288