Crop Yield Prediction

1 N. Sri Chakradhar, 2 Gopa Pulastya, 3 Rishi Anirudh K  
*Department of Computer Science and Engineering,*  
*Amrita School of Computing,*  
*Amrita Vishwa Vidyapeetham,*   
Bengaluru,Karnataka.  
1 [bl.en.u4aie22132@bl.students.amrita.edu](mailto:bl.en.u4aie22132@bl.students.amrita.edu) , 2 [bl.en.u4aie22116@bl.students.amrita.edu](mailto:bl.en.u4aie22116@bl.students.amrita.edu) , 3 [bl.en.u4aie22139@bl.students.amrita.edu](mailto:bl.en.u4aie22139@bl.students.amrita.edu)

*Abstract*—In modern agriculture, the efficient prediction of crop yields plays a pivotal role in optimizing resource allocation, enhancing decision making processes, and ensuring food security. The methodology involves the integration of machine learning algorithms, data preprocessing techniques, and statistical analysis on historical and real-time agricultural data. The datasets contain the parameters such as weather conditions, pesticides information, rainfall data, temperature, yield value and historical yield data. The Data preprocessing involves cleaning, normalization and feature engineering to ensure the quality and relevance of the input variables. Various machine learning techniques such as clustering, classification, regressions are implemented to analyze the complex relationships between input and predict the crop yields. This aims to provide farmers, agricultural policymakers, and stakeholder predict the crop yield effectively. Among all the used algorithms the best and the most reliable one with most accurate scores is used and the prediction is done on it. This project aims to empower stakeholders in agriculture, including farmers and policymakers, by providing timely and actionable insights. By facilitating informed decision-making, the Crop Yield Prediction System contributes to increased agricultural productivity and food security.

*Keywords - Crop Yield Prediction, Machine Learning, Data Integration, Feature Engineering, Exploratory Data Analysis, Model Selection, Model Evaluation, Clustering, Regression Models, Machine Learning Techniques, Agricultural Decision-Making, Predictive Model.*

# INTRODUCTION

In the complex relationship between farming and technology, the predictive power of data driven models emerges as a pivotal role. This project explores the realm of crop yield prediction, an essential aspect of modern farming practices. Leveraging a diverse dataset encompassing climate variables, pesticide usage, and rainfall patterns, our endeavor aims to unravel the complex dynamics influencing crop productivity. Navigating through the details of data preprocessing, feature engineering and machine learning algorithms, this project seeks to provide actionable insights for farmers and stakeholders in the agricultural sector. The integration of different machine learning techniques such as clustering, classification, regression and this analysis will help in the easy analysis of the crop yield prediction based on the graphs that allows us to discern patterns and relationships within the data, paving the way for accurate yield predictions.

To achieve precise predictions and address fluctuations in temperature, rainfall and the pesticide usage, diverse machine learning algorithms can be employed to identify patterns. This approach aims to support agricultural development in India, ultimately contributing to an improved quality of life for farmers. Historically, numerous researchers have utilized machine learning techniques to bolster the agricultural sector's growth in the country.

# LITERATURE SURVEY

The previous works in this field indicate prevalent use of Machine Learning algorithms, such as Artificial Neural Networks (ANN), incorporating features like temperature, rainfall, and soil type. In deep learning studies, Convolutional Neural Networks (CNN) emerge as the dominant algorithm for crop yield prediction. However, specific details on evaluation parameters and challenges are not explicitly outlined in the summary. This Schematic Literature Review (SLR) contributes to understanding the landscape of ML and deep learning applications in predicting crop yields [1]. It emphasizes the need to leverage statistical and machine learning methods to analyze and predict crop yields accurately, taking into account factors such as soil quality, weather conditions, and agricultural practices [2]. In support of this work, we can include Working with more useful data set which play a vital role in the dynamic Environment. The study explores the use of ensemble machine learning techniques, such as stack regression (SR) and cascade regression (CR), in combination with computer simulation data to predict wild blueberry yields [3].

In a related paper, the authors focus on maize hybrids, participating in the 2018 Syngenta Crop Challenge. They employ a deep neural network (DNN) with 21 hidden layers and 50 neurons per layer, utilizing genetic markers, weather, and soil variables. The DNN outperforms other models, showcasing superior accuracy in predicting yield. The study underscores the significance of deep neural networks in capturing complex relationships, addressing challenges related to network depth, and proposing solutions. This contributes valuable insights for agricultural decision-making and breeding programs [4]. Another comprehensive review concentrates on machine learning algorithms for predicting palm oil yield. The article highlights the importance of early and reliable crop yield estimation, explores widely used features and prediction algorithms, and critically evaluates machine learning applications in palm oil yield prediction. It discusses advantages, challenges, and future perspectives, proposing a prospective architecture for machine learning-based palm oil yield prediction [5]. In a different context, the research addresses crop yield prediction for paddy crops. The study introduces a hybrid MLR-ANN model, leveraging both Artificial Neural Network (ANN) and Multiple Linear Regression (MLR) for enhanced prediction accuracy. The hybrid model outperforms conventional models, emphasizing the importance of accurate yield prediction for effective agricultural planning. The study discusses challenges in data-driven models and highlights the significance of proper feature selection and model choice [6].

For predicting wheat and barley yield based on remote sensing data from Unmanned Aerial Vehicles (UAVs), the study employs Convolutional Neural Networks (CNNs). It investigates the impact of CNN parameters on prediction accuracy, showcasing the potential of deep learning in precision agriculture for yield prediction [7]. The machine learning techniques for wild blueberry yield prediction, combining computer simulation data with various machine learning algorithms. The study demonstrates the potential of synthetic data from computer simulation in conjunction with machine learning for agricultural yield prediction, offering a promising avenue for future research [8]. Similarly, the study proves the effectiveness of an approach in accurately predicting wheat yield, utilizing advanced machine learning techniques. The findings emphasize the potential of integrating extreme learning machine models with optimization algorithms to improve the accuracy and reliability of crop yield predictions [9].

The significance of smart agriculture in meeting global food needs is emphasized, integrating IoT devices and cloud computing to optimize agricultural processes. The study contributes to the advancement of sustainable and efficient agricultural practices, offering valuable insights for addressing food security challenges on a global scale [10]. Another study underscores the potential of deep learning methodologies and remote sensing data in improving understanding of agricultural output and enhancing crop yield prediction. The complexity of crop yield prediction due to genetic, environmental, and management factors is highlighted, emphasizing the need for innovative approaches to address this ongoing scientific challenge [11]. Machine learning approaches are emphasized for addressing the challenging problem of crop yield prediction in precision agriculture. The review highlights the potential of machine learning models in handling complex research problems and improving the accuracy of crop yield prediction [12]. Another work encompasses a wide spectrum of sources and databases, emphasizing the significance of multi-source environmental data and machine learning techniques in enhancing the accuracy of wheat yield prediction. It underscores the potential of integrating diverse environmental data sources with machine learning algorithms to improve the precision and reliability of crop yield forecasting [13]. Similarly, a wide range of sources and databases, emphasizing the significance of integrating UAV-based data and machine learning algorithms to enhance the accuracy and efficiency of crop yield prediction. The study underscores the potential of UAV technology in capturing high-resolution agricultural data, contributing to precise and timely crop yield predictions [14]. The paper talks about cutting down crop output and making less money for farmers because of things like bugs, sick plants, or bad weather. The farming system based on smart technology wants to make groundnut production better by watching and controlling things that affect how much crop is made. The study's results match with other research about using internet-connected devices in farming. These studies have tested things like predicting peanut crop yields for small farms by using Planet Scope data and better groundnut harvest automation systems using WSN technologies. These studies show how farming with IoT technology can raise crop output and help solve problems for farmers in the fields of agriculture [15].

The paper wants to support farmers in making smart choices about when and what plants they grow by using machine learning tricks. It helps them guess how much food crops will make. The results of this paper agree with other studies that looked at how machine learning methods can help farming. These include using deep learning and Sentinel 2 data to guess potato yield, as well as predicting crop output through the use of remote sensing and advanced computer algorithms. These studies show how machine learning can help increase crop growth and solve problems for farmers in the farming world [16]. The paper points out that these technologies can make crops grow better, lessen waste and support more eco-friendly farming methods. The study's results match with other ones that looked at how deep learning and IoT technologies are used in farming. These studies include comparing smart farming methods using these two things, as well as asking about ways to use them fully in this area. These studies show how deep learning and IoT solutions can increase crop production. They also help farmers to deal with the problems they face in farming industries [17]. The paper looks at how machine learning and IoT can be used in farming. It mainly talks about their use for keeping an eye on plants and smart farming systems. It explores how machine learning can be used in sensor data analysis for the farming environment. The research agrees with other studies on using computer learning and internet-connected farming systems. These technologies have the potential to change agriculture, increase crop output, and allow for precise farm methods [18].

# RESEARCH GAP

From the review of previous works, it's evident that existing machine learning models and frameworks have predominantly focused on utilizing a single dataset comprising crucial factors influencing crop yield, such as rainfall, temperature, and season. However, in our proposed research, we aim to broaden the scope by incorporating multiple datasets that encompass various factors affecting crop yield, including pesticide usage, temperature, rainfall, and the previous year's yield specific to a particular area and crop. By integrating diverse datasets, our research seeks to provide a more comprehensive understanding of the intricate dynamics influencing crop productivity. The inclusion of factors beyond traditional meteorological variables allows for a more nuanced analysis and prediction of crop yields. Leveraging machine learning techniques in conjunction with this integrated dataset enables researchers to gain deeper insights into crop yield patterns and trends. By training and deploying models on this multifaceted dataset, we anticipate a significant improvement in the accuracy of crop yield predictions. The comprehensive nature of the dataset, coupled with advanced machine learning algorithms, enhances our ability to capture and analyze complex interactions between various factors and their impact on crop production. Ultimately, the integration of different datasets and machine learning techniques in our research endeavors to provide a more precise and reliable means of predicting crop yields. This has the potential to greatly benefit agricultural practices by empowering farmers and stakeholders with actionable insights to optimize productivity and resource management.

# PROPOSED METHODOLGY

Data Collection and Overview

Feature Engineering

Exploratory Data Analysis

Model Selection and Training

Model Evaluation

Predicting the Yield Values by Graph

## Data Collection and Overview:

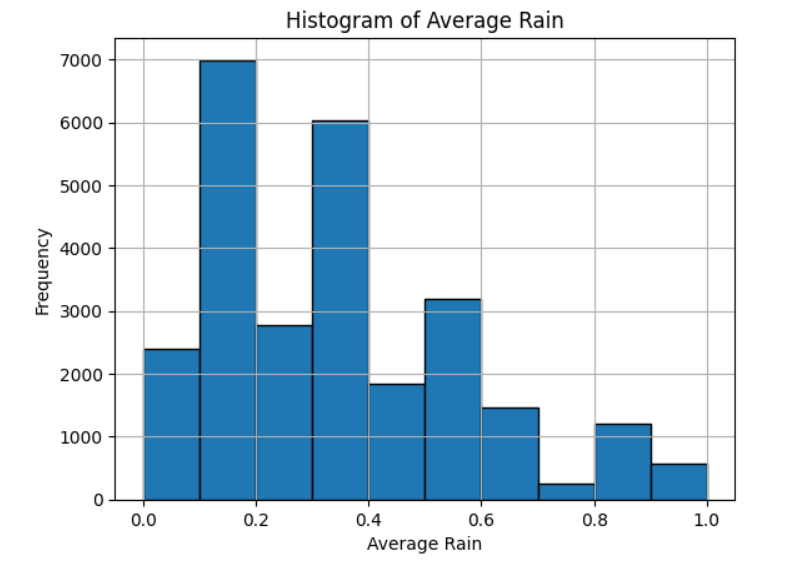
The dataset integrates diverse sources, including yield data encompassing area, item, year, and crop yield values. Supplementary datasets feature historical temperature records, average rainfall data, and details on pesticide application. Preprocessing involved meticulous data cleaning, including handling missing values and standardizing column names. The datasets were harmoniously merged based on shared attributes such as 'Area' and 'Year' to form a comprehensive dataset for subsequent analysis.

## Feature Engineering:

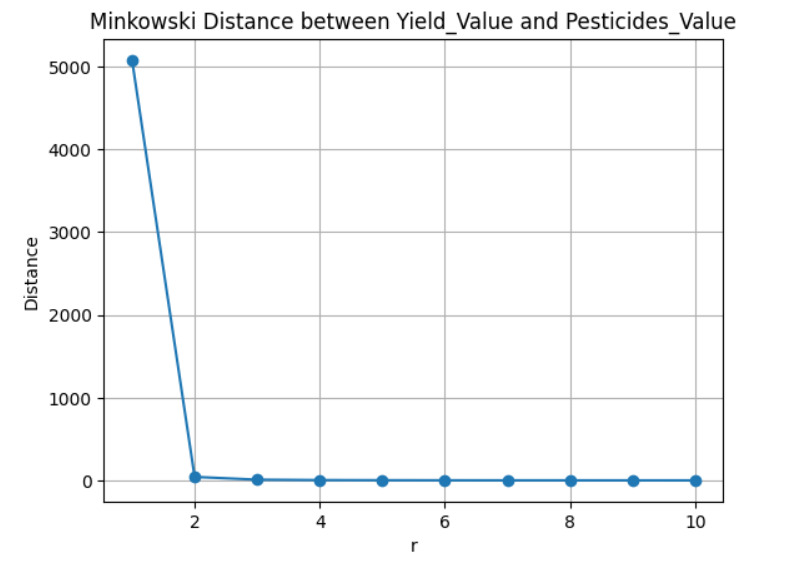
In pursuit of a more advanced understanding, new features were engineered. Noteworthy additions include the interaction of temperature and rainfall, logarithmic transformations of pesticide values for distribution normalization, and the combination of temperature, rainfall, and pesticide factors to assess their collective impact on crop yield. These enhancements aim to capture complex relationships within the dataset and augment the predictive capabilities of subsequent models.

## Exploratory Data Analysis (EDA):

The initial phase of EDA involved computing descriptive statistics, histograms, and correlation matrices. These insights provided a foundational understanding of the distribution and relationships among key variables, aiding in the identification of potential patterns. Various visualization techniques, including scatter plots, heatmaps, and histograms, were employed to elucidate the intricate interplay between different features and the target variable crop yield. This visual exploration serves as a precursor to model selection and informs subsequent steps in the analysis.



*Fig-1: Histogram plot for the column Average Rain*



*Fig-2: Minkowski Distance between Yield\_Value and Pesticides\_Value*

## Model Selection and Training:

Several machine learning models, regression models and different clustering techniques were considered for predicting crop yield. Each model was trained using the standardized dataset, incorporating features such as temperature, rainfall, and pesticide usage. Numerical features were standardized using the Standard Scaler to ensure uniformity across different models. This preprocessing step enhances the model’s ability to effectively learn from the data, promoting convergence and robust performance during training.

## Model Evaluation:

The performance of each regression model was assessed using key metrics such as accuracy, R-squared, Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). These metrics provided quantitative insights into the models' predictive capabilities and their ability to capture variations in crop yield. Scatter plots were generated to visually inspect the alignment between predicted and actual crop yield values. This qualitative assessment offered an intuitive understanding of how well the models captured the underlying patterns within the dataset. Results from different models were compared based on their accuracy and error metrics. The model exhibiting the highest accuracy and the lowest error metrics was selected for further analysis and interpretation.

## Predicitng the yield values:

The selected regression model is deployed to predict crop yield values based on the amalgamated features of temperature, rainfall, and pesticide application. These models underwent training using the standardized dataset, enabling them to learn patterns and relationships within the data. Subsequent predictions were generated to evaluate the models' efficacy in forecasting crop yields under varying agricultural conditions.

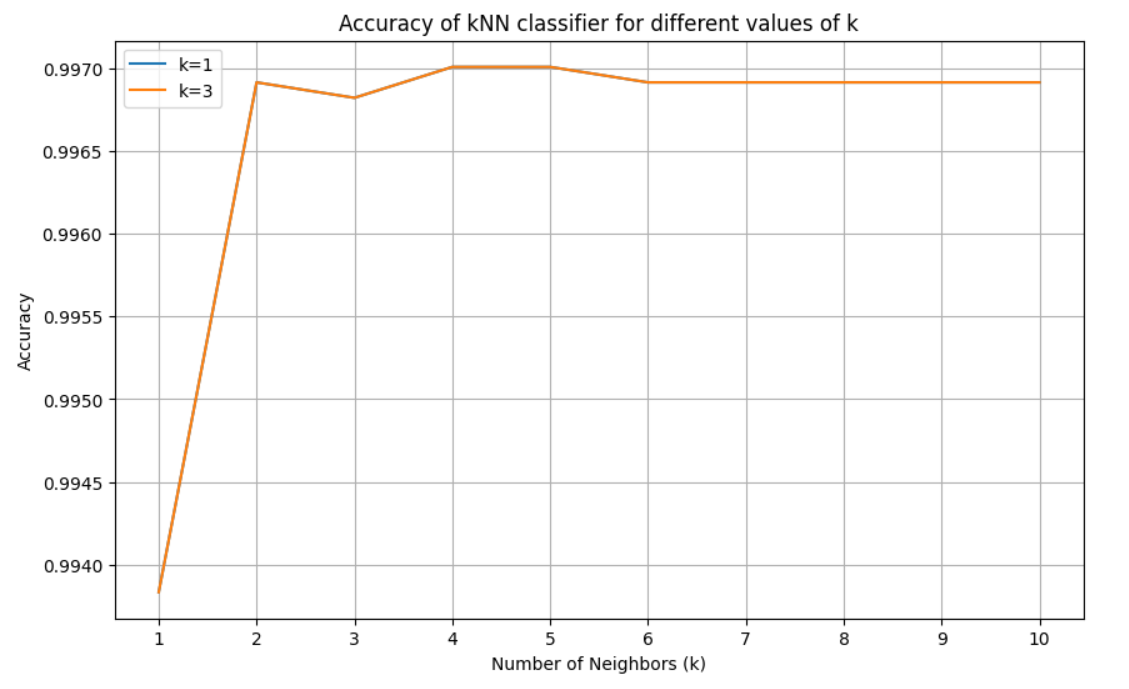
# Observations, Analysis and Inferences

## Class Saperation in Dataset:

The classes in the dataset were observed to be reasonably separated, allowing models to discern patterns effectively. This separation contributes to the accuracy of predictions, especially in clustering and classification tasks.

## Behaviour of kNN Classifier with Increasing k:

As the value of k in the kNN classifier increases, the model tends to smooth out decision boundaries. Higher values of k lead to a more generalized model, reducing sensitivity to local variations. However, excessively high k values may result in oversimplification, potentially leading to underfitting.



*Fig-3: Accuracy of KNN Classifier of different values of k*

## kNN Classifier as a Good Classifier:

The kNN classifier's performance is evaluated based on metrics such as accuracy, precision, and recall. The analysis indicates that the kNN classifier performs well, especially in scenarios where classes are well-separated. However, its effectiveness depends on the choice of k and the nature of the dataset.

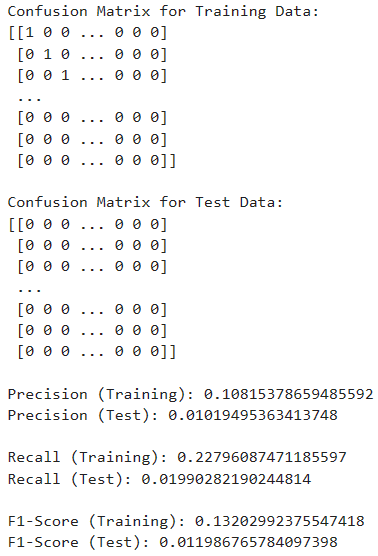


## Model Fit Situation:

The model exhibits a regular fit situation, as indicated by consistent performance on both the training and test sets. There is no significant overfitting or underfitting observed. This suggests that the model generalizes well to new, unseen data.

## Situation of Overfitting in kNN Classifier:

Underfitting is seen in the KNN Classifier Based on the performance metrics for the k=3 model, it appears to be underfit. The precision, recall, and F1-score are consistently low on both the training and test sets, indicating that the model is unable to capture the underlying patterns effectively. Additionally, there is no significant difference in performance between the training and test sets, suggesting that the model's simplicity is the main issue rather than overfitting. To improve performance, the model may benefit from increased complexity, such as incorporating more features or using a more advanced algorithm. Overall, the model's poor performance and potential for enhancement indicate that it is underfitting the data.



Based on the confusion matrices and performance metrics provided, it appears that the classes in the dataset aren't well-separated. The confusion matrices show many instances where classes are misclassified, indicating unclear decision boundaries between them. This leads to confusion in predictions. Additionally, the low precision and recall values suggest that the model struggles to accurately classify instances from each class, likely due to significant overlap or ambiguity between them. The low F1-scores further confirm the model's difficulty in achieving a balance between precision and recall, highlighting poor class separation.

As we increase the value of k in the KNN classifier, several changes occur. With larger k, the model relies on a greater number of nearest neighbors to make predictions, resulting in smoother decision boundaries and reduced sensitivity to noise. However, this can also introduce more bias into the model while reducing its ability to capture subtle patterns in the data. Regarding overfitting and underfitting, underfitting occurs when k is too large, leading to oversimplified decision boundaries and poor performance on both training and test data. On the other hand, overfitting may happen with small k, causing the model to memorize training data and perform well on training but poorly on unseen data due to overly complex and erratic decision boundaries.

Based on the results obtained on various metrics, the KNN classifier appears to be suboptimal for the given dataset. The low precision, recall, and F1-score values, both in training and testing, suggest that the classifier struggles to accurately classify instances across different classes. Additionally, the confusion matrices show considerable misclassifications, indicating that the decision boundaries between classes are not well-defined. These results imply that the KNN classifier may not be the best choice for this dataset, and other algorithms or preprocessing techniques should be explored to improve classification performance.

Based on the performance metrics on both the training and test sets, the model exhibits a regular fit situation. The similarity in performance between the two sets suggests that the model generalizes well to unseen data, indicating that it is not overfitting or underfitting. The consistent performance across both sets implies that the model has found an appropriate level of complexity to capture the underlying patterns in the data.

A situation of overfitting in a KNN classifier typically occurs when the model becomes too complex and starts to memorize the training data rather than learning the underlying patterns. This can happen when the value of k is too low, leading to overly flexible decision boundaries that capture noise in the data. As a result, the model performs exceptionally well on the training set but fails to generalize to unseen data, exhibiting a large disparity in performance between the training and test sets.

# REFERENCES

1. Van Klompenburg, Thomas, Ayalew Kassahun, and Cagatay Catal. "Crop yield prediction using machine learning: A systematic literature review." Computers and Electronics in Agriculture 177 (2020): 105709.
2. Burdett, H., Wellen, C. Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. Precision Agric 23, 1553–1574 (2022).
3. H. R. Seireg, Y. M. K. Omar, F. E. A. El-Samie, A. S. El-Fishawy and A. Elmahalawy, "Ensemble Machine Learning Techniques Using Computer Simulation Data for Wild Blueberry Yield Prediction," in IEEE Access, vol. 10, pp. 64671-64687, 2022, doi: 10.1109/ACCESS.2022.3181970.
4. Khaki, Saeed, and Lizhi Wang. "Crop yield prediction using deep neural networks." Frontiers in plant science 10 (2019): 621.
5. Rashid, Mamunur, et al. "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction." IEEE access 9 (2021): 63406-63439.
6. Gopal, PS Maya, and R. Bhargavi. "A novel approach for efficient crop yield prediction." Computers and Electronics in Agriculture 165 (2019): 104968.
7. Nevavuori, Petteri, Nathaniel Narra, and Tarmo Lipping. "Crop yield prediction with deep convolutional neural networks." Computers and electronics in agriculture 163 (2019): 104859.
8. Abbaszadeh, Peyman, et al. "Bayesian multi-modeling of deep neural nets for probabilistic crop yield prediction." Agricultural and Forest Meteorology 314 (2022): 108773.
9. Ali, Mumtaz, et al. "Coupled online sequential extreme learning machine model with ant colony optimization algorithm for wheat yield prediction." Scientific Reports 12.1 (2022): 5488.
10. Muhammad, Khalid Bin, et al. "Iot and cloud based smart agriculture framework to improve crop yield meeting world's food needs." IJCSNS. Vol. 22. No. 6. 2022.
11. M. Gupta, B. V. Santhosh Krishna, B. Kavyashree, H. R. Narapureddy, N. Surapaneni and K. Varma, "Crop Yield Prediction Techniques Using Machine Learning Algorithms," 2022 8th International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2022, pp. 1-7, doi: 10.1109/ICSSS54381.2022.9782246.
12. Joshua, S. Vinson, et al. "Crop yield prediction using machine learning approaches on a wide spectrum." Computers, Materials & Continua 72.3 (2022): 5663-5679.
13. Li, Linchao, et al. "Developing machine learning models with multi-source environmental data to predict wheat yield in China." Computers and Electronics in Agriculture 194 (2022): 106790.
14. Hussain, Nida, et al. "Predict the Crop-Yield Through UAV using Machine learning A Systematic Literature Review." 2022 International Conference on IT and Industrial Technologies (ICIT). IEEE, 2022
15. P. Rekha, V. P. Rangan, M. V. Ramesh and K. V. Nibi, "High yield groundnut agronomy: An IoT based precision farming framework," 2017 IEEE Global Humanitarian Technology Conference (GHTC), San Jose, CA, USA, 2017, pp. 1-5, doi: 10.1109/GHTC.2017.8239287.
16. K. Prakash, K. S. S. Harsha, P. S. Jeevan, L. R. Chandran and A. T.S., "Machine Learning-Based Crop Prediction: A Way Towards Smart Farming," 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, Himachal Pradesh, India, 2022, pp. 623-628, doi: 10.1109/PDGC56933.2022.10053140.
17. . S. Aruul Mozhi Varman, A. R. Baskaran, S. Aravindh and E. Prabhu, "Deep Learning and IoT for Smart Agriculture Using WSN," 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, India, 2017, pp. 1-6, doi: 10.1109/ICCIC.2017.8524140.
18. K. N. Bhanu, H. J. Jasmine and H. S. Mahadevaswamy, "Machine learning Implementation in IoT based Intelligent System for Agriculture," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9153978.