Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 18 March 2026
Revised 24 April 2026
Accepted 25 May 2026

Natural Resources for Human Health 2026, Vol. 6 (2s): 259–276

KEYWORDS:
Precision Agriculture, M-Learning, Soil micronutrients, Imbalance Dataset, Crops. SMOTE

Data Analytics for Crop Recommendation towards Sustainable Agriculture Development

Semmalar V I[1], Dr. R. A. Roseline[2]

[1]Research Scholar PG & Research Department of Computer Applications, Government Arts College, Coimbatore-641018, India,
E-mail: – semmalar441@gmail.com
[2]Head & Associate Professor, PG & Research Department of Computer Applications, Government Arts College, Coimbatore-641018, India,
E-mail: – roselinera@yahoo.com

ABSTRACT: Traditionally, farmers used the natural signs of fertile soil, rain and temperatures and sunlight to direct the planting and harvesting. Nevertheless, traditional agronomic systems of weather prediction, seasonal patterns and soil fertility have been seriously compromised since the general pollution that commenced in the 1950s. The world today thus requires the solutions of the present day. Class imbalance, i.e., some categories of crops or soils are underrepresented, is a problem that can be found in the dataset of agriculture. Such imbalance can be reduced with the help of methods Synthetic Minority Over-sampling Technique [SMOTE], which helps to increase the number of minority classes. Results indicate that the quality of classification may be advanced in a great way due to the methods of data mining and Machine Learning [ML], especially when SMOTE is used. Modern agricultural science and technology have supplied the farmers with effective mechanisms to overcome the issue of low productivity through soil testing, crop monitoring and making prudent decisions. Finally, this study will aim at helping farmers to overcome the agricultural problem and embrace the best practices to help both the small scale farmers and the farming managers to choose the crops with the highest yield.

I INTRODUCTION

The increasing global demand for diverse agricultural products—such as plantation crops, vegetables, fruits, cereals, fiber crops, flowers, and pulses—poses a significant challenge in the face of rapidly changing environmental conditions. Natural environments often fail to provide optimal conditions for crop cultivation, adversely affecting yield quantity and quality. Key factors such as sunlight variability, pest infestation, soil degradation, plant diseases, and inconsistent growth patterns contribute to these challenges in Rosero-Montalvo, P.D., Erazo-Chamorro. (2020) [1].

Intelligent solutions that could help farmers make knowledgeable choices on crop selection and cultivation techniques are desperately needed to maintain sustainable crop output and food security. Machine Learning[ML] offers a promising approach by leveraging historical data and predictive modeling to optimize crop growth and productivity Alhnaity, B., Pearson, S., (2020). [2]. Machine Learning models can support decision-making by accurately predicting ideal environmental conditions, recommending suitable crops, and estimating potential yield, even under unfavourable soil and weather conditions in Maureira, F., Rajagopalan, K., (2022). [3]. These systems can extend the growing season, stabilize yield consistency, and enable effective resource utilization—such as water and land—thereby enhancing agricultural efficiency in Shorten, C., & Khoshgoftaar, T. M. (2019). [4].

Generating an effective M-Learning-based crop recommendation system algorithms to boost crop yield and guarantee optimal production is the goal of this project. The technique was developed to help farmers choose the best crops depending on environmental and soil factors, hence improving agricultural planning and productivity. One of the primary problems this study

aims to address is the class imbalance issue, which is commonly found in agricultural statistics. We address this by using the SMOTE, a well-liked dataset balancing method that is especially helpful for enhancing the performance of classification algorithms in unbalanced conditions. In Chawla et al., 2002; He & Garcia, 2009 [5].

Machine Learning techniques are particularly valuable in agricultural applications because of their capacity to manage datasets with multiple variables and high dimensions. They drive data by making it easier to extract useful insights from complicated, unstructured agricultural data.-informed methods for making decisions in P.V.S. Reddy, (2021) [6]

A.J.B. Suarez, B. Singh, (2022), [7], G.I. Gadotti, C.A. Ascoli, (2022) [8], B.D. Bisandu, D.D. Datiri, (2022). [9]. Moreover, Machine Learning methods have evolved to accommodate both rule-based and pattern-based learning, making them well-suited for dynamic agricultural environments where conditions and requirements frequently change.

In this study, we propose an ML-based system capable of automatically classifying crop types using a set of geometric and environmental features. The system integrates advanced methods for feature selection in order to decide the most relevant variables contributing to crop classification and yield prediction, thereby improving the model's accuracy and generalizability Raja SP, Sawicka B,(2022)[10]. This ensures a streamlined dataset, reduces computational complexity, and improves the model's predictions' interpretability.

SVM work efficiently for tasks involving regression and classification, especially when working with high-dimensional data or when there is a noticeable difference between classes. They are resistant to overfitting in certain situations because they seek to choose the best hyperplane that maximizes the margin between various classes. By using different kernel functions, SVM may handle data that is linearly and non-linearly separable.

DT are easily understood and interpretable models that classify or forecast results by constructing a tree-like structure of decisions based on features. They offer precise guidelines for making decisions and are simple to comprehend and picture. Compared to certain other algorithms, RFs provide greater robustness and typically attain higher accuracy than individual decision trees by leveraging the characteristics of several trees, particularly when dealing with huge datasets and intricate feature connections. Compared to single decision trees, they are also less likely to overfit.

Notation	Description
$\boldsymbol{x}$	Feature vector (input data point)
$\boldsymbol{w}$	Weight vector (normal to hyperplane in SVM)
$\boldsymbol{b}$	Bias term (offset of hyperplane from origin)
f(x)	Decision function in SVM: signed distance from hyperplane
Y	Predicted class label (+1 or -1)
Entropy(S)	Measure of impurity in dataset S
IG(S, A)	Information Gain of attribute A in dataset S
Gini(S)	Gini Index: another impurity measure for dataset S
$\mathbf{d(x_i, x_j)}$	Euclidean distance between two samples x_i and x_j
x_synthetic	Synthetic sample generated using SMOTE
Λ	Random number between 0 and 1 used in SMOTE interpolation

$y_1, y_2, \ldots, y_n$	Predictions from individual decision trees in Random Forest
Mode(y_1, y_n)	Final class prediction in Random Forest (majority vote)
(1/N) $\sum y_i$	Average prediction for regression in Random Forest
FI(A)	Feature importance of attribute A in Random Forest
N	Number of synthetic samples to generate in SMOTE
K	Number of nearest neighbors in SMOTE
T	Total number of synthetic samples required

II. LITERATURE REVIEW

The development of Artificial intelligence and Machine Learning has supplemented the environment of contemporary agriculture, especially in development of efficient systems of yield predictions and crop forecasting. Other research studies have also emerged with new ideas of maximization of crop choice, the application of various types of fertilizers and resource management to meet the growing food production and environmental sustainability in food production.

A kind of hybrid prediction model called Wrapper-PART-Grid, to assist crops recommendation systems, was proposed in this article, Garg, D., and Alam, M. (2023). [11]. They were wrapper-based feature selection algorithms (applied in this model), the Grid Search [GS] algorithm with the purpose of optimization of hyperparameters and PART algorithm. With the combination of these parts it was possible to tune and measure predictive accuracy with precision and the PART classifier could achieve a great accuracy of 99.31. This approach proved to be quite convenient in extracting features that can be utilized, and optimizing decisions within the smart farming setting, such as the case observed in Maharashtra, India.The Musanase, C., Vodacek, A., (2023), [18] presented a context-specific model of the agricultural system of Rwanda where a Crop and Fertilizer Recommendation System [CFRS] was developed based on the available information. In Gopi, S. R., and Karthikeyan, M. (2023),[12], better yield prediction models were adopted which they named the HMFO-ML model and proved to be more effective in the aspects of accuracy, balance and efficiency as compared to the conventional ML models. It is an application of a hybrid algorithm which was premised on Harris Hawks Optimization inspired HMFO algorithm, a Probabilistic Neural Network [PNN] and Extreme Learning Machine [ELM]. The optimization led to much gain in the rate and accuracy of prediction, which offered effective advice on the timely crop advisory. Comparative studies revealed the effectiveness of this method in comparison with the traditional models.

A model of crop recommendation corresponding to the Indian agriculture case was created in Gosai, D., Raval, (2021). [13]. It has put into consideration the most important environmental and soil parameters, which includes of temperature, humidity, precipitation, pH, nitrogen and potassium. The comparison of three algorithms, Decision Tree, Random Forest, and Support Vector Machine, has shown that the algorithm of the tree is highest accurate, that is why the use of the algorithm can be successfully applied to a complex agricultural environment.

The study in Hasan, M., Marjan, (2023). The aim of [14] was to forecast crop yields in Bangladesh using ensemble ML models. According to the data of the meteorological and agricultural research institutions, a novel method of composer was proposed, which is known as the Kernel Ridge Regression [KRR]. The KRR model was the one with the highest R 2 and the minimum error measures. Diebold-Mariano [DM] test demonstrated that it has a superior predictive capability further. It would also allow the system to forecast the appropriate crops to be planted in future seasons and the system would be able to see the trends as more of rice and potatoes and less wheat.

The integration of ML and the IoT and a precision agriculture architecture using sensors were discussed in Islam, M. R., Oliullah, K.,(2023) [15]. The IoTs collected real-time data of soil and surrounding and analyzed it using the Random Forest and Cat Boost algorithms to obtain recommendations on crops and fertilizers. This is a developing ML-based IoT system that showed a high potential in enhancing the field-level decision-making and the real-time monitoring of crops. It is a Jadhav, R., and Bhaladhare, P. (2022) Study. Also [16] was concerned with the issue of lightweight and efficient models in agricultural context, where SVM is suggested to crop recommendation Due to their small processing requirements. It also favored application of Geospatial Data, seasonal variation and social-economic variables such as landholding and investment potential to come up with more inclusive and understandable recommendation models.

Lastly, Mahale, Y., Khan,. (2024). To achieve accurate climate-sensitive predictions [17] suggested a great decision-support system that relied on an integration of the Random Forest and Long Short-Term Memory networks [LSTM]. This was the approach, implemented in other areas in Maharashtra, to show how traditional ML and deep learning may be integrated to handle dynamic climate factors to help farmers to maximize crop and resource management. All these papers demonstrate that general tendency of hybrid, ensemble and deep learning models is evident in precision agriculture. They also refer to a trend towards localized, data-driven recommendations, sensor integration, and live decision support Musanase, C., Vodacek, A., (2023), [18] and all these are the basics of the sustainable and intelligent farming systems

Table 1: Comparison table on Crop Recommendation

Study	Focus	Key Findings	Acc	Challenges
Reddy et al. (2019)[19]	Developing crops according to the needs of the soil will boost output.	Highlighted the importance of improving datasets for better predictions.	97.2	Dataset needs expansion with more attributes for enhanced predictions.
Parameswari et al. (2021)[20]	Rule-based algorithms for crop recommendation.	PART algorithm outperformed others in classification.	97.4	Did not explore environmental variations such as climate changes extensively.
Thilakarathne (2022)[21]	AI and IoT-enabled precision farming and crop recommendation platform.	Demonstrated a cloud-enabled ML-driven platform for precise farming decisions.	96.8	Requires more exploration of real-world application under varying conditions.
Elbasi et al. (2023)[22]	Modern agriculture uses IoT sensors and Using Machine Learningto reduce waste and increase agricultural yields.	Random Forest and Bayes Net showed highest accuracy in analyzing agricultural data.	97.3(RF), 97(BN)	Challenges in deploying ML approaches; real-time data collection and preprocessing remain key challenges.
Prity et al. (2024)[23]	ML-based personalized crop recommendation system.	Random Forest achieved highest accuracy for crop recommendations.	96.3	Dataset lacked land quality, climate change, and historical crop data.
Shams et al. (2024)[24]	Explainable AI (XAI) for crop recommendation (XAI-CROP).	XAI facilitates data-driven decisions, increasing crop yields and improving food security.	95.1	Needs further research for scalability, accuracy, and regional adaptation.

Natural Resources for Human Health
eISSN: 2583-1194

III METHODOLOGY

Architecture (ACRC)

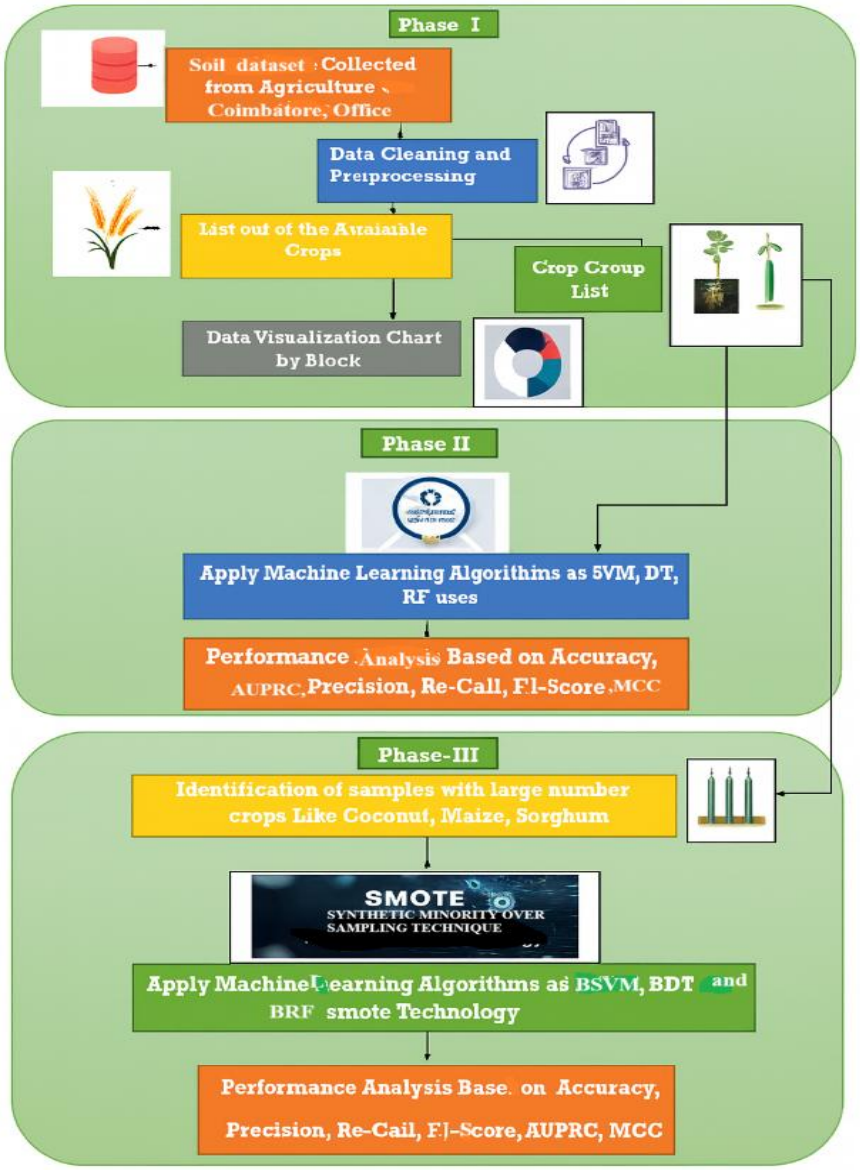

Figure 1: Architecture (ACRC)

The process starts with Dataset Collection, in which important agricultural information of the soil parameters, and crop specifics are gathered. Followed by Dataset Preprocessing where the data cleansing is to be executed and then structured accordingly to be analyzed, then Feature Selection which indicates the cardinal variables to be elevated to enhance the efficiency of the model.

The following flowchart in figure 1 (ACRC) demonstrates a three-stage model of Machine Learning-based crop recommendation system development using soil data that is available in the Coimbatore Agriculture Office.

Checking and pre-processing to achieve quality and uniformity, in Phase-I, the process starts with gathering the raw soil data that undergoes cleaning and pre-processing before being delivered. The crops that are available are displayed and classified into

groups according to similar features like coconuts, maize and sorghum. This stage ends with the visualization of the data at the block level, which makes the stakeholders perceive the spatial pattern of soil and crop distribution.

Phase-II is devoted to the implementation of machine-learning based algorithms, In Phase-III, the system identifies samples based on the crop groups with a high number of entries in order to evaluate its performance in predicting the crop suitability. In order to overcome class imbalance, the SMOTE technique is used and synthetic samples are produced relying on the nearest neighbours.

This improved dataset is then retrained on the Machine Learning models and their performance is again assessed on the same measures. This last step will make sure that the recommendations are precise and are fair to all types of crops making system more reliable in real-world agricultural decision-making.It implies that sensitivity-specificity tradeoff is better, particularly when data is imbalanced, and the models are more probable to be applied in real-world contexts.

Accuracy is a performance statistic that measures the frequency of true predictions of a model. It is obtained by the ratio of the correctly categorized occurrences to the total number of occasions. The criterion of precision is used to determine the accuracy of a model in making future predictions. The primary aim of it is to reduce the number of false negatives, or when the model fails to remember the positive situation.

The F1-score is a statistic that is used to balance the importance to evaluate the performance of a classification model. The Matthews Correlation Coefficient [MCC] is one of the statistics that have been used to evaluate the performance of binary [two-class] classification models. AUPRC, the Area Under the Precision-Recall Curve is one such useful statistic which can be useful in a classification model when the dataset is unbalanced.

It determines the level of classification at different threshold levels of two classes or spam and non-spam of a model.These algorithms assess the inputs and anticipate the best harvest based on the available characteristics. With phase I, II and III, the algorithm provides an Output Crop Recommendation, allowing farmers to make better crop selection.

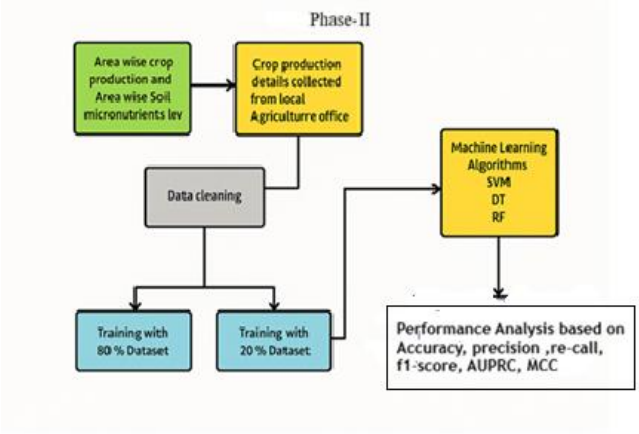

Figure 2: Phase-II Flow Chart

In this figure 2 following flow chart is phase II steps the Machine Learning Algorithm is exercised to identify the suitable crops for provided soil type under certain climatic conditions in the areas of Coimbatore. This proposal uses Machine Learningalgorithms like a SVM, DT, and RF that is designed to be more precise. This denotes the refinement of the base pre-trained model which works towards the process of deriving accurate data to support effective classification. After collecting the information, the dataset underwent several pre-processing steps, such as feature scaling, outlier, duplicate, and missing value testing, as well as SMOTE's implementation to address data imbalance. A range of Machine Learning models were used to train the dataset follows a 70:30 train-test split. Algorithms for Machine Learning are sets of instructions that, without explicit programming, enable computers to learn from data, generate predictions, and constantly improve their performance.

The preliminary phase begins with the data accumulation regarding the type of crop and soil type along with the programmed detail obtained from local governing bodies. The data collected so far will go under cleaning and preprocessing which occurs ahead of training and testing with dataset. Later the chosen Machine Learningalgorithms (SVM, DT, RF) will finalize the output considering the accurateness, precision, recall and F1-Score.

SVM tackles grouping and prediction issues with ease. It is a model or method for supervised Machine Learning. Nonetheless, it truly shines at resolving categorization issues. These are typically displayed as training information in regions that are clustered by a comprehensible gap, as large as feasible. The value of a specific location in the n-dimensional space that the support vector machine (SVM) method uses to display all of the data is a feature. Finding the hyper-plane that best divides the two groups is the next step in the classification process Rajak, R. K., Pawar, (2017). [25].

Equation $f(x) = w^T x + b$ ---------- [1]

The decision formula $f(x)$ represents the signed distance of x from the hyperplane. The hyperplane's orientation is defined through the point of the product of the weight vector (w) and the feature vector (x).

Mishra et al. [5]. In Support Vector Machines, the formula defines the decision limit. Where the sign dictates the class (+1 or -1) and the magnitude shows confidence in the classification, it computes the signed distance of a data point x from the hyperplane.

$$\hat{y} = \begin{cases} +1, if\ f\ (x) \geq 0 \\ -1, if\ f(x) < 0 \end{cases}$$ ---------- [2]

$\hat{y}$ Predicted class label Equation (+1 or -1).

$f(x)$:The signed distance between the data point x and the hyperplane is calculated using the

decision

function $f(x) = w^T x + b$

b. The orientation of the hyperplane is determined by the weight vector (w)

x: represents the feature vector (input data points).

b: The bias term determines the hyperplane's location relative to the origin.

Classifiers with decision tree often uses greedy algorithms. Using a tree to encode characteristics and class labels is a supervised learning method. To impart a model to predict the target variables' class or value, Decision Tree is mostly employed. When we feed the training data, this model will figure out how to make decisions Patil, D., Badarpura, (2020). [26]. A decision tree primarily consists of two parts: the decision nodes and the leaves. The leaves represent the outcomes. Each node in the tree represents a possible attribute test scenario, and each edge represents a possible answer to that test case. For each child tree that takes its cues from the newly added nodes, this Equation cyclical process is repeated.

$$IG(S, A) = Entropy(S) - \sum_{u \in Values(A)} \frac{|S_u|}{|S|} \cdot Entropy(S_u)$$

----------[3]

The formula calculates the decrease in entropy (uncertainty) after dividing dataset S S by attribute A. A greater IG suggests that A is a superior characteristic for data division, Equation resulting in more pure subsets.

$$Entropy(S) = -\sum_{i=1}^{c} p_i log_2\ p_i$$ ---------- [4]

The formula evaluates impurity or disorder in dataset S, where p_i is the probability of class i. Higher entropy suggests more uncertainty, Equation while lower entropy denotes a more uniform collection.

$$Gini(S) = 1 - \sum_{i=1}^{c} p_i^2$$ ---------- [5]

The formula determines the impurity of dataset S S, where p_i is the probability of class i. A $Gini$ Index closer to Equation 0 implies a purer dataset, while a number closer to 1 indicates more impurity Motamedi, B., & Villányi, B. (2024). [27].

Initially building a lot of decision trees is necessary while training Random Forest, a Machine Learning method. The result then has to be arranged based on regression [class prediction] and classification, or number of classes. The forecasts will be more accurate with the presence of more trees. Among those is the dataset are variables on output, perception, temperature, and rainfall. Training makes use of these dataset characteristics. In this work, half of the dataset is skipped. We run experiments using the remaining dataset. Two parameters, n tree and m tree, defines the minimum amounts of variables to take at each node split and the minimum amount of trees that must be generated, respectively, as a part of the random forest method. One may find the needed quantity of terminal node observations Mishra, T. K., Mishra, S. K., (2021,). [28].

$$\hat{y} = Mode(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$$ ---------- [6]

The Random Forest algorithm selects the most common class (mode) from Equation N decision tree forecasts to make the final prediction. This majority vote method improves accuracy and resilience in classification problems.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i$$ ----------[7]

The formula for Random Forest regression involves averaging the predictions ($\hat{y}, \hat{y}_i$). i based on N decision trees. Equation This reduces overfitting and improves overall prediction accuracy.

$$FI(A) = \frac{1}{N}\sum_{i=1}^{N} IG_i(A)$$ ----------[8]

This formula estimates the feature importance of Equation $(FI(A))$ for feature A in a Random Forest model. The algorithm averages the information gain (IG_i) from each of the N decision trees, indicating how much each feature helps to lowering impurity across the trees. Paithane, P. M. (2023). [29] represented comparative performance analysis of three Machine Learning models based on several key evaluation metrics. These metrics are crucial for assessing how well each model predicts crop suitability using soil data.

The Class imbalance was addressed by various SMOTE analysis. These variations frequently depend on recognizing the stress significance of concentrating the limited region among both the dominant and minority groups. In order to inform under/oversampling, SMOTE in specific places, emphasize important/safe areas in the minority class, and periodically look at the majority class in respect to the minority. In minority classes, all the areas are not created equal, may be more significant or secure than others, so it is essential to improving the original SMOTE method Temraz M, Keane MT. (2022) [30].

In order to balance the dataset and keep the model from being biased toward the majority class, it accomplishes this by producing artificial samples of the minority class. Furthermore, SMOTE is comparatively simple to implement and can enhance a model's capacity to generalize to unknown data. SMOTE reduces the bias present in imbalanced datasets, where the dominant class may have an undue influence on the model. It encourages the model to focus more on the minority class during training by producing synthetic minority class samples.

Natural Resources for Human Health
eISSN: 2583-1194

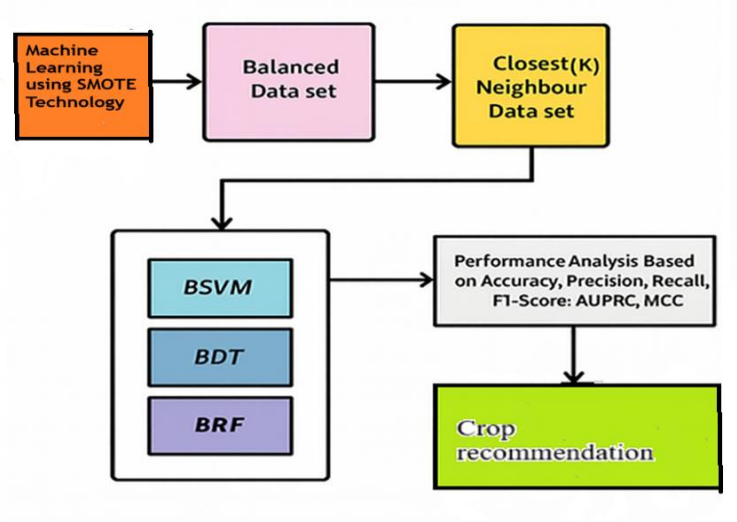

Figure 3: Phase-III Flow Chart

In this figure 3 following flowchart titled Phase-III presents a structured Machine Learning pipeline designed to enhance agricultural decision-making through soil quality analysis and crop recommendation. The process begins with the application of SMOTE, a method used to address class imbalance in agricultural datasets—such as underrepresented soil types or crop categories—by generating synthetic samples.

1.Set the amount of oversampling to use: Set (N), the desired oversampling amount, to an integer.

2. Find a random sample of the minority class: ` sample from the minority class.

3. Choose the closest neighbours: Choose the chosen sample's (K) closest neighbours.

4. Create new samples: To create new samples, choose (N) of the (K) neighbours at random for interpolation. Repeat the process until the dataset is balanced.

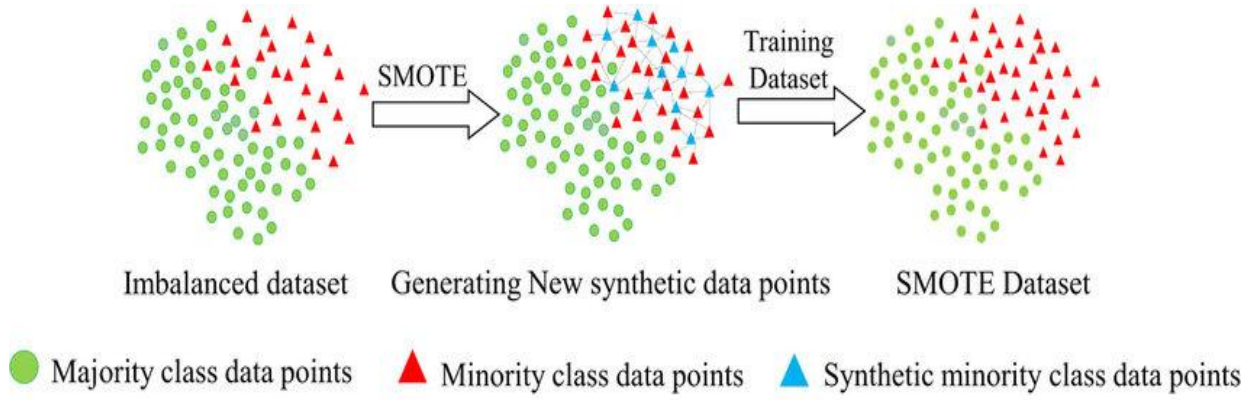

Figure 4: Imbalanced dataset

In this figure 4 Compared to models trained on extremely imbalanced data, balanced datasets produced with SMOTE have a tendency to generalize better to unseen data. This is as a result of training the model with a more representative distribution of classes. SMOTE is not limited to any particular model type and can be used with a variety of Machine Learningalgorithms.

Natural Resources for Human Health
eISSN: 2583-1194

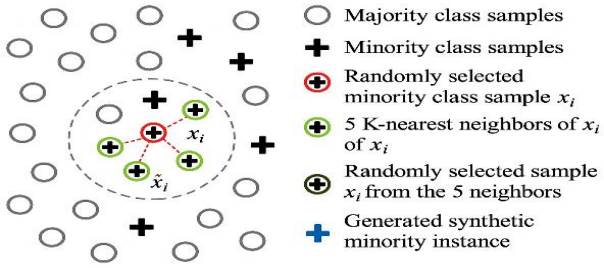

Figure 5: Find out the nearest neighbors

3.1 Implementing Machine LearningAlgorithms after SMOTE Technique

A data preparation method called SMOTE is primarily used to address class imbalance in datasets where there are not enough samples of the minority class Senapaty, M. K., Ray, A., (2024). [31]. The idea is to synthesize samples in minority class from existing minority class instances through interpolation between two neighbor data point. This unifies the class distribution, which makes classification task of Machine Learning models more effective.

In this figure 6 Before SMOTE training data key parameters such as the number of neighbors and quantification of the synthetic data to be created must be finely tuned. It works but may also introduce some noise, that is why sometimes it is used in combination with other techniques, like feature selection or ensemble methods to achieve better results.

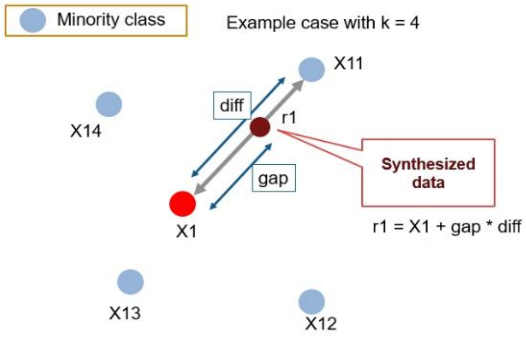

Figure 6: Classification of Imbalanced dataset

After computing samples in each class, find and fix the under-represented (minority class). Describe how many synthetic samples are necessary for approaching target ratio in the dataset or to balancing dataset.

To every minority class sample, find its k-nearest neighbors using an algorithm like Euclidean Distance.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \text{ ---------- [1]}$$

Where $x_i$and x_j are two samples, and n is the number of features. Equation [1] determines the proximity of points in the feature space.

It facilitates the process of randomly determining the k-nearest neighbors of a given minority class sample. Using the following formula, interpolate between the original sample x_i and its neighbor x_ neighbor to create a synthetic sample:

$$x_{synthetic} = x_i + \lambda \times (x_{neighbor} - x_i) \text{ ---------- [2]}$$

Here,

A random number between 0 and 1 is called λ ensuring the synthetic an example lies between $x_i$and $x_{neighbor}$. Equation [2] combines a minority class sample and its neighbor to create a new sample. The randomness λ ensures variability in synthetic

samples, preventing overfitting. Though SMOTE can enhance the accuracy of minority class to certain extent, it carries the potential risk of creating noisy instances and overfitting issues, due to the exclusion of adjacent sample distribution.

Algorithm Figure 5 and 6 illustrates the class imbalance is accommodated by the SMOTE algorithm that makes Calculating between current minority samples and the closest neighbors of each yields synthetic samples for the minority class. An augmented dataset is created for these two scenarios which removes bias in the predictive models thus enhancing classification performance for minority class. Overfitting.The rationale is that random oversampling doesn't enrich the dataset with new data. Existing data is duplicated to create the new samples.

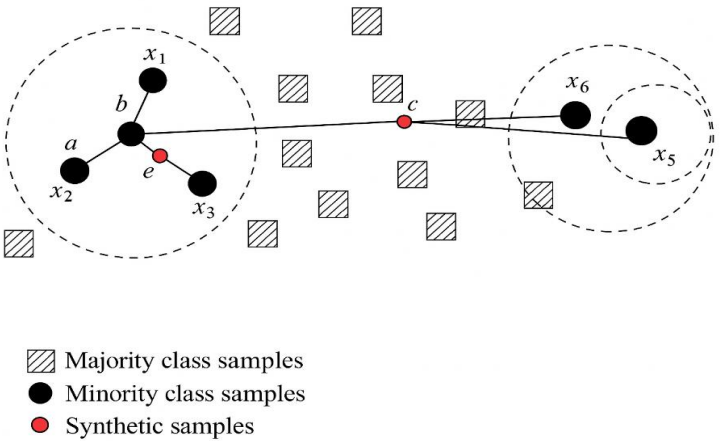

Figure 7: Generate new Dataset

In this figure 7 By oversampling the minority class, the SMOTE approach creates synthetic data for every data point in this underrepresented class. These artificial data points are created by subtracting the sample's feature vector from its closest neighbour.

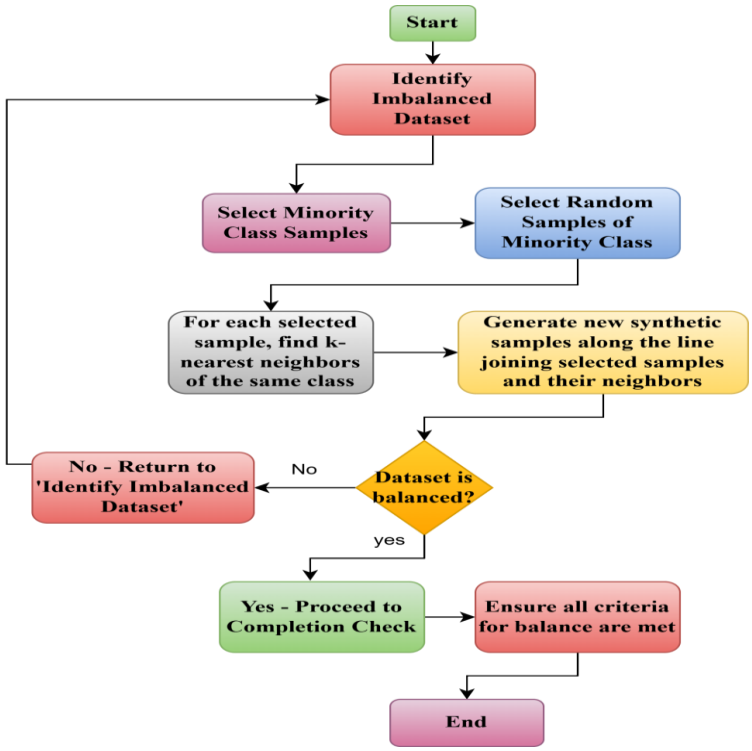

Figure 8: Flowchart of SMOTE Technique

In This figure 8 for addressing class imbalance and in this following algorithm in Machine Learning datasets. It visually guides the user through a synthetic sampling process—starting from identifying imbalance, selecting minority class samples, finding their k-nearest neighbours, and generating synthetic data points. The loop continues until the dataset achieves balance, ensuring all criteria are met before concluding.

This method is especially useful in improving model performance for classification tasks where one class is underrepresented.

The example of smote technique Algorithm is following steps:-

Algorithm 1: SMOTE

Input:

Dataset D containing N samples with m features.

k: The number of nearby neighbors.

T: Total number of collected samples required for the minority class

The Steps:

1. Identify the Minority Class

Find out how many examples there are in each class. The minority class should be given fewer samples.

2. Define the Oversampling Ratio

Determine the synthetic examples are needed to balance the classes.

3. Assign sample in the Minority Class:

Let x_i be a sampling from the minority class.

4. Determine x is k-Nearest Neighbors:

Compute the distance between x_i and every other sample in the minority class using a distance metric

5. Generate Synthetic Samples(k)

To create a synthetic sample:

Choose one neighbor at random from the k-nearest neighbors.

Make a synthetic sample

6. Repeat steps 3–5 until T Synthetic Samples Are Generated:

If required, iterate through the minority class samples several times to produce the needed amount of synthetic samples.

7. Integrate Synthetic Samples with the Original Dataset

To create the augmented dataset, incorporate the recently produced synthetic samples into the dataset.

Output:

Added synthetic samples to the augmented dataset.

3.2 Results and Analysis

The method in question has been applied using python programming The data preprocessing and modeling were carried out using Python in Jupyter Notebook (3.12.4) version, Anaconda Navigator and, applying various Machine Learning algorithms.. Machine Learning methods were evaluated with and without SMOTE, employing common performance metrics like shows the effectiveness of SMOTE improving the performance of all algorithms for Machine Learning based on the metrics. SMOTE helps models deal with imbalanced datasets, allowing models to maintain their predictive power.

In this following Table 3 SMOTE leads to the significant increases in classifiers with all the models yielding the greatest accuracy (99%) and fairly balancing the performance across all other metrics. Here, SMOTE proves to offer a solution to class imbalance, making models learn and generalize better on these complex algorithms.

Table 2: Contrasting performance metrics with and without SMOTE

metrics	Algorithm	Without SMOTE	With SMOTE
Accuracy	SVM	0.62	0.99
	Decision Tree	0.81	0.99
	RandomForest	0.77	0.99
Precision	SVM	0.66	0.100
	Decision Tree	0.98	0.100
	RandomForest	0.83	0.99
Recall	SVM	0.66	0.100
	Decision Tree	0.83	0.99
	RandomForest	0.83	0.99
F1 Score	SVM	0.66	0.100
	Decision Tree	0.90	0.98
	RandomForest	0.83	0.99
AUPRC	SVM	0.66	0.94
	Decision Tree	0.94	0.96
	RandomForest	0.90	0.95
MCC	SVM	0.60	0.94
	Decision Tree	0.84	0.93
	RandomForest	0.91	0.95

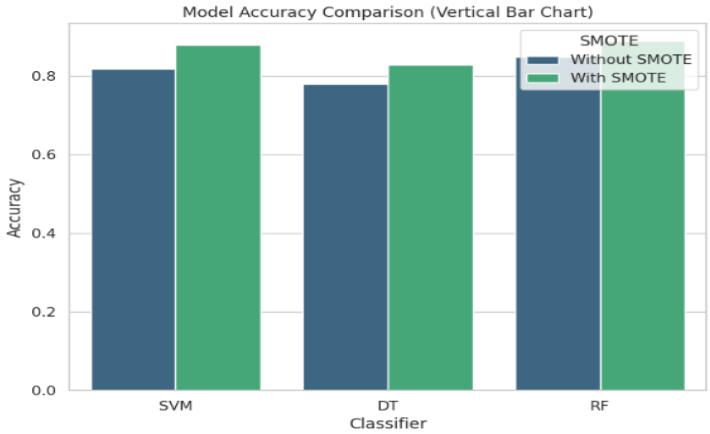

Figure 9: Accuracy Classifier

This figure 9 presents a comparison of Accuracy scores

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

for three classifiers, Without SMOTE and With SMOTE. Both SVM and RF have the same accuracy (0.95 and 0.85 respectively), whereas DT has a marginally lower accuracy (0.95 to 0.90) with SMOTE. The visualization emphasizes the impact of SMOTE on the sensitivity of the model to positive cases.

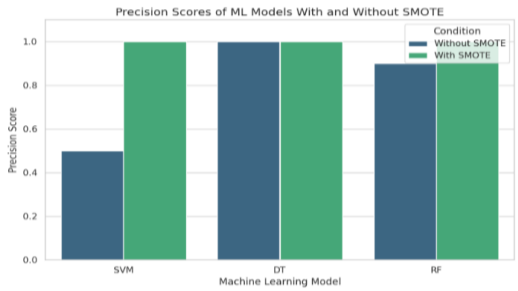

Figure 10: Precision Classifier

This figure 10 is a comparison of Without SMOTE and With SMOTE. SMOTE is of great significance in improving the accuracy of SVM (0.45 to 1.0), RF (0.75 to 0.95) and DT has a consistent accuracy of 0.95. The graphic illustrates the enhancement of the model performance with the use of SMOTE to solve the issue of the imbalance of classes.

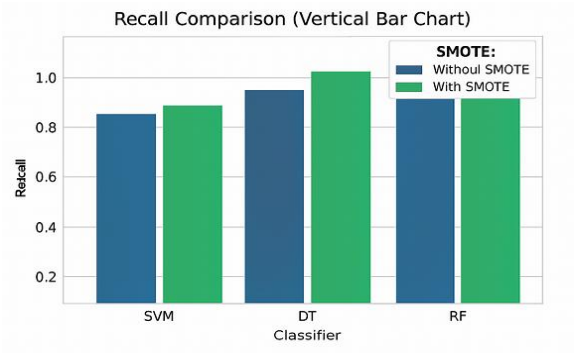

Figure 11: Recall Classifier

The recall models figure 11 varies the recall scores Recall=TP/TP+FN of models with and without SMOTE. It demonstrates that SMOTE enhances the recall of all the classifiers, with SVM gain of 0.85 to 0.90, and both DT and RF gain of 0.95 to a maximum of 1.0, indicating that SMOTE is a good model to increase sensitivity to positive cases.

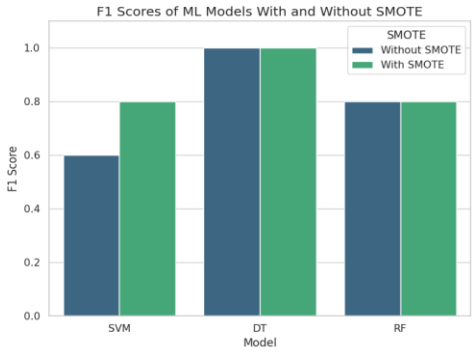

Figure 12: F1-Score Classifier

This figure 12 is the comparison , Recall/SMOTE-Precision Recall/SMOTE-Random Forest (RF) model. It demonstrates that SMOTE increases F1 values of SVM and RF and DT has a high score in both scenarios. The visualization shows the improvement of SMOTE in the balance of the model.

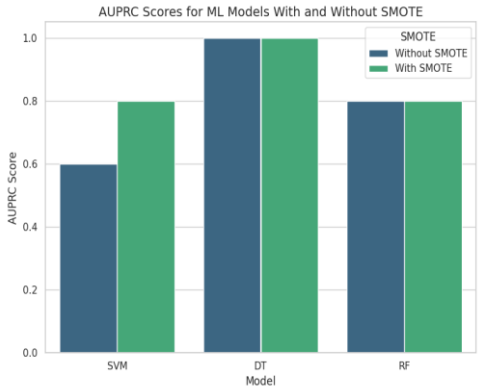

Figure 13: AUPRC Classifier

This integral represents the area under the curve formed when precision is plotted against recall. Figure 13 shows how SMOTE improves AUPRC scores for SVM (from ~0.60 to ~0.75) and Random Forest (from ~0.75 to ~0.85), enhancing their ability to detect minority class instances. Decision Tree remains stable at ~0.95 in both cases, indicating strong baseline performance regardless of sampling.

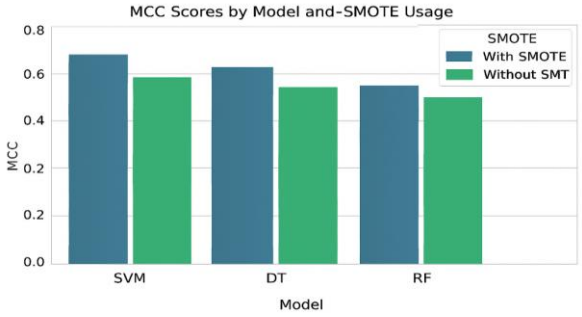

Figure 13: MCC Classifier

It improve for all models—SVM (from ~0.45 to ~0.65), Decision Tree (~0.55 to ~0.70), and Random Forest (~0.60 to ~0.75)—when SMOTE is applied. This indicates better overall classification performance and stronger handling of imbalanced data.

This suggests enhanced sensitivity-specificity tradeoff, especially in imbalanced datasets, making the models more reliable for real-world deployment.

A performance statistic called accuracy quantifies the frequency with which a model's predictions come true. The ratio of correctly categorized occurrences to all instances is used to determine it. The precision criterion assesses how well a model can predict the future. Its main goal is to decrease false negatives, or instances where the model is unable to recall a positive scenario. A statistic called the F1-score balances the significance assess how well a classification model works. It evaluates how well a model can differentiate between two classes—such as spam and non-spam—at various threshold levels.

IV.CONCLUSION

Machine Learning is an effective tool in agricultural practices in recent years, particularly crop-management and classification. This study aims to develop a classification model based on Machine Learning methods to identify three major crops, which include sorghum, maize, and coconut. It is aimed to develop an accurate and efficient model to be able to classify these crops

in respect to numerous factors. The performance of each method is measured by such common measures . as accuracy, precision, recall, F1-score, MCC, and AUPRC. We have determined that the model is able to effectively differentiate the three crops and this provides essential data on optimization of resources, crop management and precision agriculture. This research contributes to the growing body of knowledge of agricultural technology through the provision of a practical tool to enhance crop production efficiency and decision-making.

FUTURE WORK

To continue on this study, the future research could be done on improving the classification model to cover a broad range of crops, thus making it more applicable in different agricultural regions. Adding multi-modal data (satellite imagery, drone-based data collection, and IoT sensors data: soil moisture, pH, and temperature) would provide more features, which would further enhance accuracy of the classification. It would be prudent to create an interactive and easy to use decision support system or mobile application through the trained model so that the farmer has access to location specific and timely recommendations. Moreover, it would be wise to consider the explainable AI methods, such as SHAP or LIME to provide transparency in predictions, which would enhance trust and adoption. The scalability and robustness of the model in the different climates and soils might be guaranteed by validating it in cross-regional conditions and training it with the help of transfer learning. In general, all these developments would lead to precision agriculture and data-driven farming habits at a larger scale.

REFERENCE

[1] Rosero-Montalvo, P. D., Erazo-Chamorro, V. C., López-Batista, V. F., Moreno-García, M. N., & Peluffo-Ordóñez, D. H. (2020). Environment monitoring of rose crops greenhouse based on autonomous vehicles with a WSN and data analysis. *Sensors*, *20*(20), 5905.

[2] Alhnaity, B., Pearson, S., Leontidis, G., & Kollias, S. (2020). Using deep learning to predict plant growth and yield in greenhouse environments. *Acta Horticulturae*, *1296*, 425–432.

[3] Maureira, F., Rajagopalan, K., & Stöckle, C. O. (2022). Evaluating tomato production in open-field and high-tech greenhouse systems. *Journal of Cleaner Production*, *337*, 130459.

[4] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 1–48. https://doi.org/10.1186/s40537-019-0197-0

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[6] Reddy, P. V. S. (2021). Data mining and fuzzy data mining using map reduce algorithms. In Data Mining: Methods, Applications and Systems (Vol. 3, pp. 1–25).

[7] Suarez, A. J. B., Singh, B., Almukhtar, F. H., Kler, R., Vyas, S., & Kaliyaperumal, K. (2022). Identifying smart strategies for effective agriculture solution using data mining techniques. Journal of Food Quality, 2022, Article 6600049.

[8] Gadotti, G. I., Ascoli, C. A., Bernardy, R., Monteiro, R. C. M., & Pinheiro, R. M. (2022). Machine soybean seeds lots classification. Engenharia Agrícola, 42(1).

[9] Bisandu, B. D., Datiri, D. D., Onokpasa, E., Thomas, G., Haruna, M. M., Aliyu, A., & Yakubu, J. Z. (2022). Diabetes prediction using data mining techniques. International Journal of Research and Innovation in Applied Science (IJRIAS), 4(6).

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

[11] Garg, D., & Alam, M. (2023). An effective crop recommendation method using Machine Learning techniques. International Journal of Advanced Technology and Engineering Exploration, 10(102), 498.

[12] Gopi, S. R., & Karthikeyan, M. (2023). Effectiveness of crop recommendation and yield prediction using hybrid moth flame optimization with M-Learning. Engineering, Technology & Applied Science Research, 13(4), 11360–11365.

[13] Gosai, D., Raval, C., Nayak, R., Jayswal, H., & Patel, A. (2021). Crop recommendation system using M-Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(3), 558–569.

[14] Hasan, M., Marjan, M. A., Uddin, M. P., Afjal, M. I., Kardy, S., Ma, S., & Nam, Y. (2023). Ensemble M-Learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. Frontiers in Plant Science, 14, 1234555.

[15] Islam, M. R., Oliullah, K., Kabir, M. M., Alom, M., & Mridha, M. F. (2023). Machine Learning enabled IoT system for soil nutrients monitoring and crop recommendation. Journal of Agriculture and Food Research, 14, 100880.

[16] Jadhav, R., & Bhaladhare, P. (2022). A Machine Learning based crop recommendation system: A survey. Journal of Algebraic Statistics, 13(1), 426–430.

[17] Mahale, Y., Khan, N., Kulkarni, K., Wagle, S. A., Pareek, P., Kotecha, K., & Sharma, A. (2024). Crop recommendation and forecasting system for Maharashtra using Machine Learning with LSTM: A novel expectation-maximization technique. Discover Sustainability, 5(1), 134.

[18] Musanase, C., Vodacek, A., Hanyurwimfura, D., Uwitonze, A., & Kabandana, I. (2023). Data-driven analysis and M-Learning-based crop and fertilizer recommendation system for revolutionizing farming practices. Agriculture, 13(11), 2141.

[19] Reddy, D. A., Dadore, B., & Watekar, A. (2019). Crop recommendation system to maximize crop yield in Ramtek region using M-Learning. International Journal of Scientific Research in Science and Technology, 6(1), 485–489.

[20] Parameswari, P., Rajathi, N., & Harshanaa, K. J. (2021, October). Machine Learning approaches for crop recommendation. In 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) (pp. 1–5). IEEE.

[21] Thilakarathne, N. N., Bakar, M. S. A., Abas, P. E., & Yassin, H. (2022). A cloud-enabled crop recommendation platform for M-Learning-driven precision farming. Sensors, 22(16), 6299.

[22] Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., ... & Saker, L. (2023). Crop prediction model using Machine Learningalgorithms. Applied Sciences, 13(16), 9288.

[23] Prity, F. S., Hasan, M. M., Saif, S. H., Hossain, M. M., Bhuiyan, S. H., Islam, M. A., & Lavlu, M. T. H. (2024). Enhancing agricultural productivity: A Machine Learning approach to crop recommendations. Human-Centric Intelligent Systems, 1–14.

[24] Shams, M. Y., Gamel, S. A., & Talaat, F. M. (2024). Enhancing crop recommendation systems with explainable artificial intelligence: A study on agricultural decision-making. Neural Computing and Applications, 36(11), 5695–5714.

[25] Rajak, R. K., Pawar, A., Pendke, M., Shinde, P., Rathod, S., & Devare, A. (2017). Crop recommendation system to maximize crop yield using Machine Learning technique. International Research Journal of Engineering and Technology, 4(12), 950–953.

[26] Patil, D., Badarpura, S., Jain, A., & Gupta, A. A. (2020). Rainfall prediction using linear approach & neural networks and crop recommendation based on decision tree. International Journal of Engineering Research & Technology (IJERT), 9(1). ISSN: 2278-0181.

[27] Motamedi, B., & Villányi, B. (2024). A predictive analytics model with Bayesian-optimized ensemble decision trees for enhanced crop recommendation. Decision Analytics Journal, 12, 100516.

[28] Mishra, T. K., Mishra, S. K., Sai, K. J., Alekhya, B. S., & Nishith, A. R. (2021, December). Crop recommendation system using KNN and random forest considering Indian dataset. In 2021 19th OITS International Conference on Information Technology (OCIT) (pp. 308–312). IEEE.

[29] Paithane, P. M. (2023). Random forest algorithm use for crop recommendation. ITEGAM Journal of Engineering and Technology for Industrial Applications (JETIA), 9(43), 34–41.

[30] Temraz, M., & Keane, M. T. (2022). Solving the class imbalance problem using a counterfactual method for data augmentation. Machine Learning with Applications, 9, 100375.

[31] Senapaty, M. K., Ray, A., & Padhy, N. (2024). A decision support system for crop recommendation using Machine Learning classification algorithms. Agriculture, 14(8), 1256.

Semmalar V. I, the Corresponding Author, is a Research Scholar, PG, and Research in the Computer Science Department. At Bharathiar University, she completed her M.Phil. in Computer Science with a focus on data mining. About two of her papers have been published in international journals. Data mining, Machine Learning, and wireless sensor networks are some of her areas of interest. There have been two international conferences ,one patent ,and one more scopus journal, two conference where she has presented papers. The UGC-Care list has one published work.

Co-Author. Dr. R A Roseline is the head of the PG and Research Department of Computer Applications and an associate professor. She has worked in higher education for twenty-five years. At Bharathiar University, she completed her doctoral studies in computer science with a focus on wireless sensor networks. Periyar University awarded her an M.Phil. in Computer Science with a focus on Mobile Adhoc Networks. About twenty-one of her publications have been published in foreign journals. With 1.7 lacs in funding, she finished a UGC Minor Project in Pollution Monitoring Using Sensor Networks. She has given presentations at international conferences in Malaysia and Australia.