10.48047/jocaaa.2024.33.08.273

Potato Crop Yield Prediction Using Environmental and Soil Parameters

Yukti Kesharwani[1*], Dr. Amrita Verma[2], Dr. Rajesh Tiwari[3]

[1]Research Scholar, [2]Associate Professor, [3]Professor
[1,2] Department of Engineering(CSE),[3] Department of Computer Science & Engineering
[1,2] Dr. C.V. Raman University,Kota, Bilaspur, Chhattisgarh, India
[3] CMR Engineering College, Hyderabad, India

Corresponding Email : yuktikesharwani20@gmail.com,
amrita.85024@gmail.com, drrajeshtiwari20@gmail.com

Abstract—

Modern agricultural systems heavily depend on precise crop yield prediction to optimize farm management, boost productivity, and safeguard food security. Given the increasing unpredictability of climatic patterns and fluctuating soil conditions, traditional farming methods rooted in farmer intuition are no longer adequate for effective decision-making. Consequently, data-driven predictive modeling has emerged as a promising methodology for achieving precision agriculture. This research details the development of a machine learning-based framework for forecasting potato crop yield, leveraging environmental and soil-based parameters. A publicly accessible dataset from Kaggle, comprising 36,521 records, was utilized; it includes crucial attributes such as soil pH, temperature, humidity, wind speed, soil type, and nutrient levels, all of which significantly influence potato growth and yield. The study focuses on the implementation and comparative analysis of three algorithms—Random Forest Regression, XGBoost Regression, and Multi-Layer Perceptron—for yield prediction. Each model underwent training and testing using an 80:20 data split, with performance evaluation primarily conducted using Mean Absolute Error due to its robustness in yield-based prediction. The experimental findings indicate that the Random Forest Regression model achieved the lowest MAE, thereby demonstrating superior predictive accuracy when compared to XGBoost and MLP. This study concludes that machine learning-based predictive modeling can substantially assist farmers in making informed cultivation decisions. Future research can build upon these findings by incorporating hybrid models, hyperparameter optimization, and feature importance analysis to further enhance prediction accuracy.

Keywords- Machine Learning, Deep Learning, Potato Crop Growth, Atmospheric Parameters, Predictive Modelling

1. INTRODUCTION

The potato is a major food crop in India, cultivated extensively across the country. Several factors, including soil pH, soil type, and the availability of essential nutrients such as nitrogen (N), phosphorus (P), and potassium (K), can influence potato crop yields. These factors often limit the crop's potential productivity. Moreover, the increasing instability of climate patterns contributes to reduced potato production, highlighting the need for robust and effective technological interventions to safeguard yields against adverse climatic effects.

10.48047/jocaaa.2024.33.08.273

Conventional models such as DSSAT and AquaCrop have long served as the backbone for yield prediction. However, these rely on complex calibration processes and often lack the flexibility to adapt across changing climatic conditions and heterogeneous datasets. Moreover, they are limited in their capacity to model nonlinear dependencies or integrate large-scale real-world data without domain-specific configurations.

To address these limitations, this study adopts a modern, data-centric methodology that utilizes real-world, historical crop yield data with integrated soil and weather parameters. The dataset encompasses 36,521 entries with diverse crop types—this study focuses specifically on the potato subset—and includes key agronomic and climatic variables such as Crop Type, Date, Month, Soil pH, Temperature, Humidity, Wind Speed, and Nutrient Content (N, P, K).

Three predictive models are employed and benchmarked:

1. Random Forest Regression –utilizes multiple decision trees, resulting in a higher degree of accuracy compared to a single decision tree.
2. XGBoost Regression –is a robust gradient boosting algorithm in machine learning. It primarily functions to forecast a target value from a given set of input values, enhance predictive accuracy, and effectively manage large-scale datasets.
3. Multi-Layer Perceptron (MLP) – as their name implies, incorporate multiple hidden layers, making them suitable for addressing highly complex problems.

Objective:

- To develop a model that can accurately and precisely evaluate potato crop yield under different environmental conditions and soils.
- Compare the effectiveness of classical ensemble vs deep neural models,
- Identify the atmospheric and soil features most impactful for yield prediction.

The following sections of this manuscript are organized as follows: section 2 offers an in-depth review of the current literature on yield prediction and emphasizes the contributions made by earlier researchers in this area. Section 3 outlines the methodology, detailing the data set, model development, and experimental framework. Section 4 showcases the results and engages in a thorough discussion that evaluates the performance of the implemented models. Lastly, section 5 wraps up the study with a conclusion of the main findings.

2. LITERATURE REVIEW

Shan et al. [1] identified that the Aqua Crop model is a good model which gives accurate results for measuring the growth of maize crops. They compared biomass and yield using the model under different soil parameters. This research revealed that the Aqua Crop model gave the best performance in coastal saline environment.

Moroozeh et al.[2] has focused his research on the growth of two crops: soybean and maize, and used two techniques to measure their growth: the DSSAT model and the Aquacrop model. After evaluating both the crops on the basis of different parameters, it was found that the result of DSSAT model for soybean was excellent and the result of crop model for maize was also excellent. During the study, it was found that each model can evaluate a particular crop better on the basis of different parameters.

10.48047/jocaaa.2024.33.08.273

Adeboye et al.[3] evaluated maize crops at two different levels using the AquaCrop model. The model was calibrated using a wetter season's data and successfully predicted canopy cover, biomass, and grain yield, showing strong accuracy ($R^2 \geq 0.88$ for canopy cover, $R^2 = 0.99$ for yield). While it slightly over- or underestimated soil water storage and evapotranspiration, predictions remained within acceptable ranges. The study concluded that applying 150% of the recommended NPK rate improved both yield and water productivity. However, additional testing across cultivars and locations is needed for broader validation.

Addu et al. [4] explored various ML techniques like Nearest Neighbour, Random Forest and Voting Classifiers for predicting crop productivity. as Agriculture in India is highly affected by climate change.Their aim is to help farmers and policymakers in taking crop related decisions.These classifier have predicted the crop yield using various parameters like rainfall, humidity, soil pH etc.In one way, the objective of this research is also to eliminate the growing gap between agriculture and technology.

Suruliandi et al. [5] analyzed the efficacy of various machine learning methods in accurately forecasting agricultural productivity considering distinct soil and climatic factors. This research addressed the drawbacks of all the traditional agricultural practices that are affected by climate change on agriculture and focused on some specific features for accurate prediction.In this research, Recursive Feature Elimination with Adaptive Bagging Classifier, working together with these two techniques could accurately predict the suitable crops for specific land.

Raja et al. [6] investigated the application of machine learning for crop prediction. Their research underscored the necessity of employing optimal feature selection to refine model precision through the removal of extraneous data, which also enhances computational effectiveness. The application of various feature selection methodologies and classification algorithms demonstrated that ensemble-based models outperformed conventional classification approaches in prediction accuracy, thereby underscoring their utility in agricultural decision support systems.

Kuradusenge et al. [7] investigated the effects of climate change on agriculture. Farmers were also encouraged to predict crop production early to mitigate potential crop losses. The study used machine learning techniques, including random forests, polynomial regression, and support vector machines, to forecast the yields of Iris potato and maize crops using historical meteorological data. The random forest model outperformed the other models, achieving an R^2 value of 0.875 for Iris potatoes and 0.817 for maize.

Cedric et al. [8] utilized machine learning techniques to enhance the productivity of six key West African crops—rice, maize, cassava, seed cotton, yams, and bananas—by generating early predictions to help prevent agricultural output losses. The researchers developed models using k-nearest neighbor and multivariate logistic regression, incorporating parameters such as weather, climate, yield, and chemical data. Cross-validation and hyperparameter tuning were also implemented. The decision tree model demonstrated the highest accuracy, with an R^2 value of 95.3%, followed by k-nearest neighbor with 93.15% R^2 and logistic regression with 89.78% R^2.

Shankar et al. [9] investigated machine learning strategies for improving the accuracy of crop projections affected by climate and soil characteristics in India. The research encompassed four categorization algorithms: the Random Forest, the Logistic Regression, the Decision Tree, and the Support Vector

10.48047/jocaaa.2024.33.08.273

Machine. These algorithms identified potential crops based on certain environmental and soil parameters. The performance of all of these methods was assessed using ROC-AUC, precision, recall, F1 score, and support matrix.

Kalhotra et al. [10] stressed the importance of advanced crop output forecast using machine learning approaches, taking into account population expansion, which will boost demand for crops. The study elucidates machine learning methodologies and their applications in the agricultural sector, underscoring the significance of these techniques in aiding farmers with crop-related decision-making.

Padia and Sarvaiya [11] investigated the role of computer science in agriculture to enhance crop productivity and found that machine learning techniques and artificial intelligence play a significant role in enhancing crop productivity, So that accurate prediction of crop production can be made. Different techniques were used in the study - like regression model, decision tree, neural network and hybrid approaches. The model was created by analyzing historical climatic and soil data, and it was also evaluated through experiments.

Patil et al. [12] investigated the application of machine learning to improve agricultural productivity by predicting crop yields and selecting suitable crops to produce. The study also addressed the limits of traditional approaches, which are unable to account for many environmental variables and hence produce erroneous results. The researchers created models based on several parameters, and when tested, the random forest algorithm produced the most accurate predictions (MAE = 0.64, $R^2 = 0.96$). Whereas the naive base algorithm correctly detected the crop type with 99.39%.

Borekar et al.[13] investigated how machine learning can help handle crop productivity concerns. In which it helped farmers in selecting crops by evaluating the model based on different parameters like temperature, rainfall, land area etc. using historical data. In the research, a comparative analysis of the decision tree algorithm along with other algorithms was done to find a better model. Along with this, the technique also proved to be helpful to farmers in taking decisions regarding crop productivity.

Van Klompenburg et al. [14] studied various predictive machine learning techniques and applications for improving crop forecasting. The research analysed several variables and methodologies employed in multiple studies to forecast crop yield. In machine learning, precipitation, weather, and variety of soil are important variables, as well as deep learning approaches including LSTM, convolution neural networks, and algorithms based on deep neural networks are extensively researched.This work will give researchers with a comprehensive roadmap for reliably predicting crop yield using both deep learning and machine learning.

Reddy and Kumar [15] examined the function and applications of machine learning techniques in accurately forecasting improved crop yields, highlighting the necessity of integrating machine learning into Indian agriculture. According to the research, climate variability, weather conditions, and other factors have a significant impact on crop productivity.These parameters were assessed, and various predictive machine learning method were investigated to aid in properly predicting agricultural output. The research also pointed up the limitations of neural network technology. In which neural networks have low prediction

10.48047/jocaaa.2024.33.08.273

efficiency and high error rates. Aside from that, the study included a comparative comparison of machine learning methods, which will facilitate the creation of an appropriate model to enhance agricultural output.

El-Kenawy et al. [16] studied different algorithms using deep learning and machine learning for reliably estimating potato crop yield. In addition, the emphasis is on precise prediction using appropriate methods for optimum resource utilisation in agriculture and food security. The research assessed several algorithms used for machine learning, such asthe K-Nearest-Neighbors (KNN), the Gradient Boost (GB), the XGBoost, and the Multi-Layer Perceptron (MLP), as well as method of deep learning techniques that include the Graph Neural Networks (GNNs), the Long Short Term Memory (LSTM), and the Gated Recurrent Units. MSE, RMSE, and R^2 were used to evaluate various approaches. GNNs outperformed the other models, with an MSE of 0.02363 and R^2 of 0.51719. Gradient boosting and LSTM algorithms both performed better.

Kurek et al. [17] studied how machine learning algorithms could forecast the production increase of a French fry potato variety. Data from five growth seasons in 36 distinct fields were collected for the study between 2018 and 2022. The collection comprises data from five seasons, which is quite extensive in itself. These datasets were utilised to generate 3 models: non-satellite model, satellite model, and hybrid model. The combination of the models had the highest accuracy, with a mean absolute percentage error of 5.85%.

Wang & Su [18] examined the potato is crucial for global food security, and deep learning is central to intelligent agricultural practices. The integration of deep learning in potato cultivation holds significant potential for optimizing both yield and economic returns. Consequently, investigating efficient deep learning models for potato production is critically important. Key applications of deep learning across the potato production chain, designed to boost yield, encompass the detection and diagnosis of pests and diseases, monitoring plant health, predicting yield, assessing product quality, developing irrigation strategies, managing fertilization, and forecasting prices. This review primarily aims to synthesize current research progress on deep learning within various stages of potato production and to guide future research endeavors. Specifically, this paper classifies the applications of deep learning in potato cultivation into four distinct categories, facilitating a discussion of their respective advantages and disadvantages in these areas, and it outlines prospective research directions.

Zinzinhédo et al. [19] conducted a systematic and critical review on machine learning-based root and tuber yield prediction, utilizing the Preferred Reporting Items for Systematic Reviews and Meta-analyses approach. The findings indicate that root and tuber yield can be forecasted either pre-season or during the growing season. Among the root and tuber species, potato received the most considerable research focus. Temperature, precipitation, and vegetation indices emerged as the most frequently considered predictors. A notable 65.38% of the studies contrasted Linear Regression with other learning algorithms. Random Forest and Support Vector Machine were identified as the most prevalent machine learning algorithms, while the coefficient of determination and root mean squared error were the commonly applied evaluation metrics.

Archana & Kumar [20] examined precise crop yield forecasting models offer farmers essential decision-making instruments for optimized choices. Key factors influencing crop yield include fluctuations in temperature and rainfall, the prevalence of plant diseases and pests, fertilizer application, and soil quality.

10.48047/jocaaa.2024.33.08.273

This paper examines the factors impacting crop yield, investigates the features employed, and analyzes deep learning methodologies and performance metrics used in crop yield prediction.

3. METHODOLOGY

3.1 Overview

This study aimed on creating a model capable of properly predicting potato yield. Three distinct models were employed: the Random Forest Regression Model, the XGBoost Regression Model, and the Multi-Layer Perceptron Model. These models were trained, tested, and evaluated on a real-world agricultural dataset that included crop yield and weather-related information. Figure 1 shows the outline of the work.

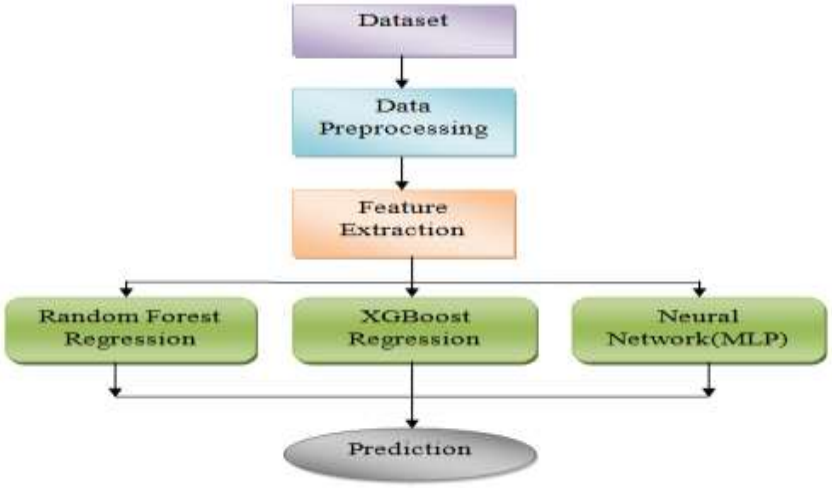

Figure 1. Outline of the work

3.2 Dataset

The dataset employed in this study was sourced from Kaggle and comprises 36,521 occurrences and 12 attributes, including:

- Atmospheric factors include temperature,humidity, and wind speed.
- Soil parameters include soil type, pH, quality, N, Pand K.
- Crop information includes crop kind, date, and month.
- Cropyield is the target.Records related to potato crops were selected from this multi-crop dataset to enable targeted prediction.

3.3 Data Pre-processing

- Filtering: Select only *Potato* entries from the multi-crop dataset.
- Missing Values: Impute nulls using statistical methods (mean/mode).
- Encoding: To encode categorical variables (Crop_Type, Soil_Type etc.) use Label.
- Scaling: Normalize features for better model convergence (especially for MLP).
- Feature Engineering: Derive additional useful features if needed (e.g., seasons from Date).

10.48047/jocaaa.2024.33.08.273

3.4 Feature Selection

Utilise a correlation matrix, Random Forest feature importance, or domain expertise to ascertain the most pertinent factors influencing potato yield.

3.5 Model Development

3.5.1 Random Forest Regression

The Random Forest is an ensemble learning method based on decision trees. It produces many trees during training and provides the average prediction of each.

- Hyper parameters have been tweaked. Number of estimators, maximum depth, and minimum sample split.
- Supports non-linearity, resilient to outliers, and offers feature importance.

3.5.2 XGBoost Regression

XGBoost is a robust boosting methodology that incrementally constructs trees while optimising a loss function.

- Hyperparameters have been tweaked. Learning rate, n_estimators, maximum_depth, and subsample
- Advantages include high performance, regularization to prevent overfitting, and efficient handling of big datasets.

3.5.3 Neural Network (MLP)

MLP is a sort of feed-forward deep neural network made up of several dense layers.

- Architecture: Input layer → 2 hidden layers (ReLU) → Output layer
- Hyperparameters: No. of neurons, epochs, batch size, optimizer (Adam), learning rate
- Advantages: Captures complex nonlinear interactions and is scalable to huge datasets.

4. RESULTS AND DISCUSSION

Comparing the performance of MAE with classifier

The Mean Absolute Error (MAE) was employed to evaluate the predictive accuracy of the Random Forest Regression, XGBoost Regression, and Multilayer Perceptron (MLP) Neural Network models. The Random Forest model exhibited a minimalMAE of 2.34, suggesting the greatest consistent success in reducing average prediction error. The MLP Neural Network followed closely behind with an MAE of 2.48, indicating competitive accuracy in identifying the fundamental trends in the data. On the other hand, XGBoost Regression had the greatest MAE of 2.99, indicating less accurate predictions. According to the results presented in the table below, the random forest regression model based on MAE outperformed all other models in accurately forecasting potato production. Table 1 shows the performance comparison of the model. Figure 2 shows the performance comparison.

10.48047/jocaaa.2024.33.08.273

Table 1.Performance Comparison

Model	MAE
Random Forest Regression	**2.34**
XGBoost Regression	**2.99**
MLP Neural Network	**2.48**

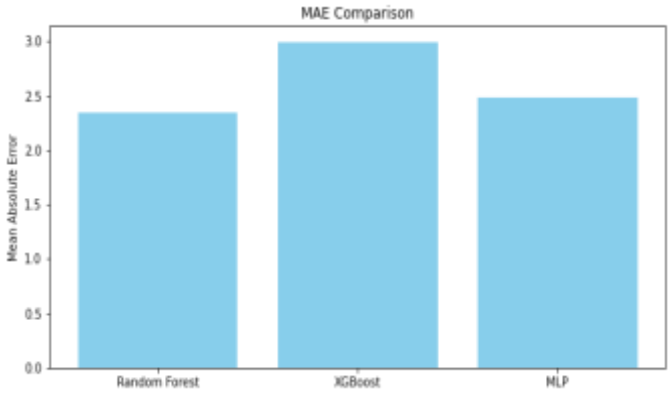

Figure 2. Performance Comparison

5. CONCLUSION

The principal objective of this study is to evaluate the efficacy of three models in accurately forecasting potato crop yield:Random Forest Regression, XGBoost Regression, and Multilayer Perceptron. The performance of these three models has been identified using MAE. Among the three models, Random Forest Regression exhibits the lowest Mean Absolute Error (MAE) of 2.34, signifying superior performance relative to the other models. Following that, Multi-Layer Perceptron scored an MAE of 2.48, followed by XGBoost at 2.99. The results indicate that the Random Forest Model demonstrates superior efficiency in predicting potato production compared to other models, as evidenced by its lowest MAE.

REFERENCES

10.48047/jocaaa.2024.33.08.273

[1] Shan, Y., Li, G., Su, L., Zhang, J., Wang, Q., Wu, J., **Mu, W.,** & Sun, Y. (2022). Performance of AquaCrop model for maize growth simulation under different soil conditioners in Shandong Coastal Area, China. *Agronomy*, *12*(7), 1541.

[2] Moroozeh, A. D., Bansouleh, B. F., Ghobadi, M., & Ahmadpour, A. (2023). Assessment of DSSAT and AquaCrop models to simulate soybean and maize yield under water stress conditions. *Spanish Journal of Agricultural Research*, *21*(3), e1201-e1201.

[3] Adeboye, O. B., Schultz, B., Adeboye, A. P., Chukalla, A., & Shittu, K. A. (2024). Assessment of the AquaCrop model to simulate the impact of soil fertility management on evapotranspiration, yield, and water productivity of maize (Zea May L.) in the sub-humid agro-ecology of Nigeria. *Discover Agriculture*, *2*(1), 28.

[4] Addu, S., Sheelam, S., Mekala, S., Sulthana, N., Mekala, L., & Alsalami, Z. (2024). Assessing Environmental Impact: Machine Learning for Crop Yield Prediction. In *E3S Web of Conferences* (Vol. 529, p. 03008). EDP Sciences.

[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.

[6] Raja, S. P., Sawicka, B., Stamenkovic, Z., & Mariammal, G. (2022). Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers. *IEEE Access*, *10*, 23625-23641.

[7] Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., Uwitonze, C., Ngabonziza,J.& Uwamahoro, A. (2023). Crop yield prediction using machine learning models: Case of Irish potato and maize. *Agriculture*, *13*(1), 225.

[8] Cedric, L. S., Adoni, W. Y. H., Aworka, R., Zoueu, J. T., Mutombo, F. K., Krichen, M., & Kimpolo, C. L. M. (2022). Crops yield prediction based on machine learning models: Case of West African countries. *Smart Agricultural Technology*, *2*, 100049.

[9] Shankar, P., Pareek, P., Patel, M. U., & Sen, M. C. (2022). Crops Prediction Based on Environmental Factors Using Machine Learning Algorithm. *Center for Development Economic Studies*, *9*(11), 127-137.

[10] Kalhotra, S. K., Prakash, K. C., Mishra, M. K., Kumar, M. A. K., & Annapurna, M. S. (2023). A Study of Crop Yield Prediction Using Machine Learning Approaches. *Journal of Advanced Zoology*, *44*(S-5), 1260-1263.

[11] Padia, N., & Sarvaiya, M. (2021). Computer Science Applied in Agriculture: Crop Yield Prediction. *Available at SSRN 5105356*.

[12] Patil, P., Athavale, P., Bothara, M., Tambolkar, S., & More, A. (2023). Crop selection and Yield Prediction using machine learning approach. *Current Agriculture Research Journal*, *11*(3).

[13] Borekar,T.,Yadav, A., Damdhar, A., Shriwas, P., & Wasankar, S.(2022). Crop Yield Estimation Using Machine Learning Algorithms.*International Journal of Creative Research Thoughts*.

[14] 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.

[15] Reddy, D. J., & Kumar, M. R. (2021, May). Crop yield prediction using machine learning algorithm. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1466-1470). IEEE.

[16] El-Kenawy, E. S. M., Alhussan, A. A., Khodadadi, N., Mirjalili, S., & Eid, M. M. (2024). Predicting potato crop yield with machine learning and deep learning for sustainable agriculture. *Potato Research*, 1-34.

[17] Kurek, J., Niedbała, G., Wojciechowski, T., Świderski, B., Antoniuk, I., Piekutowska, M., Kruk, M., & Bobran, K. (2023). Prediction of potato (solanum tuberosum l.) yield based on machine learning methods. *Agriculture*, *13*(12), 2259.

[18] Wang, R. F., & Su, W. H. (2024). The application of deep learning in the whole potato production Chain: A Comprehensive review. *Agriculture*, *14*(8), 1225.

[19] Zinzinhédo, M. L., Salako, V. K., & Glèlè Kakaï, R. (2024). Roots and tubers yield prediction using machine learning: a systematic and critical review. *Available at SSRN 4915786*.

[20] Archana, S., & Kumar, P. S. (2023). A Survey on Deep Learning Based Crop Yield Prediction. *Nature Environment & Pollution Technology*, *22*(2).