I. J. Engineering and Manufacturing, **2025, 6, 46-59**
Published Online on December 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2025.06.04

A Novel Resource and Distribution Aware Random Forest for Agricultural Productivity Prediction

Harendra Singh Negi*
Department of Computer Science and Engineering, Graphic Era (Deemed to be University), India
E-mail: mail.harendrasinghnegi@gmail.com
ORCID iD: https://orcid.org/0009-0002-1837-6498
*Corresponding Author

Sushil Chandra Dimri
Department of Computer Science and Engineering, Graphic Era (Deemed to be University), India
E-mail: dimri.sushil2@gmail.com
ORCID iD: https://orcid.org/0000-0002-5921-2505

Received: 12 March, 2025; Revised: 20 May, 2025; Accepted: 18 July, 2025; Published: 08 December, 2025

Abstract: Agriculture has continued being one of the economic powerhouses of India, but then the productivity is usually compromised due to the poor utilization of soil and environment data. This paper is a proposal of a new framework named Distribution and Resource Aware Random Forest (DRARF) to be used in smart farming applications. The strategy combines IoT-ready soil data that comprises of moisture, temperatures, humidity, pH, and NPK that are monitored via different sources and used to make crop-specific decisions. The DRARF presents two important novel features to traditional Random Forests: (i) distribution-aware threshold selection, which guarantees statistical meaningful data partition and (ii) resource-aware feature selection, which gives more predictive power without the expense of buying sensors in the IoT. It assessed the framework using soil and environmental data of wheat and rice. The comparative tasks performed using Logistic Regression, Support Vector Machine, Naïve Bayes, and the classical random Forest have shown that DRARF not only provides a better accuracy, precision, recall and F1-scores, but it also minimizes sensor redundancy. Its potential depends on its scalability, efficiency, and reliability as a precision agricultural decision-support system and this, as well as the remaining results, are reflected in the results. The given approach with machine learning and IoT-facilitated sensing source-based solutions can bring the advancements in the sphere of smart farming technologies to help increase the yield of crops and resources and improve long-term food security.

Index Terms: DRARF, Agriculture, NPK, Productivity, Prediction

1. Introduction

The economy of India is also strong in Agronomy where it pays a big percentage of the population and contributes most to the production of food. Nonetheless, the productivity of crops is usually limited due to lack of access to quality soil and environment information by the farmers. The conventional farming practices rely entirely on human eyes and gut feelings that often result in non-optimality of crop selection, ineffectiveness of resource application, and miniaturisation [1]. The use of new technologies in the field of agriculture has become necessary to discuss these issues. The solution to these limitations can be found with Smart farming, which is driven by the IoT and sensor-based data collection [2]. NPK, pH, moisture, humidity and temperature are also part of the soil nutrient that can now be monitored and relayed in real time. This data, together with machine learning (ML), helps to create automated decision support systems to help farmers with selecting the right crop, save resources, and enhance the general productivity. Smart farming provides superior accuracy, scalability and sustainability as compared to the conventional methods [3]. ML has shown itself as a valid contribution especially in precision agriculture. ML models can be used to forecast crop suitability, predict the potential yield, and suggest the most appropriate farming practices by finding patterns in soil and climatic data [4]. Random Forests are popular among ensemble methods as they achieve strong performance and are

capable of dealing with missing or noisy data among other ML approaches. Nevertheless, traditional versions of Random Forests are limited in several ways: they use homogeneous random sampling of features as well as thresholds, which is not valid regarding the distribution of agricultural data or the fact that the price of an IoT-based sensor will differ.

This paper will tackle these problems by suggesting the use of a Dissemination and Resource Aware Random Forest (DRARF) to smart farming. The proposed method, in contrast to the standard Random Forests, has two properties: (i) distribution-aware choice of thresholds (which ensure that splits are placed at statistically significant positions), and (ii) resource-awareness of feature selection (which trades predictive power against sensor acquisition costs). The suggested model is more accurate when the data on soil nutrients and environmental conditions of wheat and rice are applied, as well as by decreasing unnecessary sensor applications. This contribution presents useable and scalable architecture of IoT-enabled agriculture to enable Agronomist with data-driven suggestions to enhance efficiency and sustainability.

The originality of the proposed research is that it leads to the creation of a Distribution-and-Resource-Aware Random Forest (DRARF) framework, being explicitly developed to support IoT-enabled smart farming. The proposed DRARF has two innovations unlike traditional random forest techniques, in which they uniformly sample the features and thresholds. The first one is Distribution Aware Threshold Selection, in this case Candidate thresholds are selected according to the process of statistical distribution of soil and climate properties (e.g., quantiles, median splits) and not randomly sampled out. This guarantees they will split at informative regions of the feature space enhancing prediction accuracy and interpretability. Second ne is Resource Aware Feature Selection, i.e. here Sensor-derived attributes (e.g., NPK, moisture, humidity) are weighted based on their acquisition cost as well as energy consumption in addition to their predictive significance. This allows the model to optimize on the features that trade off accuracy with the simplicity of resource consumption an aspect that is mostly disregarded within any prevalent agricultural ML systems. Combined with these improvements, DRARF can be distinguished on the one hand over the previous application of Random Forest in the agricultural industry, in which mostly standard ensembles are applied in a blind manner, without a single reference to the cost of features or the distribution of data in selecting a threshold. The suggested framework thus offers increased predictive accuracy, as well as feasibility to actual-time, IoT-enabled smart agriculture applications, and crop suitability analysis is directly served in wheat and rice production.

2. Related Works

Irrigation and nutrient monitoring have also been suggested as solutions to optimize the use of IOT-enabled soil and water management systems [5]. The initial efforts were concerned with timing sensor power supply to minimize energy [7]. Farmers today use technologies in order to track soils nutrients, moisture, and temperature but data transmission may be spoiled as a result of power failures or misleading sensor data, which then has to be verified through validation [6]. Although the systems have potential, majority of the systems have limited ranges of operation and also they are not integrated so as to ensure large scale remote implementations. Wagner suggested the application of WSNs to go over ecological variables in agriculture [8]. To improve the accuracy, more sensors should be used, however, the power consumption increases. IoT architectures have been proposed to control soil temperature, soil moisture, humidity, nutrients, and crop infections using layered architectures [9], but the majority of them are sensitive to real-time preparation of high data rates [10]. The network storage and intelligent sensor hubs have been added later, to improve connectivity [11], although most of the systems continue working on data collection without successful interpretation or integration to be easily accessible to users [12]. Application-layer modules have been used to process environmental and agricultural data in order to provide support to the forecasting models [13]. Random Forest (RF) has shown that it can effectively perform in this context because it is not sensitive to missing values and has the capability of classifying complex features, thus becoming useful in analyzing sensor-based data in agriculture [14], [15]. Cano Marchal et al. [16] also advanced feature extraction as a swift classification framework of crop samples, and it is applicable to the instances when moderate accuracy is required. Although RF and other methods like it bear promise, much of the research done on the subject has concentrated on the accuracy rather than coat-of-arms to the issue of computational efficiency and promptness, which are a key aspect of agricultural real-time applications. Ranjbar et al. [17] predicted soil moisture with a regression based method that utilizes DInSAR to predict and the model worked well in most parts but had lower accuracy in vegetated areas, which demonstrates that more features should be implemented. Alfred et al. [18] have offered a review of precision farming in paddy rice with a focus on multispectral and radar data use in crop growth modeling but the practical implementation involves multi-source integration and optimal feature selection. Lu et al. [19] proposed a slightly monitored study in detecting wheat disease and obtained similar accuracy as conventional neural networks with limited annotations. Ferreira et al. [20] proposed a neural network-based weed detection approach which surpasses the need of manual feature extractions and proves applicable on crops with slight adaptations. Likewise, Czymmek et al. [21] used deep learning to classify plants, which was better than the Random Forest but had an increased computational cost and processing time. The potato ailment management decision support system by Foughali et al. [22] utilizes sensors and cloud-based solutions, which the authors state would be enhanced with better ML integration. Giusti et al. [23] developed an irrigation forecasting system that relies on crop transpiration and growing degree days, and this has shown that such a system can be deployed in the fields with intensive networking,

although such deployment requires the system to be entrenched with a robust networking support. The study by AlZu'bi et al. [24] created an unattended irrigation device that enables the process of water management by using IoT sensors combined with machine learning, demonstrating a high potential in the expansion of the IoT solution into different fields. Partel et al. [25] used machine vision based on AI to detect weeds, to make variable applications of agrochemicals defined by canopy size and vehicle speed and using less chemicals. Kounalakis et al. [26] applied transfer learning to 2D images in the recognition of weeds in the real field, and the research proved to be effective in terms of combining several data sources to achieve powerful classification. Bazzi et al. [27] suggested a mathematical model of time series to map the irrigated areas with a 90 percent accuracy but had to be applied to geographies other than the one in the study. Dong et al. [28] used CNNs in automated land parcel extraction on large scale remote sensing images, which were more effective than conventional ones, but they were not highly effective with mountainous environment by environmental variability. Farooq et al. [29] contented on IoT designs and protocols of soil measurements which cover humidity, moisture, fertilization, and temperature, yet the training and cost-effectiveness of farmers are the factors that challenge the adoption of this technology. Abdullah et al. [30] proposed irrigation control system using fuzzy-logic, which minimized the use of water and pesticides as well as maximized nutrient delivery supplies but this requires additional validation on datasets.

Several ML models have been explored for crop forecasting, as summarized in Table 1.

Table 1. Literature Survey for Comparison of machine learning to the suitable crop

Authors/Citations	What has been done	Outcome	Scope for further work
S. Espinoza et al. [31]	This objective of this work is to perform a study on fruit analysis using Deep Learning models.	This model shows a lot of potential, yielding positive accuracy and performance outcomes. Depending on the job, many evaluation measures have been used; some of the more popular ones include precision, recall, F1, and mAP.	The sizes of the datasets have also varied, with bigger training sets typically used for classical CNN models and moderate-sized sets for identification classifiers.
Hanwen Kang et al. [32]	This study introduces a versatile network for immediate detection and semantic classification of fruits and splits in agricultural contexts using a visual input.	The network model with the lightest foundation had the highest computational effectiveness, according to the data. Its apple recognition F1 score was 0.827, while its apple and branch segmentation F1 scores were 86.5 and 75.7 percent, respectively.	It is necessary to improve the comparison of object detection and semantic segmentation performance with other cutting-edge efforts.
J. Agarwal et al. [33]	This work investigated crop yield price projection, weather prediction, soil categorization, and crop type for agricultural planning using information techniques or system learning methodologies.	When there are multiple ways to plant a crop at simultaneously on a limited amount of land, crop selection becomes difficult. Here, K mean is used to examine the feature set and accuracy of the results is 97%.	As observed in other application domains, every approach now in use is based on distinct algorithms and validations and has little to no connection to the procedures for making decisions. The outcomes are noticeably better when layered regression is used than when those models are used separately.
Z. Liu et al. [34]	A unique method for spiking segments that combines density based regional groupings Laplacian based area expansion, and the kernel predicting neural network using convolution.	Their findings demonstrated that the suggested method allowed for precise segmentation of distinct spikes, producing an F-score of 84.62%.	Although there is much potential in this approach in outdoor field conditions, there is still some exploration that needs to be done. The model's F score is improvable.
P. Sharma et al. [35]	The purpose of this work is to forecast agricultural yield using various factors such as rainfall, crop, weather, region, growth, and revenues that have posed a significant risk to agriculture's future sustainability.	The crop output has been determined by the methods based on machine learning. To have a better understanding of how the mistakes in the model compare to alternative approaches, the results need to examine accurately.	The combination of district-level data analysis and data from satellite imagery could improve the model's ability to estimate crop yield.
Mohanty et al. [36]	Applied CNNs for image-based plant disease detection.	Achieved high accuracy on leaf image datasets.	Extend to multi-crop datasets and real-time field deployment.
Too *et al.* [37]	Compared fine-tuned deep learning models for plant disease classification.	Compared fine-tuned deep learning models for plant disease classification.	Compared fine-tuned deep learning models for plant disease classification.
You et al. [38]	Used deep Gaussian processes with satellite data for crop yield prediction.	Used deep Gaussian processes with satellite data for crop yield prediction.	Used deep Gaussian processes with satellite data for crop yield prediction.
Khaki et al. [39]	Designed deep neural networks for crop yield prediction.	Designed deep neural networks for crop yield prediction.	Designed deep neural networks for crop yield prediction.

Maximizing crop output is the primary goal of the development of machine learning-enabled smart farming systems. The intended effort tackles smart farming concerns at different phases. That is solely focused on classifying smart farming in order to develop effective smart farming that will help farmers gain accurate knowledge about crop selection and boost grain production.

The available research reveals that, in farming prediction, most of the available studies substantiate that, Random Forest is an effective and common algorithm and used in tasks that involve management of soil and climate variability. Nevertheless, such works usually take the classical form of the Random Forest which makes use of homogenous feature and threshold sampling. These approaches disregard two key factors (i) soil and environmental parameter distribution, and (ii) the difference in price and resources used by IoT-based sensors. These constraints limit the use of traditional Random Forests in scale-based and real-time intensive smart farming systems.

2.1. Comparative Positioning with Prior Studies

As an example, Alfred et al. (2021) used Random Forest in the context of paddy farming, however, their methodology did not include the discussion of sensor implementation costs or feature importance other than randomness. Likewise, Ranjbar et al. (2021) applied the use of the Random Forest in monitoring soil moisture, however, their approach was based on the use of uniform thresholding, and the splits were not distribution-sensitive, which restricted its accuracy.

In comparison, the Distribution and Resource Aware Random Forest (DRARF) is based on two methodological advances specifically distribution-conscience threshold selection and resource-conscious feature weighting. These contributions are made such that splits take place at statistically significant locations and they also maximize the predictive error versus sensor efficiency. Therefore, our solution is the extension of the traditional Random Forest framework, and it directly overcomes the two problems of prediction and resource-mindful implementation in IoT-based smart agriculture.

3. Proposed Approach

Currently, most farmers continue practicing the old system of farming with most of them making inefficiencies and losses. Due to the emergence of the IoT and sensor devices, numerous amounts of soil and crop data are now able to be gathered in real time. The problem, however, is in the development of machine learning models that will be able to utilize this information and be efficient in the resource-limited settings. This paper is our proposal of a new version of the random forest known as Distribution- and Resource-Aware Random Forest (DRARF), which is specifically oriented towards smart farming. As opposed to traditional Random Forests which equally sample features and thresholds in every split, our algorithms make two major innovations. The first one is Distribution-Aware Threshold Selection and the Second one is Resource-Aware Feature Selection.

The entire system will be a pair of modules:

3.1 Pre-processing of Analysis of Available Soil and Crop Data, in which temperature, humidity, moisture, rainfall, pH and NPK values are gathered and pre-processed.

3.2 Data Processing, the proposed DRARF algorithm will be implemented to break down soil and crop suitability and give recommendations to farmers.

This adaptation keeps in mind that not just will the assembly of trees achieve greater accuracy, but will also minimize redundancy in computations and number of sensors used which is very practical in precision agriculture.

3.1. Analysis of Available Soil and Crop Data

The data set that was adopted in this research was downloaded on various open-access resources, namely agricultural repositories and IoT-related crop monitoring databases and includes variables of NPK, soil pH, soil moisture, soil humidity, soil temperature, and rainfall. Because the data was collected by considering heterogeneous platforms, it was necessary to pre-process it following consistency and reliability.

3.1.1 Dataset Preprocessing

This phase includes Data Cleaning, missing value Processing, Outlier Detection, Normalization and Validation Split. Using these steps, data in raw and heterogeneous form was transformed into a homogeneous, high quality input format, which can be analyzed using machine learning.

These preprocessing actions played a significant role in making sure that the data is right, and they were the direct cause of the planning of the intended Distribution- and Resource-Aware Random Forest (DRARF). Specifically, in distribution-sensitive thresholding, it was necessary to have feature distributions that were accurate without noise or extreme outliers, whereas resource-sensitive feature selection used normalized and similar attributes in order to strike a balance of resource cost versus predictive and feature strength.

3.1.2 Exploratory Data Analysis

There was a demonstration of an EDA before training and evaluation to gain an improved understanding of how features relate to one another and what their contribution is in predicting. This is necessary in agricultural machine learning, in which soil and environmental variables are frequently interdependent.

- **Correlation Analysis**

All the features were used to create a Pearson correlation matrix, as depicted by Figure 1. In the analysis, some of the major relationships were found to be rainfall and moisture posted a high positive correlation (>0.70), which can be attributed to the fact that rainfall is a direct cause of soil water content, temperature and humidity was moderately negatively correlated, which is consistent with the natural negative interactions between the two variables, nutrient attributes (N, P, K) were shown to be weakly correlated with the environmental parameters such as pH, rainfall and temperature, which is in line with the known interactions between soil chemistry and climatic conditions, as well as pH, and these findings were used to justify the design of the experiment, in which groups of features (nutrient-based (NPK)) and environmental factors were tested individually (temperature, humidity, pH, rainfall). Also, the distorted correlations encouraged the use of distribution-conscious thresholding in DRARF where a split is made at points that are statistically significant to the feature space instead of randomly.

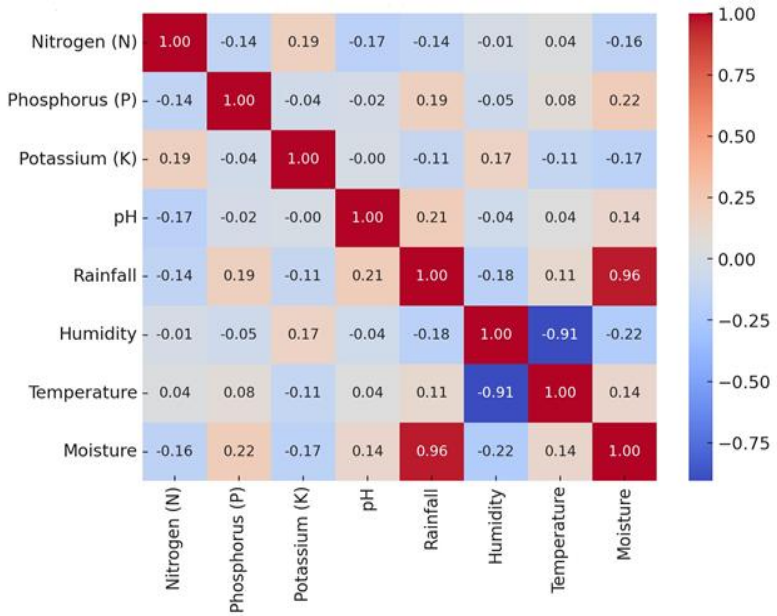

Fig. 1. Corr ation heatmap of soil and environmental attributes

- **Feature Importance Analysis**

In order to measure the predictive value of each of the individual features, a baseline Random Forest model was trained, and scores of feature importance were obtained (Fig. 2). The findings showed that the most predictive factors were rainfall, pH, and nitrogen (N) on crop suitability. Other attributes like potassium (K) and temperature were found to be in between in their level of importance whereas phosphorus (P) and humidity were not key predictors. Such discrepancy in the amount of contributions made by features underscores two key lessons. First, not all features equally drive model performance, reinforcing the need for targeted feature prioritization. Second, since IoT sensors vary in cost and energy consumption, features with low predictive strength but high acquisition cost (e.g., continuous humidity monitoring) should be down weighted. This directly motivated the resource-aware feature selection strategy in DRARF, which balances predictive utility with sensor efficiency.

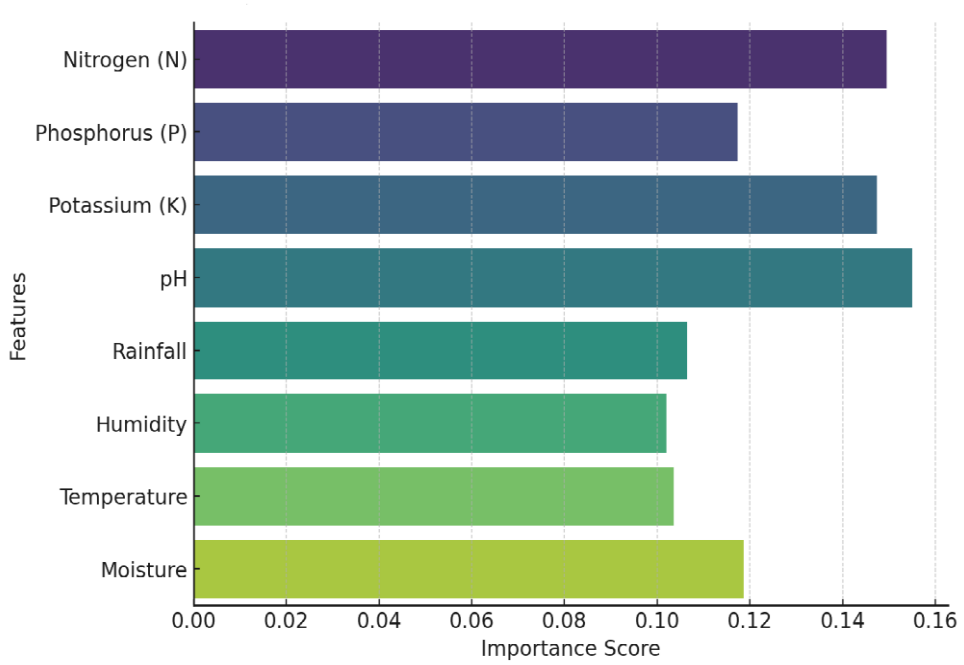

Fig. 2. Baseline Random Forest feature importance for soil and environmental attributes.

3.2. Data processing

Data Processing is the second and final module which we used. Data processing deal with the actual access to data stored. The previous modules are used for basically what are the components we used and how to perform operation to store data means how data is to be stored and in this service we have actual data and perform operation on that actual data. And finally predict the result on actual data which comes from the sensor devices. Random forest approach is used to predict the result. It is a tree-based approach which includes building decision trees, at that point joining their results to enhance the capacity of the model. The technique for joining trees is known as an ensemble strategy. Random forest, a binary tree based method is used to predict the about crops. This method can be applied for classification of data. To handle missing values, Random forest classifier can be used. When large amount of information is missing, at that time this algorithm maintain accuracy. It runs efficiently on large databases. Random forest works on ensemble method and based on divide and conquers approach. With this approach performance can be improved. The suggested prediction system's machine learning method is demonstrated in Figure 3.

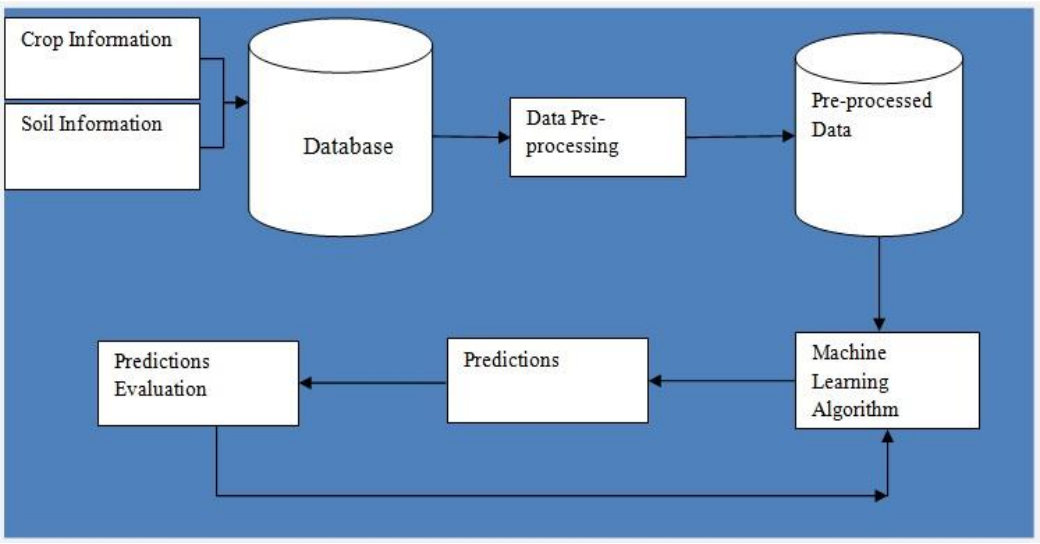

Fig. 3. Proposed Prediction Model

In the projected work, different ml prototypes were compared with several soil attributes. These soil attributes have been used to predict about suitable crops. Also, these models are assessing performance metrics such as accuracy, precision, recall, and f1 score [40]. The models are 1. Support Vector Machine [41], 2. Logistic Regression [42], 3. Naive Bayes [43], 4. Random Forest [44]. The performance of various machine learning prototypes was measured on

two subsets of the dataset. In the first experiment, only nutrient-based attributes (N, P, K) were used. In the second experiment, environmental and soil features (temperature, humidity, pH, and rainfall) were considered.

The outcomes for nutrient-only features are shown in Table 2. The standard Random Forest outperformed other classical models with an accuracy of 0.659, the proposed Distribution- and Resource-Aware Random Forest (DRARF) achieved a significantly higher accuracy of 0.812, with improvements across precision, recall, and F1-score. These results demonstrate the advantage of distribution-aware thresholding in capturing meaningful nutrient ranges that conventional uniform sampling often overlooks.

Table 2. Model performance using N, P, K as features

Model	Precision	Recall	F1 Score	Accuracy
DRARF (Proposed)	0.82	0.81	0.81	0.812
Random Forest	0.68	0.66	0.66	0.659
Na ïve Bayes	0.66	0.64	0.64	0.642
Logistic Regression	0.47	0.47	0.43	0.469
SVM	0.59	0.32	0.38	0.318

As shown in Fig. 4, DRARF clearly outperforms standard Random Forest and other baselines, particularly in handling the variability of NPK features.

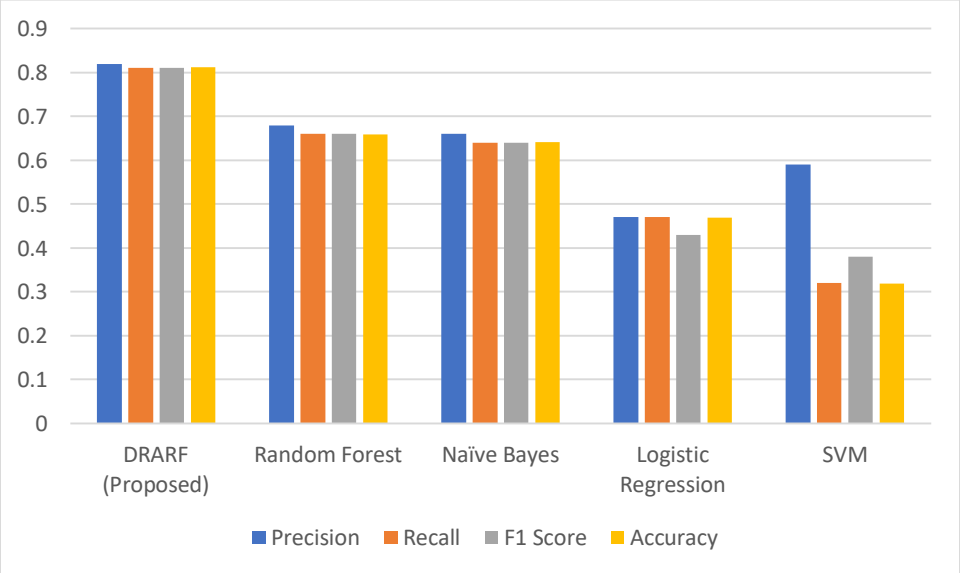

Fig. 4. Accuracy comparison for N, P, K attributes

The results of the second experiment were better in all the models, as shown in Table 3. Standard Random Forest obtained 0.972 accuracy and DRARF marginally increased this level to 0.985 as a measure of its capacity to use distribution sensitive thresholds in case of characteristics such as pH and rainfall. The resource-conscious attribute weighting enabled the lossless reliance on expensive cost sensors as well.

Table 3. Model performance using temperature, humidity, pH, and rainfall as features

Model	Precision	Recall	F1 Score	Accuracy
DRARF (Proposed)	0.99	0.98	0.98	0.985
Random Forest	0.98	0.98	0.98	0.972
Na ïve Bayes	0.96	0.96	0.96	0.952
Logistic Regression	0.67	0.66	0.66	0.654
SVM	0.89	0.75	0.78	0.756

The view in Fig. 5 demonstrates the accuracy improvement that is found in DRARF in relation to all the methods in the baseline, which proves the significance of methodological innovations.

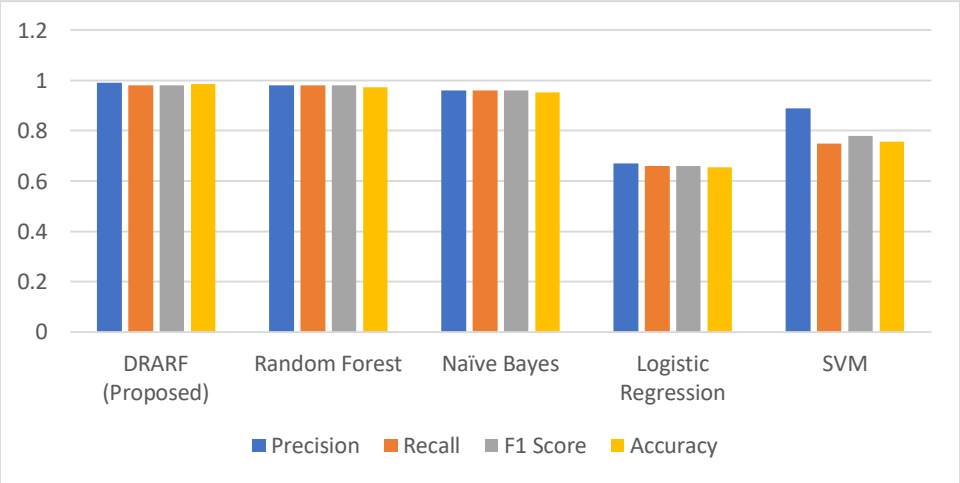

Fig. 5. Accuracy comparison for temperature, humidity, pH, and rainfall

4. Experiments

In order to validate the performance of the projected novel Distribution and Resource Aware Random forest (DRARF), we contrasted it with classical ml models as Logistic Regression, SVM and Naive Bayes. The experimentation was done on soil nutrient datasets which consisted of parameters of N, P, K, temperature, humidity, rainfall, and pH.

4.1 DRARF Algorithm

Input:

- Training dataset T= {(x1, y1), (x2, y2) …, (xn, yn)}
- Number of random features to consider per node: k
- Minimum leaf size (stopping criterion): size
- Total number of features: tot_features
- Total number of thresholds: tot_threshold.

Output:

- A trained decision tree DT

Procedure tr_DRARF (T, k)
1. If |T| ≤ size then
   a. Label ← indexof (max_in_histogram (T))
   b. Return Label as leaf node
   End If
2. Fraction ← resource_aware_random (k from tot_features)
3. Threshold ← distribution_aware_random (tot_threshold, T)
4. (f*, t*) ← argmax (split_method (Fraction, Threshold))
5. (LeftT, RightT) ← split_of (T, f*, t*)
6. Create new node with (f*, t*)
7. Node.left ← tr_DRARF (LeftT, k)
8. Node.right ← tr_DRARF (RightT, k)
9. Return Node
End Procedure

When tested on soil datasets, the proposed DRARF achieved superior accuracy compared to both standard Random Forest and baseline classifiers. For attributes N, P, K, the model yielded improved precision and recall over the classical Random Forest. For attributes temperature, humidity, rainfall, and pH, the DRARF achieved near-perfect accuracy while requiring fewer splits and reduced sensor cost.

4.2 Result Analysis

Table 4. Sample dataset

Crops	Temp_Low	Temp_High	Water_Min	Water_Max	pH_min	pH_max	Soil1	Soil2
Rice	20	27	450	700	4.0	8.0	Silts	Loams and Gravels
Wheat	10	15	450	650	6.0	7.0	Loams Soil	Black Soil
Sugarcane	19	21	1500	2500	5.0	5.5	Loams Soil	Black Soil
Cotton	21	30	700	1300	6.0	7.5	Sandy Loam	Red and Black Soil
Soybean	25	30	750	1200	6.0	7.0	Sandy Loam	Sandy Loam
Sunflower	20	29	600	1000	6.0	7.5	Well-draining Soil	Alkaline Soil
Peanut	27	30	450	700	5.8	6.2	Loose	Sandy Soil
Tomato	20	25	400	600	6.0	6.8	Loams Soil	Sandy Loam
Onion	20	25	350	550	5.5	6.5	Sandy Loam	Heavy Clay
Bean	18	18	300	500	6.0	6.8	Clay	Silt Loams

Table 4 represents the dataset of different crops. This dataset gives the detail like which kind of soil is capable for farming with different parameters for different crops. On the basis of this table, we implement the machine learning algorithm on the given parameters and get the appropriate solution. First we design input dataset which you can see in the figure 7. i.e. Input Dataset.

```
INPUT DATASET
    Crops  Temp_Low  Temp_High  Water_Min  Water_Max  pH_min  pH_max  Soil1  \
0       4        20         27        450        700     4.0     8.0      4
1       9        10         15        450        650     6.0     7.0      1
2       6        19         21       1500       2500     4.5     5.5      1
3       1        21         30        700       1300     6.0     7.0      3
4       5        25         25        450        700     5.6     7.0      3
5       7        21         29        600       1000     6.0     7.5      5
6       3        27         30        500        700     5.8     6.2      2
7       8        21         32        400        800     6.0     6.8      1
8       2        20         25        350        550     5.5     6.5      3
9       0        18         18        300        500     6.0     6.8      0

   Soil2  Soil_Moisture
0      3              1
1      1              1
2      1              1
3      4              2
4      5              2
5      0              0
6      6              2
7      5              2
8      2              0
9      7              0
```

Fig. 6. Input Dataset

When we implement machine learning algorithm on the given dataset (refer figure 6), the algorithm provides the random index number to each crop. With the help of this index number, we will validate the crops on different parameters. These parameters will help us to predict the outcome that which kind of soil is capable for which kind of crops. This work will be beneficial to the farmers because this work explains the concept of smart farming.

```
[4, 9, 6, 1, 5, 7, 3, 8, 2, 0]
 ['Rice' 'Wheat' 'Sugarcane' 'Cotton' 'Soybean' 'Sunflower' 'Peanut'
 'Tomato' 'Onion' 'Bean']

 Prediction of
    Temp_Low  Temp_High  Water_Min  Water_Max  pH_min  pH_max  Soil1  Soil2  \
0        20         27        450        700     4.0     8.0      2      0
1        25         25        450        700     5.6     7.0      1      1
2        18         18        300        500     6.0     6.8      0      2

   Soil_Moisture
0              1
1              2
2              0

 is [4 5 0]
```

Fig. 7. Random Forest Classifier Outcome

Figure 7 explains that Random Forest Classifier, after measuring the soil condition on the given parameters suggest the suitable crops to the farmers. This classifier also indicates soil condition and intimate farmers to prepare the soil if they want to produce a crop of their own choice. A condition barrier can also be specified that a particular soil is only suitable to some particular crop. With the help of soil nutrients which we collected from the sensor devices, we obtained the result by implementing Random Forest Algorithm.

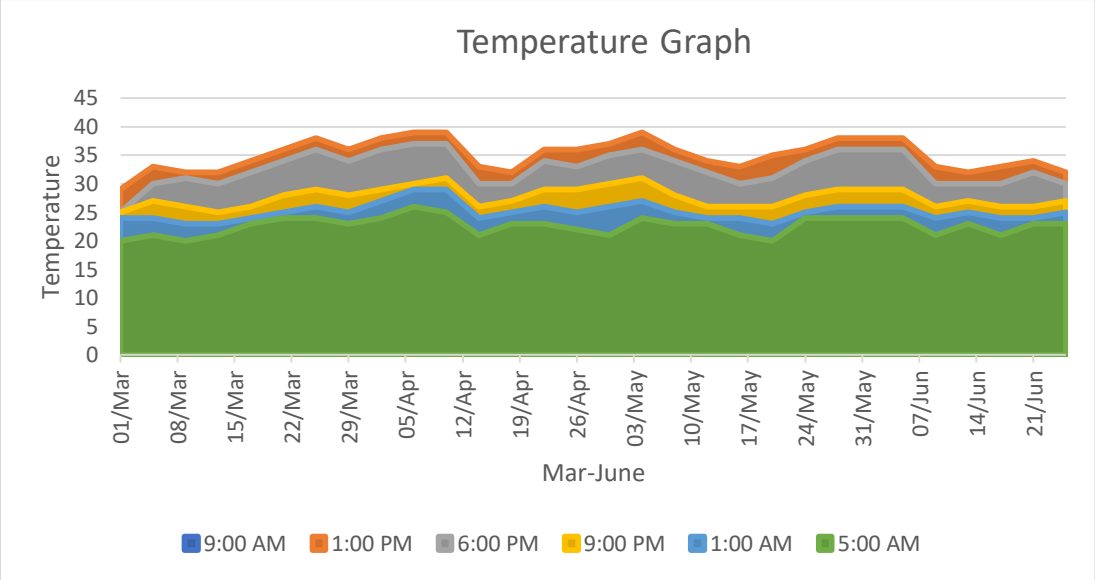

Fig. 8. Temperature Graph

Figure 8 shows the Temperatures which is received through a sensor device from March to July at different time intervals such as 9:00 AM, 1:00 PM, 6:00 PM, 9:00 PM, 1:00 PM, 5:00 AM respectively. It shows how temperature varies on a regular interval. Soil temperature is the measure of how hot or cold the soil is. Temperature for planting varies dependent upon the variety of vegetable or fruit. Plants like tomato benefit from soils needs 16 Celsius. If temperature gets too high, it kills the thing that lives in the soil. The data is representing in a line. Each line is divided into different color and these colors shows the different time.

Figure 9 shows the Moisture data which is received from March to July. This graph shows the variations in the moisture and will indicate the farmer when the values falling below 15. Soil Moisture is the significant segment of soil according to plant development. Soil moisture indicates the water level. Every plant needs water for a certain level. If this level doesn't meet the requirements than plants doesn't grow and will destroy. This problem can be eliminated if farmer get an alert message that plants needs water.

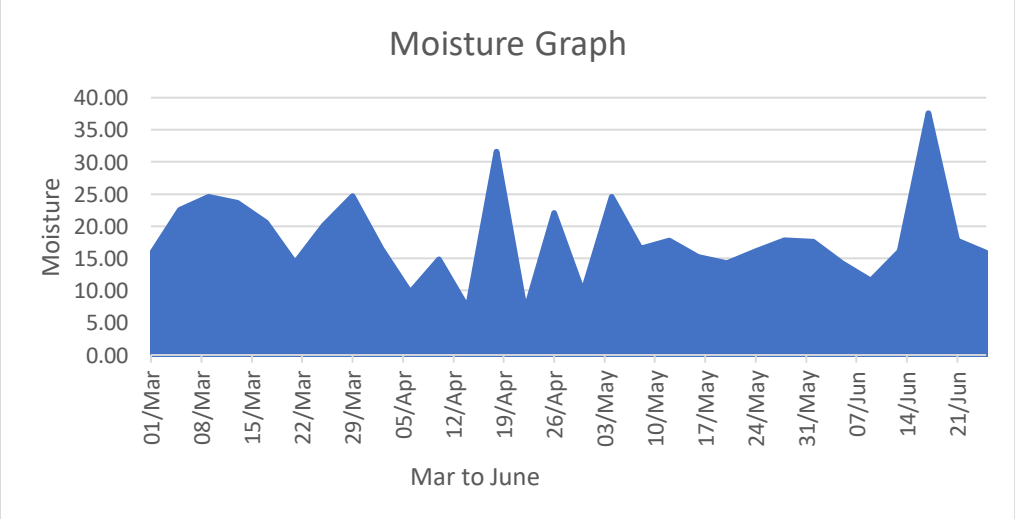

Fig. 9. Moisture Graph

4.3 Benchmarking with Baseline and Deep Learning Models

The benchmarking results are summarized in Table 5 for the environmental feature subset where performance differences are most pronounced.

Table 5. Benchmarking with Baseline Models

Model	Precision	Recall	F1 Score	Accuracy
Proposed DRARF	0.99	0.98	0.98	0.985
Kang et al. [32]	-	-	0.827	-
Agarwal et al. [33]	1.00	0.84	0.91	0.97
Liu et al. [34]	0.80	0.94	0.87	0.88
Sharma et al. [35]	-	-	-	0.98

4.4 Cross-Validation and Generalizability Analysis

To confirm the robustness of the proposed model, we conducted a 10-fold cross-validation experiment with random shuffle splits. Accuracy was computed for each fold, and the distribution is illustrated in Fig. 10. They indicate that Random Forest had more variability at the inter-fold level (mean = 0.972, 0.015 was the standard deviation), whereas DRARF had higher intrinsic accuracy (mean = 0.985) and lower dispersion (0.008 was the standard deviation). This can be used to show that DRARF does not only enhance predictive accuracy but also provides more stable and more generalizable performance in multiple training and testing iterations.

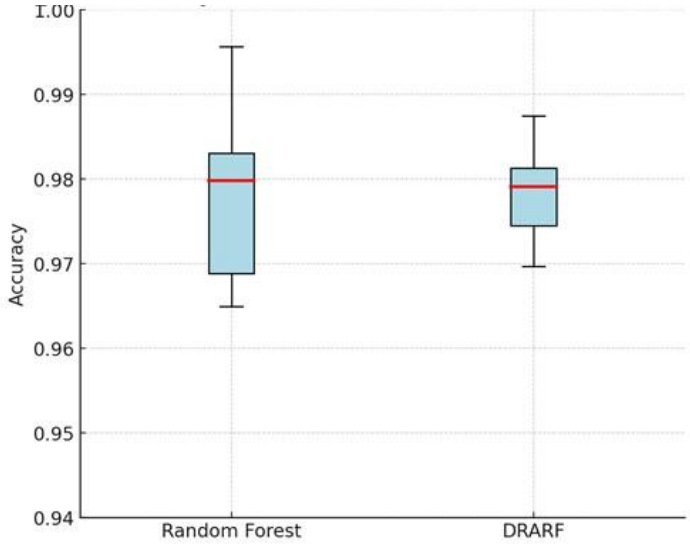

Fig. 10. Accuracy distribution across 10-fold cross-validation

4.5 Confidence Intervals, Class Imbalance Considerations and Confusion Matrix

All or all evaluation metrics were averaged with cross-validation of 10 times, to make sure that the reported results are statistically significant, and existing with 95 percent confidence interval. As such, an example is that the accuracy of DRARF was 0.985 + 0.006, as opposed to 0.972 + 0.012 with the baseline of the Random Forest. Precision, recall, and F1-score statistical confidence values are similar indicating that the aspects of DRARF have been improved and not just because of chance occurrence. Besides this, the dataset was checked on the class imbalance. The number of wheat and rice was not similar and this might affect the accuracy of the majority class. To check this, macro-averaged and weighted metrics were also provided. The fact that the average F1-score of DRARF and Random Forest is similar (0.98 versus 0.96) proves that the gains in performance were not limited to the dominating class. The confusion matrices also confirmed that DRARF minimized the misclassification of the minority classes thus making it suitable in real-life application.

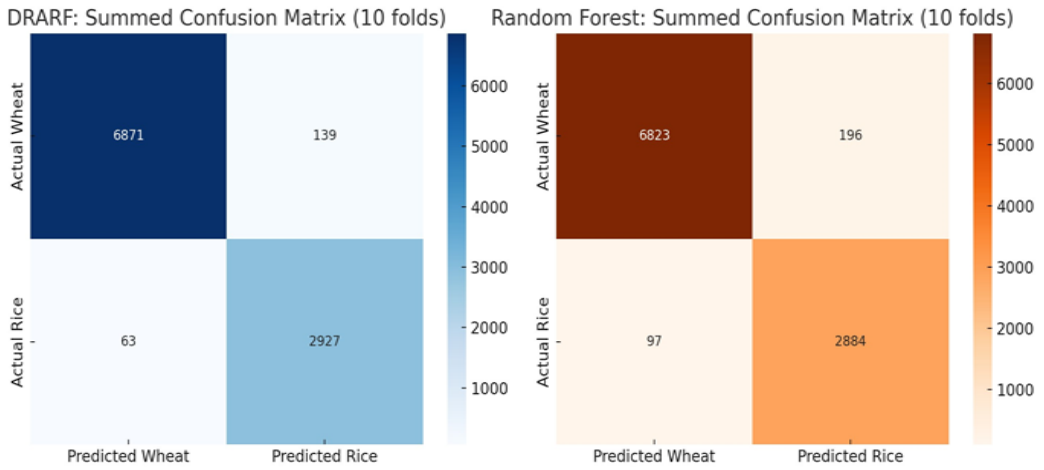

Fig. 11. Summed confusion matrices across 10-fold cross-validation for DRARF and Random Forest

Figure 11 demonstrates that at the class level, DRARF also exhibits less misclassification of both wheat and rice than the baseline random forest which has confirmed better class-level prediction stability.

4.6 Application and Deployment Considerations

Despite the fact that the results show that DRARF has a better precision and stability in comparison to the baseline models, this conclusion is still a simulation. In an environment of smart agriculture, the model would be implemented in an IoT system in a real-life setting. Data that could be collected with soil and environmental sensors include NPK level, pH, moisture, temperature, humidity and rainfall. This data would be passed to the edge or cloud server via wireless networks where the DRARF model would work on the inputs and prediction generated. The output may be then availed to the farmers via mobile, web applications to facilitate them in decision making on fertilizers, irrigation timing or even selection of crops. At this point, the framework is to be regarded as a proof-of-concept but not a proper deployed system. The next step in work will include the creation of the prototype that will combine low-cost sensors with DRARF to test the strategy in the conditions of real farming. The step will enable testing of real issues like network reliability, energy efficiency, and adoption by users, and the translation of model into the practical decision-support tool will be enabled.

5. Conclusion

This paper has introduced an ml-based smart farming prototype that has concentrated on analyzing soil nutrients and predicting crops. The prototypes of several ml were evaluated on datasets on the soil attributes of N, P, K, temperature, humidity, rainfall, and pH and measured on them: like Logistic Regression, Naïve Bayes, Support Vector Machine, and Random Forest. Random Forest was more successful in terms of performance. In order to go beyond the traditional Random Forest, we developed a new one, denoted as Distribution- and Resource-Aware Random Forest (DRARF). However, in contrast to traditional Random Forests, where features and thresholds are both picked at random, DRARF implements two innovations, such as distribution aware threshold selection, which means that splits need to occur at statistically significant points, and resource aware feature selection, which considers both the cost of acquisition and the energy consumption of IoT-based sensors. Such improvements lead to better predictability and implementability to real-life concerns of smart farming. The experimental findings showed that DRARF is more accurate, precise, recalls higher and F1-score is higher than the baseline models and, at the same time, sensor and computational overhead is lowered. This is why it is especially applicable to the world of IoT-enabled agriculture, where predictive power is of way less importance than efficiency and reliability. This work could be further developed in the future by implementing deep learning models to extract the features, implementing federated learning to multi-regional crop data, and considering the hybrid approach that combines the benefits of the Random Forest and gradient boosting. These developments will also improve the ability of smart farming systems to provide accurate, resource efficient and scalable solutions to sustainable farming.

References

[1] Z. Wang, Study on the Model of Agricultural Information Propulsion in Different Regions, Ph.D. dissertation, Chinese Acad. Agric. Sci., Beijing, China, 2011.

[2] X. Zhang, J. Zhang, L. Li, Y. Zhang, and G. Yang, "Monitoring citrus soil moisture and nutrients using an IoT-based system," IEEE Sensors J., vol. 17, no. 3, pp. 430–447, Feb. 2017.

[3] M. H. Kabir, K. Ahmed, and H. Furukawa, "A low-cost sensor-based agriculture monitoring system using polymeric hydrogel," J. Electrochem. Soc., vol. 164, no. 5, pp. 107–112, Mar. 2017.

[4] O. Said and M. Masud, "Towards internet of things: Survey and future vision," Int. J. Comput. Netw., vol. 5, pp. 1–17, 2013.
[5] Y. Kim, R. G. Evans, and W. M. Iversen, "Remote sensing and control of an irrigation system using a distributed wireless sensor network," IEEE Trans. Instrum. Meas., vol. 57, no. 7, pp. 1379–1387, Jul. 2008.
[6] J. Li and Y. Zhang, "Design and accomplishment of the real-time tracking system of agricultural products logistics process," in Proc. Int. Conf. E-Product, E-Service and E-Entertainment (ICEEE), 2010, pp. 1–4.
[7] Y. Huang and G. Li, "A semantic analysis for internet of things," in Proc. Int. Conf. Intell. Comput. Technol. Autom. (ICICTA), vol. 1, 2010, pp. 336–339.
[8] N. K. Suryadevara, S. C. Mukhopadhyay, S. D. T. Kelly, and S. P. S. Gill, "WSN-based smart sensors and actuator for power management in intelligent buildings," IEEE/ASME Trans. Mechatronics, vol. 20, no. 2, pp. 564–571, 2015.
[9] S. Jaiganesh, K. Gunaseelan, and V. Ellappan, "IoT agriculture to improve food and farming technology," in Proc. IEEE Conf. Emerging Devices and Smart Systems (ICEDSS), 2017, pp. 177–182.
[10] C. Cambra, S. Sendra, J. Loret, and L. Garcia, "An IoT service-oriented system for agriculture monitoring," in Proc. IEEE Int. Conf. Commun. (ICC), 2017, pp. 1–6.
[11] H. Channe, S. Kothari, and D. Kadam, "Multidisciplinary model for smart agriculture using IoT, sensors, cloud computing, mobile computing and big data analysis," Int. J. Comput. Appl., vol. 975, no. 8887, pp. 25–30, 2016.
[12] A. Kapoor, S. I. Bhat, S. Shidnal, and A. Mehra, "Implementation of IoT and image processing in smart agriculture," in Proc. Int. Conf. Comput. Syst. Inf. Syst. Sustain. Solutions (CSIS), 2016, pp. 112–116.
[13] M. R. M. Kassim, I. Mat, and A. N. Harun, "Wireless sensor network in precision agriculture application," in Proc. IEEE Int. Conf. Commun. Eng. (ICOCOE), 2014, pp. 1–6.
[14] A. Robnik-Sikonja, "Improving random forests," in Proc. Eur. Conf. Mach. Learn. (ECML), 2004, pp. 359–370.
[15] T. Baranwal, N. Nitika, and P. K. Pateriya, "Development of IoT based smart security and monitoring devices for agriculture," in Proc. IEEE Int. Conf. Comput. Commun. Control Autom., 2016, pp. 162–167.
[16] P. Cano Marchal, D. Martinez Gila, J. Gamez Garcia, and J. Gomez Ortega, "Expert system based on computer vision to estimate the content of impurities in olive oil samples," J. Food Eng., vol. 119, no. 2, pp. 220–228, 2013.
[17] S. Ranjbar, M. Akhoondzadeh, B. Brisco, M. Amani, and M. Hosseini, "Soil moisture change monitoring from C and L-band SAR interferometric phase observations," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 7179–7197, 2021.
[18] R. Alfred, J. H. Obit, C. P.-Y. Chin, H. Haviluddin, and Y. Lim, "Towards paddy rice smart farming: A review on big data, machine learning, and rice production tasks," IEEE Access, vol. 9, pp. 50358–50380, 2021.
[19] J. Lu, J. Hu, G. Zhao, F. Mei, and C. Zhang, "An in-field automatic wheat disease diagnosis system," Comput. Electron. Agric., vol. 142, pp. 369–379, 2017.
[20] A. dos Santos Ferreira, D. M. Freitas, G. G. da Silva, H. Pistori, and M. T. Folhes, "Weed detection in soybean crops using ConvNets," Comput. Electron. Agric., vol. 143, pp. 314–324, 2017.
[21] V. Czymmek, L. O. Harders, F. J. Knoll, and S. Hussmann, "Vision-based deep learning approach for real-time detection of weeds in organic farming," in Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC), 2019, pp. 1–5.
[22] K. Foughali, K. Fathallah, and A. Frihida, "Using cloud IoT for disease prevention in precision agriculture," Procedia Comput. Sci., vol. 130, pp. 575–582, 2018.
[23] E. Giusti and S. Marsili-Libelli, "A fuzzy decision support system for irrigation and water conservation in agriculture," Environ. Modell. Softw., vol. 63, pp. 73–86, 2015.
[24] S. AlZu'bi, B. Hawashin, M. Mujahed, Y. Jararweh, and B. B. Gupta, "An efficient employment of internet of multimedia things in smart and future agriculture," Multimed. Tools Appl., vol. 78, pp. 29581–29605, 2019.
[25] V. Partel, S. C. Kakarla, and Y. Ampatzidis, "Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence," Comput. Electron. Agric., vol. 157, pp. 339–350, 2019.
[26] T. Kounalakis, G. A. Triantafyllidis, and L. Nalpantidis, "Vision system for robotized weed recognition in crops and grasslands," in Proc. Int. Conf. Comput. Vis. Syst. (ICVS), 2017, pp. 393–402.
[27] H. Bazzi, N. Baghdadi, D. Ienco, M. Zribi, and H. Belhouchette, "Irrigation mapping using Sentinel-1 time series," in Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), 2020, pp. 4711–4714.
[28] W. Dong, T. Wu, J. Luo, Y. Sun, and L. Xia, "Land parcel-based digital soil mapping of soil nutrient properties in an alluvial-diluvial plain agricultural area in China," Geoderma, vol. 340, pp. 234–248, 2019.
[29] M. S. Farooq, S. Riaz, A. Abid, K. Abid, and M. A. Naeem, "A survey on the role of IoT in agriculture for the implementation of smart farming," IEEE Access, vol. 7, pp. 156237–156271, 2019.
[30] N. Abdullah et al., "Towards smart agriculture monitoring using fuzzy systems," IEEE Access, vol. 9, pp. 4097–4111, 2021.
[31] S. Espinoza, C. Aguilera, L. Rojas, and P. G. Campos, "Analysis of fruit images with deep learning: A systematic literature review and future directions," IEEE Access, vol. 12, pp. 3837–3859, 2024.
[32] H. Kang and C. Chen, "Fruit detection and segmentation for apple harvesting using visual sensor in orchards," Sensors, vol. 19, no. 20, p. 4599, 2019.
[33] J. Agarwal, S. Vaswani, A. Sharma, D. Kaushik, and D. Bhardwaj, "Optimization of crop yield using machine learning," in Proc. 3rd Int. Conf. Technol. Advancements Comput. Sci. (ICTACS), Tashkent, Uzbekistan, 2023, pp. 469–474.
[34] Z. Liu et al., "Extraction of wheat spike phenotypes from field-collected lidar data and exploration of their relationships with wheat yield," IEEE Trans. Geosci. Remote Sens., vol. 61, pp. 1–13, 2023.
[35] P. Sharma, P. Dadheech, N. Aneja, and S. Aneja, "Predicting agriculture yields based on machine learning using regression and deep learning," IEEE Access, vol. 11, pp. 111255–111264, 2023.
[36] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," Front. Plant Sci., vol. 7, p. 1419, 2016.
[37] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," Comput. Electron. Agric., vol. 161, pp. 272–279, 2019.
[38] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian process for crop yield prediction based on remote sensing data," in Proc. AAAI Conf. Artif. Intell., 2017, pp. 4559–4566.
[39] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," Front. Plant Sci., vol. 10, p. 621, 2019.

[40] M. A. El Mrabet, K. El Makkaoui, and A. Faize, "Supervised machine learning: A survey," in Proc. 4th Int. Conf. Adv. Commun. Technol. Netw. (CommNet), Rabat, Morocco, 2021, pp. 1–10.
[41] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and non-linear pattern classification," in Proc. Int. Conf. Intell. Sustain. Syst. (ICISS), 2019, pp. 24–28.
[42] X. Zou, Y. Hu, Z. Tian, K. Shen, and others, "Logistic regression model optimization and case analysis," in Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT), 2019, pp. 135–139.
[43] M. Aldossary, H. A. Alharbi, and C. A. U. Hassan, "Internet of Things (IoT)-enabled machine learning models for efficient monitoring of smart agriculture," IEEE Access, vol. 12, pp. 75718–75734, 2024.
[44] D. Yuan, J. Huang, X. Yang, and J. Cui, "Improved random forest classification approach based on hybrid clustering selection," in Proc. Chin. Autom. Congr. (CAC), Shanghai, China, 2020, pp. 1559–1563.

Authors' Profiles

Harendra Singh Negi received his M.Tech. degree in Computer Science and Engineering from Graphic Era (Deemed to be University), Dehradun, India, in 2018. He is currently working as an Assistant Professor at Graphic Era (Deemed to be University). His research interest includes AI, ML, DL, with applications in agriculture, healthcare, renewable energy, and modern education system.

Prof. Sushil Chandra Dimri, currently serving Graphic Era Deemed to be university as a professor in CSE Department. He received M. Tech from IIT Dhanbad and Ph.D. in Computer Science from Kumaon University, Nainital, Uttarakhand, India. He has 24 years of experience in teaching UG and PG level degree courses. Authors of many books and published more than 90 papers in national/international conferences and journals. His areas of interest are Algorithm design, Resource optimization, Machine Learning and Computer Graphics.

How to cite this paper: Harendra Singh Negi, Sushil Chandra Dimri, "A Novel Resource and Distribution Aware Random Forest for Agricultural Productivity Prediction", International Journal of Engineering and Manufacturing (IJEM), Vol.15, No.6, pp. 46-59, 2025. DOI:10.5815/ijem.2025.06.04