International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

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

Interpretable Crop Selection for Optimized Farming Decisions

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

[1]*Computer Science Department, Mohamed Khider University, Biskra, Algeria*

Abstract: Agricultural success hinges on strategic crop selection, directly influencing yield, financial stability, and risk management for farmers. Despite integrating machine learning techniques, many current systems function as opaque "black boxes," leading to reluctance among farmers who need both precision and transparency in crop recommendations. This study introduces a novel, interpretable approach for crop selection using climate and soil data, employing the AdaBoost classifier, renowned for its high accuracy and ability to prioritize misclassified data points.

To enhance transparency and foster trust among farmers, we incorporate SHapley Additive Explanations (SHAP) to elucidate the model's decision-making process. Our system analyzes diverse parameters such as nitrogen, phosphorus, potassium, pH, temperature, humidity, and rainfall to suggest suitable crops for cultivation. Evaluated on a comprehensive dataset of 22 crops, our approach achieves exceptional accuracy (99.77%) compared to conventional and boosted models, with rapid processing times (0.5 seconds per prediction). SHAP interpretations clarify the impacts of various climate and soil factors on crop suitability, offering farmers clear justifications for the recommendations provided.

By combining accuracy with transparency, our system empowers farmers to make informed decisions about their land, leading to improved yields and increased profitability. This interpretable system represents a significant advancement in developing efficient and reliable AI tools for sustainable crop selection in agriculture. We envision a future where farmers can embrace AI-driven tools with confidence, fostering a more sustainable agricultural landscape.

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

1. INTRODUCTION

Agriculture, the cornerstone of human sustenance and prosperity, extends beyond mere food production, profoundly impacting the global economy, livelihoods, and survival itself [1], [2]. A pivotal challenge in agriculture is selecting the appropriate crops for cultivation, a decision that can significantly influence a farmer's success. Suboptimal choices can lead to substantial economic losses and disrupt the entire agricultural landscape. While traditional methods and emerging AI-driven approaches aim to aid farmers, limitations in accuracy and transparency hinder widespread adoption.

Previous studies have made notable contributions to this evolving landscape. For instance, an Automated Crop Selection Model (ACRM) utilizing an optimized convolutional neural network (CNN) achieved an accuracy of 98.2% for crops such as maize, wheat, and rice in Egypt, with maize and rice attaining accuracies of 98.7% and 98.1%, respectively [3]. Another method proposed advising farmers on crop selection based on weather characteristics, soil features, and market prices using the ARIMA model and logistic regression, achieving a 2.25 RMSE and 94.24% accuracy [4]. Research targeting arid regions with machine learning techniques highlighted the random forest model's remarkable accuracy of 99.45%, effectively suggesting appropriate crops based on diverse environmental parameters [5].

Further advancements include a crop selection system for small-scale farmers, integrating weather, soil, and crop prices with an effective ARIMA weather model (RMSE: 2.254) and a multi-logistic regression model, which outputs 94.24% accuracy [6]. Another study presented a platform combining machine learning and the Internet of Things (IoT) to forecast crop yield, suggest crops, and identify diseases, achieving 99.2% accuracy in disease detection and 99% accuracy in crop selection using the ResNet model and random forest classifier, respectively [7]. Additionally, an ensemble model using majority voting demonstrated 99.4% accuracy for crop selection [8], while an IoT framework for precision agriculture using multilayer perceptron, JRip, and decision table classifiers reached 98% accuracy [9].

Despite these advancements, a significant barrier persists – trust. Many existing systems function as opaque "black boxes," causing reluctance among farmers who lack confidence in the selections provided. Farmers need not

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

only precise crop selection but also an understanding of why a particular crop is suggested. This knowledge instills confidence and empowers them to make informed decisions.

This study addresses this gap by introducing an interpretable machine learning-based crop selection system tailored for 22 different crops. Our system analyzes diverse parameters such as nitrogen, phosphorus, potassium, pH, temperature, humidity, and rainfall to suggest suitable crops for cultivation. Crucially, we leverage the AdaBoost classifier known for its accuracy and ability to prioritize misclassified instances. More importantly, our system incorporates SHapley Additive Explanations (SHAP), enabling it to provide explainable insights into the decision-making process. This transparency fosters trust and empowers farmers to make informed decisions about their land.

By combining accuracy with transparency, our system aims to empower farmers to cultivate success. We envision a future where farmers can embrace AI-driven tools with confidence, leading to improved yields, increased profitability, and ultimately, a more sustainable agricultural landscape.

The paper is structured as follows: Section 2 delves into the materials and methods employed in our study, detailing the data used and the specific machine-learning techniques implemented. Section 3 presents the results, showcasing the system's performance and key findings. This section also discusses these 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

This study proposes an interpretable AdaBoost classifier-based crop selection system aimed at achieving accurate 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 focuses on leveraging the trained AdaBoost model to provide real-time selection for farmers. Additionally, SHAP is used to analyze the trained model, identifying how specific climate and soil factors contribute to the selected crop for each prediction. This provides farmers with clear, understandable explanations.

A. Data Sources and Exploratory Analysis

This study utilized a publicly available dataset retrieved from Kaggle [10]. The dataset comprises 2,200 observations, each representing a specific crop. This translates to 100 data points for each of the 22 different crops considered in the study. The dataset offers valuable information about several parameters crucial for crop selection, including nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, pH, and rainfall.

1) Univariate Analysis

Univariate analysis, as described by [11], 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). 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 understand 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 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

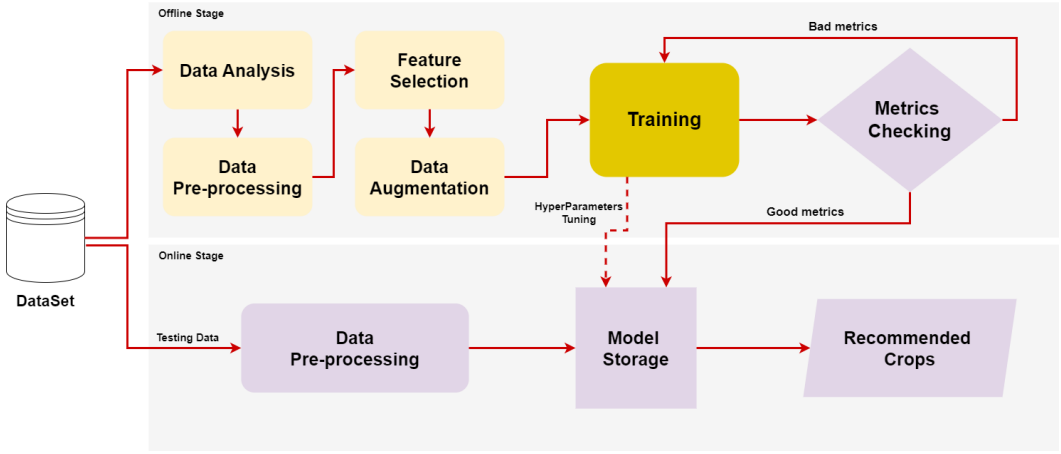

Figure 1. The general architecture of the proposed approach.

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

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 [11], 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 2 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 [12]. 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 [13] 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 [14]. 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 [15] 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.

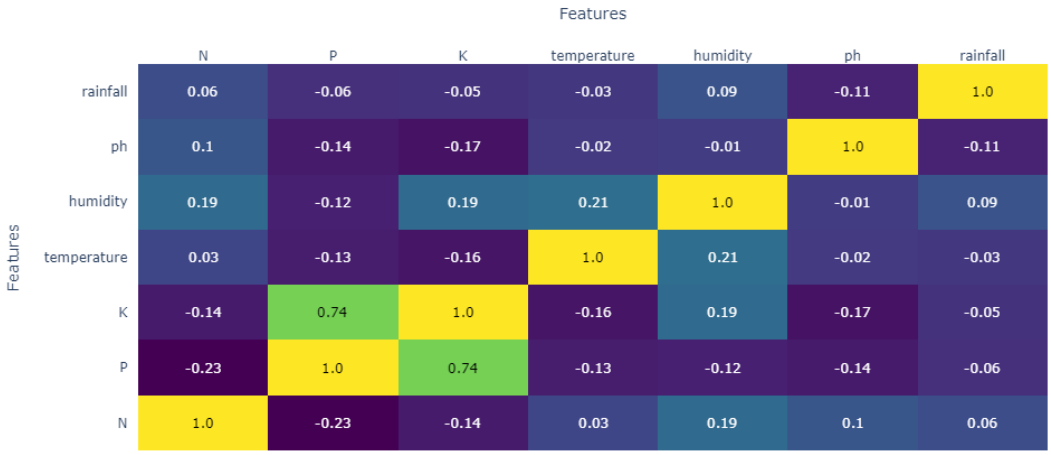

Figure 2. Correlation coefficients among variables.

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 [16], [17].

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:

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

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 and Google Colab. To achieve optimal performance, we employed an iterative trial-and-error approach to fine-tune various training options and model parameters. The chosen AdaBoost classifier configuration includes the following:

- **n_estimators = 50:** Number of weak learners contributing to the final prediction.
- **base_estimator = RandomForestClassifier:** Tree-based model used as the base learner.
- **Learning_rate = 0.001:** Controls the influence of individual weak learners on the ensemble's output.
- **random_state = 0:** Ensures reproducibility of results across different runs.

Evaluating the performance of our proposed model is crucial. We use metrics such as accuracy, precision, recall, and F1-score to assess how well the model identifies suitable crops and avoids selecting 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 this, we leverage SHapley Additive ExPlanations (SHAP) [18], 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.

SHAP assigns a SHAP value to each feature in a prediction, representing its fair share of the predicted crop class. These values are calculated by comparing the original model's prediction to predictions made on feature subsets, resembling 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)

1: **for** $k \leftarrow 1$ **to** K **do**
2: $explainer_k \leftarrow$ Initialize SHAP explainer for class k
3: $shap\_values_k \leftarrow$ Compute SHAP values for X and class k using Eq. (7)
4: **Combine** the $shap\_values_k$ **with the existing SHAP values (the specific method depends on library/framework)**
5: **end for**
6: **for** $k \leftarrow 1$ **to** K **do**
7: $Feature\_importance_k \leftarrow$ Compute feature importance for class k using individual class SHAP values
8: **end for**
9: **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:

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

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 3 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 4 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.

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 5) visualizes misclassifications, offering insights into potential overlaps between crop classes. For example, minor misclassifications exist between "Rice" and "Jute", and "Blackgram" and "Mothbeans". This suggests some feature similarities between these classes, potentially impacting the model's decisions. Notably, the model maintains a false positive rate (FPR) of 0, meaning that it rarely identifies negative instances (non-recommended crops) incorrectly as positive, ensuring greater accuracy in its selection.

These results solidify 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 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 6 and Figure 7 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 provide a more refined analysis, indicating "humidity" as the most influential feature, followed by "N" and "K". Furthermore, the impact of features varies across crops: "rainfall" significantly affects rice and pigeon peas but minimally impacts kidney beans. Similarly, "humidity" strongly influences "mungbean" peas but has a weaker effect on watermelon.

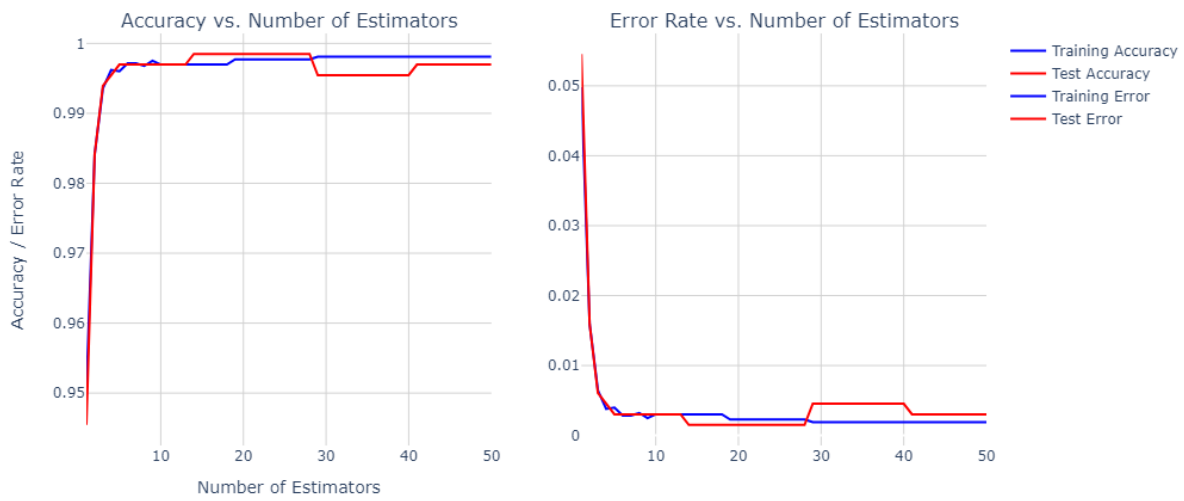

Figure 3. AdaBoost Classifier: Accuracy and Error Rate Trends.

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

We investigated the impact of various features on selected crops using SHAP values. While our dataset encompasses 22 crops, this analysis focuses on rice, maize, chickpea, and banana to illustrate the variation in feature importance across different crops. Figure 8 and Figure 9 present SHAP summary plots for each of these four crops.

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 generally 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

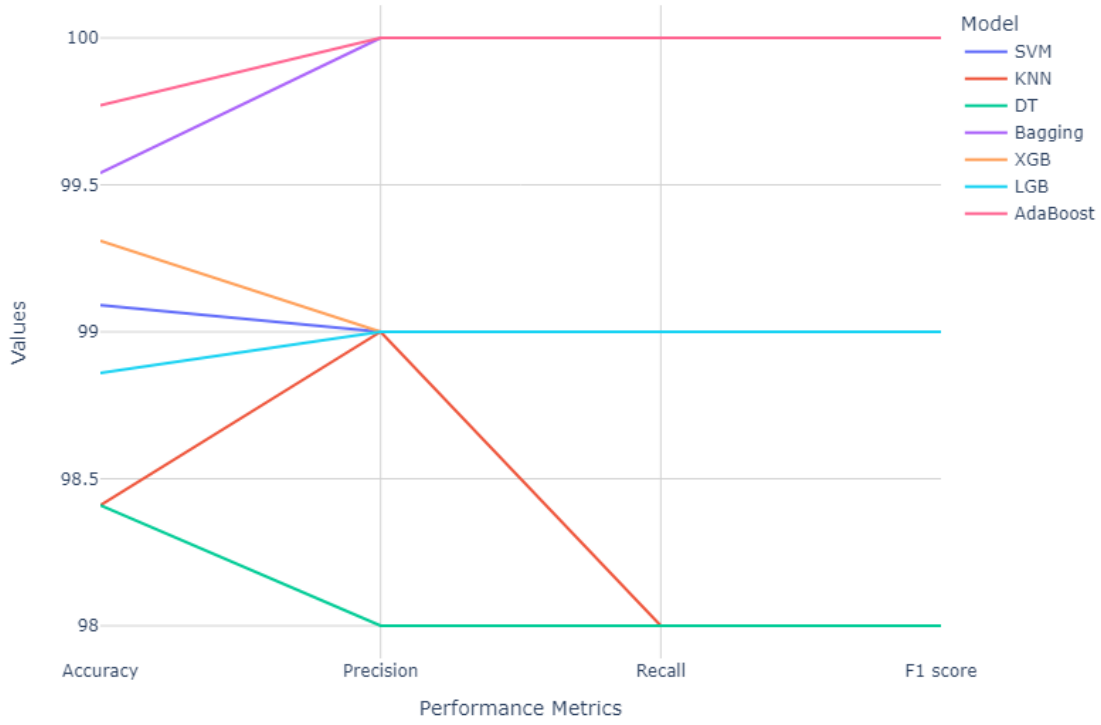

Figure 4. Comparative Analysis of Performance Metrics Across Various Models.

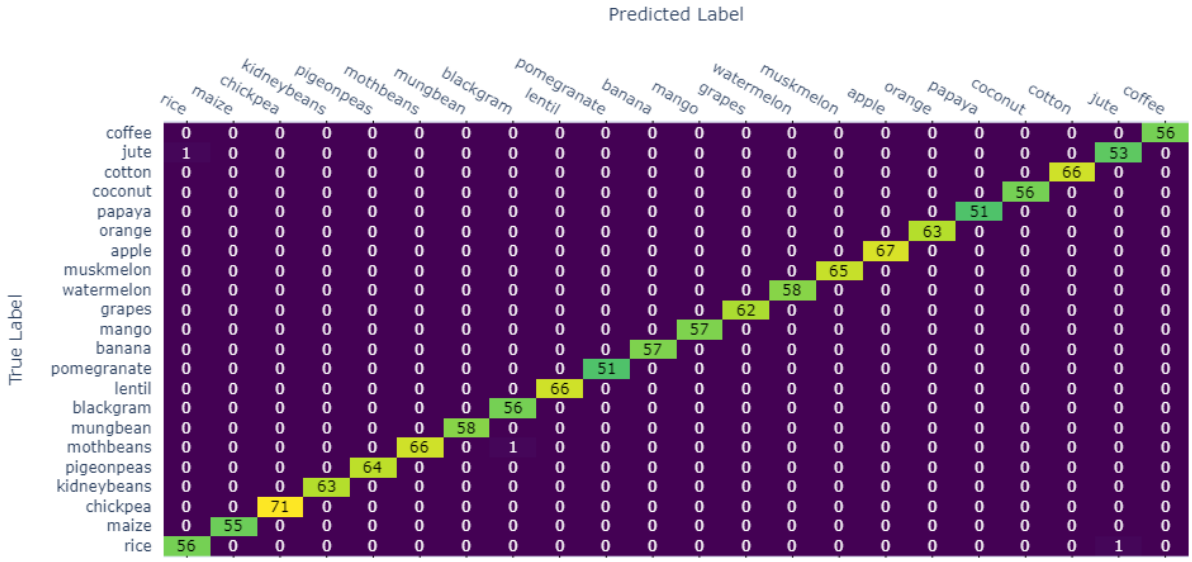

Figure 5. Confusion matrix visualization for AdaBoost classifier.

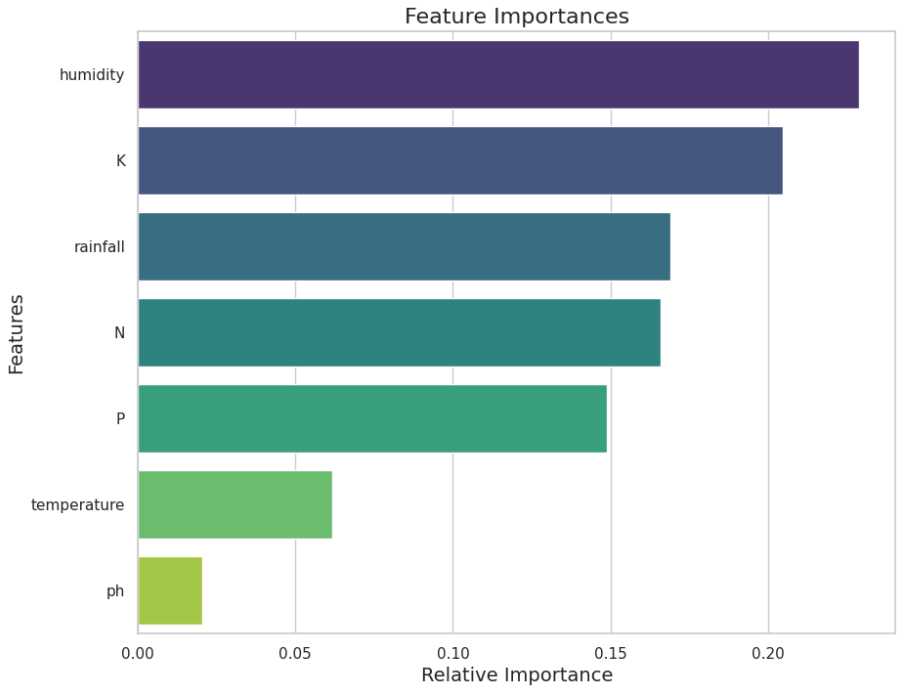

Figure 6. Feature Importance Analysis using Permutation Technique.

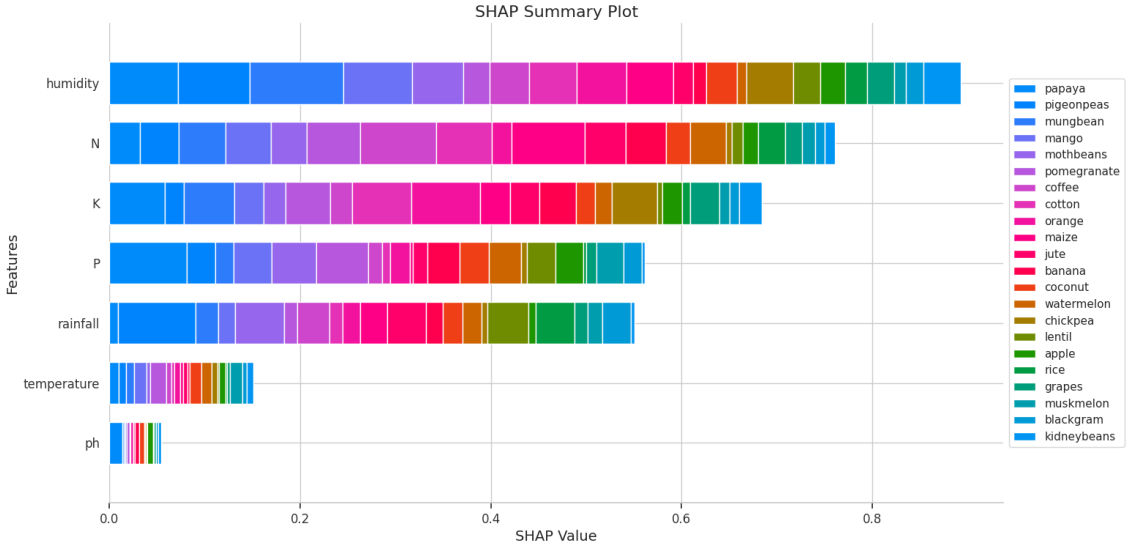

Figure 7. Feature Importance Analysis using SHAP.

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

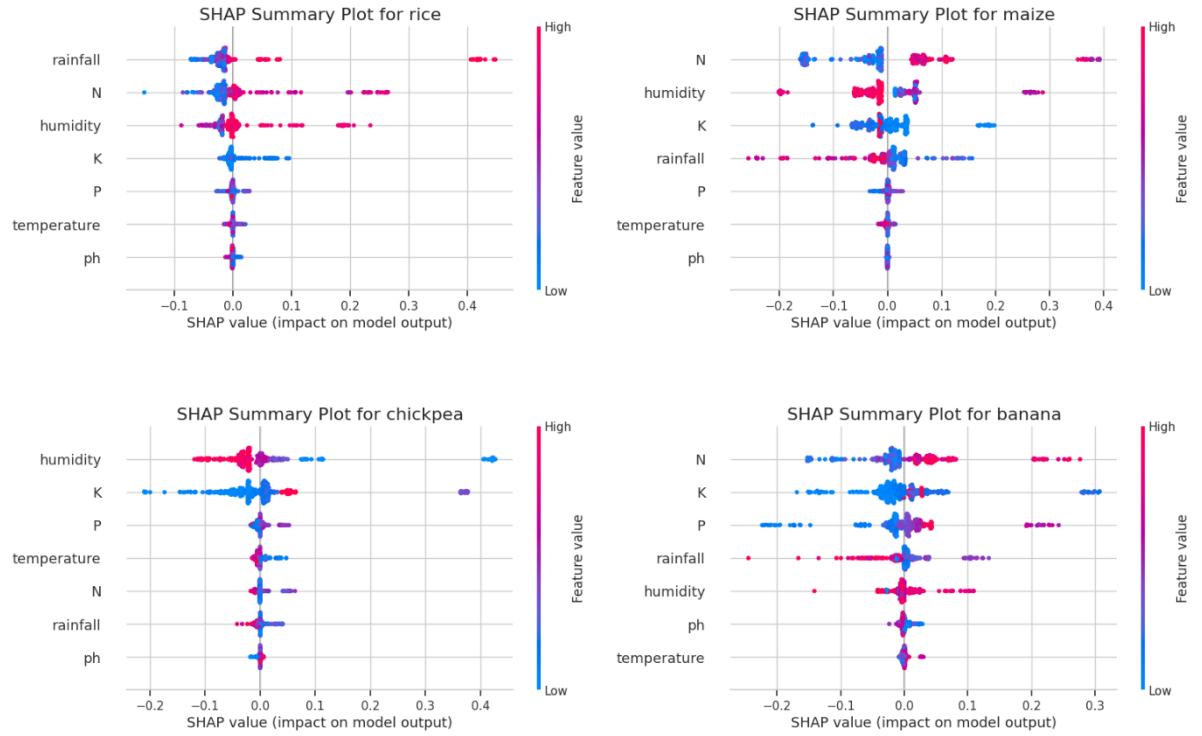

Figure 8. Feature Importance for Rice, Maize, Chickpea, and Banana Selection.

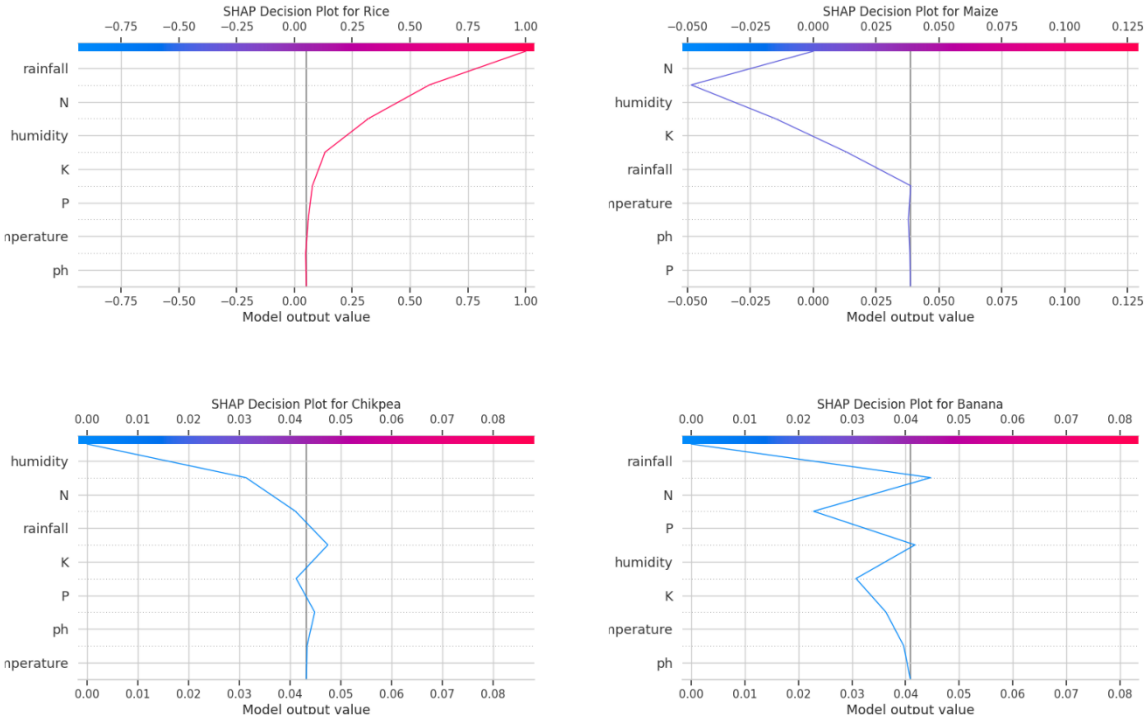

Figure 9. Decision Plot for Rice, Maize, Chickpea, and Banana Selection.

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.

The agricultural sector is increasingly turning to machine learning to harness its analytical power. These algorithms excel at processing complex datasets, uncovering insights that traditional statistical methods struggle to discern. 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) [3], others such as random forest (99.45%) [5] and an IoT-based framework (98%) [9] 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, compared to 8.05 seconds for previous models such as the MLP [9]. 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 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.

While this research demonstrates the potential of interpretable AI for crop selection, it is important to acknowledge its limitations. The current dataset might not fully capture all regional variations or crop types. Future research could focus on enriching the dataset and exploring other interpretable AI techniques to further empower farmers with data-driven insights and contribute to the development of

sustainable and efficient agricultural practices.

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] 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.

[4] A. Kumar, "Crop recommendation for maximizing crop yield using random forest," in *International Conference On Innovations In Computational Intelligence And Computer Vision*, 2022, pp. 501–515.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

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

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

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

M'hamed Mancer PhD in Computer Science, specializing in Artificial Intelligence, from the University of Mohamed Khider Biskra. As a member of the LINFI laboratory, his focus encompasses Smart Agriculture, Machine Learning, Deep Learning, and Blockchain technologies. He is dedicated to developing practical solutions that leverage Artificial Intelligence and Information Technology to address contemporary challenges in agriculture.

Labib Terrissa Professor at Biskra University, Algeria, leads the ”CoViBio” team at LINFI Laboratory. With expertise in electronics engineering and a PhD in computer engineering, he specializes in Cloud Computing, Machine Learning, and Smart Maintenance..

Soheyb Ayad Associate professor at the Computer Science Department at the University of Biskra, Algeria, and a member of LINFI laboratory, focuses on Networking, Cloud Computing, Internet of Things, Machine learning, Smart Agriculture, and Predictive Maintenance. With several international publications, he has contributed significantly to these fields.