QUBAHAN ACADEMIC JOURNAL
VOL. 4, NO. 3, September 2024
https://doi.org/10.48161/qaj.v4n3a571

A Comparison of Classification Algorithms for Predicting Distinctive Characteristics in Fine Aroma Cocoa Flowers Using WEKA Modeler

Daniel Tineo [1,2*], Yuriko S. Murillo [3], Mercedes Marín [1], Darwin Gomez [1], Víctor H. Taboada [1], Malluri Goñas [1,2] and Lenin Quiñones Huatangari [4]

[1] Yanayacu Experimental Center, Supervision and Monitoring Directorate at Agricultural Experimental Stations, National Institute of Agricultural Innovation (INIA), Jaén San Ignacio Highway KM 23.7, Jaén 06801, Cajamarca, Peru;
[2] Institute for Research on Sustainable Development of the Ceja de Selva (INDES-CES), National University Toribio Rodríguez de Mendoza, Chachapoyas 01001, Amazonas, Peru;
[3] Biology Laboratory, Department of Basic and Applied Sciences, National University of Jaén, Jaén 00000, Peru;
[4] Institute for Data Science Research, Engineering, National University of Jaén, Jaén 00000, Peru.

Corresponding author: e-mail: dt.infolab@gmail.com.

ABSTRACT: The expression of crop functional traits is influenced by environmental and management conditions, which in turn is reflected in genetic diversity. This study employed a data mining approach to determine the functional traits of flowers that influence cocoa diversity. A total of 1,140 flowers from 228 trees were utilized in this study, with 177 representing fine aroma cocoa trees and 51 trees belonging to other commercial cultivars. Three attribute evaluators (InfoGainAttributeEval, CorrelationAttributeEval and GainRatioAttributeEval), and six algorithms (Naive Bayes, Multinomial Logistic Regression, J48, Random Forest, LTM and Simple Logistic) were employed in this study. The findings indicated that the GainRatioAttributeEval attribute generator was the most efficacious in discerning the functional trait in cocoa diversity flowers. The algorithms Simple Logistic and LMT were the most accurate and specific, while Naive Bayes was the most efficient in terms of computational complexity for model building. This research provides a comprehensive overview of the use of machine learning to analyze functional traits of flowers that most influence cocoa genetic diversity. It also highlights the need to further improve these models by integrating additional techniques to increase their efficiency and extend the data mining approach to other agricultural sectors.

Keywords: Algorithms, data mining, functional traits, machine learning, Theobroma cacao.

I. INTRODUCTION

The expression of crop functional traits (morphological, chemical, physiological and phenological properties) is influenced by environmental and management conditions, which in turn influence plant yield [1, 2] and plant-soil interactions [3]. Among the environmental conditions, changes in temperature and photoperiod have the greatest influence on the intraspecific variation of individuals of the same species [1, 3]. In contrast, fertilization is a key factor in the management conditions influencing intraspecific variation [3]. Conversely, the use of artificial selection also influences the expression of functional traits in diverse crops versus their wild ancestors [4]. In addition to functional traits, it is important to mention genetic differentiation that is caused by the local adaptation of geographically separated subpopulations of a species. This has led to the adaptation of individuals with different functional traits, including those observed in leaves, flowers, and fruits [5, 6]. These patterns of intraspecific diversity in numerous crop and wild species are a reflection of their distribution during the last glaciation period [7]. Consequently, it is probable that the impact of these climatic conditions has also influenced the distribution of cocoa (Theobroma cacao L.), confining it to a series of geographically and genetically isolated refugia [5]. This has resulted in numerous cocoa populations being constrained, thereby generating a high diversity of genetic groups [8].

The genus Theobroma L. (Malvaceae) has been the subject of numerous studies, which have yielded insights into various patterns of reproduction, gene flow and speciation. Additionally, the genus is of economic interest, particularly in the context of cultivars of T. cacao species [9]. The floral organization of T. cacao is conserved throughout the genus Theobroma yet exhibits a wide range of morphological characteristics as a natural process. This is because diverse wild cocoa populations exhibit high floral phenotypic plasticity, a high degree of vegetative reproduction, and different pollinators [7, 10, 11]. However, the actual distribution and diversity of cocoa is influenced mas for activity human rather than by natural processes [8], as there are higher recombination rates in domesticated plants (criollo population) and as a result of this, greater diversification and variability of the flowers of various cocoa individuals that adapt to adverse climatic conditions [1, 6]. The rapid diversification and distribution of cocoa has become a challenge for the differentiation of individuals from different cocoa populations, especially if they have been distributed in geographical areas with different conditions of humidity, precipitation, and temperature. These environmental factors influence said diversity, which entails that the farmers who are generally open to innovation and new technologies, must adapt to the changing conditions [12].

II. LITERATURE REVIEW

Morphological methods are currently employed to distinguish between distinct cocoa populations, relying solely on the functional traits of flowers, leaves, and fruits. However, this necessitates expertise in cultivation, even for professionals who collect a substantial amount of data that is challenging to analyze using traditional approaches. Currently, this type of analysis is conducted through data mining, which employs a variety of models to generate robust and efficient results [13]. Furthermore, techniques of combining several models are employed to enhance the accuracy of the results [13]. This approach is exemplified by the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology [14], which was developed to facilitate the utilization of any tool and the structuring of any data mining problem [15]. This technique has been employed to direct data mining (DM) operations and decision support in business intelligence (BI) by identifying pivotal performance indicators (KPIs) for cocoa production and marketing [16]. Furthermore, it is employed to classify the quality of cocoa beans during the fermentation process, distinguishing between well-fermented and over-fermented beans with an accuracy value of 92.50% [17]. Additionally, it is utilized to analyze opportunities and standardize cocoa processing based on data from omics studies conducted globally [18].

Another application of these techniques is their use in predicting sensory qualities of cocoa using optimal models of the Simple Linear Regression data mining algorithm from WEKA (Waikato Environment for Knowledge Analysis). The application of machine learning technologies has enabled the prediction of cocoa sensory quality with minimal error, in addition to the analysis of (-)-epicatechin and flavonoids [19]. Therefore, it is evident that data mining is being used in other sectors, including "e-commerce," to improve decision-making processes or generic processes [15]. It is likely that these technological advances will continue to provide comprehensive data sets with sophisticated algorithmic solutions that will enable better estimation and decision-making for crops [20], including the cultivation of cocoa, to estimate yields and other genetic diversity parameters [21].

Therefore, the cocoa improvement program at the Yanayacu Experimental Center, Jaén, is of importance for the future development of the cocoa industry in the department of Cajamarca, northeastern Peru. Under this context, this innovative study focused on examining functional traits in fine aroma cocoa flowers using data mining to determine functional traits in flowers that influence cocoa diversity. There is currently no scientific evidence regarding data mining studies on functional traits in cocoa flowers. Consequently, this work provides vital information about the collection of fine aroma cocoa, which in turn aids in the development of genetic improvement.

III. MATERIAL AND METHOD

1. STUDY AREA AND SITE DESCRIPTION

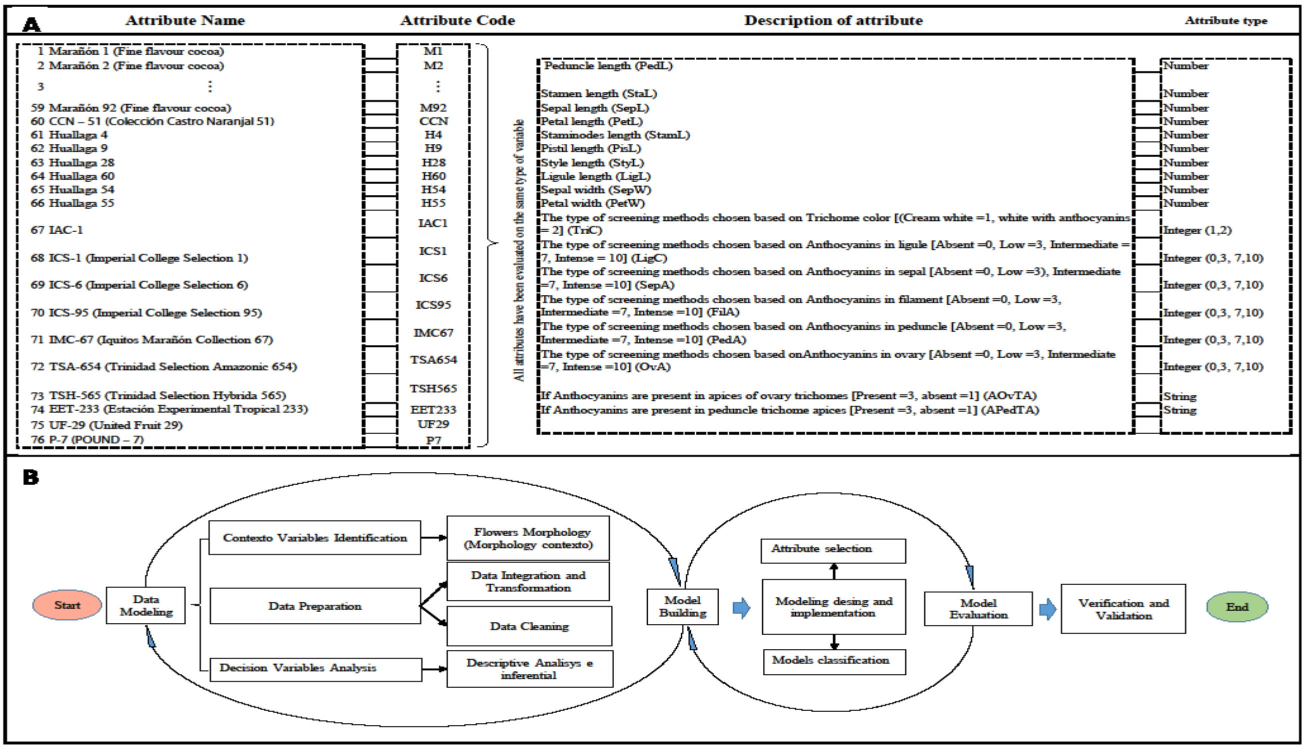
The study was conducted at the Yanayacu Experimental Center, situated in the province of Jaén, Cajamarca (5°40'35.99'' North and 78°46'27.05'' West), within an area of 5 hectares of cocoa plantations. The Experimental Center is situated at an altitude of approximately 618 meters above sea level, with temperatures reaching a maximum of 29°C to 33°C during the day and a minimum of 19°C to 23°C at night. The annual precipitation ranges from approximately 900 mm to 1200 mm, according to the National Meteorology and Hydrology Service, SENAMHI [22].

2. ACQUISITION OF THE DATA SET

A total of 228 cacao trees, approximately 30 years old, were subjected to analysis. The 228 cacao trees were found to correspond to 76 accessions. Of these, 59 accessions were identified as fine aroma cacao (Cashew genetic group), while the remaining 17 accessions were other cultivable cacao clones (Table S1). For each cacao accession, three trees were sampled, and five flowers per tree were analyzed, resulting in a total of 1,140 flowers. Floral buds were selected daily between 8:00 and 9:00 a.m. to avoid loss of turgidity and other morphological characteristics until they were transported to the biology laboratory of the National University of Jaén. Each floral structure was extracted with tweezers and placed on the slide of a LABOMED LUXEO 6Z binocular stereoscopic microscope, on which a fraction of a millimetered leaf was previously placed to be photographed. Image J software (https://imagej.net/ij/download.html) was employed to quantify the length of the peduncle (mm), stamens (mm), sepal (mm), petal (mm), staminodes (mm), pistil (mm), style (mm), and ligule (mm). Furthermore, the width of the sepal and petal of each flower was quantified. To gather further data on the various characteristics of each flower, the color of the trichomes, the presence of anthocyanin in the sepal, filament, ligule, ovary, peduncle, trichome apex of the ovary, and trichome apex of the peduncle were analyzed (S2 Table). Consequently, the dataset comprised 18 attributes and 228 instances.

3. DATA MINING PROCESS

The classification algorithms were executed using the open-source data mining tool WEKA (https://sourceforge.net/projects/weka/). The analyses were conducted in two groups. The first group consisted of samples of fine aroma cocoa (59 Marañón cocoa accessions), while the second group was formed by 17 cultivable cocoa clones and the 59 fine aroma cocoa accessions (S1 Table). The study variables were adapted according to the model proposed by Altaleb et al. [15], with modifications for the functional traits of cocoa flowers in this study. The process was conducted in three phases. The initial phase was designated as variable selection.



Attribute Name	Attribute Code
1 Marañón 1 (Fine flavour cocoa)	M1
2 Marañón 2 (Fine flavour cocoa)	M2
3 ⋮	⋮
59 Marañón 92 (Fine flavour cocoa)	M92
60 CCN – 51 (Colección Castro Naranjal 51)	CCN
61 Huallaga 4	H4
62 Huallaga 9	H9
63 Huallaga 28	H28
64 Huallaga 60	H60
65 Huallaga 54	H54
66 Huallaga 55	H55
67 IAC-1	IAC1
68 ICS-1 (Imperial College Selection 1)	ICS1
69 ICS-6 (Imperial College Selection 6)	ICS6
70 ICS-95 (Imperial College Selection 95)	ICS95
71 IMC-67 (Iquitos Marañón Collection 67)	IMC67
72 TSA-654 (Trinidad Selection Amazonic 654)	TSA654
73 TSH-565 (Trinidad Selection Hybrida 565)	TSH565
74 EET-233 (Estación Experimental Tropical 233)	EET233
75 UF-29 (United Fruit 29)	UF29
76 P-7 (POUND – 7)	P7

All attributes have been evaluated on the same type of variable

Description of attribute	Attribute type
Peduncle length (PedL)	Number
Stamen length (StaL)	Number
Sepal length (SepL)	Number
Petal length (PetL)	Number
Staminodes length (StamL)	Number
Pistil length (PisL)	Number
Style length (StyL)	Number
Ligule length (LigL)	Number
Sepal width (SepW)	Number
Petal width (PetW)	Number
The type of screening methods chosen based on Trichome color [(Cream white =1, white with anthocyanins = 2] (TriC)	Integer (1,2)
The type of screening methods chosen based on Anthocyanins in ligule [Absent =0, Low =3, Intermediate = 7, Intense = 10] (LigC)	Integer (0,3, 7,10)
The type of screening methods chosen based on Anthocyanins in sepal [Absent =0, Low =3), Intermediate =7, Intense =10] (SepA)	Integer (0,3, 7,10)
The type of screening methods chosen based on Anthocyanins in filament [Absent =0, Low =3, Intermediate =7, Intense =10] (FilA)	Integer (0,3, 7,10)
The type of screening methods chosen based on Anthocyanins in peduncle [Absent =0, Low =3, Intermediate =7, Intense =10] (PedA)	Integer (0,3, 7,10)
The type of screening methods chosen based onAnthocyanins in ovary [Absent =0, Low =3, Intermediate =7, Intense =10] (OvA)	Integer (0,3, 7,10)
If Anthocyanins are present in apices of ovary trichomes [Present =3, absent =1] (AOvTA)	String
If Anthocyanins are present in peduncle trichome apices [Present =3, absent =1] (APedTA)	String

B
Start
Data Modeling
Contexto Variables Identification
Flowers Morphology (Morphology contexto)
Data Preparation
Data Integration and Transformation
Data Cleaning
Decision Variables Analysis
Descriptive Analisys e inferential
Model Building
Attribute selection
Modeling desing and implementation
Models classification
Model Evaluation
Verification and Validation
End

FIGURE 1. A: Attributes of the dataset for detecting functional traits in fine aroma cocoa flowers. B: Structure of the proposed data mining process model for the domain of detecting functional traits in cocoa flowers

In this study, the first step was to identify the study variables (attributes). This was followed by data preparation and analysis of variable importance. The second phase, data preparation, entailed integrating the requisite data from contextual variables. During this phase, data cleansing was conducted to guarantee the integrity of subsequent steps and phases. This entailed standardizing the number of flowers per tree and ensuring the accurate assignment of attribute codes for each measurement parameter. The third phase was data modeling, which involved the use of both descriptive and inferential data. In this phase, the output variable, which was the code for each cocoa accession (S1 Table), was defined. The second phase of the process was the selection of attributes and the classification of models. In this phase, data preprocessing was conducted. The third and final phase was the verification and validation of each classifier, with particular attention paid to those that demonstrated the most promising performance (Figure 1).

4. ELECTION OF ATTRIBUTES

The selection of attributes influencing the morphological diversity of cocoa flowers was performed using attribute evaluators *InfoGainAttributeEval*, *CorrelationAttributeEval*, and *GainRatioAttributeEval* through the Ranker search method. For each attribute evaluator, a 10-fold cross-validation was employed.

A: *InfoGainAttributeEval*. *InfoGainRatio* is the Rapid-I/RapidMiner implementation. Evaluates the worth of an attribute by measuring the information gain with respect to the class. Counts are distributed across other values in proportion to their frequency. Otherwise, missing is treated as a separate value.

B: *CorrelationAttributeEval*. Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.

C: *GainRatioAttributeEval*. *GainRatioAttributeEval* is the Weka implementation of this metric. Evaluates the worth of an attribute by measuring the gain ratio with respect to the class. Counts are distributed across other values in proportion to their frequency. Otherwise, missing is treated as a separate value.

5. CLASSIFICATION TECHNIQUES

In order to assess the efficacy of classification techniques, this study has selected six distinct algorithms for analysis. The aforementioned techniques have been employed to assess the performance of classifiers in identifying the most variable feature detection dataset, which has been compiled from a wide range of cocoa clones. The dataset comprises two distinct groups. The first, designated the "marañón group" (fine aroma cocoa), encompasses 59 fine aroma cocoa clones (Table S1). The second group comprises all other cultivable clones (17). The classifiers included in the study were Naive Bayes, Multinomial Logistic Regression, J48, Random Forest, Logistic Model Trees (LTM), and Simple Logistic. A 10-fold cross-validation and Kappa statistic were employed for each classifier.

5.1. Naive Bayes

The algorithm is probabilistic in nature and is based on Bayes' theorem. The efficacy of the algorithm is contingent upon the precise existence of the probability model [23].

5.2. Multinomial Logistic Regression

It predicts categorical variables or the probability of participation in a category as a dependent variable based on multiple independent variables. Furthermore, the model employs maximum likelihood estimation to assess the probability of categorical participation [24].

5.3. J48

Is based on the ID3 algorithm. This process begins with the construction of a tree, with the initial branching based on attributes that best divide objects into appropriate classes. This is achieved by searching for rules or hypotheses that can be used to classify the objects [25].

5.4. Random Forest

The algorithm is designed to facilitate the systematic organization of voluminous data sets, extending the capabilities of classification and regression tree models [26].

5.5. Logistic Model Trees

The tool is a simple prediction that addresses regression issues through the collaborative application of linear regression and decision tree models [27].

5.6. Simple Logistic

A statistical test is employed to predict a single binary variable based on another variable. It is also used to determine the numerical relationship between two of these variables [28].

6. CLASSIFIER PERFORMANCE VERIFICATION

The accuracy and dataset scoring of the classifiers (Naive Bayes, Multinomial Logistic Regression, J48, Random Forest, Logistic Model Trees (LTM), and Simple Logistic) were simultaneously tested in WEKA Experimenter to validate the results obtained in WEKA Explorer. The dataset pertaining to fine aroma cocoa was exclusively considered. The 10-fold cross-validation test mode was executed ten times, with ten repetitions, using both the simple mode configuration and the corrected paired T-test mode. This was done to verify the performance of each classifier by comparing the accuracy and area under the curve (AUC) results of each classifier. The test employed a two-tailed significance level of 0.05.

IV. RESULTS

1. FEATURE SELECTION

The attribute evaluator GainAttributeEval determined that the attributes of OvA (0.965 ± 0.029; presence of anthocyanin in the flower ovary), SepA (0.827 ± 0.058; anthocyanin in the sepal), FilA (0.844 ± 0.006; anthocyanin in the peduncle), PetW (0.722 ± 0.011; petal width) and ApedTA (0.673 ± 0.01; anthocyanin at the apex of the peduncle) are the most representative for the purpose of determining the intraspecific morphological variability of fine aroma cocoa flowers (Table 1). Conversely, the attribute evaluator CorrelationAttributeEval also selected five attributes that most significantly influence the morphological variability of fine aroma cocoa flowers. Among these attributes, PetW (0.788 ± 0.009; petal width) and OvA (0.75 ± 0.005; anthocyanin in the ovary) stood out. ApedTA (0.673 ± 0.0; anthocyanin in the apex of the trichomes), SepA (0.636 ± 0.011; anthocyanin in the sepal) and FilA (0.579 ± 0.004; anthocyanin in the flower filament), among these, the PetW attribute exhibited the greatest statistical support (0.788 ± 0.009). Finally, the InfoGainAttributeEval evaluator was the least effective in determining the characteristic of intraspecific variability in fine cocoa aroma, in most attributes it exhibited a statistical support of less than 10% (Table 1).

Table 1. Methods to select the attributes that most influence fine aroma cocoa

InfoGainAttributeEval		GainRatioAttributeEval		CorrelationAttributeEval	
Average rank	**Attribute**	**Average rank**	**Attribute**	**Average rank**	**Attribute**
0.101 ± 0.001	ApedTA	0.965 ± 0.029	OvA	0.788 ± 0.009	PetW
0.096 ± 0.003	PetW	0.827 ± 0.058	SepA	0.75 ± 0.005	OvA
0.09 ± 0.001	FilA	0.844 ± 0.006	FilA	0.673 ± 0.01	ApedTA
0.085 ± 0.001	SepA	0.722 ± 0.011	PetW	0.636 ± 0.011	SepA
0.076 ± 0.001	AovTA	0.673 ± 0.01	ApedTA	0.579 ± 0.004	FilA
0.073 ± 0.001	SepL	0.422 ± 0.01	AovTA	0.445 ± 0.007	PedA
0.071 ± 0.001	PedA	0.409 ± 0.008	PedA	0.437 ± 0.01	AovTA
0.07 ± 0.001	LigL				
0.067 ± 0.001	StaL				
0.067 ± 0.001	OvA				
0.067 ± 0.002	SepW				
0.067 ± 0.001	PedL				
0.058 ± 0.002	PetL				
0.057 ± 0.001	PisL				
0.055 ± 0.001	StyL				
0.055 ± 0.001	StamL				

0.046 ± 0.001	LigC
0.037 ± 0.001	TriC

To provide a more complete overview, Table 2 shows the evaluations of three attribute evaluators: InfoGainAttributeEval, GainRatioAttributeEval, and CorrelationAttributeEval. The GainAttributeEval evaluator identified five attributes that were considered most representative of those needed to determine the type of cocoa clonal varieties. These attributes include OvA (0.773 ± 0.007; anthocyanin in ovary), ApedTA (0.654 ± 0.005; anthocyanin in apex of peduncle trichomes), PetW (0.651 ± 0.006; petal width), FilA (0.715 ± 0.103; anthocyanin in filament), and SepA (0.618 ± 0.01; anthocyanin in sepal) (Table 1). In contrast, CorrelationAttributeEval identified five attributes that showed the greatest precision in representing the phenotypic variability of cacao: PetW (0.773 ± 0.007; petal width), OvA (0.741 ± 0.004; anthocyanin in the ovary), ApedTA (0.657 ± 0.005; anthocyanin in the apex of the peduncle trichomes), SepA (0.619 ± 0.01; anthocyanin in the sepal), and FilA (0.578 ± 0.011; anthocyanin in the filament). The InfoGainAttributeEval evaluator was the least effective in determining the character of intraspecific variability in cacao. The majority of the attributes had a statistical support of less than 9% (Table 2).

Table 2. Methods to select the attributes that most influence on others cocoa varieties.

InfoGainAttributeEval		**GainRatioAttributeEval**		**CorrelationAttributeEval**	
Average rank	Attribute	Average rank	Attribute	Average rank	Attribute
0.088 ± 0.001	ApedTA	0.983 ± 0.028	OvA	0.773 ± 0.007	PetW
0.079 ± 0.003	PetW	0.654 ± 0.005	ApedTA	0.741 ± 0.004	OvA
0.077 ± 0.001	FilA	0.651 ± 0.006	PetW	0.657 ± 0.005	ApedTA
0.073 ± 0.001	SepA	0.715 ± 0.103	FilA	0.619 ± 0.01	SepA
0.076 ± 0.001	AovTA	0.618 ± 0.01	SepA	0.578 ± 0.011	FilA
0.068 ± 0.001	SepL	0.416 ± 0.008	AovTA	0.44 ± 0.009	PedA
0.066 ± 0.001	PedA	0.358 ± 0.12	PedA	0.409 ± 0.136	AovTA
0.064 ± 0.001	LigL				
0.064 ± 0.001	StaL				
0.062 ± 0.001	OvA				
0.062 ± 0.002	SepW				
0.058 ± 0.001	PedL				
0.057 ± 0.002	PetL				
0.054 ± 0.001	PisL				
0.055 ± 0.001	StyL				
0.049 ± 0.001	StamL				
0.044 ± 0.001	LigC				
0.03 ± 0.001	TriC				

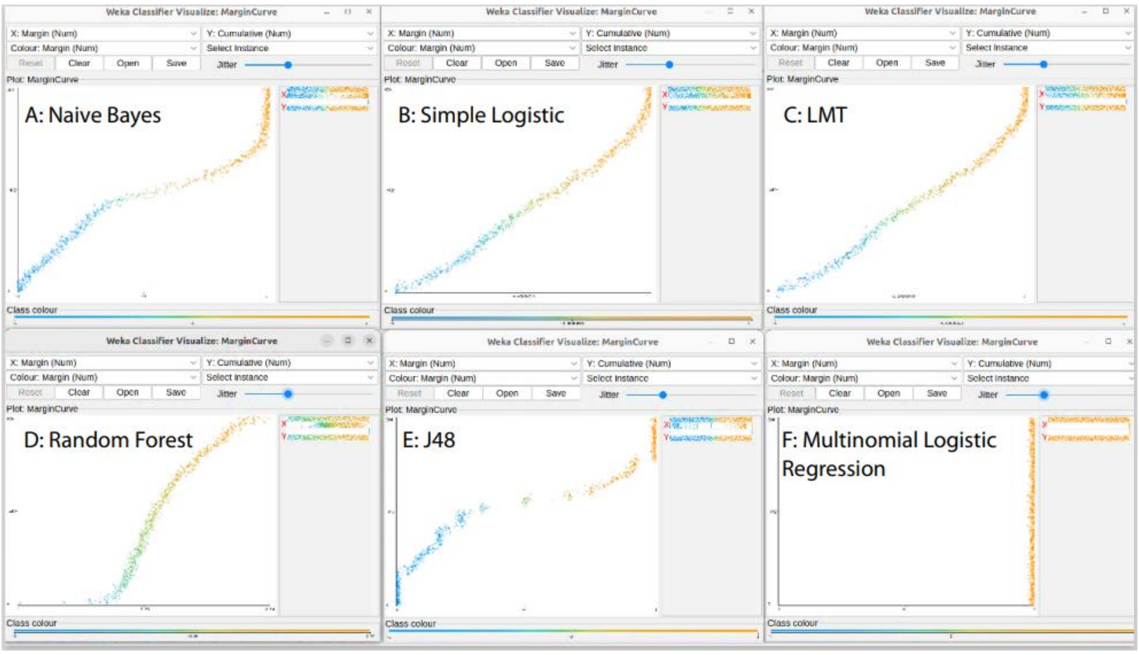
2. TESTING CLASSIFIER PERFORMANCE USING THE WEKA EXPLORER

The results obtained from the data set revealed that the algorithms exhibited a degree of inconsistency. As illustrated in Table 3, the classification results of all the algorithms tested demonstrated comparable performance, with an accuracy, sensitivity, and specificity of 63.57%. Furthermore, Random Forest demonstrated a classification performance of 61.65%, while Multinomial Logistic Regression exhibited the lowest classification performance, with a score of 52.82% (Table 3). With regard to the computational complexity required to build the classification model, Naive Bayes was the most efficient, requiring only 0.0 seconds, followed by J48 with 0.03 seconds. The time required to build the models was as follows: Logistic regression (simple) and LMT required 1.01 and 2.58 seconds, respectively, while Random Forest required 0.17 seconds. In contrast, multinomial logistic regression showed the least efficient performance, requiring 85.56 seconds to build the model. Despite the relatively lengthy run time (85.56 seconds) required for classification of the data set, the multinomial logistic regression model demonstrated an accuracy of 52.83% above that of the J48 model.

Table 3. Weighted average of classifiers for the fine aroma cocoa flower dataset.

Classifier	Training runtime (s)	Correctly classified instance % (Accuracy)	Kappa statistic
Simple Logistic	1.01	63.57	0.62
J48	0.03	42.28	0.41
Random Forest	0.17	61.65	0.6
Multinomial logistic regression	85.56	52.828	0.52
LMT	2.58	63.57	0.63
Naive Bayes	0.0	54.47	0.53

The curve margin for all algorithms is depicted in Figure 2, which elucidates the discrepancy between the predicted probability for the current class (dependent variable) and the highest predicted probability for the other classes (independent variable). Each algorithm exhibits a distinct curve, with the exception of the fine aroma cocoa data. This demonstrates that the performance of the algorithms can provide a distinct improvement in performance. In this case, the Naive Bayes (Figure 2A), Simple Logistic (Figure 2B), LMT (2C), and Random Forest (Figure 2D) algorithms present a superior curve with an optimal data distribution, indicating that they are the most suitable for use in classification. While the J48 (Figure 2E) and Multinomial Logistic Regression (Figure 2F) algorithms present an unusual curve, their performance in data classification is poor.



Weka Classifier Visualize: MarginCurve

X: Margin (Num)
Y: Cumulative (Num)
Colour: Margin (Num)
Select Instance
Plot: MarginCurve
Class colour

FIGURE 2. Margin curve for A: Naïve Bayes, B: Simple Logistic, C: LMT, D: Random Forest, E: J48 and F: Multinomial Logistic regression for the fine aroma cocoa flower dataset.

The experiment yielded disparate results among the algorithms, as illustrated in Table 4. The classification results of all the algorithms tested revealed that the Simple Logistic and LMT exhibited the highest accuracy, sensitivity, and specificity, with an accuracy rate of 57.74%. The Random Forest and Naive Bayes algorithms exhibited noteworthy performance, with accuracy rates of 58.33% and 54.47%, respectively. In contrast, the multinomial logistic regression demonstrated a performance of 52.67%. Finally, the J48 algorithm exhibited the lowest classification performance, with an accuracy of 42.28% (Table 4). In terms of computational complexity to build the classification model, the Naive Bayes algorithm proved to be the most efficient, requiring only 0.0 seconds, followed by J48 with 0.03 seconds, while Random Forest and Simple Logistic required 0.65 and 1.4 seconds, respectively, while LMT required only 8.83 seconds. In contrast, multinomial logistic regression exhibited the least efficient performance, requiring 133.43 seconds to build the model. Notwithstanding the relatively lengthy run time, the model demonstrated an accuracy of 52.67%, which was higher than that of J48 (42.28%).

Table 4. Weighted average of classifiers for the dataset of fine aroma cocoa clone flowers and other clonal varieties.

Classifier	Training runtime (s)	Correctly classified instance % (Accuracy)	Kappa statistic
Simple Logistic	1.4	59.74	0.59
J48	0.03	42.28	0.41
Random Forest	0.65	58.33	0.57
Multinomial logistic regression	133.43	52.67	0.52
LMT	8.83	59.74	0.59
Naive Bayes	0.0	54.47	0.53

Figure 3 illustrates the margin of the curve for all algorithms when applied to broader data sets, in this case, data pertaining to fine aroma cocoa and other clonal varieties of cocoa. The Naive Bayes (Figure 3A), Simple Logistic (Figure 3B), LMT (Figure 3C), and Random Forest (Figure 3D) algorithms present a superior curve with an optimal distribution of data, suggesting that they may be employed with data from a single population of cocoa (fine aroma cocoa, Figure 2A-D). The J48 (Figure 3E) and Multinomial Logistic Regression (Figure 3F) algorithms also present an unusual curve, resulting in poor performance in total data classification.

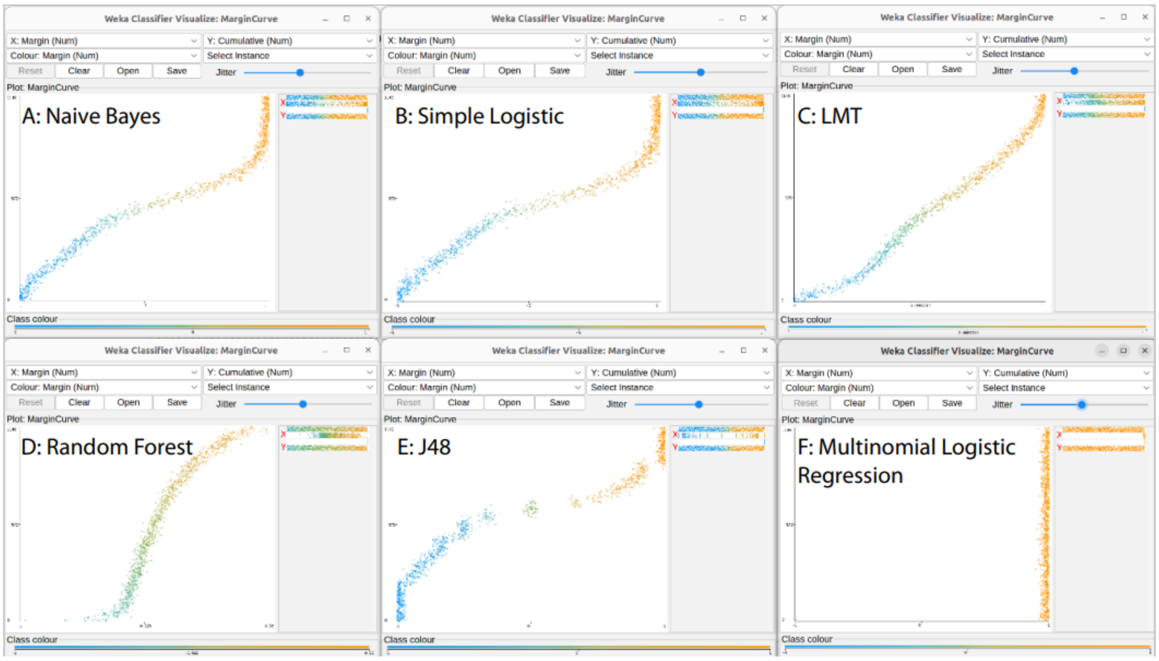


FIGURE 3. Margin curve for A: Naïve Bayes, B: Simple Logistic, C: LMT, D: Random Forest, E: J48 and F: Multinomial Logistic regression, for the dataset of fine aroma cocoa clone flowers and other clonal.

3. VERIFICATION OF CLASSIFIER PERFORMANCE USING THE WEKA EXPERIMENTER

The accuracy and area under the curve (AUC) of each algorithm were compared when tested in both WEKA Explorer and WEKA Experimenter. The results demonstrated that the Simple Logistic and LTM algorithms exhibited the highest accuracy, with 60.45% and 60.50%, respectively. In contrast, the J48 algorithm exhibited the lowest accuracy (43.28%) and AUC accuracy (0.84), while the other algorithms exhibited an AUC read accuracy above 0.98 (Table 5).

Table 5. Accuracy and score of all classifiers tested simultaneously in WEKA Experimenter.

Classifier	Correctly classified instance % (accuracy)	AUC
Native Bayes	52.65	0.98
Multinomial logistic regression	52.68	0.99
Simple Logistic	60.45	0.99
J48	43.28	0.84
LMT	60.50	0.99
Random Forest	58.50	1.00

V. DISCUSSION

It is of paramount importance to ensure that machine learning is further incorporated and expanded in the agricultural sector, with particular emphasis on cocoa (T. cacao), an economically important crop that requires special attention in terms of its diversity. Machine learning approaches can be used to identify patterns in the functional traits of flowers. This machine learning technique would be a fast and practical alternative to resolve uncertainties about cocoa varieties. It is not uncommon for farms that have not been renovated for a long period to encounter a common problem such as graft-to-rootstock substitution. The misidentification of cocoa trees poses a challenge in accurately determining the correct identity and structure of individuals within a population, as gene flow plays an important role in diversification [29]. This is especially evident in cultivated cocoa, which exhibits a higher level of genetic diversity than wild cocoa. Furthermore, the prevalence of self-pollination and constant hybridization makes it challenging to accurately identify cocoa trees [30].

As the quantity of data continues to expand at an unprecedented rate due to technological advances [31], it has become standard practice to categorize it in a way that allows for straightforward analysis using machine learning technologies [32]. This study employed an approach to elucidate the variables that influence flower traits in a cocoa population. The results indicated that the presence of ovary anthocyanin (OvA) was identified as the most influential factor in the diversity of each cocoa individual, either in intra- or interspecific variation. These variations in anthocyanin define the color of cocoa fruit and are probably due to the high rate of cross-pollination, where several cocoa populations share a close geographical area. Consequently, this study demonstrates the importance of employing data mining classification algorithms to predict the morphological traits of flowers that most influence cocoa diversity. The Simple Logistic and LMT algorithms were found to be the most effective in elucidating such diversity. This implies that each model exhibits a distinct behavior contingent upon the approach and type of data. It is evident that this approach can be applied to other agricultural sectors, including crop yield prediction, decision-making in the area of fertilization, and pest and disease identification. Furthermore, the incorporation of deep learning for analysis or multispectral images can be considered. For instance, machine learning approaches are currently being utilized to analyze various characteristics of cocoa crops, including their physicochemical properties, quality, fermentation processes, patterns, yield losses, and sales [16,17,19]. Additionally, they are employed in the processing of multiple images [17] and the classification of rice varieties [33]. However, the degree of precision will depend on the nature of the data and the algorithm employed. These approaches are of interest for the future, as they have the potential to improve existing applications and develop new ones in a way that minimizes production costs and saves time in certain agricultural activities. Consequently, the relationship between the agricultural sector and data science is being consolidated with the aim of analyzing data in a simpler and more reliable way.

Furthermore, although it has been observed that the accuracy of each algorithm will depend on the type of attributes and the target sample, it has been demonstrated that algorithms such as Naive Bayes, J48, Multilayer Perceptron, and Support Vector Machines (SVM) generate high classification accuracy of seeds and wheat genotypes [35]. However, in other parameters such as area, season, and crop yield quality, the performance of the algorithm is significantly reduced (76.82%) [36]. Consequently, it is imperative to persist in advancing studies that employ diverse models that align more closely with the data, as these machine learning approaches present opportunities for farmers in decision-making and crop yield predictions [37]. This machine learning approach will be one of the most fundamental tools for the agricultural sector, especially cocoa cultivation, not only to determine the existing diversity but also for fertilization, maintenance, prevention, and planning issues.

Moreover, it is essential to acknowledge that the generation of extensive, unified data sets may not always be feasible in all agricultural contexts [38]. In certain cases, the combination of disparate agricultural scenarios is tacitly suggested [39, 40], and the integration of multiple data sets with different conditions is employed before training the model to encompass the full spectrum of scenarios and target variability [41], since the use of large data sets under the machine learning approach is found to be viable in different conditions and data distributions within a specific agricultural domain. However, this task is inherently endless and must be addressed through the development of new, innovative, and efficient methods, such as deep learning for analysis [42] and generative adversarial networks [43]. Consequently, a machine learning approach to cocoa diversity necessitates the sequential training of multiple models on data sets with the objective of achieving more efficient and consistent results. This objective will be achieved by proposing a pre-trained agricultural model, which will be developed with the selection of data quality in mind, in order to address the challenges associated with the diversity of fine aroma cocoa and other diverse agricultural crops.

Consequently, the results of this study, which were based on functional traits (morphology) in cocoa flowers, do not negate the importance of applying them to the agricultural sector. This leads us to propose that future studies of this nature should include disease detection, pest detection, leaf area index, yield prediction, and grading of cocoa beans. This will facilitate a more comprehensive understanding of the recognition of diverse cocoa accessions and cultivars, whether in arable or wild populations, and will also help to reduce production costs and improve decision-making.

VI. CONCLUSION

This study has demonstrated the significant impact of using different attribute selections and algorithms to identify functional traits in cocoa flowers under agricultural scenarios. The results indicate that the validation and testing of models within specific circumstances that accurately reflect the variability found when using different algorithms and attribute selection is appropriate. This is exemplified by the use of GainRatioAttributeEval, which proved to be more effective in identifying the most prevalent functional trait in cocoa diversity, namely the presence of ovary anthocyanin (OvA). Nevertheless, the integration of multiple models was also shown to enhance the performance of each algorithm. In terms of algorithms, Simple Logistic and LMT yielded the best accuracy results (>60%), although Naive Bayes proved to be the most efficient for model building (0.0 s) in terms of computational complexity. It is, however, important to note that the use of large data sets may not always be feasible in all agricultural contexts, as not all algorithms produce optimal results. This suggests that future research should investigate alternative approaches to enhance models for the agricultural sector, with a particular focus on cocoa diversity. In many cases, this can be analyzed for other traits, such as leaves, photographs for deficiency analysis, pests and diseases of cocoa. Consequently, the research provides a comprehensive overview of the floral traits that most influence cocoa diversity, which underscores the necessity for further refinement of models to obtain more robust results.

Funding statement

The authors wish to acknowledge that no specific funding or support was provided for this study.

Author contribution

All authors made an equal contribution to the development and planning of the study.

Conflict of Interest

The authors have no potential conflicts of interest, or such divergences linked with this research study.

Data Availability Statement

Data are available from the authors upon request.

Acknowledgements

We are deeply grateful to Joseph Campos, Walter M. Abarca, Jonathan Cruz, Yolmer L, Yeltsin A, Jani M, Beimer Ch., for his support in the collection of field samples.

REFERENCES

1. Gómez, J. M., Perfectti, F., Armas, C., Narbona, E., González-Megías, A., Navarro, L., DeSoto, L., & Torices, R. (**2020**). Within-individual phenotypic plasticity in flowers fosters pollination niche shift. *Nat. Commun.*, *11*(1), 4019.
2. Buchanan, S., Isaac, M. E., Van den Meersche, K., & Martin, A. R. (**2019**). Functional traits of coffee along a shade and fertility gradient in coffee agroforestry systems. *Agrofor. Syst.*, *93*, 1261-1273.
3. Isaac, M. E., Martin, A. R., de Melo Virginio Filho, E., Rapidel, B., Roupsard, O., & Van den Meersche, K. (**2017**). Intraspecific trait variation and coordination: Root and leaf economics spectra in coffee across environmental gradients. *Front. Plant. Sci.*, *8*, 1196.
4. Montazeaud, G., Violle, C., Roumet, P., Rocher, A., Ecarnot, M., Compan, F., Maillet, G., Florián, F., & Fréville, H. (**2020**). Multifaceted functional diversity for multifaceted crop yield: Towards ecological assembly rules for varietal mixtures. *J. Appl. Ecol.*, *57*(11), 2285-2295.
5. Motamayor, J. C., Lachenaud, P., Da Silva e Mota, J. W., Loor, R., Kuhn, D. N., Brown, J. S., & Schnell, R. J. (**2008**). Geographic and genetic population differentiation of the Amazonian chocolate tree (Theobroma cacao L). *PloS One*, *3*(10), e3311.
6. Schwarzkopf, E. J., Motamayor, J. C., & Cornejo, O. E. (**2020**). Genetic differentiation and intrinsic genomic features explain variation in recombination hotspots among cocoa tree populations. *BMC Genomics*, *21*, 1-16.
7. Lachenaud, P., & Zhang, D. (**2008**). Genetic diversity and population structure in wild stands of cacao trees (Theobroma cacao L.) in French Guiana. *Ann. For. Sci.*, *65*, 310.
8. Thomas, E., van Zonneveld, M., Loo, J., Hodgkin, T., Galluzzi, G., & van Etten, J. (**2012**). Present spatial diversity patterns of Theobroma cacao L. in the neotropics reflect genetic differentiation in Pleistocene refugia followed by human-influenced dispersal. *PLoS One*, *7*(10), e47676.
9. Chumacero de Schawe, C., Durka, W., Tscharntke, T., Hensen, I., & Kessler, M. (**2013**). Gene flow and genetic diversity in cultivated and wild cacao (Theobroma cacao) in Bolivia. *Am. J. Bot.*, *100*(11), 2271-2279.
10. Sereno, M. L., Albuquerque, P. S. B., Vencovsky, R., & Figueira, A. (**2006**). Genetic diversity and natural population structure of cacao (Theobroma cacao L.) from the Brazilian Amazon evaluated by microsatellite markers. *Conserv. Genet.*, *7*, 13-24.
11. Zhang, D., Martínez, W. J., Johnson, E. S., Somarriba, E., Phillips-Mora, W., Astorga, C., Mischke, S., & Meinhardt, L. W. (**2012**). Genetic diversity and spatial structure in a new distinct Theobroma cacao L. population in Bolivia. *Genet. Resour. Crop. Evol.*, *59*, 239-252.
12. Boadi, S. A., Olwig, M. F., Asare, R., Bosselmann, A. S., & Owusu, K. (**2022**). The role of innovation in sustainable cocoa cultivation: Moving beyond mitigation and adaptation. In *Climate-induced innovation: Mitigation and adaptation to climate change* (pp. 47-80). Cham: Springer International Publishing.
13. Subasi, A., Balfaqih, M., Balfagih, Z., & Alfawwaz, K. (**2021**). A comparative evaluation of ensemble classifiers for malicious webpage detection. *Procedia Comput. Sci.*, *194*, 272-279.
14. Sher, C. (**2000**). The CRISP-DM model: The new blueprint for data mining. *J. Data Warehousing*, *5*(4), 13-22.
15. Altaleb, M., Deeken, H., & Hertzberg, J. (**2022**). A data mining process for building recommendation systems for agricultural machines based on big data. *Lecture Notes in Informatics (LNI), Proceedings-Series of the Gesellschaft für Informatik (GI)*.
16. Mazon, B., Jaramillo, M., Romero, O., Borja, A., Aguirre, M., & Contento, M. (**2018**). Tecnologías de Inteligencia de Negocios y Minería de datos para el análisis de la producción y comercialización de cacao. *Revista Espacios*, *39*(32).
17. Angelia, R. E., & Linsangan, N. B. (**2018**). Fermentation level classification of cross cut cacao beans using k-NN algorithm. In *Proceedings of the 5th International Conference on Bioinformatics Research and Applications* (pp. 64-68).
18. Herrera-Rocha, F., Fernández-Niño, M., Cala, M. P., Duitama, J., & Barrios, A. F. G. (**2023**). Omics approaches to understand cocoa processing and chocolate flavor development: A review. *Food Res. Int.*, *165*, 112555.
19. Wood, J. E., Allaway, D., Boult, E., & Scott, I. M. (**2010**). Operationally realistic validation for prediction of cocoa sensory qualities by high-throughput mass spectrometry. *Anal. Chem.*, *82*(14), 6048-6055.
20. Chlingaryan, A., Sukkarieh, S., & Whelan, B. (**2018**). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Comput. Electron. Agric.*, *151*, 61-69.
21. Brenes, E. R., Martinez, O., Lopez, M. F., Ciravegna, L., & Pichardo, C. A. (**2023**). Cacao Oro. *Int. Food Agribus. Manag. Rev.*, *26*(5), 783-799.
22. SENAMHI. (**2020**). Mapa Climático del Perú. Retrieved from https://www.senamhi.gob.pe/?p=mapa-climatico-del-peru
23. Rrmoku, K., Selimi, B., & Ahmedi, L. (**2022**). Application of trust in recommender systems—utilizing naive Bayes classifier. *Computation*, *10*(1), 6.
24. Ragazou, K., Passas, I., Garefalakis, A., Kourgiantakis, M., & Xanthos, G. (**2022**). Youth's entrepreneurial intention: A multinomial logistic regression analysis of the factors influencing Greek HEI students in time of crisis. *Sustainability*, *14*(20), 13164.
25. Choi, L. K., Rii, K. B., & Park, H. W. (**2023**). K-means and J48 algorithms to categorize student research abstracts. *IJCITSM*, *3*(1), 61-64.
26. Rigatti, S. J. (**2017**). Random forest. *J. Insur Med.*, *47*(1), 31-39.
27. Li, N., Zare, M., Yi, C., & Jimenez, R. (**2022**). Stability risk assessment of underground rock pillars using logistic model trees. *Int. J. Environ. Res. Public Health*, *19*(4), 2136.
28. Gouda, M., Lugnan, A., Dambre, J., van den Branden, G., Posch, C., & Bienstman, P. (**2023**). Improving the classification accuracy in label-free flow cytometry using event-based vision and simple logistic regression. *IEEE J. Sel. Top. Quantum Electron.*, *29*(2), 1-8.
29. Fekner, S., Austerlitz, F., Cuguen, J., & Arnaud, J. F. (**2007**). Long distance pollen-mediated gene flow at a landscape level: The weed beet as a case study. *Mol. Ecol.*, *16*(18), 3801-3813.
30. Ha, L. T. V., Hang, P. T., Everaert, H., Rottiers, H., Anh, L. P. T., Dung, T. N., & Messens, K. (**2016**). Characterization of leaf, flower, and pod morphology among Vietnamese cocoa varieties (Theobroma cacao L.). *Pak. J. Bot.*, *48*(6), 2375-2383.
31. Jordan, M. I., & Mitchell, T. M. (**2015**). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.
32. Martinez, I., Viles, E., & Olaizola, I. G. (**2021**). Data science methodologies: Current challenges and future approaches. *Big Data Res.*, *24*, 100183.
33. Cinar, I., & Koklu, M. (**2019**). Classification of rice varieties using artificial intelligence methods. *Int. J. Intell. Syst. Appl. Eng.*, *7*(3), 188-194.
34. Sabancı, K., & Akkaya, M. (**2016**). Classification of different wheat varieties by using data mining algorithms. *Int. J. Intell. Syst. Appl. Eng.*, *4*(2), 40-44.

35. Golcuk, A., & Yasar, A. (**2023**). Classification of bread wheat genotypes by machine learning algorithms. *J. Food Compos. Anal.*, *119*, 105253.
36. Ismael, H. R., Abdulazeez, A. M., & Hasan, D. A. (**2021**). Comparative study for classification algorithms performance in crop yields prediction systems. *Qubahan Acad. J.*, *1*(2), 119-124.
37. Agrawal, D., & Dahiya, P. (**2018**). Comparisons of classification algorithms on seeds dataset using machine learning algorithm. *Compusoft*, *7*(5), 2760-2765.
38. León, L., Campos, C., & Hirzel, J. (**2024**). Deep learning for broadleaf weed seedlings classification incorporating data variability and model flexibility across two contrasting environments. *Artif. Intell. Agric.*, *12*, 29-43.
39. Dyrmann, M., Karstoft, H., & Midtiby, H. S. (**2016**). Plant species classification using deep convolutional neural network. *Biosyst. Eng.*, *151*, 72-80.
40. Makanapura, N., Sujatha, C., Patil, P. R., & Desai, P. (**2022**). Classification of plant seedlings using deep convolutional neural network architectures. *J. Phys. Conference Series*, *2161*, No. 1, p. 012006.
41. Olsen, A., Konovalov, D. A., Philippa, B., Ridd, P., Wood, J. C., Johns, J., Banks, W., Girgenti, B., Kenny, O., Whinney, J., Calvert, B., Azghadi, M. R., & White, R. D. (**2019**). DeepWeeds: A multiclass weed species image dataset for deep learning. *Sci. Rep.*, *9*(1), 2058.
42. Sener, O., & Savarese, S. (**2017**). Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.
43. Zhang, W., Chen, K., Wang, J., Shi, Y., & Guo, W. (**2021**). Easy domain adaptation method for filling the species gap in deep learning-based fruit detection. *Hortic. Res.*, *8*.