International Journal of Computing and Digital Systems
2025, VOL. 17, NO.1, 1-14

http://dx.doi.org/10.12785/ijcds/1571036228

Interpretable Crop Selection for Optimized Farming Decisions

M'hamed Mancer[1], Labib Sadek Terrissa[1] and Soheyb Ayad[1]

[1] *LINFI Laboratory, Computer Science Department, Mohamed Khider University, Biskra, Algeria*
Received 5 June 2024, Revised 1 February 2025, Accepted 2 February 2025

Abstract: The success of agriculture depends on effective crop selection, which influences yield, profitability, and risk management for farmers. Although machine learning tools are increasingly used for crop recommendations, many current models operate as opaque "black boxes", causing farmers to hesitate due to a lack of transparency. This study introduces an interpretable crop selection model that leverages the AdaBoost classifier, using soil and climate data to predict crop suitability. To ensure transparency and foster trust, we incorporate SHapley Additive Explanations (SHAP) to break down the model's decision-making process. SHAP plots visually illustrate how each input such as nitrogen, phosphorus, potassium, pH, temperature, humidity, and rainfall contributes to crop predictions. These visual aids offer farmers practical, actionable insights, helping them understand the rationale behind the system's recommendations. Our model was evaluated on a dataset of 22 crops, achieving outstanding accuracy (99.77%) with a rapid prediction time of 0.5 seconds per query. Transparency is provided not only through SHAP visualizations but also through clear, user-friendly interfaces that display feature contributions in an accessible manner. This combination of high predictive performance and easy-to-interpret explanations empowers farmers to make informed, confident decisions, leading to improved crop yields and greater profitability.

Keywords: Interpretable Crop Selection, AdaBoost Classifier, SHAP Explanations, Sustainable Agriculture, Decision Support System

1. INTRODUCTION

Agriculture is the backbone of global sustenance and economic stability, influencing everything from food security to livelihoods [1], [2]. Recent advancements in smart agriculture are reshaping traditional farming practices by incorporating technologies such as Artificial Intelligence (AI), the Internet of Things (IoT), and machine learning (ML). These technologies facilitate precise crop management, yield forecasting, and optimal resource allocation [3], [4].

Amid this technological revolution, crop selection remains a pivotal challenge. The right crop choices can increase profitability and sustainable practices, while poor selections result in financial losses and environmental risks. Though helpful, traditional methods for crop selection are often limited by inefficiency, while AI-driven systems promise to revolutionize decision-making. However, a persistent issue hinders the widespread adoption of these technologies: the lack of interpretability. Farmers are hesitant to trust systems that provide recommendations without clear explanations, regardless of their accuracy.

Interpretability refers to the ease with which a human user can understand the reasons behind a model's predictions or decisions. In agriculture, this means providing farmers with clear and comprehensible insights into why a particular crop is recommended for cultivation. Transparency, on the other hand, refers to the openness of the model's decision-making process, allowing users to see how inputs (e.g., soil and climate data) contribute to the model's final recommendation. While both concepts are related, transparency focuses on revealing the internal workings of the model, while interpretability focuses on making those workings easily understandable. Without such concepts, even highly accurate systems face barriers to real-world adoption, as farmers need more than just accurate predictions.

Several studies have advanced the field of AI-based crop selection. For instance, an Automated Crop Recommendation Model (ACRM) using Convolutional Neural Networks (CNNs) has achieved high accuracy rates for wheat (98.2%), maize (98.7%), and rice (98.1%) in Egypt [5]. The model uses climate data to provide strategic crop recommendations. Despite the ACRM success in predictive performance, CNNs are often perceived as "black boxes" due to their complexity, limiting user trust and understanding.

Similarly, a Random Forest-based crop selection system for arid regions has achieved an impressive accuracy of 99.45%, outperforming algorithms like SVM, KNN, and

E-mail address: mohamed.mancer@univ-biskra.dz, terrissa@univ-biskra.dz, s.ayad@univ-biskra.dz

Naïve Bayes [6]. While Random Forest provides some level of transparency through feature importance, it remains insufficient for explaining detailed decision pathways. Without enhanced interpretability, even highly accurate systems face barriers to real-world adoption, as farmers need more than just accurate predictions.

Some studies have explored more transparent methods. For instance, the system proposed in [7] integrates soil, weather, and profitability data using logistic regression and ARIMA for weather forecasting, providing a 94.24% accuracy in crop recommendations. Although logistic regression offers greater transparency compared to complex models, the study fell short of leveraging modern Explainable AI (XAI) techniques that could further enhance decision-making clarity, particularly in scenarios where clear justifications for recommendations are vital for adoption.

Furthermore, the integration of IoT data into a Random Forest-based model showed promising adaptability with a 99% accuracy rate for real-time crop recommendations [8]. While the integration of IoT data adds real-time adaptability, the absence of detailed interpretability in real-time systems again limits their practicality in agricultural settings. Another ensemble-based approach combining Decision Trees, KNN, and Random Forest achieved a notable 99.4% accuracy, but the complexity of the ensemble system undermines its transparency, making it difficult for farmers to trust the outputs [9].

In a related study, crop prediction using a machine learning approach combined with IoT data was investigated, focusing on a dataset of 2,200 instances covering 22 different crops [10]. This study utilized models like multilayer perceptron, JRip, and Decision Table classifiers in WEKA to predict high-yield crops, with the multilayer perceptron achieving a maximum accuracy of 98.23%. While the model's accuracy is high, the use of complex neural network structures introduces interpretability challenges, particularly for end-users such as farmers.

Moreover, another study proposed an ensemble-based crop recommendation system that utilized Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), combined through a Voting Classifier [11]. Using environmental data such as nitrogen, phosphorus, potassium, and rainfall, the model achieved a high accuracy of 99.31%. Although effective, the system lacks a focus on interpretability.

Finally, study [12] introduced a crop selection model that integrates Long Short-Term Memory (LSTM) for weather prediction and Random Forest for crop selection, applied to a dataset from Telangana, India. LSTM achieved solid performance with an RMSE of 5.02% for temperature and 8.24% for rainfall, while Random Forest showed an accuracy of 97.23% in crop selection. However, this sophisticated model also lacks interpretability, a recurring challenge in AI for agriculture.

According to these studies, high accuracy is often prioritized at the expense of interpretability, a critical factor for building user trust. Studies have demonstrated the effectiveness of algorithms such as Random Forest, Voting Classifiers, and Deep Reinforcement Learning (DRL) in delivering high predictive accuracy. For instance, the Voting Classifier achieved 99.31% accuracy in crop recommendations, and Random Forest reached 97.23%, outperforming other models. Additionally, DRL models excel in adaptability and precision, particularly in real-time decision-making contexts. This lack of interpretability has practical consequences: if farmers cannot grasp how or why decisions are made by these AI systems, they may hesitate to adopt them, despite their high accuracy.

Such a gap between predictive accuracy and model interpretability represents a key shortcoming in existing research and practice. Despite advances in AI, most models function as opaque systems that fail to offer the interpretability necessary for real-world decision-making in agriculture. The absence of clear explanations for recommendations can discourage farmers from using AI-based systems, no matter how accurate they may be.

To face such limitations, our study proposes an interpretable crop selection system that balances high accuracy with transparency. We utilize the AdaBoost classifier, renowned for its capability to combine weak learners and focus on misclassified instances, thereby ensuring robust performance across diverse environmental conditions. More importantly, we incorporate SHapley Additive Explanations (SHAP) to provide clear, feature-level explanations for crop recommendations. By detailing the contribution of features like nitrogen, phosphorus, potassium, pH, temperature, humidity, and rainfall, our system provides an understanding of decision-making processes, fostering greater trust among farmers.

By combining accuracy with interpretability, our approach addresses the dual challenges of performance and transparency, empowering farmers to make informed, data-driven decisions with confidence. This work contributes to the field by providing a comprehensive system that meets the practical needs of end users while advancing the current research landscape on explainable AI in agriculture.

The paper is structured as follows: Section 2 delves into the materials and methods employed in our study, detailing the used data and outlining the specific implementation of machine-learning techniques. Section 3 presents the results, showcasing the system's performance and key findings. As well as, it depicts a comprehensive discussion of the obtained results, exploring their implications and limitations. Finally, Section 4 concludes the paper by summarizing the key contributions and outlining future research directions.

2. Materials and Methods

Our study proposes an interpretable AdaBoost classifier-based crop selection system aimed at achieving accurate

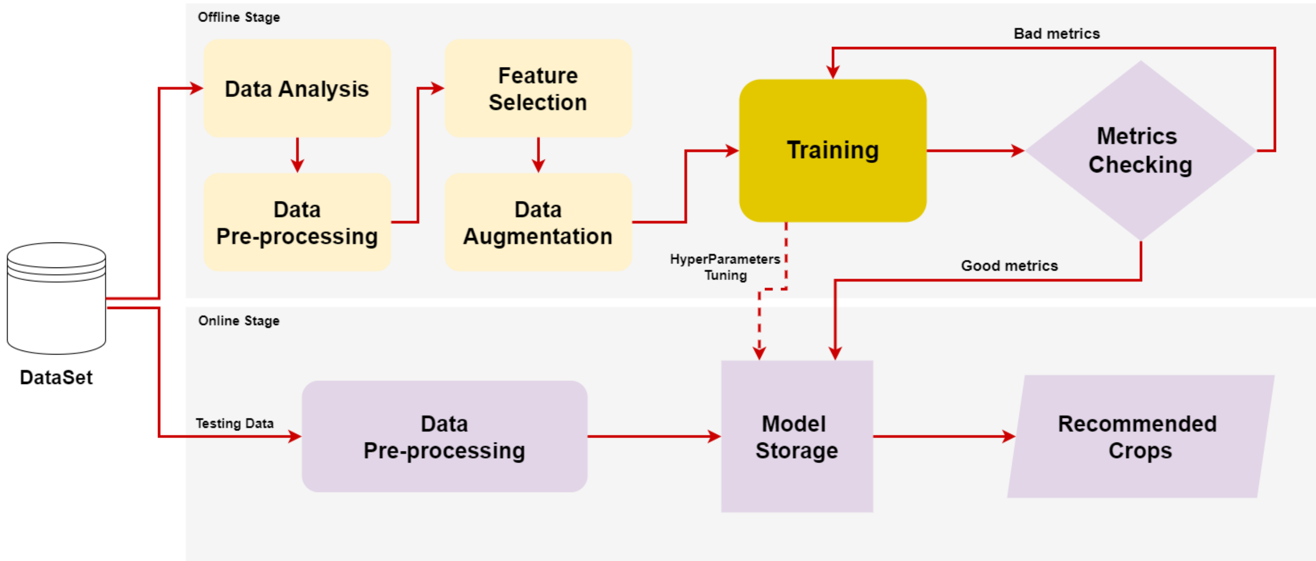

Figure 1. The general architecture of the proposed system.

selection while providing farmers with clear explanations of the decision-making process. The proposed approach, outlined in Figure 1, consists of two main stages: offline and online. The offline stage involves constructing the proposed model, which includes data preprocessing, feature selection, data augmentation, and training the AdaBoost classifier. The online stage leverages the trained AdaBoost model to provide farmers with real-time selection. Additionally, SHAP is used to analyze the trained model, identifying how specific climate and soil factors contribute to the selected crop for each prediction. Thus, farmers will be provided with clear and understandable explanations.

A. Data Sources and Exploratory Analysis

This study utilized a publicly available dataset retrieved from Kaggle [13]. The dataset consists of 2,200 observations, with 100 observations for each of the 22 crops considered in the analysis. Each observation includes critical parameters that are vital for effective crop selection, such as nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, pH, and rainfall. These parameters are key for determining crop suitability based on environmental and soil conditions.

However, it is important to note certain limitations of such a dataset. While it includes a variety of crops and environmental data, it may not fully represent all possible growing conditions. For example, the dataset does not account for other potentially important factors, such as soil type, crop diseases, or local pest pressures.

1) Univariate Analysis

Univariate analysis, as described by [14], examines the characteristics and distribution of individual variables within a dataset. By analyzing each variable separately, we gain insights into its central tendency (average value), spread (variability), and shape (distribution of values). Figure 2 presents violin plots for key variables, visually demonstrating the data distribution, including skewness and spread.

Descriptive statistics provide a summary of the data distribution, including:

- **Quantile statistics:** Minimum, maximum, and median values provide basic information about the data spread.
- **Descriptive statistics:** Skewness, kurtosis, and standard deviation offer deeper insights:
  - **Skewness:** Measures the asymmetry of a distribution, indicating whether it leans to one side (positive) or the other (negative).
  - **Kurtosis:** Describes the shape of the distribution tails, indicating if they are peaked (more extreme values), flat (fewer extreme values), or similar to a normal distribution.
  - **Standard deviation:** Measures the spread of data points around the mean, indicating how variable the data is.

Table I summarizes the quantile and descriptive statistics for each variable.

- **Median values:** Analyzing median values alongside minimum and maximum values helps understanding the central tendency and potential concentration of data points. For example, high median values close to minimum values for N, P, and K suggest a higher concentration of low values in these variables.
- **Data dispersion:** High standard deviation values for

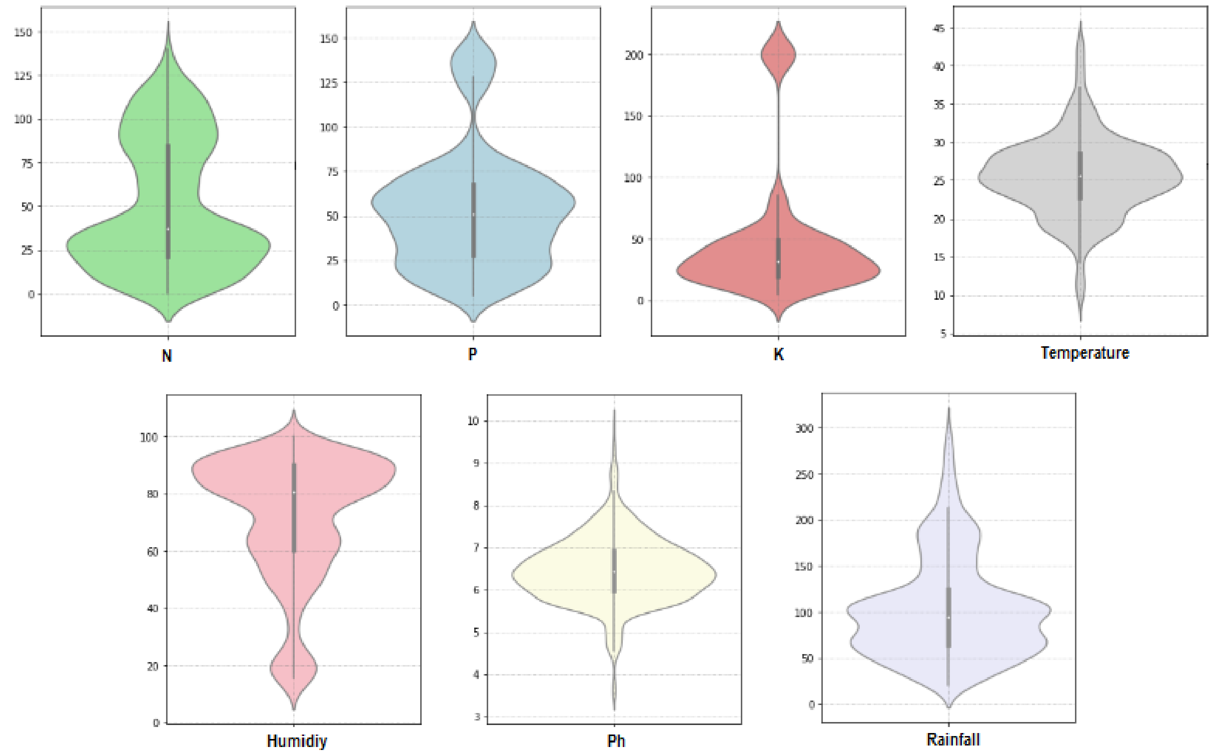

Figure 2. Violin plots illustrating the distribution of key variables in the dataset.

TABLE I. Descriptive statistics.

Features	Min	Max	Median	Standard Deviation	Mean	Skewness	Kurtosis
N (mg/kg)	0	140	37	36.9	50.55	0.5	-1.05
P (mg/kg)	5	145	51	33.05	53.36	1.01	0.85
K (mg/kg)	5	205	32	50.6	48.14	2.4	4.4
Temperature (°C)	8.8	43.7	25.6	5.06	25.61	0.18	1.2
Humidity (%)	14.3	100	80.5	22.3	71.48	-1	0.3
pH	3.5	9.94	6.43	0.774	6.46	0.3	1.6
Rainfall (mm)	20	299	95	55	103.46	0.96	0.6

N, P, K, humidity, and rainfall indicate greater data spread, while low values for temperature and pH suggest that their data points are clustered closer to the mean.

- **Data symmetry:** Positive skewness values for N, P, K, and rainfall indicate right-skewed distributions. The negative skewness for humidity indicates a left-skewed distribution. The temperature and pH have near-zero skewness, suggesting nearly symmetrical distributions.
- **Distribution shape:** Kurtosis values close to 0 indicate normal distributions, while values between 0 and 3 suggest heavy tails close to normal. A negative kurtosis (N) indicates a short tail, while high values (> 3) for K indicate a more peaked distribution.

Understanding the data distribution is crucial for identifying potential relationships and patterns within the data. For example, the normal distributions of pH and temperature suggest their values are relatively independent of other variables. Conversely, the skewed and dispersed distributions of other features might be linked to the diversity of crops and potential outliers present in the data.

2) Bivariate Analysis

Bivariate analysis, as described by [14], explores the relationships between two variables within a dataset. It evaluates their correlation, which can be positive (variables increase together), negative (one increases while the other decreases), or zero (no linear relationship). Correlation coefficients quantify the strength and direction of this relationship.

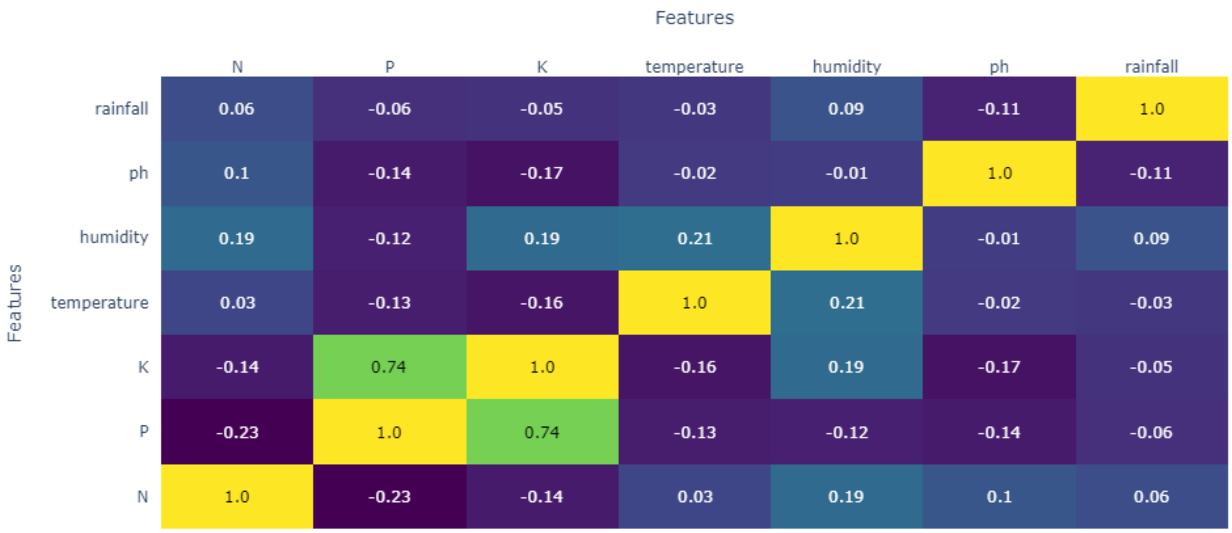

Figure 3. Correlation coefficients among variables.

Figure 3 presents a correlation matrix that visually depicts the correlation coefficients between each pair of variables. The results indicate a strong positive correlation (0.74) between phosphorus (P) and potassium (K). This suggests that higher levels of P in the soil are often accompanied by higher levels of K, and vice versa. This finding might be attributed to factors such as the application of fertilizers containing both nutrients or the natural co-occurrence of these elements in certain soil types.

Other pairs in the matrix exhibit weaker or negligible correlations, suggesting less pronounced or absent linear relationships between those variables. These findings can inform further investigations into the factors influencing crop growth and guide the development of targeted crop selection strategies.

B. *Data Preprocessing*

The initial phase of our data preprocessing involves mitigating missing data using median imputation [15]. This method replaces missing values with the median value of the corresponding feature, effectively filling the gaps in the dataset.

Next, we address outliers by employing the z-score technique [16] to identify and manage data points that significantly deviate from the norm. Outlier management techniques can involve removing outliers or transforming them to reduce their influence on the analysis.

Following outlier management, we perform numerical data normalization using a MinMax scaler [17]. This ensures that all numerical features are on a standardized scale, typically between 0 and 1. Normalization improves model convergence during training and often leads to better performance.

Finally, we address categorical data, representing different crop types. We use label encoding [18] to transform them into numerical representations. This conversion facilitates the seamless integration of categorical features into machine learning models for tasks such as prediction and classification.

C. *Data Augmentation*

While our dataset contains 2,200 observations representing 22 different crop types, each class contains only 100 data points. This can hinder the effectiveness of machine learning model training. To address this challenge, we implemented data augmentation, a technique that artificially expands the dataset size while preserving its inherent characteristics.

Our augmentation strategy focused on increasing the number of data points per class from 100 to 300. This ensures a balanced representation of each crop type within the dataset. Importantly, the augmentation process targeted individual classes to avoid introducing biases or distorting the overall data distribution. The following equation mathematically represents the augmentation process:

$$R_{\text{augmented}} = N_{\text{original}} + (R_{\text{target}} - R_{\text{original}}) \times C \qquad (1)$$

where:
$R_{\text{augmented}}$: Total number of rows in the augmented dataset (6600 rows)
N_{original}: Original number of rows in the dataset before augmentation (2200 rows)
R_{target}: Desired number of rows per class after augmentation (300 rows)
R_{original}: Number of rows per class in the original dataset before augmentation (100 rows)
C: Number of unique classes or crops (22 classes).

D. AdaBoost for Crop Selection

Crop selection tasks in agriculture often involve complex datasets with numerous features representing climate, soil characteristics, and other factors. AdaBoost, a powerful ensemble learning algorithm, has demonstrated success in handling such challenging classification tasks, making it well-suited for our purposes [19], [20].

Our AdaBoost-based model leverages climate and soil characteristics data to select suitable crops. The algorithm builds a "strong" classifier by iteratively combining multiple "weak" classifiers. Each iteration focuses on data points misclassified by previous iterations, assigning them higher weights to guide the learning process. This approach leads to a robust and accurate model for crop selection.

Let's denote the dataset as $D = (X, Y)$, where X represents the N features and Y represents the target crop labels. AdaBoost iteratively updates the weights w_i assigned to each data point (x_i, y_i) based on the model's error at each iteration t. Here, $G_t(x)$ is the weak classifier at iteration t. The final AdaBoost model is a weighted combination of these weak learners:

$$F(x) = \sum_{t=1}^{T} \alpha_t G_t(x) \tag{2}$$

where $F(x)$ is the final "strong" classifier, α_t is the contribution weight of the weak classifier $G_t(x)$, and T is the total number of iterations.

This AdaBoost-based approach offers several advantages. First, it effectively addresses high-dimensional data with potentially nonlinear relationships. This is because AdaBoost utilizes multiple weak learners, each capable of capturing different aspects of the data, ultimately leading to a more robust and flexible model. Second, AdaBoost assembles multiple weak learners into a stronger and more accurate classifier. By combining the predictions of individual learners, AdaBoost reduces the overall error rate and improves the model's ability to generalize to unseen data.

E. Implementation and Optimization

We developed the model using Python 3.7 in the Google Colab environment. To achieve optimal performance, we employed an iterative trial-and-error approach to fine-tune various training options and model parameters. The configuration of the chosen AdaBoost classifier includes the following key parameters:

- **n_estimators = 50:** The number of weak learners contributing to the final prediction, allowing for a robust ensemble.
- **base_estimator = RandomForestClassifier:** The tree-based model selected as the base learner, leveraging its strength in handling complex datasets.
- **Learning_rate = 0.001:** This parameter controls the influence of individual weak learners on the ensemble's overall output, balancing the trade-off between model complexity and performance.
- **random_state = 0:** This ensures the reproducibility of results across different runs, allowing for consistent evaluation and comparison.

To ensure the robustness of our model, the dataset was split into a training set (80%) and a test set (20%). The AdaBoost model was trained on the training data, while the test set was utilized to evaluate the model's performance. Accuracy was measured on the test set to assess the model's ability to generalize to unseen data. We employed additional metrics such as precision, recall, and F1-score to provide a comprehensive evaluation of how well the model identifies suitable crops while minimizing the selection of unsuitable ones.

F. Interpretable Crop Selection with SHAP

Understanding the factors influencing crop selection is crucial for both *interpretability* and *building trust* in the model. To achieve that, we leverage SHapley Additive ExPlanations (SHAP) [21], a powerful technique for explaining individual predictions made by complex models such as our AdaBoost classifier. SHAP helps us identify the key drivers behind each selection for a specific crop.

In the case of AdaBoost, which is an ensemble of weak learners, SHAP values are computed by analyzing the contribution of each feature across the ensemble. AdaBoost assigns different weights to each weak learner, and SHAP integrates these weighted contributions to explain the final prediction. SHAP helps us understanding how the ensemble model collectively arrives at a decision by assigning a SHAP value to each feature, representing its contribution to the predicted crop class.

SHAP values are computed by comparing the original model's prediction to predictions made on feature subsets, akin to a cooperative game where each feature "explains" a portion of the prediction (Algorithm 1).

Algorithm 1 Interpretable Crop Selection with SHAP

Require: Machine learning model f (AdaBoost), dataset X, number of classes K (22 crops)

for $k \leftarrow 1$ **to** K **do**
  $explainer_k \leftarrow$ Initialize SHAP explainer for class k
  $shap\_values_k \leftarrow$ Compute SHAP values for X and class k using Eq. (7)
  Combine the $shap\_values_k$ **with the existing SHAP values (the specific method depends on library/framework)**
end for
for $k \leftarrow 1$ **to** K **do**
  $Feature\_importance_k \leftarrow$ Compute feature importance for class k using individual class SHAP values
end for
return $feature\_importance_k$ ▷ Return interpretable feature importance for each crop

For our model f predicting one of 22 crop classes for a specific instance x, SHAP values are calculated using:

$$\mathrm{SHAP}(f, x_i, k) = \phi_k \times \sum_{S \subseteq F \setminus \{x_i\}} [f_k(x_S \cup \{x_i\}) - f_k(x_S)] \quad (3)$$

where:

- x_i is an individual feature.
- k represents the specific crop class (1 to 22).
- S is a subset of features in the model (f) excluding x_i.
- f_k denotes the model's prediction for class k.
- $f(x_S \cup \{x_i\})$ and $f(x_S)$ are the model's predictions on instances containing only features in S with and without x_i, respectively.
- ϕ_k is the normalizing factor specific to class k, calculated similarly to the single-class case:

$$\phi_k = \frac{1}{|F|!} \sum_{S \subseteq F} [f_k(x_S) - f_k(\emptyset)] \quad (4)$$

SHAP values provide insights into the influence of features on the predicted crop class. Here, how to interpret them:

- **Higher positive SHAP values:** These features push the prediction toward a specific crop class. In other words, instances with higher values for these features are more likely to be predicted as that specific crop.
- **Lower negative values:** These features push the prediction away from that class. Conversely, instances with higher values for these features are less likely to be predicted as that specific crop.
- **The magnitude of the SHAP value:** This reflects the relative importance of the feature in influencing the selection. Larger absolute values (positive or negative) indicate a stronger influence on the predicted crop class compared to features with smaller SHAP values.

By analyzing SHAP values, we obtained valuable insights into the factors driving crop selection. This allows us to:

- Understand the *rationale* behind each prediction.
- Identify *critical features* influencing crop suitability under different scenarios.
- Assess the model's *fairness* and potential biases based on feature contributions.
- Improve model *interpretability* and build trust in the selection of stakeholders.

3. Results and Discussion

This section presents the findings of our study, evaluating the proposed model's performance for crop selection. Moreover, we applied XAI methods such as SHAP to the analysis output.

A. *Evaluation of AdaBoost Performance*

We evaluated the proposed model for crop selection, focusing on both its efficiency and predictive ability. We used key metrics such as accuracy, precision, recall, and F1 score to assess how well the model could make accurate predictions.

Figure 4 depicts the AdaBoost Classifier's accuracy and error rate trends during training and testing. The error rate steadily decreases from 0.06 to 0.003, indicating efficient learning. This improvement extends to the testing error, decreased from 0.054 to 0.004, demonstrating strong generalizability to unseen data. Conversely, both training and testing accuracy increase from 0.95 and 0.945 to nearly 0.998 and 0.999, respectively, signifying effective misclassification minimization and high accuracy without overfitting.

We evaluated the effectiveness of the AdaBoost classifier for crop selection by comparing it to several other models (SVM, DT, KNN, XGBoost, LightGBM, and Bagging). Table II and Figure 5 present this comparison.

AdaBoost achieved the highest accuracy (99.77%), surpassing other models by up to 0.46%, such as Bagging (99.54%) and XGBoost (99.31%). The Precision, Recall, and F1-score metrics all achieve perfect scores of 100%, indicating AdaBoost's ability to identify positive instances while accurately minimizing false positives.

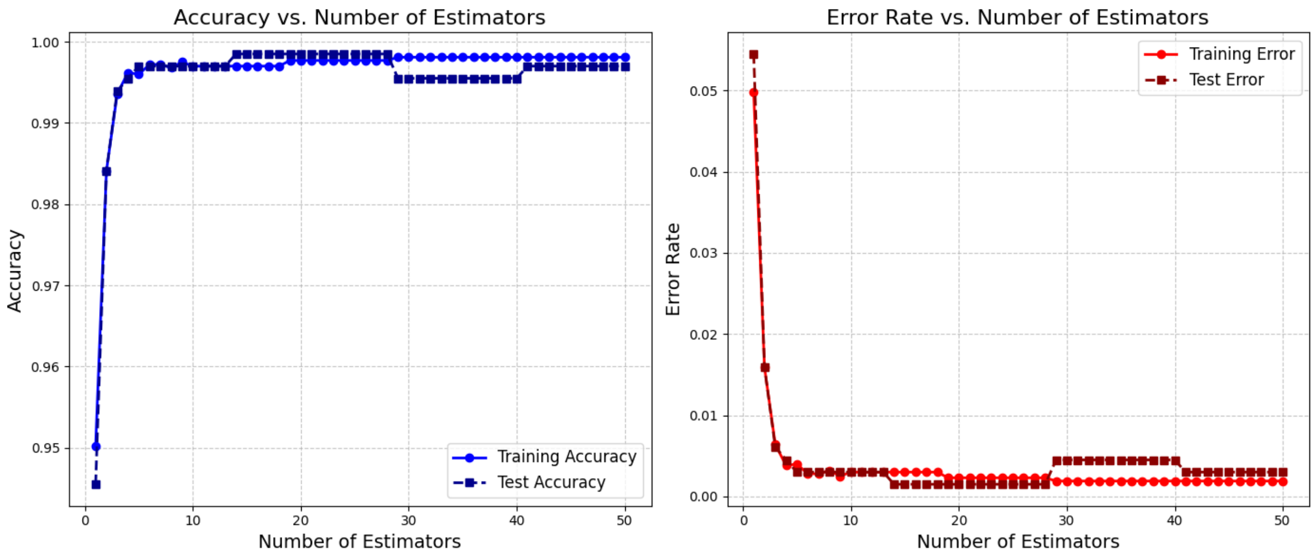

Figure 4. AdaBoost Classifier: Accuracy and Error Rate Trends.

TABLE II. Comparative Analysis of Performance Metrics Across Various Models.

Models	Correctly instances	Incorrectly instances	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Fit time (s)
SVM	1308	12	99.09	99	99	99	0.07
KNN	1299	21	98.41	99	98	98	0.003
DT	1299	21	98.41	98	98	98	0.037
Bagging	1314	6	99.54	100	100	100	9.7
XGB	1311	9	99.31	99	99	99	12.3
LGB	1305	15	98.86	99	99	99	4.5
AdaBoost	1317	3	**99.77**	100	100	100	0.57

AdaBoost had the lowest number of misclassified instances (3) compared to the other models (Table II). Moreover, AdaBoost exhibits commendable computational efficiency, boasting a fit time of 0.57 seconds (Table II), making it highly practical for crop selection applications.

The confusion matrix (Figure 6) provides an overview of the model's classification performance across all 22 crops. For the majority of crops, the model demonstrates near-perfect classification. However, there are a few notable exceptions where the model struggles slightly.

- **Rice vs. Jute:** There is a minor misclassification between these two crops, with one instance of "Rice" being misclassified as "Jute." This suggests a potential overlap in feature space or similarities in environmental factors that affect the two crops, leading the model to occasionally confuse them.
- **Blackgram vs. Mothbeans:** Another instance of misclassification is seen between "Blackgram" and "Mothbeans." This can likely be attributed to similarities in the crops' growing conditions, as they share common environmental parameters, such as soil nutrient requirements or climate preferences.

Overall, the AdaBoost classifier maintains strong performance across all crops, with minimal misclassifications. The false positive rate (FPR) is effectively 0 for most crops, indicating the model's reliability in avoiding the incorrect classification of non-suitable crops as suitable.

These results settle AdaBoost as a strong candidate for real-world crop selection, especially in Tationally limited settings, due to its exceptional accuracy and efficiency

B. SHAP Values: Interpretable Crop Selection

Understanding which features in our proposed model contribute most to its predictions is crucial. Adaboost feature importance utilizes a permutation technique to assess the impact of individual features. However, it can be susceptible to biases. When features are highly correlated (e.g., "P" and "K" with a correlation of 0.74), their importance might be overestimated or underestimated, leading to potentially misleading results. Additionally, it does not capture the direction and magnitude of a feature's influence, meaning

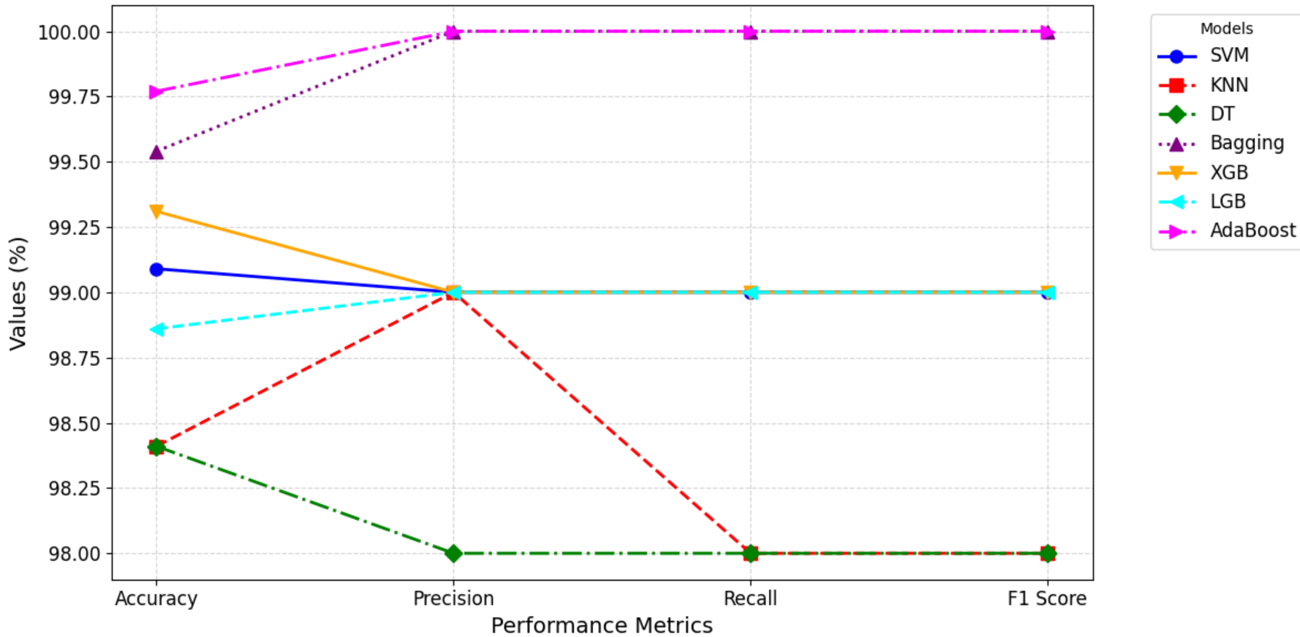

Figure 5. Comparative Analysis of Performance Metrics Across Various Models.

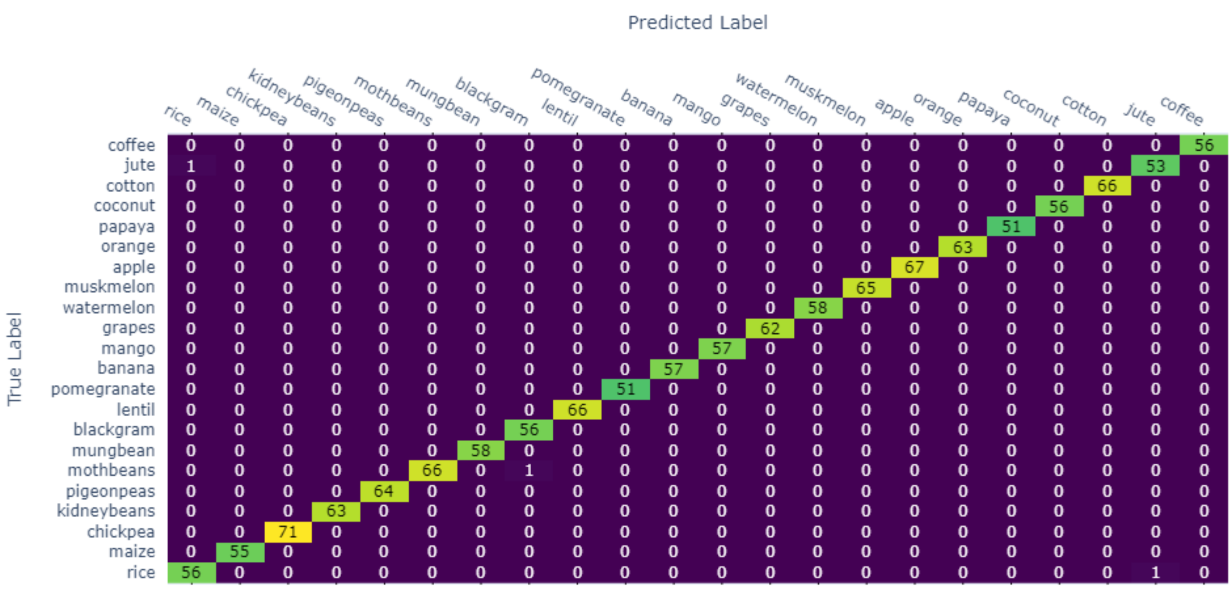

Figure 6. Confusion matrix visualization for AdaBoost classifier.

it cannot distinguish between features with positive or negative contributions.

SHAP values address these limitations by employing a game theory approach to calculate a feature's specific contribution to a prediction. This allows SHAP to:

- Account for dependencies between features, providing a more accurate picture of individual importance.
- Capture the direction and magnitude of influence, revealing whether a feature has a positive or negative impact on the prediction and its relative strength.

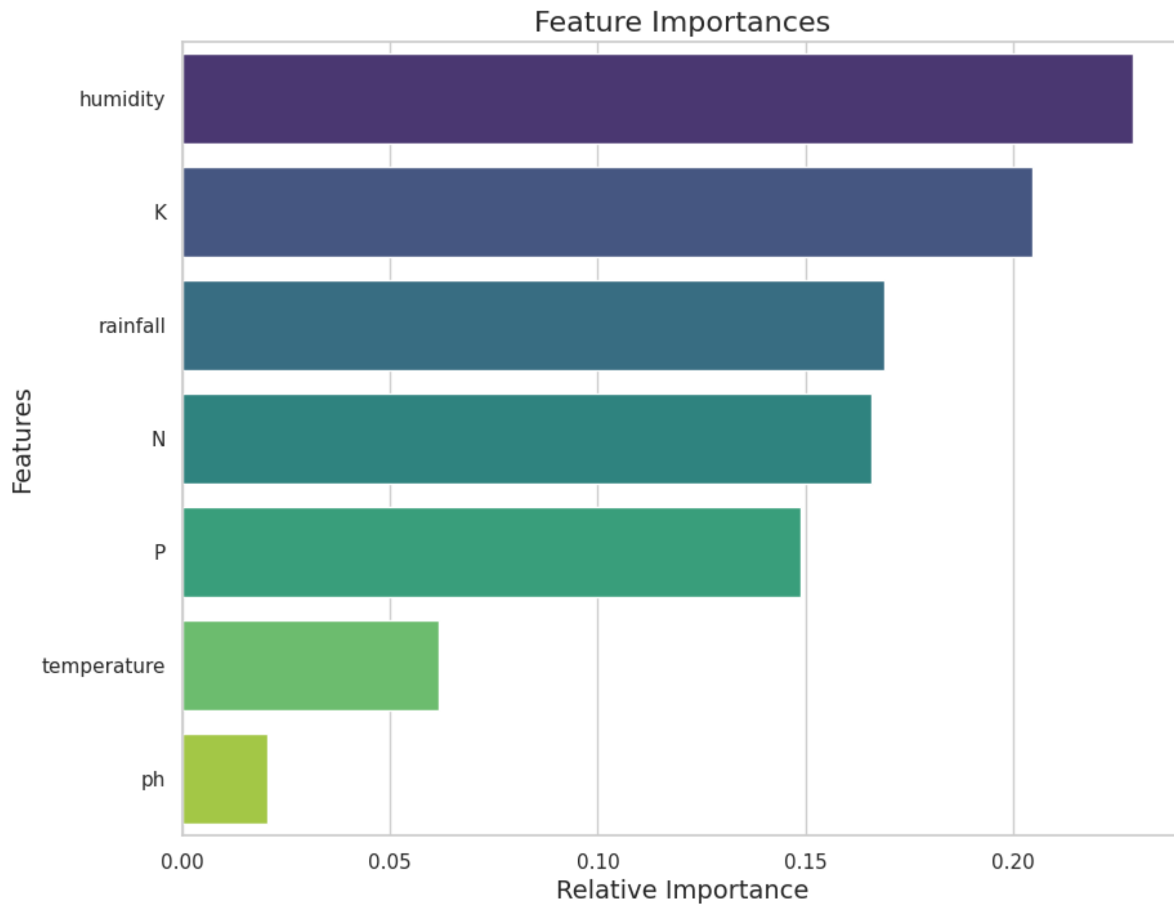

Figure 7. Feature Importance Analysis using Permutation Technique.

Figure 7 and Figure 8 visually represent the differences between the methods. We observe discrepancies in the ranking of features, highlighting the potential biases of feature importance. For example, the strong correlation between "P" and "K" might inflate their importance in the feature importance plot.

SHAP values allow us to interpret the contribution of each feature to the model's predictions. For instance, as shown in Figure 8, "humidity" emerges as the most influential feature across the dataset, followed closely by "nitrogen (N)" and "potassium (K)." These features directly affect the model's predictions for crop suitability. However, the impact of features varies significantly between different crops. For example, "rainfall" plays a crucial role in predicting the suitability of crops like rice and pigeon peas, while it has minimal impact on crops such as kidney beans. Likewise, "humidity" has a stronger influence on mungbean peas but is less significant for watermelon.

To provide a more concrete example, we conducted SHAP analysis on four selected crops: rice, maize, chickpea, and banana. Figure 9 and Figure 10 present SHAP summary plots for these crops, illustrating how the importance of features differs between them. This analysis highlights the interpretability benefits of SHAP, as it helps us understanding which factors drive the model's decisions for each crop. Through these visual examples, we demonstrate the varying influence of environmental features, reinforcing the transparency of our model's predictions.

Crop Specific Interpretations:

- **Rice:** Rainfall is the most important factor for rice selection, with a strong positive SHAP value. This translates to areas receiving more rainfall being more suitable for rice cultivation due to their water intensive nature. Conversely, low rainfall regions might be discouraged by the model due to insufficient water availability, potentially leading to poor crop growth and yield. However nitrogen also has a positive influence, it plays a less significant influence than rainfall. Adequate nitrogen levels are still crucial for rice growth, and soils lacking nitrogen might not be suitable for rice planting. Humidity exhibits a positive influence, suggesting that humid environments gener-

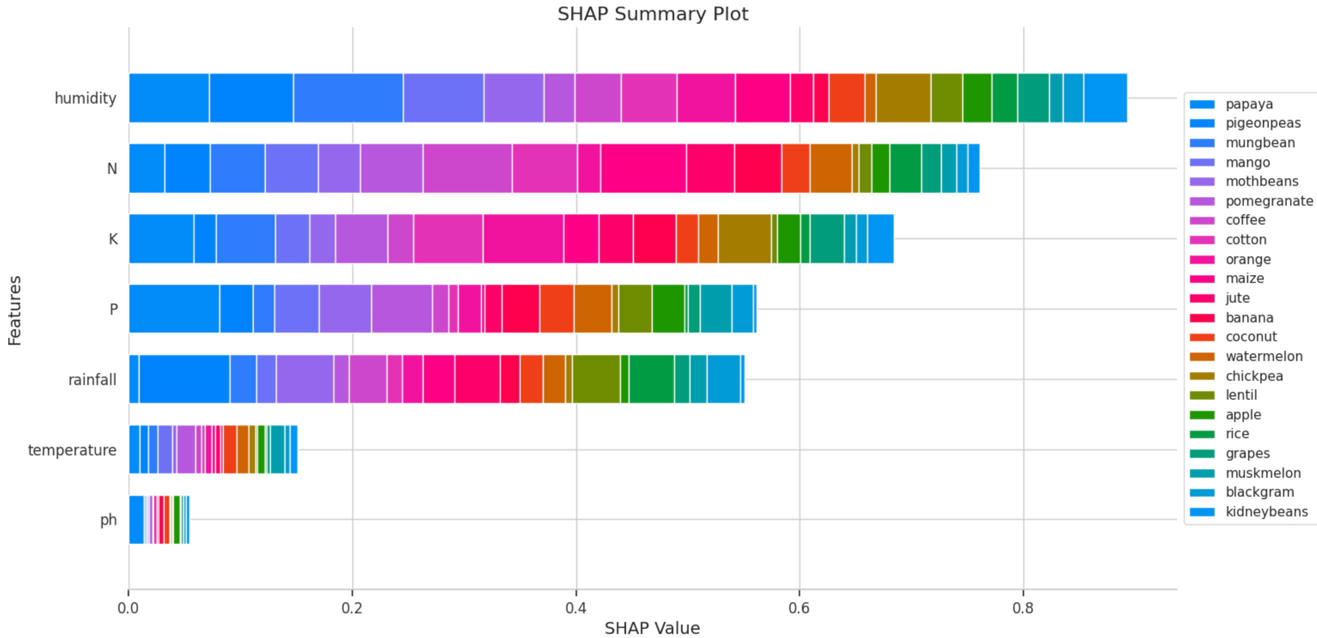

Figure 8. Feature Importance Analysis using SHAP.

ally favor rice growth. However, high humidity can become detrimental, potentially increasing the risk of disease outbreaks.

- **Maize:** Similar to rice, nitrogen plays a crucial role in maize selection, with a positive SHAP value. Low nitrogen levels could negatively impact maize yield and quality, potentially leading the model to discourage maize cultivation in such areas. While the influence of humidity is weaker than that of rice, it still exhibits a positive influence on maize selection, suggesting that maize can tolerate a wider range of humidity levels than rice. However, excessively high humidity can still be detrimental. Adequate potassium availability is also crucial for maize, as indicated by the positive SHAP value. Low potassium levels could hinder maize growth and development. Rainfall generally has a positive influence on the model selection, similar to rice. However, excessively high rainfall can also be detrimental, potentially leading to waterlogging and reduced crop yield.

- **Chickpea:** The SHAP plot reveals a positive influence of humidity on chickpea selection. This suggests that moderate humidity levels are suitable for chickpea growth. However, excessively high humidity can still be detrimental, similar to the other crops discussed. Potassium emerges as another crucial factor, with a positive SHAP value indicating the importance of adequate potassium availability for optimal chickpea growth and yield. While nitrogen, temperature, rainfall, and pH also have positive SHAP values, their influence is less significant compared to potassium. Insufficient levels or unsuitable values of these features could still negatively impact chickpea growth and yield.

- **Banana:** The SHAP plot reveals a positive influence of nitrogen on banana selection, highlighting the importance of sufficient nitrogen availability for banana growth and fruit production. However, excessively high nitrogen levels could also be detrimental, potentially leading to issues such as compromised fruit quality or increased disease susceptibility. Both potassium and phosphorus exhibit positive SHAP values, indicating that adequate levels of these nutrients are also important for banana selection. Rainfall had a slightly positive influence, suggesting that moderate rainfall is beneficial for banana cultivation. However, excessively high or low rainfall can be detrimental, potentially leading to waterlogging or drought stress, respectively.

Our approach underscores the significance of both accuracy and explainability in crop selection systems. By integrating SHAP values, we not only enhance the predictive capability but also offer transparent insights into the features that steer the model's decisions for various crops. This transparency provides farmers and agricultural professionals with a deeper understanding of the decision-making process, fostering trust and potentially catalyzing broader adoption of these AI-powered tools.

C. Discussion

The agricultural sector is increasingly turning to machine learning to make use of its analytical power. These

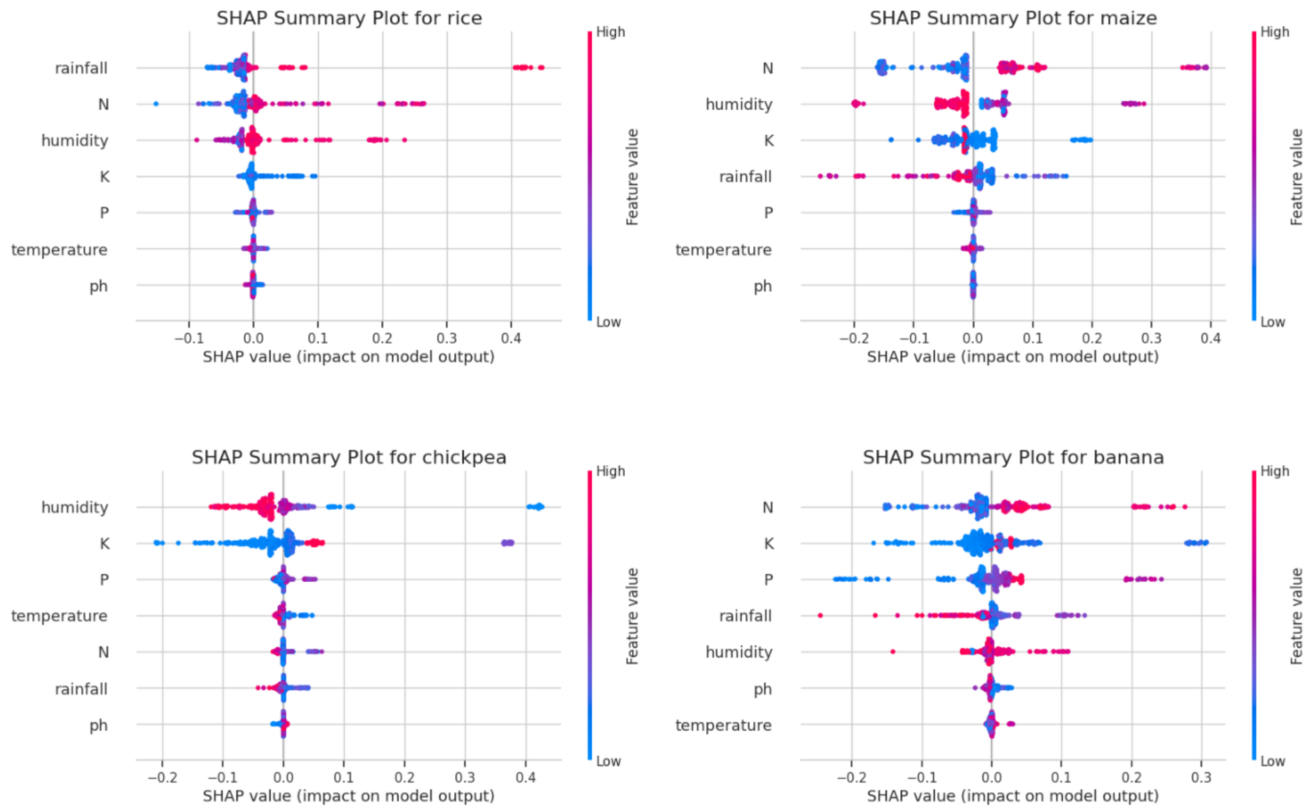

Figure 9. Feature Importance for Rice, Maize, Chickpea, and Banana Selection.

algorithms excel at processing complex datasets, uncovering insights that traditional statistical methods struggle to perceive. Our study aimed to develop a highly accurate and interpretable crop selection model, leveraging the AdaBoost algorithm to minimize false positives and optimize prediction accuracy. This dual emphasis on accuracy and interpretability sets our work apart from previous studies, offering farmers valuable insights alongside reliable crop selection.

Accurate crop selection relies heavily on understanding the intricate interplay of climate and soil characteristics. Our model was evaluated on a diverse dataset encompassing 22 crops. Rigorous data cleaning addressed missing values and outliers, followed by a crucial feature selection step. By employing correlation coefficients, we identified the most influential factors for model training, focusing our attention on the most relevant information to enhance performance.

Our AdaBoost model achieved outstanding results: 99.77% accuracy, 100% precision, recall, and F1-score. This represents a significant improvement over existing models. For example, while ACRM achieved high accuracy for specific Egyptian crops (98.7% for maize and 98.1% for rice) [5], others such as random forest (99.45%) [6] and an IoT-based framework (99%) [8] displayed lower performance. These enhancements translate to tangible benefits for farmers, with minimized false positives leading to more reliable predictions and ultimately, better decision-making. In the context of crop selection, the significance of minimizing false positives cannot be overstated, as any misclassification poses substantial risks and potential losses for farmers.

In time-sensitive agricultural scenarios, model efficiency is equally crucial. Our AdaBoost model boasts a rapid training time of 0.57 seconds. This efficiency translates to optimized resource utilization, making AdaBoost a compelling choice for real-time decision support. Faster training times pave the way for practical applications, empowering farmers with quicker and more efficient decision-making tools.

Bridging the gap between model predictions and actionable insights for farmers is essential. We utilize SHAP values, a powerful interpretability technique, to determine how climate and soil factors influence crop selection. Our analysis reveals humidity as the most influential factor, underscoring its substantial impact on model predictions. This aligns with established agricultural knowledge, as humidity significantly affects plant health, water use efficiency, and overall productivity. Understanding this key driver

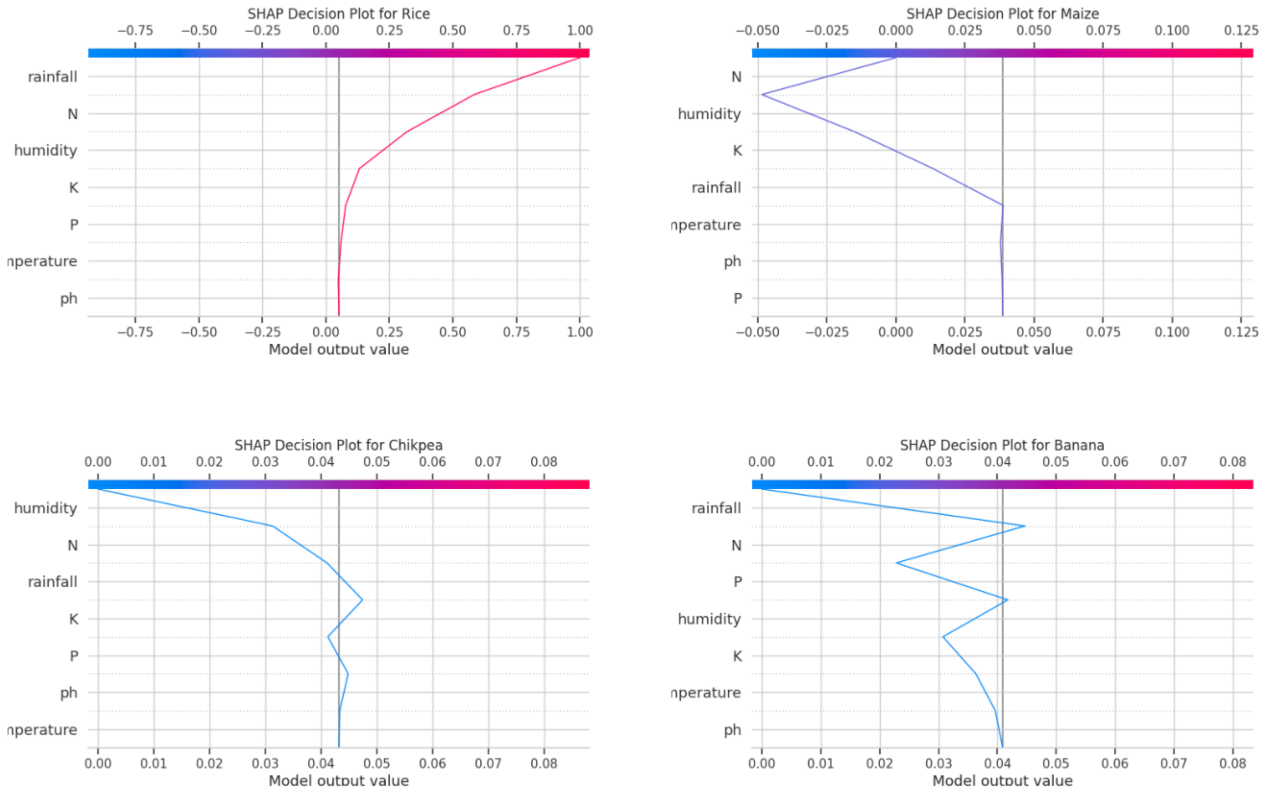

Figure 10. Decision Plot for Rice, Maize, Chickpea, and Banana Selection.

empowers farmers to optimize irrigation strategies based on expected rainfall and humidity levels, or adjust planting schedules accordingly. Nitrogen (N) follows closely as a crucial factor, highlighting its importance for various plant processes such as photosynthesis and protein synthesis. Potassium (K) emerges as another significant factor impacting various plant functions. A moderate influence is observed for rainfall, emphasizing the importance of adequate soil moisture management. Additionally, temperature and pH have a moderate influence, playing a role in the model's decision-making process by affecting nutrient availability and diverse plant functions. These SHAP results not only aid in comprehending our model's decision-making process but also offer valuable insights into crop selection. They enhance the interpretability and understanding of the model's predictions for stakeholders and farmers alike. By demystifying the model's inner workings, farmers can grasp its reasoning and feel more confident in its selection. This transparency builds trust and encourages wider adoption of AI in agriculture, ultimately leading to the development of even more interpretable and effective AI models for diverse agricultural applications.

The system's interface incorporates SHAP explanations, presented to users through intuitive visual plots. These SHAP plots help users understand how key features such as nitrogen levels, rainfall, and humidity impact crop selection decisions. For instance, a farmer can easily visualize how the model weighs the significance of "rainfall" when recommending a crop like rice, or how "humidity" influences the choice of mung beans. This visual approach effectively bridges the gap between the model's complex internal decision-making process and the farmer's ability to interpret and trust the system's recommendations.

While the SHAP-based explanations are designed to enhance interpretability, it is essential to acknowledge the system's current limitations. The dataset used may not fully capture the diversity of regional variations or the full range of crop types. Additionally, although the SHAP visualizations provide insight into the model's reasoning, a formal usability study is required to assess their practical impact. Such a study would involve testing with a diverse group of farmers, evaluating how well they interact with the system, how clearly they understand the SHAP-based explanations, and how much they trust the recommendations. This study will focus on gathering feedback on the clarity of SHAP plots, ease of use, and farmers' grasp of how model features relate to outcomes. Insights from this testing will be crucial for refining the system in future iterations to better meet the practical needs and expectations of farmers.

4. Conclusions

In conclusion, this research underscores the effectiveness of interpretable machine learning in developing highly accurate and efficient crop selection systems. By leveraging the AdaBoost algorithm, our system achieved an impressive 99.77% accuracy and a rapid fit time, rendering it suitable for real-time decision support in agriculture. By minimizing false positives and enhancing predictive capabilities, this system significantly mitigated financial risks for farmers and enhanced their decision-making processes. Moreover, the incorporation of SHAP values provided invaluable insights into the model's reasoning, allowing farmers to comprehend how climate and soil factors influence crop selection. Notably, humidity emerged as the most critical factor, emphasizing the significance of considering water availability in crop selection decisions.

While this research work primarily focused on a specific dataset and model, it lays the groundwork for further research exploring diverse data sources, advanced interpretability techniques, and user-friendly decision support tools. By combining high accuracy, interpretability, and efficiency, this approach heralds the advent of AI-powered tools that empower farmers and contribute to sustainable agricultural practices.

References

[1] S. Velten, J. Leventon, N. Jager, and J. Newig, "What is sustainable agriculture? a systematic review," *Sustainability*, vol. 7, pp. 7833–7865, 2015.

[2] M. Mancer, L. Terrissa, S. Ayad, and H. Laouz, "A blockchain-based approach to securing data in smart agriculture," in *2022 International Symposium On INnovative Informatics Of Biskra (ISNIB)*, 2022, pp. 1–5.

[3] M. Mancer, L. S. Terrissa, S. Ayad, and H. Laouz, "Tomato crop forecasting: A comparative analysis of regression models," in *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*. IEEE, 2024, pp. 649–653.

[4] M. Mancer, L. S. Terrissa, and S. Ayad, "Machine learning-based prediction of tomato yield in greenhouse environments," in *International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS 2024)*. Atlantis Press, 2024, pp. 117–128.

[5] S. Elghamrawy, A. Vasilakos, A. Darwish, and A. Hassanien, "An intelligent crop recommendation model for the three strategic crops in egypt based on climate change data," in *The Power Of Data: Driving Climate Change With Data Science And Artificial Intelligence Innovations*, 2023, pp. 189–205.

[6] B. Alsowaiq, N. Almusaynid, E. Albhnasawi, W. Alfenais, and S. Sankrayananarayanan, "Crop recommendation assessment for arid land using machine learning," in *International Conference On ICT For Sustainable Development*, 2023, pp. 323–332.

[7] S. Palle and S. Raut, "Crops recommendation system model using weather attributes, soil properties, and crops prices," in *Sentiment Analysis And Deep Learning: Proceedings Of ICSADL 2022*, 2023, pp. 323–338.

[8] A. Varghese and I. Mamatha, "A unified system for crop yield prediction, crop recommendation, and crop disease detection," in *International Conference On Robotics, Control, Automation And Artificial Intelligence*, 2022, pp. 1025–1035.

[9] R. Bandi, M. Likhit, S. Reddy, S. Bodla, and V. Venkat, "Voting classifier-based crop recommendation," *SN Computer Science*, vol. 4, p. 516, 2023.

[10] K. Bakthavatchalam, B. Karthik, V. Thiruvengadam, S. Muthal, D. Jose, K. Kotecha, and V. Varadarajan, "Iot framework for measurement and precision agriculture: predicting the crop using machine learning algorithms," *Technologies*, vol. 10, p. 13, 2022.

[11] M. Mancer, L. S. Terrissa, S. Ayad, H. Laouz, and N. Zerhouni, "Advancing crop recommendation systems through ensemble learning techniques," in *The Proceedings of the International Conference on Smart City Applications*. Springer, 2023, pp. 45–54.

[12] S. Rani, A. K. Mishra, A. Kataria, S. Mallik, and H. Qin, "Machine learning-based optimal crop selection system in smart agriculture," *Scientific Reports*, vol. 13, no. 1, p. 15997, 2023.

[13] A. Ingle, "Crop recommendation dataset," Kaggle, 12 2020. [Online]. Available: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

[14] T. Runkler, *Data analytics*. Springer, 2020.

[15] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals Of Translational Medicine*, vol. 4, 2016.

[16] V. Aggarwal, V. Gupta, P. Singh, K. Sharma, and N. Sharma, "Detection of spatial outlier by using improved z-score test," in *2019 3rd International Conference On Trends In Electronics And Informatics (ICOEI)*, 2019, pp. 788–790.

[17] B. Deepa and K. Ramesh, "Epileptic seizure detection using deep learning through min max scaler normalization," *Int. J. Health Sci*, vol. 6, pp. 10 981–10 996, 2022.

[18] Z. Lin, G. Ding, M. Hu, and J. Wang, "Multi-label classification via feature-aware implicit label space encoding," in *International Conference On Machine Learning*, 2014, pp. 325–333.

[19] R. Schapire, "Explaining adaboost," in *Empirical Inference: Festschrift In Honor Of Vladimir N. Vapnik*, 2013, pp. 37–52.

[20] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics And Its Interface*, vol. 2, pp. 349–360, 2009.

[21] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan *et al.*, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, pp. 1–33, 2023.