Incorporating Meteorological Data and Pesticide Information to Forecast Crop Yields Using Machine Learning

Ms. M. Lakshmi [1], Ms. S. Sahaja [2],

Assistant Professor [1][2], Department of Computer Science and Engineering, Geethanjali Institute of Science and Technology, Nellore, Andhra Pradesh-524137

Abstract: The agricultural sector faces growing challenges due to climate change and the overuse of pesticides, which threaten global food security. Accurately forecasting crop yields is critical to addressing these challenges and promoting sustainable farming practices. Leveraging a comprehensive crop yield dataset, this research integrates meteorological data and pesticide usage information to develop a robust predictive framework for crop yield estimation. The approach evaluates multiple machine learning techniques and uses key performance metrics such as R2-Score, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) to assess the models. The analysis identifies significant patterns between environmental factors, pesticide applications, and yield outcomes, offering actionable insights for optimizing agricultural productivity. Among the models applied, the Voting Classifier demonstrated the best performance with an error rate R2-Score of 86.3%, underscoring its reliability for predictive tasks. The results highlight the potential of machine learning to enhance agricultural decision-making, reduce dependency on harmful practices, and ensure food security in the face of evolving climate conditions.

Index Terms *- Agriculture, crop yield prediction, machine learning, Deep learning.*

1. INTRODUCTION

Agriculture is a cornerstone of the global economy, heavily influenced by meteorological conditions [1]. Seasonal agriculture, often referred to as rainfed agriculture, depends predominantly on prevailing weather patterns. Covering nearly 80% of the world's cropland, rainfed agriculture achieves favorable yields when meteorological conditions align with crop requirements [2]. However, this dependence on natural rainfall and other weather-related factors makes agricultural productivity inherently vulnerable. Variations in rainfall, whether scarcity or excess, can significantly affect farmers' ability to achieve anticipated yields [3][4]. This dynamic underlines the challenge of accurately predicting crop production within the field of precision agriculture [5].

The impacts of climate change exacerbate these challenges, threatening the agricultural sector with adverse outcomes such as food insecurity, poverty, and famine [6]. Climatic variables, particularly precipitation and temperature, play pivotal roles in determining agricultural productivity. These factors not only directly influence crops but also impact secondary elements like soil moisture and solar irradiance, further complicating yield predictions [7]. Focusing on key meteorological variables provides critical insights for improving agricultural practices and safeguarding food security amid evolving climatic conditions.

Numerous studies have demonstrated the profound effects of climate indicators, both globally and regionally, on crop yields and food security [8], [9]. For instance, Javadinejad et al. [10] identified strong correlations between reduced crop yields and two environmental factors: elevated temperatures and excessive precipitation. Extreme temperatures can adversely affect crop production through increased evapotranspiration and respiration rates, as well as heightened susceptibility to pest infestations. Similarly, excessive precipitation can lead to amplified water flow patterns, causing floods and increasing the risk of crop failure. Additionally, rising temperatures exacerbate water demand for crops, further challenging sustainable agricultural practices [11].

It is crucial to acknowledge that while climatic factors may remain consistent within specific regions, their impact varies significantly across different crops and growth stages [12]. Each crop exhibits unique levels of resilience to meteorological conditions, with extreme variations in temperature or precipitation often leading to substantial reductions in yield [13]. These complexities underscore the need for accurate prediction models that account for variations in climatic conditions and their interplay with specific crop requirements. By incorporating meteorological data and other critical variables, precision agriculture can develop effective strategies to mitigate climate-induced risks, ensuring more reliable crop yields and promoting sustainable agricultural practices.

2. RELATED WORK

The literature on crop yield prediction highlights the interplay between climate variability, agricultural practices, and the application of advanced machine learning and deep learning techniques. Burrows [6] emphasizes the necessity of understanding crop yield behavior under climate change, pointing out that extreme weather events, such as prolonged droughts and excessive rainfall, critically affect crop productivity [15]. These events amplify the complexity of managing agricultural systems, where key climatic factors such as precipitation and temperature play vital roles in shaping crop outcomes. Liu and Basso [11] underline the importance of developing adaptation strategies that consider the impacts of climate variability on both crop yields and soil organic carbon. Their findings focus on the US Midwest, showcasing the global relevance of region-specific insights into mitigating adverse climatic effects. Ahmad et al. [14] extend this perspective to South Asia, analyzing the impact of climate variability on irrigation water demand and crop yields, further demonstrating the intricate interdependence between water resources and agricultural productivity.

Jhajharia et al. [7] explore the utility of machine learning and deep learning for predicting crop yields, highlighting their effectiveness in integrating diverse datasets, including meteorological data, soil properties, and crop-specific information. They advocate for the integration of robust predictive models to enhance decision-making in precision agriculture. Paudel et al. [8] further reinforce this argument, illustrating how machine learning can forecast crop yields on a large scale by effectively analyzing complex interactions between climatic factors and crop physiology. Their study underscores the scalability of machine learning models in agricultural systems, making them an indispensable tool for tackling food security challenges.

In the context of India, Reddy and Kumar [18] and Nishant et al. [17] explore crop yield prediction using machine learning techniques, demonstrating how these methods can optimize agricultural

productivity by accurately forecasting yields based on historical and real-time data. They emphasize the critical role of meteorological variables such as temperature and precipitation, as well as region-specific conditions like soil quality and irrigation practices, in shaping the effectiveness of these predictive models. Kumar et al. [19] expand on this by proposing supervised learning approaches tailored for the Indian agricultural sector, addressing the unique challenges posed by its diverse climatic zones and cropping patterns.

Chakraborty et al. [16] focus on the usability of weather forecasts for mitigating climatic variability and its effects on maize yields in India's northeastern regions. Their study highlights the importance of integrating localized weather data with predictive algorithms to improve yield forecasts. Similarly, Javadinejad et al. [10] identify temperature and precipitation as pivotal factors affecting global agricultural yields, emphasizing their correlation with pest infestations and flood-induced crop failures. These findings align with the broader consensus that extreme climatic events, whether due to rising temperatures or intensified precipitation, substantially disrupt agricultural systems.

The collective insights from these studies underscore the significance of leveraging advanced technologies, such as machine learning and deep learning, to address the multifaceted challenges of crop yield prediction in the face of climate change. By integrating diverse datasets and developing region-specific models, these approaches offer a path toward more resilient and sustainable agricultural practices.

3. MATERIALS AND METHODS

The proposed system aims to forecast crop yields by integrating meteorological data and pesticide usage information into a comprehensive predictive framework. Utilizing a detailed crop yield dataset, the system employs machine learning algorithms, including Linear Regression, K-Nearest Neighbors (KNN), Gradient Boosting, and a Voting Regressor that combines a Bagging Regressor with Random Forest Regressor (RFR) and Decision Tree Regressor (DTR). Each algorithm is optimized using techniques like K-Fold Validation, Cross-Validation Scores, and GridSearchCV to determine the best hyperparameters, ensuring high-performance predictions. The system focuses on analyzing the relationships between environmental factors, pesticide applications, and crop yields to provide actionable insights for sustainable agriculture. By leveraging multiple algorithms and a robust evaluation approach, the framework offers a scalable and adaptable solution for improving agricultural decision-making and optimizing productivity in varying climatic conditions.

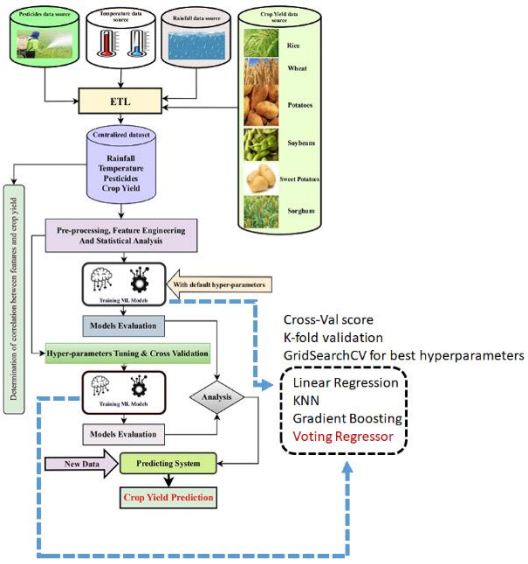


Fig.1 Proposed Architecture

This system predicts crop yields by integrating meteorological data and pesticide usage. It utilizes machine learning models like Linear Regression, KNN, Gradient Boosting, and a Voting Regressor. Hyperparameter tuning is performed using K-Fold

Validation and GridSearchCV for optimal performance. The system analyzes the relationship between environmental factors and crop yields to provide actionable insights for sustainable agriculture.

i) Dataset Collection:

The dataset used for this study comprises 28,242 entries and includes seven features: Area, Item, Year, hg/ha_yield (crop yield in hectograms per hectare), average_rain_fall_mm_per_year (average annual rainfall), pesticides_tonnes (pesticides used in tonnes), and avg_temp (average temperature). This data integrates key agricultural and environmental factors crucial for crop yield [20] prediction. The dataset offers a comprehensive view of crop production trends across different regions and years, enabling detailed analysis of meteorological and pesticide influences on agricultural productivity.

	Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes
0	Albania	Maize	1990	36613	1485.0	121.0
1	Albania	Potatoes	1990	66667	1485.0	121.0
2	Albania	Rice, paddy	1990	23333	1485.0	121.0
3	Albania	Sorghum	1990	12500	1485.0	121.0
4	Albania	Soybeans	1990	7000	1485.0	121.0

Fig.2 Dataset Collection Table

ii) Pre-Processing:

Pre-processing involves data cleaning, visualization, feature selection, and label encoding to prepare the dataset for machine learning, ensuring accuracy, relevance, and interpretability of the predictive framework.

a) Data Processing: The first step in pre-processing is data processing, which involves cleaning the dataset to ensure the quality and accuracy of the data used for modeling. The removal of duplicate data ensures there are no redundant entries that could skew the results. This step helps maintain the integrity of the dataset. Next, drop cleaning is performed, which entails removing any rows or columns containing missing or irrelevant values. This ensures that the dataset is well-structured and consistent, reducing potential noise and errors that could affect model performance.

b) Data Visualization: Data visualization is an essential step in understanding the relationships between different variables. By using Area and Item widgets, the dataset is visualized across different geographical regions and crop types. This visualization provides insights into the distribution of crop yields, rainfall, temperature, and pesticide usage for various regions and crop items. In addition, a correlation matrix is created to analyze the relationships between different numeric features like crop yield, rainfall, temperature, and pesticide usage. This helps in identifying strong correlations, aiding in the understanding of how various factors influence crop yields, which in turn supports more informed feature selection and modeling decisions.

c) Feature Selection: Feature selection is crucial for identifying the most relevant predictors for crop yield forecasting. Based on insights from the base paper and domain knowledge, key parameters such as average rainfall, pesticides usage, and average temperature are chosen as the primary features for prediction. These parameters are expected to have a significant impact on crop yields. This step eliminates any irrelevant or redundant features that do not contribute meaningfully to the prediction task, improving model efficiency and reducing overfitting.

d) Label Encoding: Since some features, such as Area and Item, are categorical in nature, label

encoding is applied. This technique converts categorical variables into numerical values that machine learning algorithms can easily process. By assigning a unique numerical label to each category, label encoding helps transform the dataset into a format suitable for algorithms like KNN, Gradient Boosting, and Voting Classifiers. This step ensures that the model can handle categorical data without introducing biases or errors during training.

iii) Training & Testing:

The training and testing phase involves preparing the dataset by splitting it into features and labels. The features (meteorological data, pesticide usage, and other factors) are separated from the target variable (crop yield). The dataset is then divided into training and testing sets, ensuring that the model is trained on one portion and validated on another to evaluate its performance. This step is essential to ensure the model generalizes well to new, unseen data and avoids overfitting.

iv) Algorithms:

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation. In our project, Linear Regression [17] will be utilized to predict crop yields based on meteorological data and pesticide usage. We will implement Cross-validation scores to assess model performance, employing K-Fold Validation to ensure robust evaluation by dividing the dataset into training and testing subsets. Additionally, GridSearchCV will be applied to optimize hyperparameters, enhancing the model's accuracy and effectiveness in forecasting crop yields.

K-Nearest Neighbors (KNN) is a non-parametric algorithm used for classification and regression tasks based on the proximity of data points. In our project, KNN [18] will be employed to predict crop yields by considering the closest data points in the feature space. Cross-validation scores will help evaluate the model's performance, while K-Fold Validation will ensure that our findings are consistent across different data splits. To fine-tune the model, we will use GridSearchCV to identify the best hyperparameters, improving the model's accuracy and adaptability to variations in data.

Gradient Boosting is an ensemble machine learning technique that builds models sequentially, with each new model attempting to correct errors made by previous ones. In our project, we will use Gradient [19] Boosting to predict crop yields based on integrated data sources. The model will leverage Cross-validation scores to measure its effectiveness, and K-Fold Validation will help in assessing performance across different subsets of data. To optimize the model's hyperparameters and enhance accuracy, we will employ GridSearchCV, ensuring that the Gradient Boosting algorithm is well-tuned for reliable yield predictions.

The **Voting Regressor** is an ensemble method that combines the predictions from multiple regression models to improve accuracy and robustness. In our project, we will implement a Voting Regressor that integrates predictions from a Random Forest Regressor (RFR) and a Decision Tree Regressor (DTR). This approach leverages the strengths of different algorithms, enhancing overall predictive performance. By using K-Fold Validation, we will evaluate the effectiveness of the ensemble method, ensuring its reliability. This technique will ultimately contribute to more accurate crop yield predictions, supporting better agricultural decision-making.

4. RESULTS & DISCUSSION

R2 Score: The sum squared regression is the sum of the residuals squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} (1)$$

MSE: Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \ (2)$$

RMSE: The root mean square error (RMSE) measures the average difference between a statistical model's predicted values and the actual values. Mathematically, it is the standard deviation of the residuals. Residuals represent the distance between the regression line and the data points.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}||y(i) - \hat{y}(i)||^2}{N}} \ (3)$$

MAE: Absolute Error is the amount of error in your measurements. It is the difference between the measured value and "true" value. For example, if a scale states 90 pounds but you know your true weight is 89 pounds, then the scale has an absolute error of 90 lbs – 89 lbs = 1 lbs.

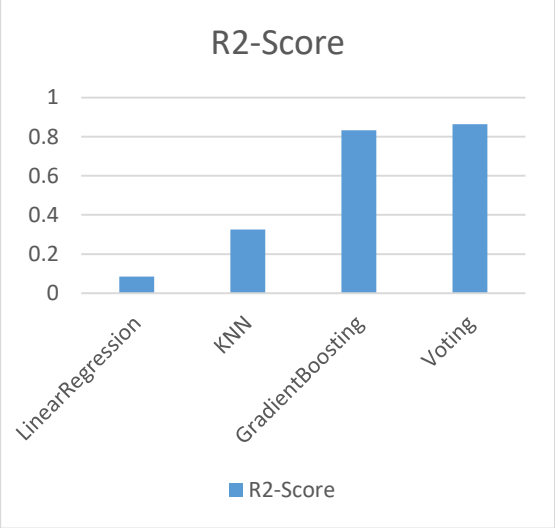
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \ (4)$$

In Table 1, the performance metrics—R2-Score, MSE, RMSE, and MAE—are evaluated for each algorithm. The Voting Classifier achieves the best scores, with all metrics. Other algorithms' metrics are also presented for comparison.

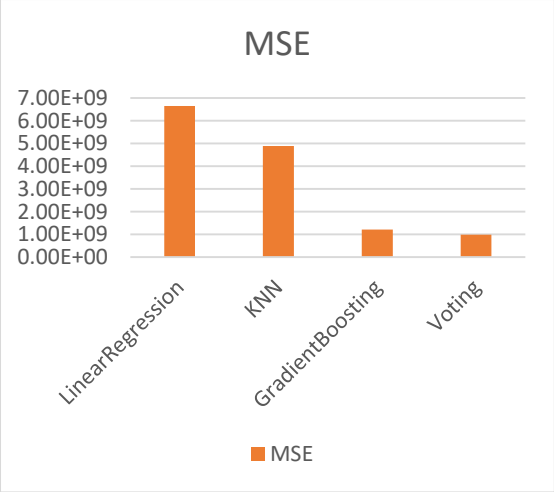
Table.1 Performance Evaluation Metrics

Model	R2-Score	MSE	RMSE	MAE
Linear Regression	0.084	6.642538e+09	81501.764	62444.311
KNN	0.326	4.886253e+09	81501.764	47819.209
Gradient Boosting	0.833	1.209219e+09	34773.823	21805.263
Voting	**0.863**	**9.956320e+08**	**31553.636**	**21438.149**

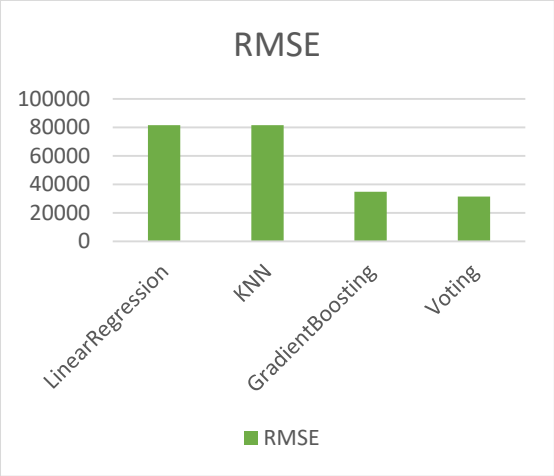
Graph.1 Comparison Graph – R2-SCORE



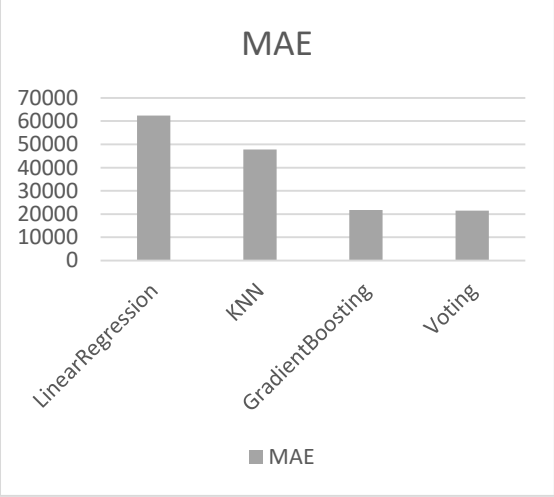
Graph.2 Comparison Graph – MSE



Graph.3 Comparison Graph – RMSE



Graph.4 Comparison Graph – MAE



In Graphs (1, 2, 3, & 4) R2-Score is represented in light blue, MSE in orange, RMSE in green, and MAE in grey. The Voting Classifier outperforms the other algorithms in all metrics, with the highest values compared to the remaining models. These details are visually represented in the above graph.

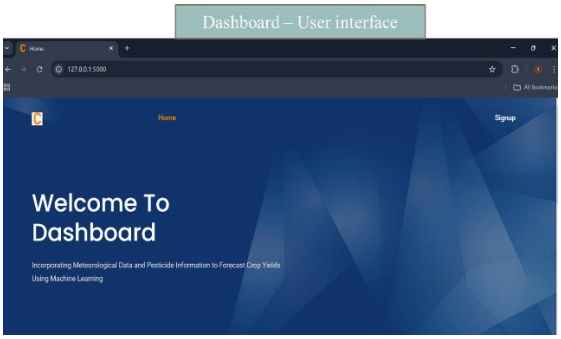


Fig.3 Home Page

In the above figure 3, this is a user interface dashboard, it is a welcome message for navigating page.

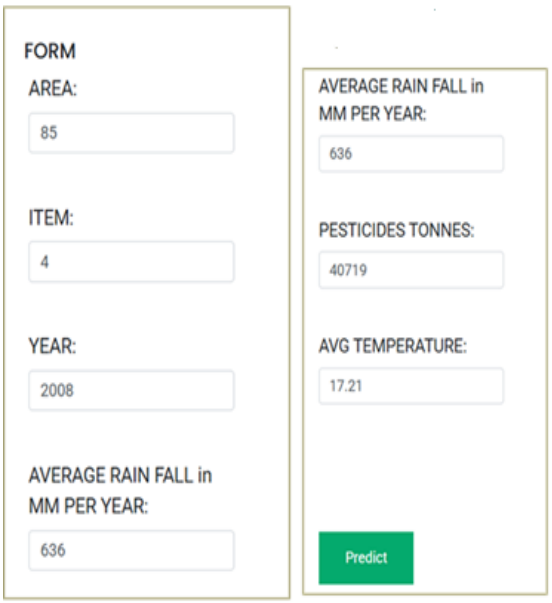


Fig.4 User input Page

In the above figure 4, this is a user input page, using this user can upload data for testing.

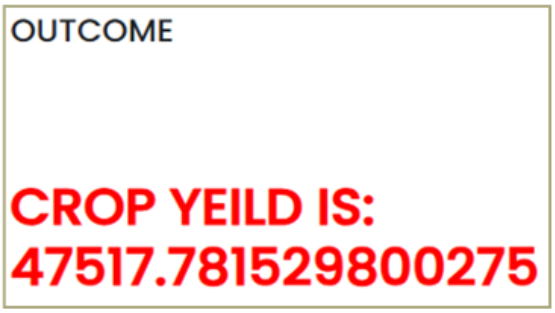


Fig.5 Classification result

In the above figure 5, this is a result screen, in this user will get output for loaded input data.

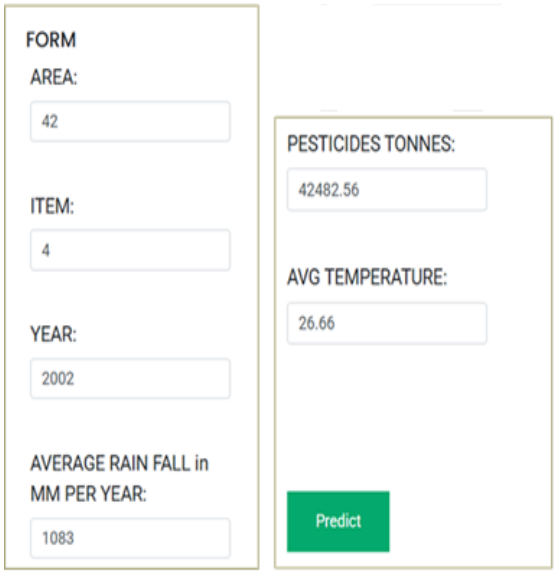


Fig.5 User input Page

In the above figure 5, this is a user input page, using this user can upload data for testing.

Fig.6 Classification result

In the above figure 6, this is a result screen, in this user will get output for loaded input data.

5. CONCLUSION

The agricultural sector faces growing challenges due to climate change and excessive pesticide use, which jeopardize global food security. Accurately predicting crop yields is crucial to mitigate these risks and promote sustainable agricultural practices. Using a comprehensive crop yield dataset, meteorological data, and pesticide usage information, a predictive framework was developed employing multiple machine learning algorithms. The analysis identified significant relationships between environmental factors, pesticide applications, and crop yield outcomes. Among the algorithms tested, the Voting Classifier demonstrated the best performance with an error rate R2-Score of 86.3% demonstrating its potential as a reliable tool for crop yield prediction. This result underscores the effectiveness of machine learning in enhancing agricultural decision-making, optimizing crop management, and ensuring food security in the face of evolving climate conditions. By leveraging such predictive models, farmers and policymakers can better anticipate challenges, reduce dependency on harmful practices, and improve overall productivity, contributing to sustainable agriculture and long-term food security.

The ***future scope*** of this project can be enhanced by incorporating advanced machine learning techniques such as deep learning models like LSTM and CNN to capture complex temporal patterns in meteorological data. Additionally, integrating satellite imagery and remote sensing data can improve the model's precision. Incorporating feature engineering techniques and domain-specific data like soil quality and crop rotation history can further enhance prediction accuracy. These

improvements will make the model more adaptable and robust in addressing evolving agricultural challenges.

REFERENCES

[1] M. Kavita and P. Mathur, ‘‘Crop yield estimation in India using machine learning,’’ in Proc. IEEE 5th Int. Conf. Comput. Commun. Autom. (ICCCA), Oct. 2020, pp. 220–224.

[2] M. A. A. Osman, J. O. Onono, L. A. Olaka, M. M. Elhag, and E. M. Abdel-Rahman, ‘‘Climate variability and change affect crops yield under rainfed conditions: A case study in Gedaref state, Sudan,’’ Agronomy, vol. 11, no. 9, p. 1680, Aug. 2021.

[3] M. N. Thimmegowda, M. H. Manjunatha, L. Huggi, H. S. Shivaramu, D. V. Soumya, L. Nagesha, and H. S. Padmashri, ‘‘Weather-based statis tical and neural network tools for forecasting rice yields in major growing districts of Karnataka,’’ Agronomy, vol. 13, no. 3, p. 704, Feb. 2023.

[4] C.Song, W.Ma, J.Li, B.Qi, and B.Liu, ‘‘Development trends in precision agriculture and its management in China based on data visualization,’’ Agronomy, vol. 12, no. 11, p. 2905, Nov. 2022.

[5] M. Chandler. (2023). How Does Climate Change Affect Agriculture? Accessed: May 12, 2023. [Online]. Available: https://impakter.com/how climate-change-affects-agriculture/

[6] L. Burrows. (Sep. 2022). A Better Understanding Of Crop Yields Under Climate Change. Accessed: May 12, 2023. [Online]. Available: https://seas.harvard.edu/news/2022/09/better-understanding-crop-yieldsunder-climate-change,

[7] K. Jhajharia, P. Mathur, S. Jain, and S. Nijhawan, ‘‘Crop yield prediction using machine learning and deep learning techniques,’’ Proc. Comput. Sci., vol. 218, pp. 406–417, Jan. 2023.

[8] D. Paudel, H. Boogaard, A. de Wit, S. Janssen, S. Osinga, C. Pylianidis, and I. N. Athanasiadis, ‘‘Machine learning for large-scale crop yield forecasting,’’ Agricult. Syst., vol. 187, Feb. 2021, Art. no. 103016.

[9] R. Affoh, H. Zheng, X. Zhang, W. Yu, and C. Qu, ‘‘Influences of meteorological factors on maize and sorghum yield in Togo, West Africa,’’ Land, vol. 12, no. 1, p. 123, Dec. 2022.

[10] S.Javadinejad, S.Eslamian, and K.O.A.Askari, ‘‘The analysis of the most important climatic parameters affecting performance of crop variability in a changing climate,’’ Int. J. Hydrol. Sci. Technol., vol. 11, no. 1, pp. 1–25, 2021.

[11] L. Liu and B. Basso, ‘‘Impacts of climate variability and adaptation strategies on crop yields and soil organic carbon in the US midwest,’’ PLoS ONE, vol. 15, no. 1, Jan. 2020, Art. no. e0225433.

[12] A.Wegrzyn, A.Klimek-Kopyra, E.Dacewicz, B.Skowera, W.Grygierzec, B.Kulig, and E.Flis-Olszewska, ‘‘Effect of selected meteorological factors on the growth rate and seed yield of winter wheat—A case study,’’ Agronomy, vol. 12, no. 12, p. 2924, Nov. 2022.

[13] J. Cao, Z. Zhang, F. Tao, L. Zhang, Y. Luo, J. Zhang, J. Han, and J. Xie, ‘‘Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches,’’ Agricult. Forest Meteorol., vol. 297, Feb. 2021, Art. no. 108275.

[14] Q.-U.-A. Ahmad, H. Biemans, E. Moors, N. Shaheen, and I. Masih, ‘‘The impacts of climate variability on crop yields and irrigation water

demand in South Asia,'' Water, vol. 13, no. 1, p. 50, Dec. 2020.

[15] V. Geethalakshmi, R. Gowtham, R. Gopinath, S. Priyanka, M. Rajavel, K. Senthilraja, M. Dhasarathan, R. Rengalakshmi, and K. Bhuvaneswari, ''Potential impacts of future climate changes on crop productivity of cereals and legumes in Tamil Nadu, India: A mid-century time slice approach,'' Adv. Meteorol., vol. 2023, pp. 1–17, Jan. 2023.

[16] D. Chakraborty, S. Saha, B. K. Sethy, H. D. Singh, N. Singh, R. Sharma, A. N. Chanu, I. Walling, P. R. Anal, S. Chowdhury, S. Hazarika, V. K. Mishra, P. K. Jha, and P. V. V. Prasad, ''Usability of the weather forecast for tackling climatic variability and its effect on maize crop yield in northeastern Hill region of India,'' Agronomy, vol. 12, no. 10, p. 2529, Oct. 2022.

[17] P. S. Nishant, P. Sai Venkat, B. L. Avinash, and B. Jabber, ''Crop yield prediction based on Indian agriculture using machine learning,'' in Proc. Int. Conf. Emerg. Technol. (INCET), Jun. 2020, pp. 1–4.

[18] D. J. Reddy and M. R. Kumar, ''Crop yield prediction using machine learning algorithm,'' in Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS), 2021, pp. 1466–1470.

[19] Y. J. N. Kumar, V. Spandana, V. Vaishnavi, K. Neha, and V. Devi, ''Supervised machine learning approach for crop yield prediction in agriculture sector,'' in Proc. 5th Int. Conf. Commun. Electron. Syst. (ICCES), 2020, pp. 736–741.

[20] R. Patel. Crop Yield Prediction Dataset. Accessed: Oct. 9, 2023. [Online]. Available: https://www.kaggle.com/datasets/patelris/crop-yieldprediction-dataset