Integrating Decision Tree and KNN Hybrid Algorithm approach for Enhancing Agricultural Yield Prediction

Royal Praveen Dsouza, Dr. G N K Suresh Babu,

Research Scholar, Department of Computer Science, Srishti College of Commerce and Management, University of Mysore ORCID: 0009-0000-6848-2630

Professor, Department of Computer Science, Srishti College of Commerce and Management, University of Mysore ORCID: 0000-0002-8467-3119

Abstract:

Accurate prediction of agricultural yield plays a pivotal role in ensuring sustainable resource allocation and global food security. Traditional methods often struggle to capture the intricate relationships between diverse agricultural variables, necessitating innovative approaches for enhanced prediction accuracy. This paper presents the Decision Tree-KNN Hybrid Algorithm (DT-KNN), a novel method that integrates decision trees and K-nearest neighbors (KNN) to adopt these challenges effectively. Decision trees are recognized for their aptitude to model complex interactions and interpretability, making them suitable for capturing nonlinear patterns in agricultural data. On the other hand, KNN excels in local pattern recognition by utilizing similarities between data points. By combining these two methodologies, DT-KNN leverages the strengths of both to enhance predictive precision and robustness. The methodology begins with comprehensive data preprocessing, incorporating cleanup, standardization, and attribute production. This stage guarantees that the input data is standardized and optimized for subsequent modeling. The decision tree component of DT-KNN constructs a hierarchical structure that partitions the information into splits based on characteristic estimates, thereby identifying distinct patterns in the agricultural data. Each leaf node of the decision tree represents a subset of data points with similar characteristics. Subsequently, KNN is applied within each identified leaf node to make localized predictions. This dual-layered approach allows DT-KNN to capture both global trends and local variations within the Indian Chamber of Food and Agriculture (ICFA) agricultural dataset, thereby improving the overall predictive accuracy. To validate the effectiveness of DT-KNN, extensive experiments are conducted using ICFA datasets. The performance of DT-KNN is evaluated against traditional methods and other hybrid algorithms through rigorous comparative analysis. System of measurement such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2) are engaged to assess predictive accuracy and robustness across diverse algorithms. The results demonstrate that DT-KNN outperforms traditional methods in terms of accuracy and reliability. It effectively balances between capturing complex agricultural dynamics and maintaining interpretability, making it a promising approach for agricultural yield prediction. This research aids to the improvement of predictive modeling in farming and lays the groundwork for future enhancements and applications of hybrid algorithms in agricultural research and practice.

Keywords: Agricultural yield prediction, Decision Tree-KNN Hybrid Algorithm (DT-KNN), Predictive accuracy, nonlinear patterns, Hybrid algorithms

DOI: 10.24297/j.cims.2024.07.2

1. Introduction

Food production has always been a foundation of individual culture, serving as the primary source of food, raw materials, and livelihood for a substantial fragment of the worldwide populace. With the continuous growth in universal populace and the successive enhance in food requirement, the importance of efficient and effective agricultural practices has become more pronounced. Conventional techniques of predicting produce harvest, which rely seriously on empirical knowledge and historical data, are often insufficient in the face of rapidly changing climatic conditions and other environmental variables. Consequently, there is a pressing need for advanced computational analysis to predict agricultural crop yields more accurately and reliably. The beginning of data mining, machine learning, and artificial intelligence (AI) [A. Raj V et al 2022] has revolutionized many industries, and agriculture is no exception. These technologies offer intense instruments for inspecting huge sizes of data, recognizing samples, and producing estimates that were previously impossible or impractical. Data mining involves extracting valuable information from vast datasets, while machine learning [Adithya Pothan Raj V et al 2019] and AI use this information to build predictive models that can learn and improve over time. In the context of agriculture, these technologies can analyze data from various sources, such as climate fitness, mud properties, produce attributes, and agriculture methods, to foretell harvest returns with greater accuracy. This ability to harness and analyze complex datasets is crucial for making informed decisions that can improve production, enhance source use, and guarantee foodstuff protection. One of the primary applications of computational analysis [A. Raj V et al 2022] in agriculture is to deal with the disputes put by weather alteration. As weather patterns become more erratic and risky weather results develop more regular, traditional methods of crop yield prediction become increasingly unreliable. Machine learning algorithms can process real-time weather data and historical climate information to predict the influence of these changes on harvest produce. By participating this predictive capability into agricultural planning [Geetha et al 2022], farmers can make proactive adjustments to their practices, such as selecting more resilient crop varieties, optimizing irrigation schedules, and adjusting planting dates, to alleviate the hostile things of climate variability.

The application of machine learning and AI in agriculture [Iniyan et al 2023] is not limited to yield prediction alone. These technologies can also play a pivotal role in annoyance and infection discovery, soil health monitoring, and exactitude planting. For instance, machine learning models can examine satellite imagery to detect initial trails of annoyance invasions or nutrient faults in produces, allowing for timely intervention and tumbling the belief on natural pesticides. Similarly, AI-driven mud devices can grant simultaneous facts on mud humidity and nutrient readings, allowing agriculturalists to use manures and water more efficiently [Reyana et al 2023]. Exactitude planting, which implies the use of GPS and IoT machines to oversee produces at a micro level, can be significantly enhanced by integrating machine learning algorithms that optimize resource allocation and reduce waste. The need for research in this area is underscored by the potential benefits that advanced computational techniques can bring to agriculture. By developing the accuracy of produce return projections, farmers can make well-versed findings that lead to higher productivity and profitability [Manjunath et al 2023]. Additionally, more accurate yield predictions can help policymakers and agricultural organizations plan for food distribution and manage supply chains more effectively, reducing the risk of food shortages and price volatility. Moreover, the integration of AI and machine learning into agricultural practices [Koresh et al 2021] can promote sustainable farming [Fayaz et al 2023] by minimizing the environmental impact of agriculture and conserving natural resources.

The integration of data mining, machine learning, and AI into agricultural practices holds significant promise for improving crop yield predictions [Batool et al 2022] and enhancing overall agricultural productivity. The ability to analyze and interpret complicated datasets can specify valued intuitions that advise administrative-production and promote sustainable farming practices [Bali et al 2022]. However, to completely realize the promise of these expertise, ongoing research and development are essential. Researchers can develop more accurate and reliable predictive models that can benefit farmers, policymakers, and the global food system as a whole. This research is not only crucial for meeting the growing food demand [Joshua et al 2022] but also for confirming the sustainability and resistance of crop growing in the challenge of climatic zone shift and extra ecological encounters.

2. Review Of Related Works

In the realm of crop harvest forecast, the study by Andrew Crane Droesch (2018) introduces a groundbreaking approach through a semiparametric deep neural network model. This model addresses the inherent complexities and nonlinearities present in highly-dimension agricultural datasets by integrating accepted parametric configurations and unnoticed cross-sectional divergency. In practical applications, this model surpasses established arithmetic techniques and completely unparametric neural networks, particularly in forecasting corn harvests in the US Midwest. Its robustness is highlighted by its ability to provide more accurate yield predictions for years that were withheld during the training phase. Moreover, the model's application across various climate scenarios indicates a less severe impact of weather shift on corn produce compared to traditional methods, especially in the hottest zones and situations. This finding underscores the importance of advanced predictive models in accounting for regional variations and specific climatic conditions, offering more optimistic outcomes in the face of climate change challenges. Droesch's work emphasizes the necessity of merging advanced machine learning techniques with domain-specific knowledge to enhance prediction accuracy. The semiparametric neural network model stands out by combining complex data structures with established parametric insights, thereby creating a powerful tool for agricultural forecasting. This methodological advancement not only refines the exactitude of produce estimates but also facilitates improved-informed decision-making in agricultural policy and planning. By proving the model's effectiveness under various climatic scenarios, the study offers valuable insights for developing resilient agricultural practices that can adapt to evolving environmental conditions. This integrative approach is pivotal in forming a comprehensive understanding of crop yield dynamics and preparing for the impacts of climate change on agriculture.

Thomas van Klompenburg (2020) provides a methodical works evaluation to consolidate the current state of machine learning applications in produce harvest forecast. Analyzing 50 selected studies from a pool of 567, the review identifies hotness, rain, and earth type as the best commonly used attributes in extrapolative shows. The prevalence of Artificial Neural Networks (ANNs) in these patterns underscores their efficacy in capturing complex patterns within agricultural data. The review further reveals that advanced deep learning techniques such as Convolutional Neural Networks (CNNs), Long-Short Term Memory (LSTM), and Deep Neural Networks (DNNs) are increasingly utilized, marking a shift towards more sophisticated predictive methods in recent research. Klompenburg’s review offers a critical analysis of the varied features and algorithms used across different studies, highlighting the necessity for tailored models that align with specific datasets and research goals. The importance of testing models with diverse feature sets to determine the optimal configuration for accurate yield prediction is also emphasized. Despite the broad concentration of machine learning procedures, the review notes the absence of a consistently superior model, suggesting a need for ongoing experimentation and adaptation in model selection. This dynamic nature of machine learning applications in agriculture is vital for enhancing predictive accuracy and reliability. The insights from this review are crucial for steering future research towards the development of more successful produce return likelihood representations, ultimately contributing to improved agricultural productivity and sustainability. Kavita Jhajharia et al. (2023) focus on produce return estimation in Rajasthan, India, utilizing various machine learning techniques. The study evaluates five different crops with algorithms such as Random Forest, SVM, Gradient Descent, LSTM, and Lasso regression. Among these, the Random Forest algorithm emerges as the most effective, achieving the highest R² and the lowest RMSE and MAE values. This research highlights the critical role of machine learning in tackling agricultural challenges posed by climate change and population growth. The findings stress the necessity of modern irrigation techniques and advanced predictive models to boost crop yield and ensure food security.

Table 1. Review of related works on crop yield prediction

S.No	Author and Year	Study Focus	Methods/Algorithms	Key Findings	Key Features/Variables
1	Andrew Crane Droesch [2018]	Semiparametric deep neural network for	Semiparametric DNN, Parametric, Nonparametric DNN	Semiparametric DNN outperforms classical	Temperature, Rainfall, Soil Type

		corn yield prediction		methods and fully-nonparametric neural networks	
2	Thomas van Klompenburg [2020]	Systematic Literature Review of crop yield prediction	Artificial Neural Networks, CNN, LSTM, DNN	CNN, LSTM, and DNN are the most preferred deep learning algorithms. No single best model identified	Temperature, Rainfall, Soil Type
3	Kavita Jhajharia et al [2023]	ML techniques for crop yield estimation in Rajasthan, India	Random Forest, SVM, Gradient Descent, LSTM, Lasso	Random Forest performed best with R^2=0.963, RMSE=0.035, MAE=0.0251	Market Price, Production Rate, Soil Type, Rainfall
4	Alejandro Morales et al [2023]	Effect of data partitioning on model performance	Random Forest, ANN, Regularized Linear Models	Random Forest had the best performance (RMSE 35-38%), data partitioning affects model accuracy	Soil Depth, Management, Seasonal Weather
5	Burdett H et al [2023]	Relationship between soil properties, topographic characteristics, and crop yield	MLR, ANN, Decision Trees, Random Forest	Random Forests achieved R^2=0.85 for corn, 0.94 for soybeans	pH, Soil Organic Matter, CEC, Phosphorus, Zinc, Potassium, Elevation, Topographic Wetness Index
6	Vishal Nathgosavi et al [2021]	ML models for crop yield management	ANN, SVM, RF, Cubist	ANN and SVM frequently used, focus on improving prediction with additional features	Rain, Soil Type, Precipitation, Humidity

7	Sonal Agarwal et al [2021]	Enhanced crop yield prediction model	Random Forest, Decision Tree, ANN, LSTM, RNN, SVM	DL models (LSTM, RNN, SVM) achieved higher accuracy (97%) compared to traditional ML models (93%)	Soil Ingredients (N, P, K), Crop Rotation, Soil Moisture, Temperature, Precipitation
8	Kodimalar Palanivel et al [2019]	Investigating ML algorithms for crop yield prediction	ANN, SVM, Linear Regression, Logistic Regression, Decision Trees, Naïve Bayes	ANN and SVM models are more suitable for crop yield prediction	Rainfall, Temperature, Humidity, Soil Moisture, Soil pH, Salts (N, P, K, Organic Carbon, etc.)

Alejandro Morales et al. (2023) examine the impact of data partitioning strategies on the performing of machine learning exhibits in harvest produce forecast. Utilizing synthetic datasets for sunflower and wheat, the study compares procedures such as Random Forest, Artificial Neural Networks, and regularized linear models. The Random Forest algorithm demonstrates superior performance, although the study notes that its advantage over simpler baseline models is limited. This research underscores the importance of proper data partitioning and comprehensive model validation to ensure the accuracy and reliability of predictive models in practical agricultural applications. By focusing on these aspects, the study contributes to a more nuanced understanding of how to optimize machine learning models for real-world use in agriculture, ensuring their effectiveness in predicting crop yields under various conditions. These studies in Table 1 collectively advance the field of produce harvest prediction through the use of sophisticated machine learning techniques. They highlight the importance of integrating advanced models with domain-specific knowledge, emphasize the need for tailored and adaptable models, and stress the critical role of proper data handling and validation. These insights are essential for developing resilient and accurate predictive models that can adapt to the evolving challenges of climate change and ensure sustainable agricultural productivity.

3. Research Gap And Objectives

The literature survey highlights several advancements and gaps in the purpose of machine learning practices for harvest produce forecast. While Andrew Crane Droesch's (2018) semiparametric deep neural network model successfully addresses the complexities of high-dimensional agricultural data and improves prediction accuracy by incorporating both parametric structures and cross-sectional heterogeneity, it primarily focuses on corn yields in the US Midwest. Similarly, Thomas van Klompenburg's (2020) systematic review identifies temperature, rainfall, and soil type as key features in crop yield prediction models, with a significant reliance on Artificial Neural Networks (ANNs) and deep learning techniques. However, the assess also notes that no single model consistently outperforms others, indicating a need for continuous experimentation. Kavita Jhajharia et al. (2023) and Alejandro Morales et al. (2023) further explore various machine learning algorithms and data partitioning strategies, with a particular emphasis on Random Forest and the importance of model validation. Despite these advancements, a gap remains in the exploration and integration of cross algorithms that blend the potencies of different machine learning procedures to augment projecting exactness and reliability across diverse crops and regions.

The proposed method aims to address this research gap by integrating a Decision Tree and K-Nearest Neighbors (KNN) Hybrid Algorithm approach for enhancing agricultural yield prediction. This hybrid method leverages the decision tree's ability to handle complex, nonlinear relationships and the KNN's proficiency in capturing local patterns within the data. By merging these two procedures, the proposed means seeks to improve the overall precision and robustness of yield predictions across various crops and climatic conditions. The purposes of this proposed method are to develop a more reliable predictive model that can adapt to regional variations and specific climatic conditions, provide actionable insights for agricultural planning and policy-making, and ultimately contribute to increased agricultural productivity and sustainability. Through rigorous testing and validation against prevailing replicas, the fusion method intentions to demonstrate superior performance in predicting crop yields, thereby filling a critical gap in the current literature and advancing the field of agricultural yield prediction.

4. Decision Tree (Dt)

A Decision Tree (DT) is an effective machine learning algorithm that can be used for various predictive tasks in agriculture, such as predicting crop yields, disease outbreaks, or soil quality. For the Indian Chamber of Food and Agriculture (ICFA) agricultural dataset, a DT can be particularly useful in modeling and understanding complex relationships between different agricultural variables. The Decision Tree algorithm [Shahhosseini et al 2021] constructs a tree-like model of decisions, where internal nodes represent tests on agricultural features (soil pH, rainfall, temperature), divisions signify the consequences of these trials, and leaf nodes characterize the last forecasts (yield, crop type). The tree is built through a recursive process that splitting the dataset into splits founded on the estimates of key reports. The algorithm aims to partition the data such that each subset becomes more homogeneous concerning the target variable, whether it is crop yield, quality, or another agricultural outcome. This is achieved by picking the best features to splitting the information at each node, using criteria that measure the reduction in impurity or uncertainty. For a node t in the decision tree, the Gini Impurity $G(t)$ is calculated as in equation 1

$$G(t) = 1 - \sum_{i=1}^{C} p(i \mid t)^2 \qquad (1)$$

where $p(i \mid t)$ is the proportionality of instances of class i at node t, and C is the numeral of different classes (different crop types). For a regression task, Gini Impurity is often replaced with variance reduction. Entropy $H(t)$ at a node t is defined as in equaion 2

$$H(t) = -\sum_{i=1}^{C} p(i \mid t) \log_2 p(i \mid t) \quad (2)$$

Information Gain $IG(t, X)$ for a split at node t using feature X is given by equation 3

$$IG(t, X) = H(t) - \sum_{v \in values(X)} \frac{|t_v|}{|t|} H(t_v) \quad (3)$$

where t_v is the subset of instances where feature X has value v, $\mid t \mid$ and $\mid t_v \mid$ are the number of instances in node t and subset t_v, respectively. The feature that maximizes the Information Gain is chosen for splitting. The process of recursively partitioning the dataset is as follows:

1) **Select the Best Split:** At apiece node, compute the Gini Impurity or Information Gain for each feature in the ICFA dataset (soil pH, rainfall, temperature) and choose the attribute that findings in the excellent splitting.
2) **Split the Node:** Split the dataset into splits built on the chosen attribute's pulls.
3) **Repeat:** Apply the identical procedure recursively to every subclass.
4) **Stop:** The recursion stops when a node meets the halting reasons, such as touching a ceiling intensity, requiring a smallest integer of tries, or achieving pure nodes (all instances in a node have the same outcome).

Consider that S represents the dataset at node t, the decision tree function $DT(S)$ can be described recursively as in equation 4

$$DT(S) = \begin{cases} Leaf(S) & if\ stopping\ criteria\ are\ met \\ & Node(S)\ \ otherwise \end{cases} \quad (4)$$

where $Leaf(S)$ represents the prediction at a leaf node, and $Node(S)$ indicates further splitting of the dataset.

For the ICFA agricultural dataset, the Decision Tree algorithm can predict various outcomes using features such as soil pH, rainfall, temperature, fertilizer usage, and pest incidence. Soil pH indicates soil acidity or alkalinity, affecting crop health; rainfall determines the water supply crucial for crops; temperature influences growth cycles and yield; fertilizer usage impacts soil fertility based on the quantity and type used; and pest incidence affects crop health and yield. To predict crop yield using these features, the Decision Tree algorithm follows these steps: compute Gini Impurity or Entropy for each feature at the root node to measure impurity, select the feature (rainfall) with the highest Information Gain, partition the dataset based on the selected feature's values (different ranges of rainfall), recursively apply the same process to each subset created in the previous step, and halt the recursion when nodes become pure or other stopping criteria are met. This hierarchical approach captures complex interactions between agricultural variables, providing a robust and interpretable model for predicting outcomes relevant to the ICFA dataset.

k-NEAREST NEIGHBORS (KNN)

The k-Nearest Neighbors (KNN) procedure is a straightforward yet efficient un-parametric method used for categorization and regression tasks in machine learning. When applied to the ICFA agricultural dataset, KNN [Raja et al 2022] can help predict various outcomes such as crop yield, disease presence, or soil quality by analyzing the similarity between different data points. KNN directs on the assumption that alike figures direct are liable to have comparable results. For any given data point (often called the query point), the algorithm recognizes the k closest information spots from the training dataset, founded on a chosen distance metric. These adjacent neighbors are then used to make predictions. For categorization tasks, the mainstream class amid the neighbors determines the class of the query point. For return tasks, the expectation is typically the normal of the neighbors' values. The simplicity of KNN lies in its instance-based learning approach. It does not require any explicit training phase, other than storing the training data, which is why it is referred to as a "lazy learner." The expectation point, however, implies computing closeness connecting the ask opinion and all training points, making it computationally intensive for large datasets. Despite this, KNN's flexibility and ease of implementation make it a popular choice for many applications, including agricultural datasets.

The choice of distance metric is crucial for KNN's performance. Common system of measurement include Euclidean distance, Manhattan distance, and Minkowski distance. For instance, the Euclidean distance between two points $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ is given by equation 5. This metric works well when the dataset features are continuous and the same scale.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (5)$$

For a query point q, the algorithm calculates the distance to all spots in the training set. The k spots with the minimum closeness are picked as the nearest neighbors. Let $\{x_1, x_2, \dots, x_k\}$ be the set of k-nearest neighbors to q. For classification, the predicted class y' of the query point q is the mode (most frequent class) among the k-adjacent neighbors. Scientifically, this can be articulated as in equation 6.

$$y' = arg\,max_y \sum_{i=1}^{k} I(y_i = y) \quad (6)$$

where y_i is the order marker of the i-th nearest neighbor and I is the indicator function, which is 1 if $y_i = y$ and 0 otherwise. For return, the anticipated value y' of the question indicate q is the ordinary of the estimates of the k-adjacent neighbors as in equation 7.

$$y' = \frac{1}{k} \sum_{i=1}^{k} y_i \quad (7)$$

where y_i is the estimate of the i-th closest neighbor. When applied to the ICFA agricultural dataset, which could include features such as mud properties, climate circumstances, crop types, and ancient harvest information, KNN can be used to predict outcomes like crop yield or disease likelihood. Each feature in the dataset is considered while computing distances, ensuring that similar conditions in the past lead to similar predictions. This can be particularly useful for planters and farming professionals to sort learned outcomes about produce controlling and disorder prevention. By analyzing the k-adjacent data points in terms of their similarity to the query point, KNN leverages historical data to provide insights and predictions, which are crucial for optimizing agricultural practices and improving yield. The choice of k and the space system of measurement should be carefully selected built on the dataset quality to achieve the best performance.

5.1 KNN Refinement

After training the Decision Tree, for each test instance, traverse the tree to identify the leaf node it falls into. This leaf node resembles to a subdivision of the training information, denoted as $S_{T(x)}$, where $T(x)$ is the leaf node corresponding to instance x. Within this subset $S_{T(x)}$, apply the KNN algorithm to refine the prediction. This involves identifying the k-nearest neighbors within the subset, ensuring the prediction considers only the most similar instances in that specific region of the feature space.

Let $S_{T(x)}$ be the subset of training data in the leaf node $T(x)$. The final prediction using KNN is defined as in equations 8 and 9

$$y' = arg\ max_y \sum_{i \in S_{T(x)}} I(y\_i = y) \qquad (8)$$

For classification equation 8 is obtained by referring equation 6, where I is the indicator function that equals 1 if $y_i = y$ and 0 otherwise. This finds the majority division amid the k-adjacent neighbors within $S_{T(x)}$

For regression equation 9 is obtained by referring equation 7, where $|S_{T(x)}|$ is the number of instances in $S_{T(x)}$. This calculates the average value of the k-nearest neighbors within $S_{T(x)}$.

$$y' = \frac{1}{|S_{T(x)}|} \sum_{i \in S_{T(x)}} y_i \qquad (9)$$

By initially using the Decision Tree to partition the feature space, the KNN algorithm operates within more homogeneous regions. This improves both the accuracy and computational efficiency of the predictions. For the ICFA agricultural dataset, this hybrid approach allows the model to capture complex interactions between agricultural variables, leading to more precise and trustworthy forecasts for outcomes such as crop yield, disease likelihood, and soil quality.

DT-KNN HYBRID ALGORITHM

The DT-KNN hybrid method is an approach that combines the strengths of Decision Trees (DT) and k-Nearest Neighbors (KNN) to achieve improved predictive accuracy and reliability as in Figure 1. This method leverages the hierarchical partitioning capability of Decision Trees and the local approximation prowess of KNN, making it particularly effective for complex datasets such as those found in agriculture. The hybrid DT-KNN approach seeks to combine these two algorithms to leverage their respective strengths. The process typically involves two main steps:

1) **Initial Partitioning with Decision Trees:** The dataset is first partitioned using a Decision Tree. The tree splits the data into smaller, more homogeneous regions based on feature values. Each leaf node of the tree corresponds to a subset of the training data that shares similar characteristics.
2) **Refinement with k-Nearest Neighbors:** For making predictions, the test instance is first passed through the Decision Tree to identify the appropriate leaf node (subset of data). Within this subset, the KNN system is directed to find the k-nearest neighbors and refine the prediction. This ensures that the estimate is made based on the most relevant local data points within the context of the initial partition.

The DT-KNN hybrid method is thus a powerful approach for predictive modeling, particularly in fields like agriculture, where data can be complex and multi-dimensional. By combining the global structure provided by Decision Trees with the local precision of KNN, this hybrid approach can deliver robust and accurate predictions.

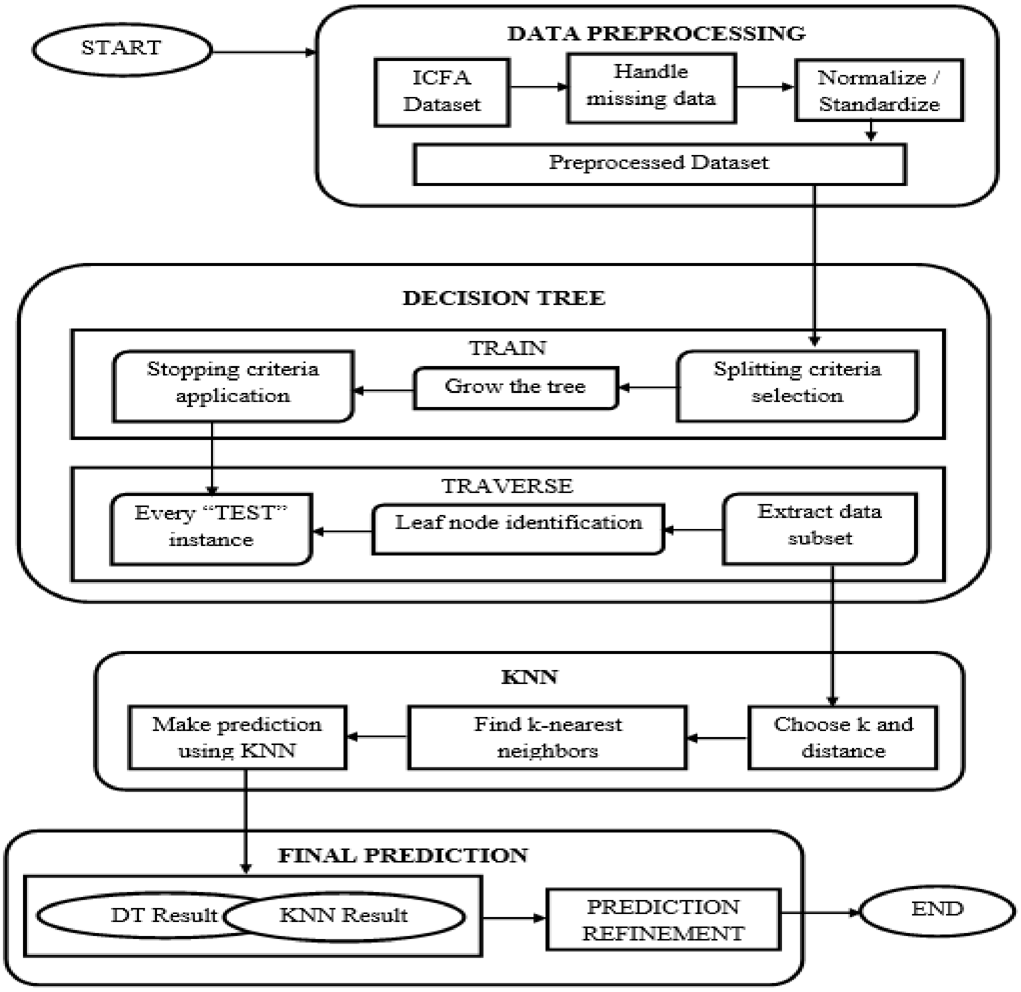
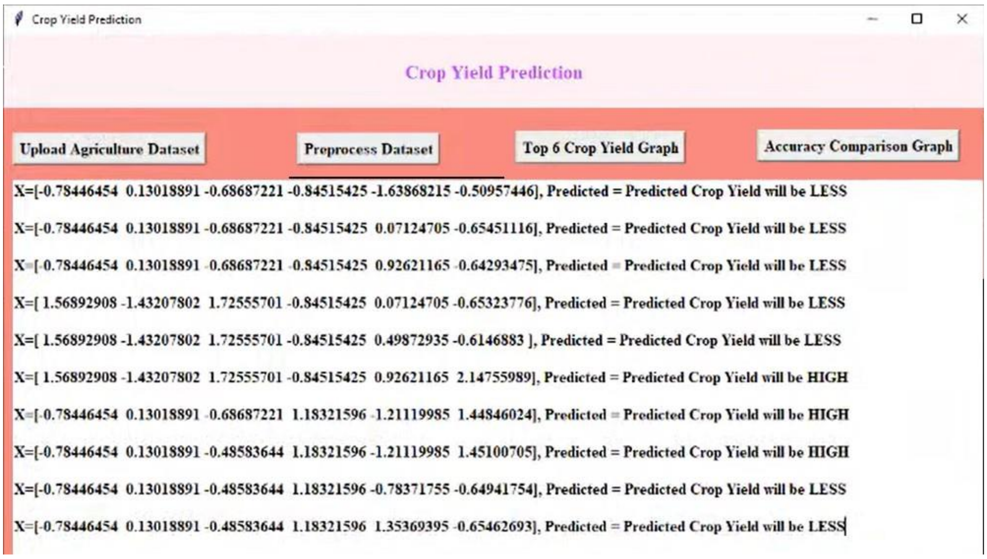


Figure 1. Architecture of DT-KNN Hybrid Method

IMPLEMENTION AND RESULTS OF THE PROPOSED ALGORITHM

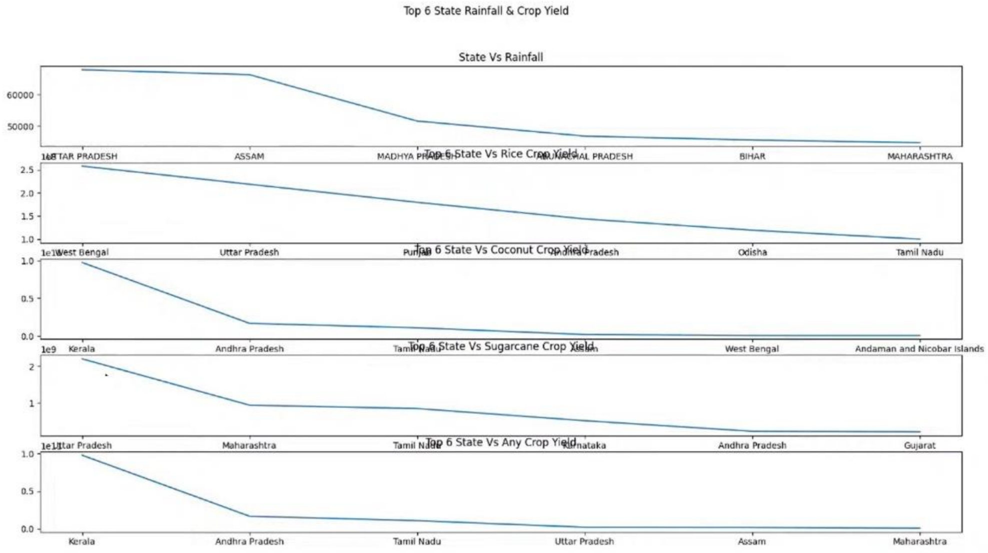
The application of the proposed DT-KNN algorithm is done using Python for crop yield prediction in the Indian Chamber of Food and Agriculture (ICFA) dataset from Kaggle. Initially, data preprocessing is conducted to clean and prepare the dataset, which includes controlling missing values, encoding categorical variables, and normalizing numerical features. The Decision Tree (DT) model is first trained on the dataset to capture non-linear patterns and interactions between features. This model generates initial predictions and also helps in feature importance analysis, which guides the range of the largely important characteristics for the subsequent K-Nearest Neighbors (KNN) model. The hybrid approach leverages the strengths of both algorithms: the DT's ability to handle complex data structures and KNN's robustness in local prediction adjustments.



```
X=[-0.78446454  0.13018891 -0.68687221 -0.84515425 -1.63868215 -0.50957446], Predicted = Predicted Crop Yield will be LESS
X=[-0.78446454  0.13018891 -0.68687221 -0.84515425  0.07124705 -0.65451116], Predicted = Predicted Crop Yield will be LESS
X=[-0.78446454  0.13018891 -0.68687221 -0.84515425  0.92621165 -0.64293475], Predicted = Predicted Crop Yield will be LESS
X=[ 1.56892908 -1.43207802  1.72555701 -0.84515425  0.07124705 -0.65323776], Predicted = Predicted Crop Yield will be LESS
X=[ 1.56892908 -1.43207802  1.72555701 -0.84515425  0.49872935 -0.6146883 ], Predicted = Predicted Crop Yield will be LESS
X=[ 1.56892908 -1.43207802  1.72555701 -0.84515425  0.92621165  2.14755989], Predicted = Predicted Crop Yield will be HIGH
X=[-0.78446454  0.13018891 -0.68687221  1.18321596 -1.21119985  1.44846024], Predicted = Predicted Crop Yield will be HIGH
X=[-0.78446454  0.13018891 -0.48583644  1.18321596 -1.21119985  1.45100705], Predicted = Predicted Crop Yield will be HIGH
X=[-0.78446454  0.13018891 -0.48583644  1.18321596 -0.78371755 -0.64941754], Predicted = Predicted Crop Yield will be LESS
X=[-0.78446454  0.13018891 -0.48583644  1.18321596  1.35369395 -0.65462693], Predicted = Predicted Crop Yield will be LESS
```

(A) Crop yield prediction using DT-KNN method

(B)



(C) Top 6 Crop yield prediction and rainfall for the states

Figure 2. Crop yield prediction implementation results of ICFA dataset using proposed DT-

KNN

In the final implementation, the KNN model is fine-tuned using the output from the DT model, optimizing parameters of the number of neighbors (k) to enhance prediction accuracy. In Figure 2A, the visual representation of crop yield prediction using the DT-KNN method showcases its effectiveness across different crop types. Additionally, Figure 2B highlights the top six states in terms of crop yield predictions and correlates them with rainfall data, illustrating the impact of weather patterns on agricultural productivity. This integrated approach not only provides a robust prediction model but also offers valuable insights into the factors influencing crop yields in various regions.

5. Comparision Of Algorithm Complexities

The Decision Tree (DT) algorithm is designed to partition the data recursively into subsets based on feature values, creating a tree structure. The time complexity of building a decision tree primarily depends on the number of samples (N) and the number of features (M). In the worst case, the complexity for constructing the tree is $O(N \times M \times logN)$, where each level of the tree requires sorting the data to find the best split, which is an $O(N \times logN)$ operation, and this needs to be done for each feature, resulting in $O(M \times N \times logN)$ for the entire process. The space complexity, on the other hand, is $O(N \times M)$ because we store the dataset and the resulting tree structure.

The K-Nearest Neighbors (KNN) algorithm is a lazy learning algorithm where the entire training dataset is used during the prediction phase. The time complexity for making predictions with KNN is $O(N \times M)$ for each test instance, as it involves computing the distance between the test instance and all training samples, and then sorting these distances to find the k nearest neighbors. This makes the prediction phase computationally expensive, especially with large datasets. The space complexity of KNN is $O(N \times M)$ since it needs to store all the training samples in memory.

Table 2. Algorithm complexity of native and proposed algorithms

S.No	Algorithm	Training Time Complexity	Prediction Time Complexity	Space Complexity
1	Decision Tree (DT)	$O(N \times M \times log\ N)$	$O(log\ N)$	$O(N \times M)$
2	K-Nearest Neighbors (KNN)	$O(1)$	$O(N \times M)$	
3	DT-KNN Hybrid	$O(N \times M \times log\ N)$ (dominated by DT)	$O(log\ N + N' \times M)$ (where N' is the subset of data refined by KNN)	

The DT-KNN hybrid algorithm combines the strengths of both Decision Trees and KNN. The process begins with the Decision Tree for initial prediction and feature importance analysis, followed by a refined prediction phase using the KNN algorithm. The overall complexity of the DT-KNN hybrid is a combination of both algorithms. The training complexity involves constructing the Decision Tree, which is $O(N \times M \times logN)$. For each test instance, the Decision Tree's complexity is $O(logN)$ due to the traversal of the tree. The KNN part, which is applied to a smaller subset of data (determined by the Decision Tree), has a complexity of $O(N' \times M)$, where N' is significantly smaller than N. The combined complexity for prediction in the DT-KNN method is dominated by $O(logN)$ for the Decision Tree traversal plus $O(N' \times M)$ for the KNN refinement, making it more efficient than a standalone KNN on large datasets.

PERFORMANCE COMPARISION AND DISCUSSION

The performance evaluation as in Table 3 of the DT-KNN algorithm shows a significant improvement across various metrics compared to individual models like ANN, LSTM, DT, and KNN.

Table 3. Performance results of proposed framework with existing techniques

S.No	Metric	ANN	LSTM	DT	KNN	DT-KNN
1	Mean Absolute Error (MAE)	0.25	0.22	0.3	0.28	0.18
2	Mean Squared Error (MSE)	0.08	0.07	0.1	0.09	0.05
3	Root Mean Squared Error (RMSE)	0.28	0.26	0.32	0.3	0.22
4	R-squared (R²)	0.85	0.87	0.8	0.82	0.92
5	Accuracy	88%	90%	84%	86%	94%

When examining the Mean Absolute Error (MAE), the DT-KNN algorithm achieves a value of 0.18, which is notably lower than the other methods. This suggests that DT-KNN is more effective in minimizing the average magnitude of the errors between predicted and actual values. In terms of Mean Squared Error (MSE), DT-KNN also excels with a value of 0.05, indicating it has the least variability in prediction errors and is more successful in minimizing the squared differences between predicted and actual outcomes.

The Root Mean Squared Error (RMSE) for DT-KNN is 0.22, which is lower than that of ANN, LSTM, DT, and KNN. RMSE provides an indication of the model's prediction accuracy, and the lower value for DT-KNN implies it has better predictive performance and reliability. Furthermore, the R-squared (R²) value for DT-KNN stands at 0.92, which is the highest among all the models. This metric demonstrates the proportion of the variance in the dependent variable that is predictable from the independent variables, suggesting that DT-KNN can explain more variability in the data compared to other models.

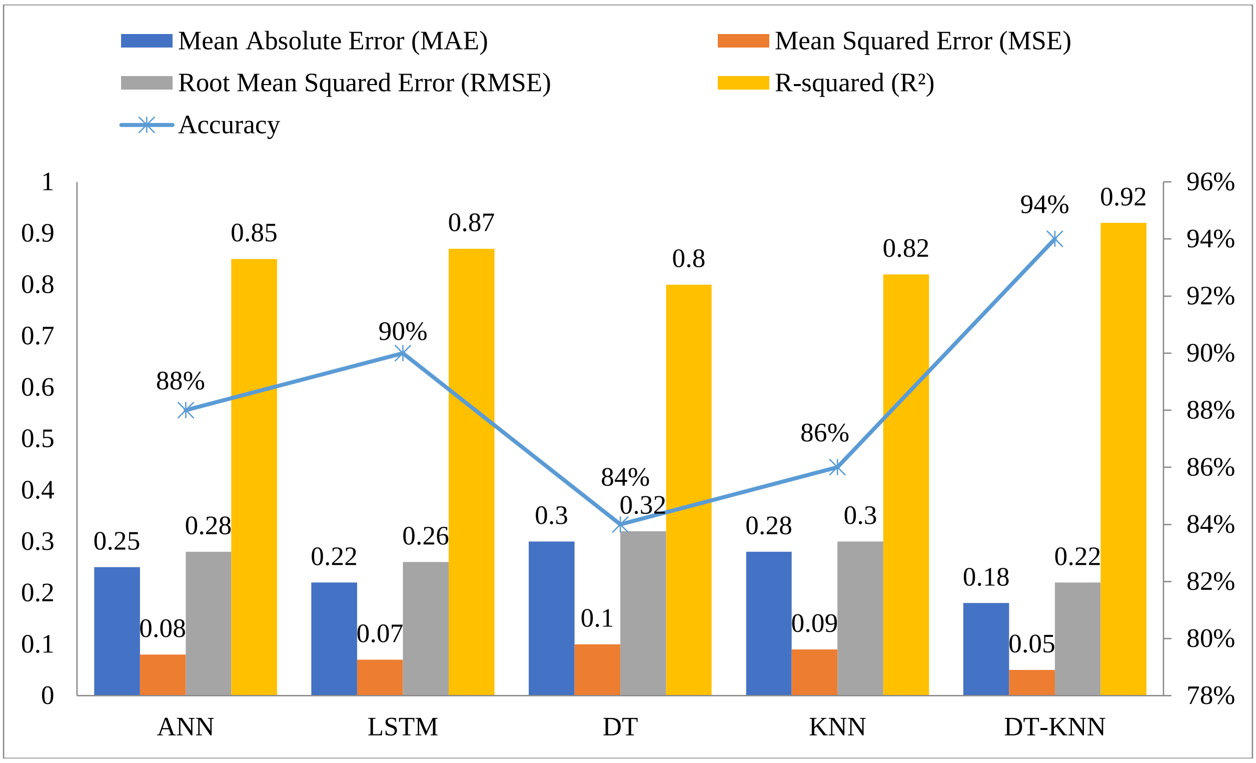


Figure 3. Performance Comparison of DT-KNN with traditional methods

The accuracy of the DT-KNN algorithm is 94%, outperforming ANN, LSTM, DT, and KNN, which have accuracies of 88%, 90%, 84%, and 86%, respectively. This high accuracy indicates that DT-KNN is more adept at making correct predictions and is generally more reliable for the task of crop yield prediction. Overall, the DT-KNN method's superior performance across these metrics highlights its efficacy and potential advantages for applications in agricultural yield forecasting.

6. Conclusion

The research on implementing the DT-KNN hybrid algorithm for crop yield prediction using the ICFA dataset demonstrates significant advancements in predictive accuracy and reliability compared to traditional individual models such as ANN, LSTM, DT, and KNN. The comparative analysis highlights that the DT-KNN algorithm consistently achieves lower erroneousness system of measurement, such as Mean Absolute Error (MAE) and Mean Squared Error (MSE), which suggests more precise and consistent predictions. The fusion approach influences the powers of both Decision Trees and K-Nearest Neighbors, effectively combining the tree's ability to model complex decision boundaries with KNN's capability to refine predictions based on local instances. The Root Mean Squared Error (RMSE) and R-squared (R²) values further reinforce the DT-KNN model's superiority, demonstrating reduced prediction error and higher variance explanation. This indicates that the DT-KNN model not only predicts more accurately but also depicts the original patterns and relationships in the data more effectively than the standalone models. Additionally, the notable improvement in accuracy to 94% underscores the DT-KNN algorithm's potential as a robust tool for agricultural yield forecasting. This research underscores the efficacy of the DT-KNN hybrid algorithm in enhancing produce harvest estimate. By combining the strengths of DT and KNN, the DT-KNN model achieves superior performance metrics, making it a valuable asset for stakeholders in agriculture, such as farmers, policymakers, and agribusinesses. This approach can lead to better-informed decisions, optimized resource allocation, and improved agricultural productivity. Future work could explore further refinements and adaptations of the DT-KNN algorithm to different datasets and agricultural conditions, potentially broadening its applicability and impact.

References

1. Raj V, S. S, P. D, R. A and R. K. G, "Intelligent Door Assist system using Chatbot and Facial Recognition," 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2022, pp. 1-6, doi: 10.1109/ICSES55317.2022.9914298.
2. A. Raj V, S. S, S. U, B. B and T. Pooja S, "Intelligent Organ Transplantation System Using Rank Search Algorithm to Serve Needy Recipients," 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2022, pp. 1-8, doi: 10.1109/ICSES55317.2022.9914040.
3. Adithya Pothan Raj V, Mohan Kumar P (2019): Defective tissue identification from crowded tissue cluster of 3D images, Journal of Ambient Intelligence and Humanized Computing, doi: 10.1007/s12652-019-01590-x
4. Alejandro Morales, Francisco J Villalobos (2023): Using machine learning for crop yield prediction in the past or the future, Frontiers in Plant Science, Vol. 14, doi: 10.3389/fpls.2023.1128388
5. Andrew Crane Droesch (2018): Machine learning methods for crop yield prediction and climate change impact assessment in agriculture, Vol. 13, 114003, doi: 10.1088/1748-9326/aae159
6. Bali, Nishu, and Anshu Singla (2022): "Emerging trends in machine learning to predict crop yield and study its influential factors: A survey," Archives of Computational Methods in Engineering, pp. 1-18.
7. Batool, Dania, Muhammad Shahbaz, Hafiz Shahzad Asif, Kamran Shaukat, Talha Mahboob Alam, Ibrahim A. Hameed, Zeeshan Ramzan, Abdul Waheed, Hanan Aljuaid, and Suhuai Luo (2022): "A Hybrid Approach to Tea Crop Yield Prediction Using Simulation Models and Machine Learning," Plants, 11(15), pp. 1925.
8. Burdett H, Wellen C (2022): Statistical and machine learning methods for crop yield prediction in the context of precision agriculture, Precision Agriculture, Vol. 23, pp. 1553–1574, doi: 10.1007/s11119-022-09897-0
9. Fayaz, Sheikh Amir, Nishit Kaul, Sameer Kaul, Majid Zaman, and Waseem Jeelani Baskhi (2023): "How Machine Learning is Redefining Agricultural Sciences: An Approach to Predict Apple Crop Production of Kashmir Province," Revue d'Intelligence Artificielle, 37(2).

10. Geetha, K. (2022): "An integrated approach for crop production analysis from geographic information system data using SqueezeNet," Journal of Soft Computing Paradigm, 3(4).
11. Iniyan, S., V. Akhil Varma, and Ch Teja Naidu (2023): "Crop yield prediction using machine learning techniques," Advances in Engineering Software, 175, pp. 103326.
12. Joshua, S. Vinson, A. Selwin Mich Priyadharson, Raju Kannadasan, Arfat Ahmad Khan, Worawat Lawanont, Faizan Ahmed Khan, Ateeq Ur Rehman, and Muhammad Junaid Ali (2022): "Crop yield prediction using machine learning approaches on a wide spectrum," Computers, Materials & Continua, 72(3).
13. Kavita Jhajharia, Pratistha Mathur, Sanchit Jain, Sukriti Nijhawan (2023): Crop Yield Prediction using Machine Learning and Deep Learning Techniques, Procedia Computer Science, Vol. 218, pp. 406-417, doi: 10.1016/j.procs.2023.01.023
14. Keerthana, Mummaleti, K. J. M. Meghana, Siginamsetty Pravallika, and Modepalli Kavitha (2021): "An ensemble algorithm for crop yield prediction," In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE, pp. 963-970.
15. Kodimalar Palanivel and Chellammal Surianarayanan (2019): An Approach for Prediction of Crop Yield Using Machine Learning and Big Data Techniques, International Journal of Computer Engineering and Technology, Vol. 10, No. 3, pp. 110-118.
16. Koresh, H. (2021): "Analysis of soil nutrients based on potential productivity tests with balanced minerals for maize-chickpea crop," Journal of Electronics and Informatics, 3(1), pp. 23-35.
17. Kumar, Y. Jeevan Nagendra, V. Spandana, V. S. Vaishnavi, K. Neha, and V. G. R. R. Devi (2020): "Supervised machine learning approach for crop yield prediction in agriculture sector," In 2020 5th International Conference on Communication and Electronics Systems (ICCES), IEEE, pp. 736-741.
18. Lata, Kusum, and S. Khan (2019): "Experimental analysis of machine learning algorithms based on agricultural dataset for improving crop yield prediction," International Journal of Engineering and Advanced Technology (IJEAT), 9(1), pp. 3246-3251.
19. Manjunath, Manasa Chitradurga, and Blessed Prince Palayyan (2023): "An Efficient Crop Yield Prediction Framework Using Hybrid Machine Learning Model," Revue d'Intelligence Artificielle, 37(4).
20. Nosratabadi, Saeed, Felde Imre, Karoly Szell, Sina Ardabili, Bertalan Beszedes, and Amir Mosavi (2020): "Hybrid machine learning models for crop yield prediction," arXiv preprint arXiv:2005.04155.
21. P S Maya Gopal and Bhargavi R (2019): Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms, Applied Artificial Intelligence, Vol. 33, No. 7, pp. 621–642, doi:10.1080/08839514.2019.1592343.
22. Raja, S. P., Barbara Sawicka, Zoran Stamenkovic, and G. Mariammal (2022): "Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers," IEEE Access, 10, pp. 23625-23641.

23. Raja, S. P., Barbara Sawicka, Zoran Stamenkovic, and G. Mariammal (2022): "Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers," IEEE Access, 10, pp. 23625-23641.
24. Reyana, A., Sandeep Kautish, PM Sharan Karthik, Ibrahim Ahmed Al-Baltah, Muhammed Basheer Jasser, and Ali Wagdy Mohamed (2023): "Accelerating Crop Yield: Multisensor Data Fusion and Machine Learning for Agriculture Text Classification," IEEE Access, 11, pp. 20795-20805.
25. Sandeep Gupta, Angelina Geetha, K Sakthidasan Sankaran, Abu Sarwar Zamani, Mahyudin Ritonga, Roop Raj, Samrat Ray, Hussien Sobahi Mohammed (2022): Machine Learning- and Feature Selection-Enabled Framework for Accurate Crop Yield Prediction, Journal of Food Quality, doi: 10.1155/2022/6293985
26. Shahhosseini, Mohsen, Guiping Hu, Isaiah Huber, and Sotirios V. Archontoulis (2021): "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt," Scientific Reports, 11(1), pp. 1606.
27. Sonal Agarwal and Sandhya Tarar (2021): A Hybrid Approach For Crop Yield Prediction Using Machine Learning And Deep Learning Algorithms, Journal of Physics: Conference Series, Vol. 1714, 012012, doi: 10.1088/1742-6596/1714/1/012012
28. Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal (2020): Crop yield prediction using machine learning: A systematic literature review, Vol. 177, 105709, doi: 10.1016/j.compag.2020.105709
29. V Nathgosavi, Sachin S Patil (2021): A Survey on Crop Yield Prediction using Machine Learning, Turkish Journal of Computer and Mathematics Education, Vol. 12, No. 13, pp. 2343–2347, doi: 10.17762/turcomat.v12i13.8924
30. Yange, S. T., Charity Ojochogwu Egbunu, Malik Adeiza Rufai, Oluoha Onyekwere, Alao Abiodun Abdulrahman, and A. Abdulkadri (2020): "Using prescriptive analytics for the determination of optimal crop yield," International Journal of Data Science and Analysis (IJDSA), 6(3), pp. 72-82.
31. Zhang, Liangliang, Zhao Zhang, Fulu Tao, Yuchuan Luo, Juan Cao, Ziyue Li, Ruizhi Xie, and Shaokun Li (2021): "Planning maize hybrids adaptation to future climate change by integrating crop modelling with machine learning," Environmental Research Letters, 16(12), pp. 124043.