

Article

Research on Factors Affecting Global Grain Legume Yield Based on Explainable Artificial Intelligence

Yadong Li, Rujia Li, Rongbiao Ji, Yehui Wu, Jiaojiao Chen, Mengyao Wu and Jianping Yang * 

College of Big Data, Yunnan Agricultural University, Kunming 650201, China; 14787825720@163.com (Y.L.); 18988097525@163.com (R.L.); jirb4532@163.com (R.J.); wuyehui0511@163.com (Y.W.); 18988097725@163.com (J.C.); 15393997081@163.com (M.W.)

* Correspondence: yangjp@ynau.edu.cn

Abstract: Grain legumes play a significant global role and are integral to agriculture and food production worldwide. Therefore, comprehending and analyzing the factors that influence grain legume yield are of paramount importance for guiding agricultural management and decision making. Traditional statistical analysis methods present limitations in interpreting results, but explainable artificial intelligence (AI) provides a visual representation of model results, offering insights into the key factors affecting grain legume yield. In this study, nine typical grain legume species were selected from a published global experimental dataset: garden pea (*Pisum sativum*), chickpea (*Cicer arietinum*), cowpea (*Vigna unguiculata*), garden vetch (*Vicia sativa*), faba bean (*Vicia faba*), lentil (*Lens culinaris*), pigeon pea (*Cajanus cajan*), peanut (*Arachis hypogaea*), and white lupine (*Lupinus albus*). Seven commonly used models were constructed for each legume species, and model performance evaluation was conducted using accuracy, AUC, recall, precision, and F1 score metrics. The best classification model was selected for each grain legume species. Employing Decision Tree analysis, Feature Importance Evaluation, and SHapley Additive exPlanations (SHAP) as explainable techniques, our study conducted both individual and comprehensive analyses of nine leguminous crops. This approach offers a novel perspective, unveiling not only the unique responses of each crop to the influencing factors but also demonstrating the common factors across different crops. According to the experimental results, XGboost (XGB) and Random Forests (RF) are the best-performing models among the nine types of grain legumes, and the classification accuracy of a specific species is as high as 87.33%. Insights drawn from the feature importance map reveal that several factors, including aerial biomass, precipitation, sunshine duration, soil conditions, growth cycle, and fertilization strategy, have a pivotal influence. However, it was found from the SHAP graph that the responses of various crops to these factors are not the same. This research furnishes novel perspectives and insights into understanding the factors influencing grain legume yields. The findings provide a robust scientific foundation for agricultural managers, experts, and policymakers in the pursuit of optimizing pulse yields and advancing agricultural sustainability.



Citation: Li, Y.; Li, R.; Ji, R.; Wu, Y.; Chen, J.; Wu, M.; Yang, J. Research on Factors Affecting Global Grain Legume Yield Based on Explainable Artificial Intelligence. *Agriculture* **2024**, *14*, 438. <https://doi.org/10.3390/agriculture14030438>

Academic Editors: Ewa Szpunar-Krok and Marcin Kozak

Received: 6 February 2024

Revised: 5 March 2024

Accepted: 6 March 2024

Published: 7 March 2024

Keywords: grain legumes; explainable artificial intelligence; feature importance analysis; Decision Trees; SHAP



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The Fabaceae family encompasses approximately 20,000 species, making it one of the largest plant families worldwide. It is extensively cultivated and utilized globally. Grain legumes, belonging to the Fabaceae family, possess numerous significant characteristics and applications, rendering them a vital component of global agriculture and food production. Firstly, grain legumes are recognized for their high protein content, establishing them as a crucial source of dietary protein for both humans and animals. Additionally, grain legumes play a pivotal role in soil improvement [1]. These plants form a symbiotic relationship with specific soil bacteria, resulting in the formation of root nodules. Through

nitrogen fixation, they convert atmospheric nitrogen into a readily available form for plants, effectively enhancing soil fertility and reducing reliance on chemical nitrogen fertilizers [2]. Therefore, analyzing the factors that influence grain legume yield can offer a scientific basis for agricultural decision makers to formulate sound agricultural policies and plans, thereby significantly contributing to sustainable agricultural development and ensuring food security [3].

Statistical analysis is a commonly used method for analyzing factors that affect grain legume yield. In addition, multivariate analysis techniques, such as factor analysis and principal component analysis [4], have been applied to explore the relationship between multiple factors and their degree of influence on grain legume yield. The utilizations of the Geographic Information System (GIS) and remote sensing technology (RS) [5] have also been widespread in the spatial analysis of factors impacting grain legume yield. Researchers can combine yield data and environmental factors by acquiring remote sensing data, such as land use, land cover, and vegetation index, to conduct spatial analysis. This enables the identification of the spatial distribution of grain legume yield and the assessment of spatial variability in influencing factors [6]. Simulation modeling [7] stands as a significant approach in studying the factors affecting grain legume yield. Mathematical models and computer simulations are employed by researchers to simulate the growth process and yield response of grain legumes, as well as to analyze the impact of various factors. These models are capable of capturing complex interactions among multiple factors, facilitating the prediction of pulse yield under different management practices and environmental conditions. Machine learning and data mining methods [8] play an important role in the analysis of factors affecting grain legume yield. Researchers can employ classification and regression algorithms, such as Decision Trees, Support Vector Machines, and Random Forests, to construct predictive models. By incorporating various environmental and management factors as input, these models enable the prediction of grain legume yield and the analysis of the importance of influencing factors.

However, it is essential to emphasize that while these methods may provide predictions or correlations regarding factors influencing grain legume yield, they do not offer detailed explanations of causality or underlying mechanisms. Our study aims to address these limitations in the existing technologies. Explainable artificial intelligence [9] aims to enhance the transparency and comprehensibility of decision making in machine learning models. By enabling users to understand how a specific prediction is derived, it enhances trust and acceptance of the model. This method not only provides prediction results, but also reveals key factors that influence grain legume yield [10]. Moreover, grain legume yield is influenced by multiple factors, and complex interactions and nonlinear relationships may exist among these factors. Explainable artificial intelligence has the capability to better capture these complexities, enabling us to present complex data relationships in a more intuitive and understandable manner, surpassing the capabilities of traditional analytical methods. This aids researchers and policymakers in gaining a better understanding and explanation of the causes of yield fluctuations [11,12].

Currently, research on grain legumes primarily focuses on analyzing individual species, overlooking the comprehensive analysis of multiple grain legumes. To bridge this information and methodological gap, this study leverages a global experimental dataset encompassing a substantial volume of grain legume yield data from numerous countries and regions [13]. The dataset includes long-term grain legume production data, allowing for an exploration of diverse factors, such as climate regions, soil types, and growing conditions. By employing this dataset, multivariate analysis can be conducted to simultaneously consider multiple potential influencing factors, enabling the identification and quantification of interactions and comprehensive effects among these factors. This approach reveals a more comprehensive network of influencing factors, facilitating the provision of comprehensive, accurate, and reliable analysis results. Based on the aforementioned datasets, in this study, we selected yield-related data from nine representative grain legume species as the dataset required for this paper. Subsequently, data normalization, numericalization of

classification labels, and addressing of category imbalance were performed on the dataset. Following preprocessing, seven models for the nine grain legume species were constructed. The optimal model was selected to elucidate and analyze the factors affecting crop yield through the utilization of explainable artificial intelligence. The specific process is depicted in Figure 1. Architecture diagram in Figure 1.

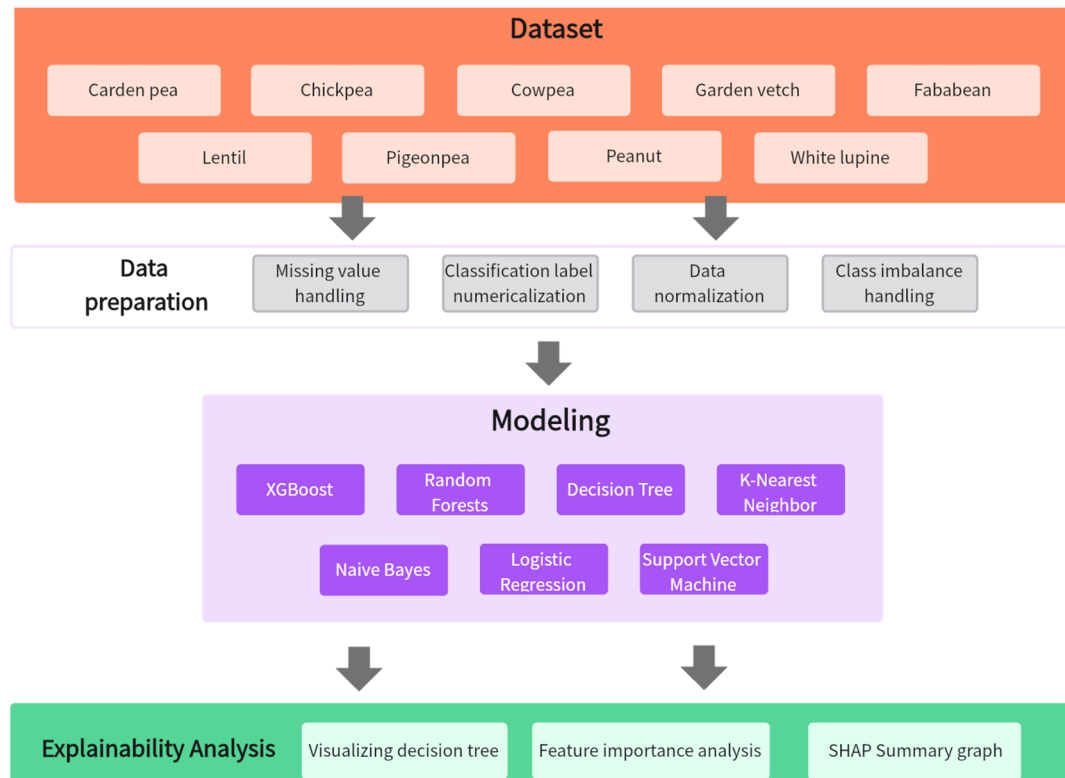


Figure 1. Architecture diagram.

2. Models and Methods

2.1. Dataset

The dataset utilized in this study was derived from the global grain legume experiment dataset published in Scientific Data by Charles Cernay et al. [13]. This dataset comprises the outcomes of field experiments documented in 173 articles. It encompasses measurement data obtained from 360 field experiment sites across 18 countries spanning five continents, encompassing 39 species of grain legumes. The dataset is composed of nine structured tables and 198 standardized attributes, making it the most comprehensive agronomic dataset for cereal grain legume crops on a global scale. After organizing, consolidating, and refining the dataset, nine specific grain legume species were ultimately selected from the nine genera of grain legumes defined by the Food and Agriculture Organization of the United Nations (FAO) [14]. These selected species constitute the dataset employed in this study. The dataset encompasses a total of 18,259 records, and the nine grain legume species included are garden pea, chickpea, cowpea, garden vetch, faba bean, lentil, pigeon pea, peanut, and white lupine. These species correspond to the following genera: *Pisum* spp., *Cicer arietinum*, *Vigna unguiculata*, *Vicia sativa*, *Vicia faba*, *Lens culinaris*, *Cajanus* spp., *Arachis*, and *Lupinus* spp. Some samples of the dataset are shown in Table 1. The dataset employed in this study encompasses 38 attributes, including grain yield, aerial biomass, crop nitrogen content, soil texture, soil nitrogen content, experimental field precipitation, experimental field temperature, fertilization management, irrigation management, and pest management.

Table 1. Partial sample of dataset attributes.

Crop_Yield_Grain	Site_Latitude	Site_Longitude	Site_Soil_Sand_Percentage	Site_Soil_Silt_Percentage	Site_Soil_Clay_Percentage	Site_Soil_pH	Site_Soil_N_Percentage	Site_Precipitation_mm	Site_Temperature_Celsius	Crop_Biomass_Aerial	Irrigation_Presence_Irrigation_Dose	Fertilization_NPK_Dose	Weed_Insect_Fungi_Presence_Treatment_Chemical_Dose
2.89	37.3	13.31	33.8	20.3	45.9	7.9	0.14	552	16.7	0	0	69	0
1.42	37.3	13.31	27	23.2	49.8	8	0.12	552	16.7	0	0	69	0
2.17	37.3	13.31	27	23.2	49.8	8	0.12	552	16.7	0	0	69	1080
2.6	37.3	13.31	48.6	24.7	26.7	8	0.07	552	16.7	5.27	8	8	1080
2.84	37.3	13.31	18.1	24.7	57.2	8	0.09	552	16.7	7.43	8	8	1080
0.31	−31.29	118.12	0	0	42	7.5	0	272	18.1	4.47	0	145	300
0.48	−31.29	118.12	0	0	40	7	0	173	19.3	2.06	0	72.5	0
1.45	−28.32	115.3	0	0	21.7	7.9	0	209	19.2	3.44	0	145	300
1.22	−28.32	115.3	0	0	26.7	8.1	0	201	20.1	3.21	0	72.5	300
2.48	50.2	−107.4	0	0		6.5	0	351	4	6.59	106	20	0
2.9	35.05	147.21	0	15	29	4.7	0.13	715	0	8.2	0	20	0
2.41	36.4	37.2	0	0	0	0	0.06	429	15.6	5.99	0	26	0
3.18	43.61	3.88	0	0	0	0	0.1	531	11.7	6.5	0	44	0
0.87	47.03	−109.57	0	0	0	6.5	0	233	11.9	2.51	0	7.3	0
1.19	47.03	−109.57	0	0	0	6.5	0	233	11.9	2.51	0	22	0
0.91	45.4	−111.9	0	0	0	7.5	0	341	0	2.43	0	5	0
0.27	47.2	−109.53	0	0	0	7.6	0	303	0	2.27	0	5	0
0.53	45.46	−111.23	0	0	0	8.1	0	308	0	1.7	0	5	0
0.08	47.52	−111.41	0	0	0	6.7	0	235	0	0.97	0	5	0
0.91	45.4	−111.9	0	0	0	7.5	0	341	0	2.43	0	5	0
1.09	39.37	22.22	50.9	20	29.1	7.5	0	461	16.2	4.02	0	30	0
0.92	21.31	70.36	77	13	10	7.6		844				10	0
2.89	37.3	13.31	27	23.2	49.8	8	0.12	479	19.6	8.51		92	0

2.2. Data Preprocessing

Given the substantial volume of data in the original dataset, this study selectively focused on the data tables pertinent to the yield of grain legumes. The seven tables—Site, Crop_Sequence_Trtr, Crop, Tillage, Fertilization, Weed_Insect_Fung, and Irrigation—were interconnected via primary and foreign keys, thereby forming the dataset required for this study. As the dataset exhibited numerous missing values and unbalanced categories, we employed data preprocessing methods such as missing value processing, classification label numericalization, data normalization, and class imbalance processing.

The dataset, characterized by a large number of missing values and zero values, necessitated certain measures to ensure data integrity and accuracy. We adopted a method of deleting missing values and filling specific values to normalize the data, eliminating data columns with excessive missing values. Missing values and zero values, which do not represent a particular attribute value, were filled as the median using the fillna() function. To ensure that all the features of the nine grain legumes data were on a similar scale, we normalized the data using the fit_transform() function, allowing for the subsequent model to consider the influence of different features in a more balanced manner, thereby enhancing the accuracy of feature importance analysis. Due to the significant disparities in yield levels across the various countries and regions covered by the dataset, we employed the Food and Agriculture Organization (FAO) global average yield of grain legumes to classify each crop into two categories: above and below the global average yield. Samples falling below the

global average yield were labeled as “0”, while those exceeding it were labeled as “1” [14]. Due to the original dataset encompassing grain legume yield from various global countries and regions, leading to significant yield discrepancies, there was a category imbalance issue following the binary classification of crop yield data. To address the category imbalance, data sampling was typically employed. Mainstream sampling methods include under-sampling, oversampling, and mixed sampling, which all aim to balance the sample size of different categories by altering the data volume. Undersampling serves as a technique to decrease the majority class’s sample size, thereby ensuring a balanced sample size [15]. Random undersampling, a classic method, achieves this balance by randomly discarding some samples from the majority class. In contrast to undersampling, oversampling aims to increase the minority class’s sample size through mathematical modeling or method synthesis, thereby balancing the sample sizes across different classes [16]. As oversampling can augment the sample size, it is more commonly applied to smaller datasets. However, this approach of sample synthesis may lead to overfitting. Mixed sampling is a technique that amalgamates undersampling and oversampling to balance the sample sizes across various categories. BATISTA et al. [17] proposed the SMOTEENN algorithm as a notable mixed sampling method. Mixed sampling compensates for the reduction in sample size induced by undersampling, concurrently optimizing the issue of sample overlap resulting from oversampling and thereby balancing the dataset without altering the data volume [18].

In this study, we employed different sampling methods to address category imbalance, tailored to the unique data scale and distribution characteristics of the nine grain legumes. As illustrated in Table 2, the Garden_pea dataset is large, with the majority class significantly outnumbering the minority class, making random undersampling a suitable approach. The White_lupine dataset is relatively small and exhibits a higher degree of imbalance compared to the other small sample datasets. By implementing the random oversampling method to augment the number of minority samples, we enhanced its representation in the dataset, enabling the model to better learn the characteristics of the minority class. The datasets for the remaining seven grain legumes were sampled using the SMOTEENN method.

Table 2. Sampling methods of 9 grain legumes. (‘1’ indicates yields above global average, ‘0’ indicates below).

Species	Genus	Number of Entries	1	0	Sampling Methods
Garden_pea	<i>Pisum</i> spp.	7093	4420	2673	Random Undersampling
Chickpea	<i>Cicer arietinum</i>	2266	1845	421	SMOTEENN
Cowpea	<i>Vigna unguiculata</i>	1263	1171	92	SMOTEENN
Garden_vetch	<i>Vicia sativa</i>	644	411	233	SMOTEENN
Fababean	<i>Vicia faba</i>	1368	823	545	SMOTEENN
Lentil	<i>Lens culinaris</i>	2072	1624	448	SMOTEENN
Pigeonpea	<i>Cajanus</i> spp.	432	256	176	SMOTEENN
Peanut	<i>Arachis</i>	2175	1520	655	SMOTEENN
White_lupine	<i>Lupinus</i> spp.	946	663	283	Random Oversampling

2.3. Evaluation Metrics

This study evaluated the classification performance of each model using metrics such as accuracy, AUC, recall, precision, and F1 score. Accuracy represents the proportion of correctly classified samples out of the total number of samples. Precision indicates the proportion of true positive samples among the samples predicted as positive. Recall represents the proportion of true positive samples among all positive samples. AUC (area under the receiver operating characteristic curve) measures the area under the ROC curve, which is used to assess the classification performance of imbalanced datasets. The F1 score is the weighted average of precision and recall, calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

where TP (True Positive) represents instances correctly identified as positive by the model, TN (True Negative) represents instances correctly identified as negative by the model, FP (False Positive) represents instances incorrectly identified as positive by the model, and FN (False Negative) represents instances incorrectly identified as negative by the model.

2.4. Model Building

Given the nature of the structured data employed in this study, we selected seven models that are well-suited for this type of data: Logistic Regression (LR), Naive Bayes (NB), Random Forests (RF), XGBoost (XGB), K-Nearest Neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM). These models represent various learning methods and algorithms, suitable for handling structured data and widely applied in binary classification tasks. Logistic Regression (LR) is a classical classification algorithm that utilizes a logistic function on top of linear regression for classification prediction. It is applicable to binary classification problems and excels in interpretability. Naive Bayes (NB) is a classification algorithm based on Bayes' theorem and the assumption of feature independence. Despite its simplicity, it performs well in handling structured data. Random Forests (RF) and XGBoost (XGB) are two ensemble learning methods that classify by constructing multiple Decision Trees and aggregating their results. They typically perform well on structured binary classification datasets. K-Nearest Neighbor (KNN) is an instance-based learning method that predicts by finding the K nearest neighbors to a new sample in feature space, often using voting to determine the sample's class. Decision Tree (DT) is a tree-based classification algorithm that categorizes samples through a series of decisions. It is widely applied in handling structured data. Support Vector Machine (SVM) is a model that classifies data by finding the optimal hyperplane, demonstrating strong generalization capabilities. We divided the training and testing sets using a 7:3 ratio. GridSearchCV was employed to optimize the parameters of all models. GridSearchCV is a widely used parameter optimization method in machine learning, which searches for the optimal parameter combination by iterating through a given parameter grid. This strategy was applied to each model for all nine crops. We selected Logistic Regression (LR) and Random Forests (RF) as example models and have detailed their parameter settings and the ranges tested, as shown in Table 3. Subsequently, we evaluated each model for the nine crops using metrics such as accuracy, AUC, recall, precision, and F1 score. Taking Fababean as an example, the model evaluation results are presented in Table 4, indicating that the Random Forest model performed the best for Fababean.

Table 3. Parameter settings and ranges tested for Logistic Regression (LR) and Random Forests (RF).

Model	Parameter	Values
LR	C	0.01, 0.1, 1, 10
	penalty	11, 12
RF	solver	liblinear, saga
	n_estimators	20, 50, 100, 200
	max_depth	6, 8, 10, 20, 30
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4, 6
	max_features	auto, sqrt, log2

Table 4. Evaluation results of 7 models of Fababean.(Best values in bold).

Model	Acc.	Auc.	Prec.	Rec.	F1
XGBoost	0.75	0.74	0.77	0.75	0.74
Random Forests	0.79	0.770	0.83	0.79	0.78
Decision Tree	0.72	0.77	0.79	0.72	0.69
K-Nearest Neighbor	0.73	0.72	0.73	0.73	0.73
Naive Bayes	0.76	0.76	0.77	0.76	0.76
Logistic Regression	0.73	0.74	0.74	0.73	0.73
Support Vector Machine	0.69	0.69	0.69	0.69	0.69

The seven machine learning models were independently applied to the nine crops, and the accuracy of each model for each crop is reported in Table 5. For example, the classification accuracy of the XGB model for Garden_pea was 76.74%, while the RF model achieved a classification accuracy of 86.76% for Chickpea and 87.33% for Cowpea. The RF model also achieved classification accuracies of 77.31% for Garden_vetch, 78.83% for Fababean, and 86.33% for Lentil. The XGB model achieved classification accuracies of 71.53% for Pigeonpea and 76.86% for Peanut, while the RF model achieved a classification accuracy of 71.83% for White_lupine. As can be seen from Table 5, among the nine grain legume species, the XGB and RF models yielded the best classification results.

Table 5. Accuracy-based performance comparison of 9 grain legume species across 7 models.(Best values in bold).

Species	Genus	XGBoost	Random Forests	Decision Tree	K-Nearest Neighbor	Naive Bayes	Logistic Regression	Support Vector Machine
Garden_pea	<i>Pisum</i> spp.	76.74	66.64	64.00	58.00	59.60	74.30	64.00
Chickpea	<i>Cicer arietinum</i>	82.06	86.76	79.26	81.32	81.02	77.20	76.76
Cowpea	<i>Vigna sinensis</i>	77.83	87.33	84.96	66.75	70.71	72.29	75.46
Garden_vetch	<i>Vicia sativa</i>	76.29	77.31	65.46	71.64	70.10	70.10	69.59
Fababean	<i>Vicia faba</i>	74.94	78.83	72.02	72.50	75.66	73.23	69.34
Lentil	<i>Lens esculenta</i>	85.36	86.33	75.24	71.86	85.85	85.85	79.10
Pigeonpea	<i>Cajanus</i> spp.	71.53	64.61	71.54	63.85	71.54	64.62	64.61
Peanut	<i>Arachis</i>	76.86	74.12	64.78	41.50	60.33	74.12	50.69
White_lupine	<i>Lupinus</i> spp.	57.75	71.83	51.06	57.75	57.75	40.85	56.70

Random Forests, a bagging-based ensemble learning method proposed by Breiman et al. [19], is a classifier composed of multiple independent Decision Trees. It boasts strong

generalization ability, rapid training speed, and the capacity to handle high-dimensional data without the need for feature selection, making it widely applicable to numerous classification problems. XGBoost, an ensemble learning algorithm based on boosting proposed by Chen T Q et al. [20], combines basis functions and weights through boosting principles. It aggregates the results of multiple trees, summing the scores from each tree to obtain the final score. Both Random Forests and XGBoost are ensemble tree models. Given that Decision Trees tend to overfit when dealing with high-dimensional samples, most practical applications utilize ensemble learning methods based on Decision Trees [21]. These methods combine multiple base learners (Decision Trees) to perform classification or regression tasks, thereby enhancing the accuracy and robustness of the model [22].

3. Results and Discussion

3.1. Visualization of Decision Trees

As evidenced in Table 3, ensemble tree models demonstrate superior classification performance in the structured data domain classification model [23] when compared to other machine learning models. Furthermore, compared to certain black box models (such as neural networks), Random Forests and XGBoost retain a degree of explainability. This is due to the fact that the decision-making processes of Random Forests and XGBoost can be understood and explained by visualizing the structure of each Decision Tree.

Figure 2, for instance, depicts a Decision Tree randomly selected from the Cowpea dataset based on the RF model. The Decision Tree reveals that each node's division condition is the feature selection condition. Features closer to the root node in the Decision Tree are of greater importance and play a pivotal role in the division of classification results. The data flow process within the model is clearly visible from the Decision Tree. The samples in the training set are divided according to the root node's value condition, *Crop_N_Quantity_Aerial* 0.068. Samples with a value less than 0.068 enter the left child node, while those with a value greater than or equal to 0.068 enter the right child node. The term "samples is 100%" indicates that the node encompasses the entire training dataset. The value [0.382, 0.618] represents the proportion of samples with category 0 and category 1 in the node, while "class = 1" signifies that the node's predicted category is 1. The left node, a leaf node, lacks characteristic conditions, indicating that no further division is required. A gini index of 0 signifies that the node's impurity is 0, and all samples belong to the same category. The right node is further divided based on the sample's "*Crop_Biomass_Aerial*" feature value: if the sample's *Crop_Biomass_Aerial* is less than 0.107, it enters the left child node; if it is greater than or equal to 0.107, it enters the right child node. The sample's category 0 proportion is approximately 35.3%, while category 1 accounts for about 64.7%, hence the node's predicted category is 1. This division process continues until the entire Decision Tree is divided, with each node assigning samples to different sub-nodes based on feature conditions and predicting the final category. Thus, the Decision Tree can determine the sample's category based on the feature value, providing an intuitive method to explain the data flow understanding process.

Although Decision Trees inherently possess a certain degree of explainability, in ensemble tree models such as Random Forests and XGBoost, explainability diminishes when the classification results of multiple Decision Trees are combined. This is due to the introduction of randomness and complex optimization processes during the construction of each Decision Tree by Random Forests and XGBoost, resulting in a single Decision Tree's weakened explainability. Consequently, ensemble learning models are still considered "black box models" with some explainability deficiencies. To provide more detailed explanations, we can utilize explainability methods, one of the most commonly used being SHapley Additive Explanations (SHAP), proposed by Lundberg and Lee in 2017 [24]. Based on the concept of Shapley value in game theory, SHAP quantifies the contribution of different features to the prediction result, revealing the degree of influence of each feature on the prediction result.

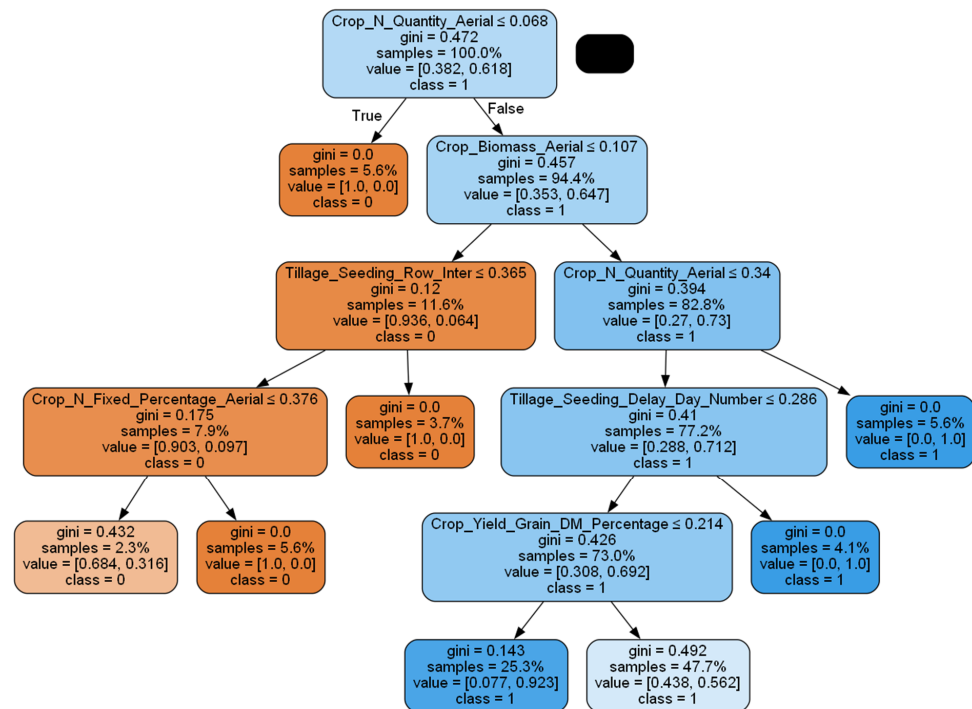


Figure 2. Decision Trees visualization: orange for class 0, blue for class 1.

For instance, in the present Decision Tree, the primary important feature is identified as “Crop_N_Quantity_Aerial”. Its location at the root node underscores its pivotal role in the classification process. The succeeding crucial features include “Crop_Biomass_Aerial”, “Tillage_Seeding_Row_Inter”, “Crop_N_Quantity_Aerial”, “Crop_N_Fixed_Percentage_Aerial”, “Tillage_Seeding_Delay_Day_Number”, and “Crop_Yield_Grain_DM_Percentage”. These features, as illustrated in the SHAP summary graph of the crop, rank within the top 20 in terms of importance. However, there exists a discrepancy between their order of feature importance and their ranking within the Decision Tree. This inconsistency primarily arises from the disparate methodologies employed by Decision Trees and the SHAP summary graph in evaluating feature importance. While Decision Trees prioritize the splitting capability of features, SHAP values quantify the average contribution of features to the anticipated outcome. Consequently, a feature might be deemed important in a Decision Tree owing to its effectiveness in segregating the sample into purer subsets, yet it might not garner equivalent importance in a SHAP summary graph if its overall impact on the prediction result is not substantial. Furthermore, the random forest model comprises multiple Decision Trees, with each tree being trained on a subset of randomly selected data and features. Therefore, the feature importance ascertained from a single Decision Tree might not wholly represent the feature importance of the entire random forest model. In contrast, SHAP values are calculated based on the entirety of the model—in other words, all trees are taken into account. Therefore, the feature importance represented by the SHAP value incorporates all data and features and, hence, presents a more holistic reflection of the model’s behavior.

3.2. Feature Importance Analysis

The SHAP method, considering the combination of different feature values, computes the average contribution of each feature to the model prediction and interprets this in relation to the influence of feature values. The SHAP method is versatile, applicable to a spectrum of model types ranging from traditional linear models to complex black box models. It not only elucidates the prediction results of a single sample but also provides a ranking and explanation of the importance of global features. This aids users in gaining a profound understanding of the model’s decision-making process and prediction results.

In this study, we utilized the SHAP value to analyze the salient feature information of the training model. The central premise of SHAP feature importance is straightforward: a feature with a larger absolute Shapley value holds significant importance. To ascertain global importance, we computed the average of the absolute Shapley values for each feature as shown in Equation (5), where I_j represents the total contribution measure of feature j to the model's prediction results; $\phi_j^{(i)}$ denotes the SHAP value of feature j for the i -th sample; and n is the total number of samples.

$$I_j = \sum_{i=1}^n |\phi_j^{(i)}| \quad (5)$$

Taking the feature importance map of Chickpea as an example, Figure 3 displays the results of the SHAP values sorted in descending order post RF model training. This figure illustrates the 20 important features selected from a total of 38 features. By comparing the SHAP values of the different features, it becomes evident that these important features exert a substantial impact on the model output. Among them, the aerial biomass of crops, the precipitation at the experimental point, and the longitude of the experimental point are the three features that most significantly influence the yield of Chickpea. The concordance of these findings with the SHAP summary graph (refer to Figure 4a) serves to reinforce the significance of these features within the model.

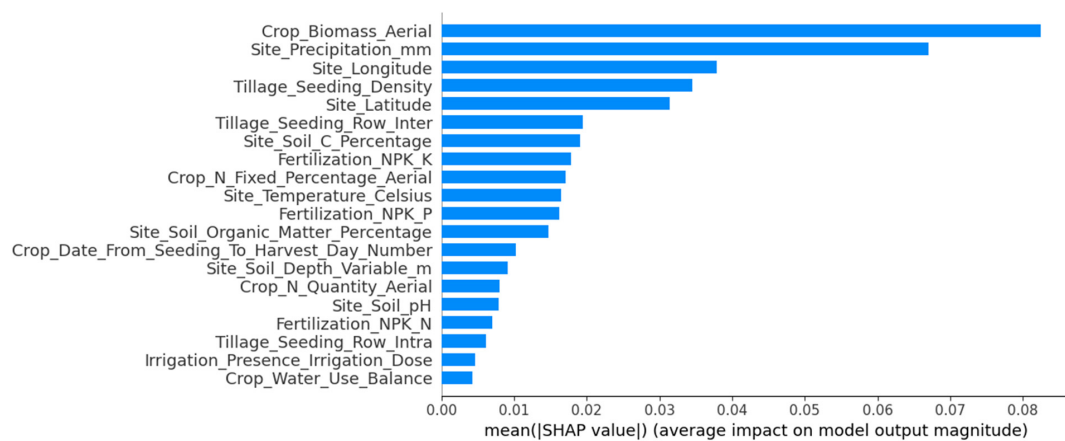
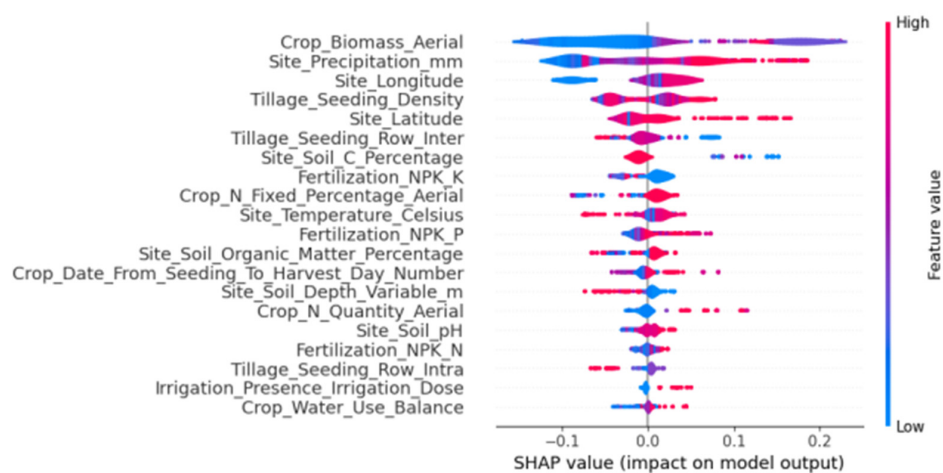
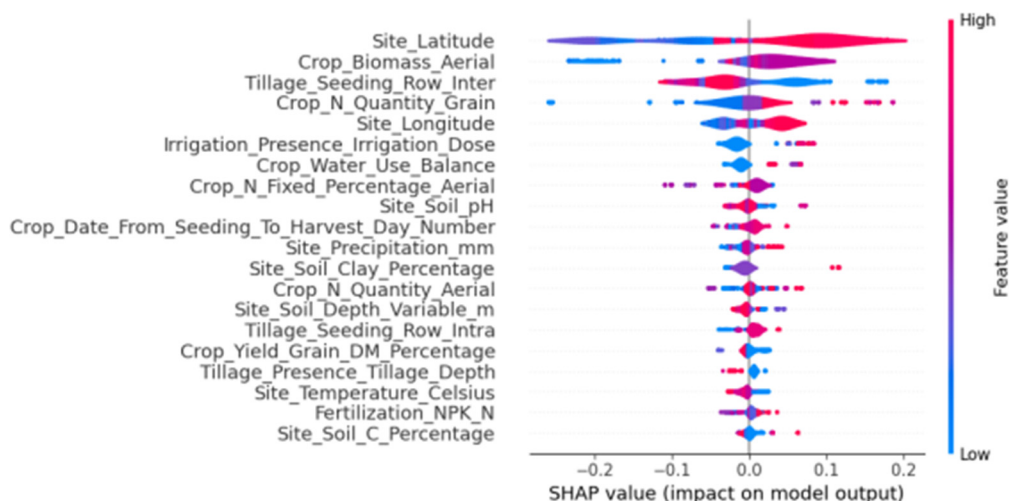


Figure 3. Feature importance analysis.

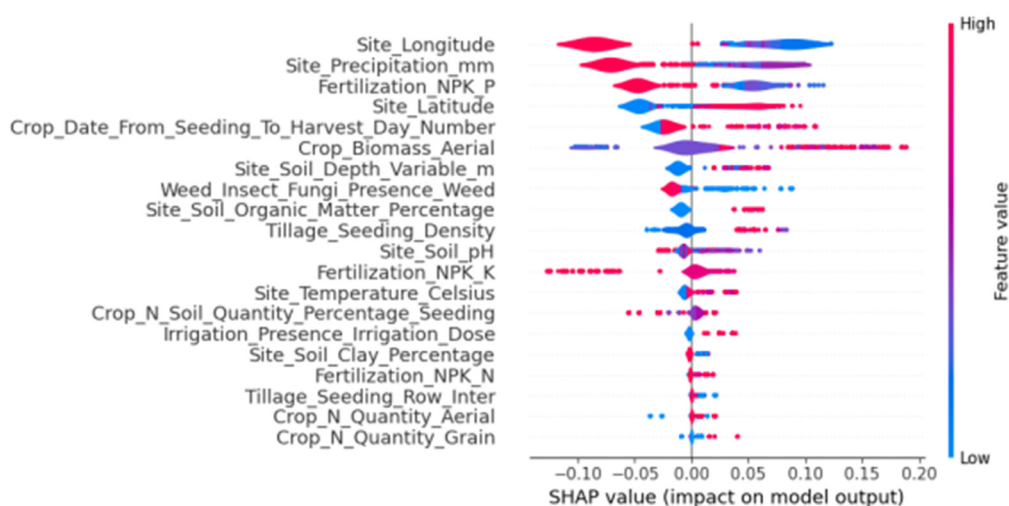


(a) Chickpea

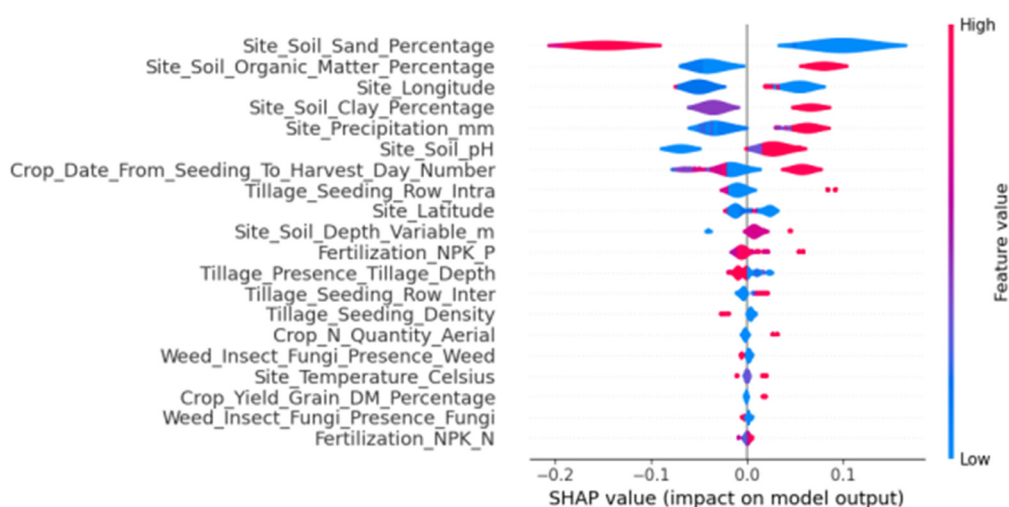
Figure 4. Cont.



(b) Cowpea

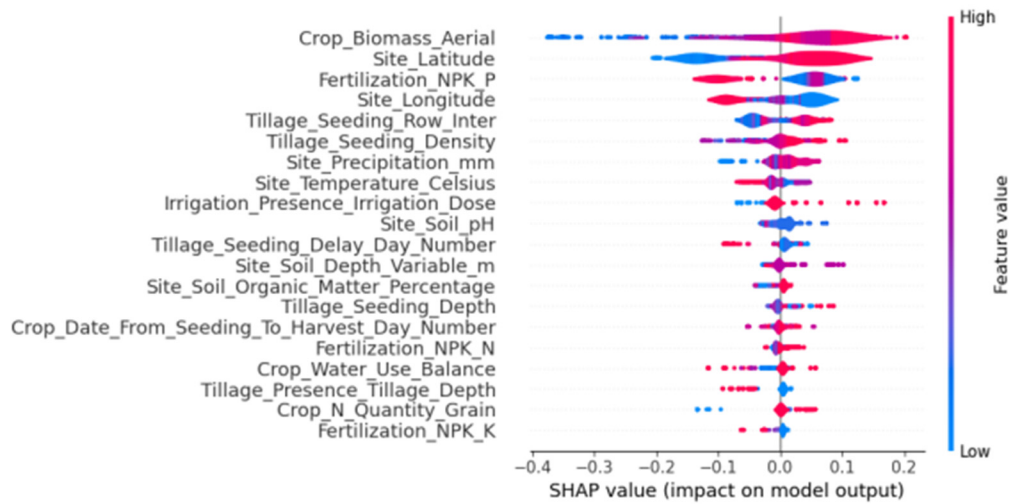


(c) Garden_pea

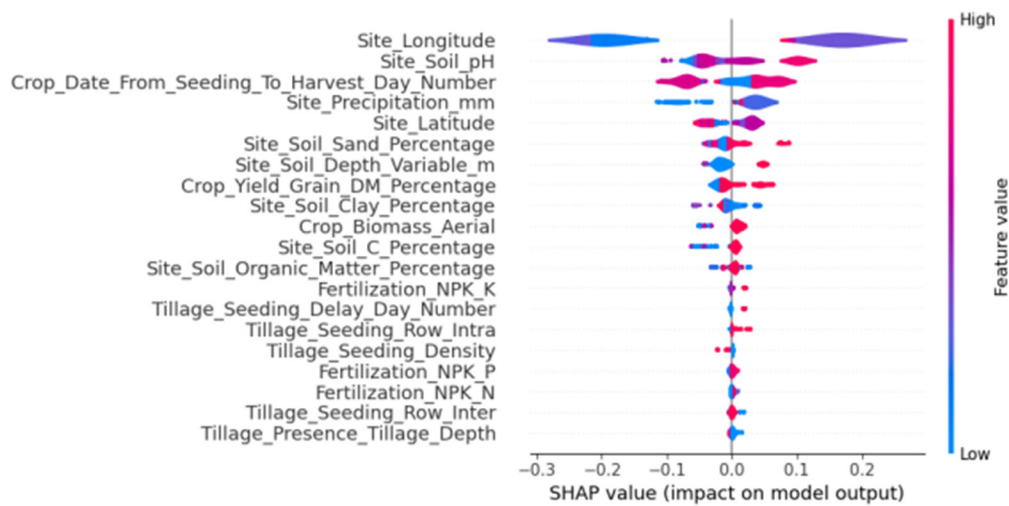


(d) Faba bean

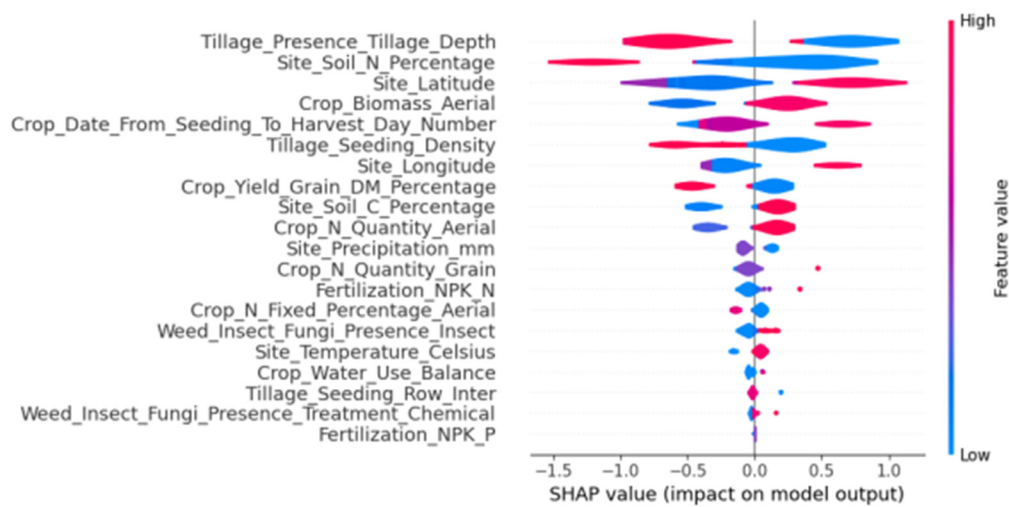
Figure 4. Cont.



(e) Lentil



(f) Peanut



(g) Pigeon pea

Figure 4. Cont.

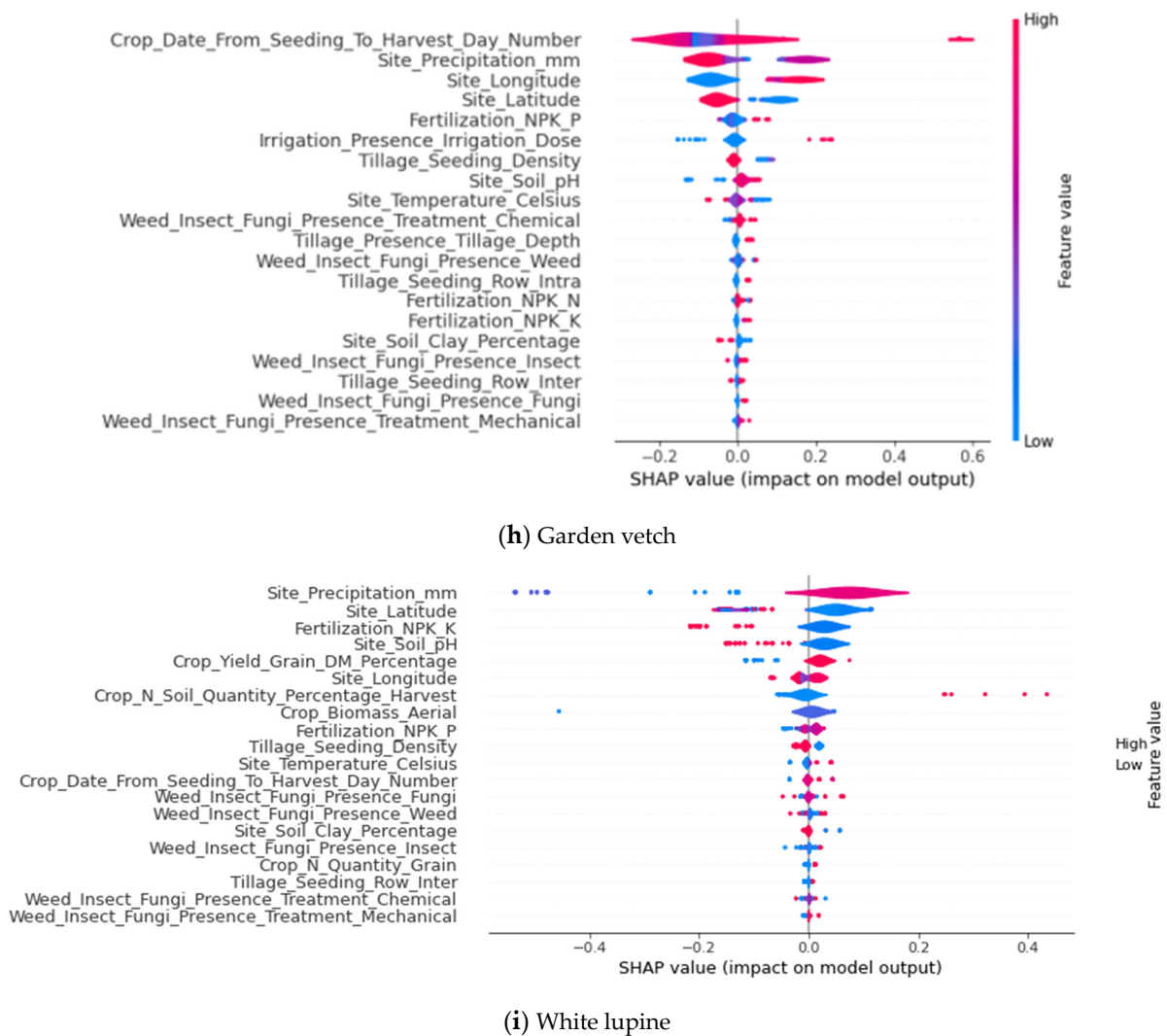


Figure 4. SHAP summary graph: x-axis shows feature contributions; line thickness reflects sample size; color transition indicates feature value change.

3.3. SHAP Graph Analysis

In an effort to further clarify the positive and negative relationships between significant features and model output results, we utilized the SHAP summary graph for analysis in this study. The SHAP summary graph amalgamates feature importance with feature effects, where each point on the summary graph represents a Shapley value for a feature and an instance. The feature determines the position on the y-axis, while the Shapley value determines the position on the x-axis. Moving rightward from the origin, a positive Shapley value indicates a positive contribution of the feature to the positive prediction result. The further to the right, the greater the contribution, and vice versa. The thickness of the line corresponds to the sample size, with a thicker line indicating a larger sample size. The color transition from blue to red represents the change in the feature value from small to large.

The graph depicting chickpea yield (Figure 4a) indicates that precipitation is the second most influential factor. An increase in precipitation generally results in a positive impact on chickpea yield, as it is a crucial environmental factor for plant growth. Adequate precipitation supplies the plant with necessary moisture, thereby promoting growth and development. The longitude of the experimental site also positively influences chickpea yield, with higher longitudes typically correlating with extended sunshine hours, particularly in the summer. Sufficient sunlight promotes ample photosynthesis in chickpeas, leading

to the synthesis and accumulation of more organic matter, thereby enhancing biomass accumulation and yield.

The cowpea graph (Figure 4b) demonstrates that higher latitudes yield greater harvests of cowpea, a summer crop that completes its life cycle within a brief growing season. Consequently, cowpea is a suitable crop for regions with shorter growing seasons, especially in high latitudes or monsoon regions. Additionally, optimal seeding spacing can enhance cowpea yield by optimizing plant density through inter-plant interaction and resource competition. The nitrogen content of cowpea crops positively correlates with yield, as legumes can symbiotically interact with rhizobia, which fix atmospheric nitrogen into a usable form. This self-sufficient nitrogen supply increases leaf area and photosynthetic intensity, thereby improving the conversion of light energy into biomass and yield.

The garden pea graph (Figure 4c) indicates a negative correlation between yield and both site longitude and precipitation. For garden peas, extended sunshine hours may result in excessive transpiration and water evaporation, negatively impacting plant growth and yield. Higher precipitation can result in overly saturated soils, especially in areas with poor drainage. Excessive moisture can adversely affect garden pea root health and growth, increase water saturation in the soil, limit oxygen access to the roots, and lead to root suffocation and rot. These factors can impede normal plant growth, thereby affecting yield.

The yield of faba beans (Figure 4d) negatively correlates with the percentage of sand in the soil, as faba beans require a relatively high water supply, particularly during the growing season and the flowering and fruiting periods. A higher sandy soil content may cause the soil to be overly permeable, leading to rapid water drainage and making it difficult to maintain an effective water supply, thereby affecting normal plant growth and yield. A higher percentage of soil organic matter may positively impact the yield of faba beans. Organic matter is a crucial component of soil and plays a significant role in plant growth and development. An appropriate amount of organic matter can enhance the soil's water retention, fertilizer retention, and nutrient supply capacity, which benefits the root development and nutrient absorption of plants.

The lentil graph (Figure 4e) shows that the latitude of the experimental sites positively impacts yield, with higher latitudes generally having shorter growing seasons. Lentils are early maturing crops with a relatively short growing period, typically around 70 to 120 days, and can complete their life cycle within a brief growing season. Additionally, the yield of lentils negatively correlates with the dosage of phosphorus fertilizer. This may be due to an excess of phosphorus fertilizer: although phosphorus is a crucial nutrient for plant growth and development, a high concentration of phosphorus fertilizer may disrupt the balance of the root system environment, interfering with the plant's root system's ability to absorb water and other nutrients, thereby affecting plant growth and yield formation.

The analysis of the peanut yield (Figure 4f) graph indicates a positive correlation with longitude. As a heliophilous plant, peanuts require ample sunlight for optimal growth and development. However, the impact of pH on the predicted results does not exhibit clear directionality. This suggests that the contribution of this feature to the predicted result does not have a definitive positive or negative relationship. It may exert a minor influence on the result, or its effects may be counterbalanced by other features. Furthermore, the duration from sowing to harvest exerts a bidirectional effect on peanut yield. Extended growth periods may have either negative or positive effects on yield, indicating a non-linear or non-monotonic relationship between yield and the growth period. This could be attributed to the strong interaction between this feature and other related features, causing changes within a certain value range to exert a minor impact on the model's output. Therefore, to enhance peanut yield, it is advisable to cultivate in regions with ample sunlight while considering the influence of other related factors on the growth cycle. This can aid in optimizing planting strategies and yield prediction.

The pigeon pea (Figure 4g) yield graph demonstrates a negative correlation with both tillage depth and soil nitrogen percentage, implying that the soil tillage depth and nitrogen content in the soil adversely impact pigeon pea yield. Excessive tillage depths may disrupt

soil structure and impede root development. Overly deep tillage may damage soil structure, disintegrate soil aggregates, lead to soil compaction, and decrease aeration, thereby constraining the growth and development of pigeon pea roots. Additionally, elevated soil nitrogen content may result in an oversupply or imbalance of nitrogen. Therefore, to optimize pigeon pea yield, it is essential to maintain an appropriate depth of soil tillage and nitrogen content in the soil. Avoiding excessive deep tillage and nitrogen supply, preserving a stable soil structure, and moderating nitrogen supply will contribute to a conducive growth environment, promote the growth and development of pigeon pea, and consequently enhance yield.

The garden vetch yield (Figure 4h) graph also exhibits a bidirectional effect on yield, similar to peanuts, indicating that garden vetch yield may interact more significantly with other characteristics. Furthermore, the yield of garden vetch is negatively correlated with both precipitation in millimeters and site longitude. Garden vetch, known for its strong adaptability, is a relatively drought-tolerant plant that can thrive under various soil types and water conditions, including drought and poor soil. Garden vetch is also shade-tolerant and exhibits a certain degree of tolerance to low-light conditions, enabling successful growth in some semi-shady areas.

White lupine (Figure 4i) yield is positively influenced by a moderate supply of precipitation, which can provide an adequate water supply. However, the latitude of the experimental site and the dosage of potassium fertilizer exert a negative impact on the yield, suggesting that white lupine requires a longer growing season to complete its life cycle at lower latitudes. Longer growing seasons allow crops more time to grow, accumulate nutrients, and develop yields, contributing to higher yields. Furthermore, white lupine has a relatively low demand for potassium fertilizer, and an appropriate supply can promote healthy plant growth, increase yield, and improve quality.

In the summary graph, we discern the relationship between the top 20 eigenvalues, which exert the most substantial impact on yield, and the predicted outcomes. However, to fully grasp the essence of this relationship, it is necessary to consult the SHAP dependency plot. SHAP feature dependencies provide a clear-cut method for globally interpreting the graphs. It becomes clear from the SHAP summary graph that a higher Biomass_Aerial positively sways the yield of various crops, including Garden_pea, Chickpea, Cowpea, Lentil, Pigeonpea, Peanut, and Narrowleaf_lupine. This suggests that as plants produce more aerial biomass during their growth, their yields correspondingly escalate. Aerial biomass includes the biomass of the stems, leaves, and fruits of plants. The increase in aerial biomass signifies that plants can engage in more photosynthesis, synthesizing and accumulating additional nutrients and organic substances, thereby fostering higher yields. To exemplify this, we consider lentil. By selecting the feature Crop_Biomass_Aerial, which has the most pronounced impact on the yield of lentil, we can plot a point for each data instance. The x-axis represents the feature value, while the y-axis represents the corresponding Shapley value. Figure 5 exhibits the SHAP feature dependence of Crop_Biomass_Aerial, illustrating that as the value of Crop_Biomass_Aerial ascends, the probability of surpassing the global average yield also rises.

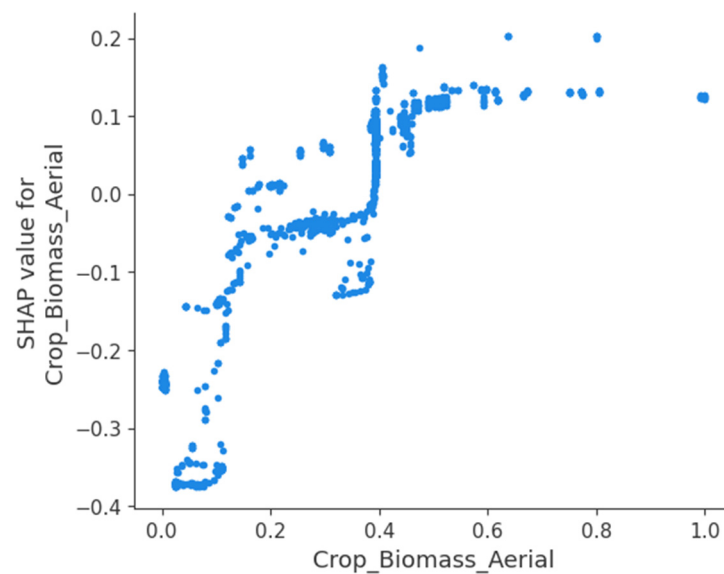


Figure 5. SHAP dependence plot.

4. Conclusions

Grain legumes, as crucial food and protein sources, play an integral role in ensuring global food security. This study offers an exhaustive perspective on understanding the diverse factors that influence grain legume yields. By assessing the yield-influencing factors at a global level, we gain a more profound comprehension of the varied responses of different grain legumes to these factors, alongside discerning the causes and trends of yield fluctuations. These insights serve as a foundation for formulating food security policies and strategies. Utilizing explainable artificial intelligence, this study probes into the factors influencing the yield of nine representative grain legumes. The analysis covers data from varied geographical regions and temporal spans, taking into account the impact of multiple factors on the yields of these legumes. It delineates the responses of different grain legumes to environmental and management factors and identifies key determinants affecting their yields. For example, chickpea yield exhibits positive correlations with precipitation and sunshine duration, suggesting its adaptability to environmental factors. Cowpea demonstrates higher productivity at elevated latitudes, indicative of its ability to complete its life cycle within a brief growing season. Garden pea and faba bean yields are significantly impacted by soil conditions, such as moisture content, organic matter content, and soil type. Lentil and peanut yields have positive correlations with latitude and longitude, underscoring the significance of their growing seasons and dependence on photosynthesis. Peanut yield demonstrates a bidirectional effect with the duration from sowing to harvest, showcasing its adaptability to environmental factors. Similarly, pigeon pea yield is influenced by soil tillage depth and soil nitrogen content, emphasizing the necessity of appropriate tillage and soil nutrient management. However, white lupine yield is contingent upon a moderate supply of precipitation and soil potassium fertilizer dosage, reflecting its sensitivity to environmental and soil nutrient conditions.

From the perspective of the factors affecting the yield of nine grain legumes, this study has ascertained that an array of environmental and management factors, encompassing aerial biomass, precipitation, geographical coordinates (latitude and longitude), soil conditions, growth cycle, and fertilizer application, collectively influence the yield of these legumes. An augmentation in aerial biomass is associated with a boost in crop yield; optimal levels of precipitation cater to the requisite hydration needs of the crops; the geographical location (latitude and longitude) determines the sunshine time and the length of the growing season, indirectly modulating photosynthesis and the accumulation of biomass. Soil conditions, inclusive of nitrogen content, organic matter content, soil type, and tillage depth, wield influence over root development and nutrient absorption; the

span of the growth cycle dictates biomass accumulation; proper fertilizer use can provide sufficient nutrients to promote crop growth. The interaction of these factors determines the yield of grain legumes. In future planting management, it is necessary to fully consider these factors and optimize planting strategies to meet the challenges brought about by climate change and ensure the stable yield of grain legumes.

In future research, we can delve deeper into the specific mechanisms of action of these factors and examine the interrelationships between different factors. Additionally, we can consider employing more advanced explainable artificial intelligence models to enhance our ability to explain grain legume yield. Furthermore, tailored planting strategies can be studied and formulated based on specific regions and planting conditions to optimize yield and quality. For instance, through efficient irrigation management and strategic fertilization measures, we can provide the appropriate water and nutrient supply to enhance the growth and yield of grain legumes. In summary, the study of factors impacting global grain legume yield through the application of explainable artificial intelligence enables us to identify and quantify the key influencers of grain legume yield. These analytical findings serve as valuable references for agricultural management and policy formulation, aiding in the improvement of crop production practices, adjustment of planting strategies, and optimization of resource utilization. Ultimately, these efforts contribute to increasing global grain legume yields, promoting sustainable agricultural development, and boosting global food production.

Author Contributions: Supervision: J.Y.; Conceptualization: Y.L.; Methodology: Y.L. and R.J.; Software: Y.L., J.C. and Y.W.; Validation: R.L., M.W. and Y.L.; Research: R.L. and Y.L.; Data Curation: Y.L., M.W., Y.W. and J.C.; Writing—Original Draft Preparation: Y.L. and R.J.; Writing—Review and Editing: R.J. and J.Y.; Visualization: Y.L.; Supervision: J.Y.; Project Administration: Y.L.; Funding Acquisition: J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Project of Yunnan Science and Technology under Project No. 202302AE09002003. The article processing charges (APC) were funded by the Major Project of Yunnan Science and Technology.

Institutional Review Board Statement: Since this study is mainly based on data analysis of factors affecting global cereal legume production and the application of interpretable artificial intelligence models and does not involve direct participation of humans or animals, this study does not require ethical review. The study was designed and conducted in compliance with relevant regulatory and ethical guidelines.

Data Availability Statement: The dataset utilized in this study is derived from the global grain legume experiment dataset published on Scientific Data by Charles Cernay et al. <https://datadryad.org/stash/dataset/doi:10.5061/dryad.mf42f> (accessed on 5 February 2024).

Acknowledgments: The authors acknowledge the provision of the global grain legume experiment dataset by Charles Cernay et al.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Whittaker, J.; Nyiraneza, J.; Zebbarth, B.J.; Jiang, Y.; Burton, D.L. The effects of forage grasses and legumes on subsequent potato yield, nitrogen cycling, and soil properties. *Field Crops Res.* **2023**, *290*, 108747. [\[CrossRef\]](#)
2. Dela, M.; Shanka, D.; Dalga, D. Biofertilizer and NPSB fertilizer application effects on nodulation and productivity of common bean (*Phaseolus vulgaris* L.) at Sodo Zuria, Southern Ethiopia. *Open Life Sci.* **2023**, *18*, 20220537. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Rawal, V.; Navarro, D.K. (Eds.) *The Global Economy of Pulses*; FAO: Rome, Italy, 2019. [\[CrossRef\]](#)
4. Salar, F.-A.; Shahram, T.; Ruijun, Q.; Christos, N.; Yanyan, L.; Suduan, G. Biochar effects on yield of cereal and legume crops using meta-analysis. *Sci. Total Environ.* **2021**, *775*, 145869.
5. Ranjan, R.; Chandel, A.K.; Khot, L.R.; Bahlol, H.Y.; Zhou, J.; Boydston, R.A.; Miklas, P.N. Irrigated pinto bean crop stress and yield assessment using ground based low altitude remote sensing technology. *Inf. Process. Agric.* **2019**, *6*, 502–514. [\[CrossRef\]](#)
6. Meraj, G.; Kanga, S.; Ambadkar, A.; Kumar, P.; Singh, S.; Farooq, M.; Johnson, B.; Rai, A.; Sahu, N. Assessing the Yield of Wheat Using Satellite Remote Sensing-Based Machine Learning Algorithms and Simulation Modeling. *Remote Sens.* **2022**, *14*, 3005. [\[CrossRef\]](#)

7. Boote, K.J.; Gerrit, H.; Srinivasulu, A.; Curtis, A.; Rajan, S.; Francis, M.R.; Kumar, H.S.; Kulbhushan, G.; Sangu, A. Adapting the CROPGRO model to simulate growth and yield of guar, *Cyamopsis tetragonoloba* L, an industrial legume crop. *Ind. Crops Prod.* **2023**, *197*, 116596. [\[CrossRef\]](#)
8. Zhou, Z.; Morel, J.; Parsons, D.; Kucheryavskiy, S.V.; Gustavsson, A.-M. Estimation of yield and quality of legume and grass mixtures using partial least squares and support vector machine analysis of spectral data. *Comput. Electron. Agric.* **2019**, *162*, 246–253. [\[CrossRef\]](#)
9. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2019. Available online: <https://christophm.github.io/interpretable-ml-book/> (accessed on 10 June 2023).
10. Mehrdad, A. AI explainability framework for environmental management research. *J. Environ. Manag.* **2023**, *342*, 118149.
11. Wang, K.; Tian, J.; Zheng, C.; Yang, H.; Ren, J.; Liu, Y.; Han, Q.; Zhang, Y. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Comput. Biol. Med.* **2021**, *137*, 104813. [\[CrossRef\]](#)
12. Moreno-Sanchez, P.A. Improvement of a prediction model for heart failure survival through explainable artificial intelligence. *arXiv* **2021**, arXiv:2108.10717. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Cernay, C.; Pelzer, E.; Makowski, D. A global experimental dataset for assessing grain legume production. *Sci. Data* **2016**, *3*, 160084. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Statistics Division of Food and Agriculture Organization of the United Nations (FAOSTAT). Available online: <http://www.fao.org/faostat/> (accessed on 1 June 2013).
15. Daewoon, S.; Euimin, L.; Yoonseok, K.; Sangyeob, H.; Jaeyul, L.; Mansik, J.; Jeehyun, K. Three-dimensional reconstructing undersampled photoacoustic microscopy images using deep learning. *Photoacoustics* **2023**, *29*, 100429.
16. Zhen, W.; Li, Z.; Lei, Z. Minority-prediction-probability-based oversampling technique for imbalanced learning. *Inf. Sci.* **2023**, *622*, 1273–1295.
17. Batista, G.; Prati, T.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explor. Newsl.* **2004**, *6*, 20–29. [\[CrossRef\]](#)
18. Lin, X.; Wu, Z.; Chen, J.; Huang, L.; Shi, Z. A Credit Scoring Model Based on Integrated Mixed Sampling and Ensemble Feature Selection: RBR_XGB. *J. Inf. Technol.* **2022**, *23*, 1061–1068. [\[CrossRef\]](#)
19. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
20. Chen, T.Q.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794.
21. Zheng, H.-L.; An, S.-Y.; Qiao, B.-J.; Guan, P.; Huang, D.; Wu, W. A data-driven interpretable ensemble framework based on tree models for forecasting the occurrence of COVID-19 in the USA. *Environ. Sci. Pollut. Res.* **2023**, *30*, 13648–13659. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 507–520.
23. Ammar, M.; Rania, K. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud Univ. Comput. Inf. Sci.* **2023**, *35*, 757–774.
24. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.