Universidade Federal Rural do Semi-Árido
Pró-Reitoria de Pesquisa e Pós-Graduação
https://periodicos.ufersa.edu.br/index.php/caatinga

ISSN 0100-316X (impresso)
ISSN 1983-2125 (online)
http://dx.doi.org/10.1590/1983-21252022v35n404rc

CLASSIFICATION OF *Phaseolus lunatus* L. USING IMAGE ANALYSIS AND MACHINE LEARNING MODELS[1]

ÉRIKA BEATRIZ DE LIMA CASTRO[2]*, RAYLSON DE SÁ MELO[2], EMANUEL MAGALHÃES DA COSTA[2], ANGELA MARIA DOS SANTOS PESSOA[3], RAMONY KELLY BEZERRA OLIVEIRA[2], CÂNDIDA HERMÍNIA CAMPOS DE MAGALHÃES BERTINI[2]

ABSTRACT - Image analysis combined with machine learning models can be an excellent tool for classification of fava (Phaseolus lunatus L.) genotypes and is a low-cost system. Fava is grown by family farmers, mainly, in the Northeast and South regions of Brazil, presenting economic and social importance. Evaluations to gather information on qualitative and quantitative characters of seeds enable the description and distinction of genotypes, allowing the evaluation of variability of plant species, which is essential in breeding programs. The use of image analysis is a fast and economic tool for obtaining large quantity of information. Machine learning techniques have been developed and implemented in the agricultural sector due to technological advances and increasing use of artificial intelligence, which enables the automatization of several processes. In this context, the objective of this work was to evaluate different machine learning models to classify fava genotypes, using data obtained through image analysis. Images of fava seeds were captured using a table scanner (HP Scanjet 2004), set to true color mode, arranged upside down inside of an aluminum box fully closed during the capture of the images for an adequate illumination and prevention of environmental noises. The K-Nearest Neighbor, Naive Bayes, Linear Discriminant Analysis, Support Vector Machine, Gradient Boosting, Bootstrap Aggregating, Classification and Regression Trees, Random Forest, and C50 models were used for the study. Linear Discriminant Analysis was the model that presented the highest efficiency for classifying the genotypes, with an accuracy of 90%.

Keywords: Artificial intelligence. Image processing. Seeds.

CLASSIFICAÇÃO DE *Phaseolus lunatus* L. USANDO TÉCNICA DE ANÁLISE DE IMAGEM E MODELOS DE APRENDIZAGEM DE MÁQUINA

RESUMO – A análise de imagem associada com modelos de aprendizado de máquina pode ser uma excelente ferramenta de classificação para genótipos de fava, além de ser um sistema de baixo custo. A produção de feijão-fava é realizada por agricultores familiares, principalmente, nas regiões Nordeste e Sul do país, apresentando importância econômica e social. A avaliação e o conhecimento de caracteres qualitativos e quantitativos das sementes, permite a descrição e distinção de genótipos, permitindo a avaliação da variabilidade desta espécie, que é fundamental em um programa de melhoramento. O uso de análise de imagem é uma das ferramentas para obtenção de uma grande quantidade de informações de forma rápida e econômica. Com os avanços tecnológicos, e o uso cada vez mais comum de inteligência artificial, as técnicas de aprendizado de máquinas vêm sendo desenvolvidas e implementadas no setor agropecuário, permitindo que vários processos sejam automatizados. Diante do exposto, objetivou-se com esse trabalho, avaliar diferentes modelos de Machine Learning para classificar genótipos de fava, por meio de dados obtidos por análise de imagem. As imagens das sementes de fava, foram capturadas por um scanner de mesa, configurado no modo "true color", adaptado de maneira invertida, dentro de uma caixa de alumínio, completamente fechada durante a captura da imagem, para ter iluminação adequada e eliminar ruídos do ambiente. Neste estudo foram usados os modelos de KNN, NB, LDA, SVM, GBM, BAGGING, CART, RF e C50. O modelo de LDA foi o que apresentou maior eficácia na classificação dos genótipos, com uma precisão de 90%.

Palavras-chave: Inteligência artificial. Processamento de imagens. Sementes.

*Corresponding author

[1]Received for publication in 02/16/2022; accepted in 07/12/2022.
Paper approved from V Symposium of the Northeast Plant Genetic Resources Network 2021.

[2]Department of Phytotechnics, Universidade Federal do Ceará, Fortaleza, CE, Brazil; erika-beatriz@hotmail.com - ORCID: 0000-0002-1396-9967, raylsonmello@gmail.com – ORCID: 0000-0001-9512-7335, emanuelmagalhaes0@gmail.com – ORCID: 0000-0001-9361-2089, ramonyoliveira06@gmail.com - ORCID: 0000-0002-1908-0560, candida@ufc.br – ORCID: 0000-0003-2949-5660.
[3]Department of Phytotechnics, Universidade Federal Rural do Semi-Árido, Mossoró, RN, Brazil; angelapessoapb@gmail.com - ORCID: 0000-0002-7393-984X.

INTRODUCTION

Fava (*Phaseolus lunatus* L.) is grown in the Northeast region of Brazil, mainly in Ceará and Paraíba, which are the largest producing states (IBGE, 2020). Fava crops impact positively family farmers in these regions due to its economic importance, since they sell the surplus production, and its social importance due to the nutritional benefits of this legume as a food, which improve their quality of life (CARMO et al. 2015). Despite the varieties used are adapted to local climate conditions, the yield is usually low due to lack of commercial cultivars and recommendation of management techniques (SILVA; DULTRA FILHO, 2018).

Information on seed physiological and morphological characteristics are important and serve as a base for breeding programs (ADVÍNCULA et al. 2015). In addition, evaluations to gather information on qualitative and quantitative characters of seeds enable the description and distinction of genotypes (PERINI et al., 2018), which are essential for the management of germplasm collections and for pre-improvement works. However, despite traditional techniques are simple and easy to understand, the process is time consuming, requiring a team of qualified people and, consequently, significant financial resources (SOUSA et al., 2015). In addition, the results can be subjective when performing visual evaluations of samples without using high-precision tools, since they may vary according to the experience and limitations of the evaluator (TORRES, 2018).

One of the tools to overcome these limitations is the use of image analysis, which enables a fast and economic obtaining of a large quantity of information, require little labor, and is a non-destructive technique (TORRES, 2018). Thus, the use of digital images and a software to assess the results allows a fast evaluation of qualitative and quantitative characters, such as area, length, width, perimeter, and integument color (ABUD et al., 2022) as well as the obtaining of accurate data (MOREIRA et al., 2022).

Technology advances and use of artificial intelligence have enabled the development of machine learning techniques and implement them in different sectors (CARLEO et al., 2019), including agriculture, allowing the automatization of several processes (MORETI et al., 2021).

Kayabasi et al. (2018) compared machine learning techniques for classification of wheat seeds based on seed quantitative characteristics through Artificial Neural Network models, Support Vector Machine, and Adaptive Neuro-fuzzy Inference System and obtained efficient results for classification and identification of seeds. Medeiros et al. (2020a) used Linear Discriminant Analysis, Random Forest, and Support Vector Machine models, combined with digital image analysis, to classify soybean seeds and seedlings according to their physiological quality and obtained results that confirmed the precision of these methods, based on seedling vigor responses.

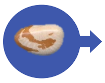
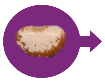
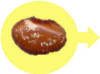
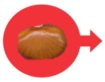
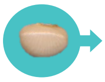
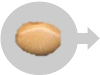
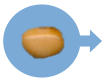
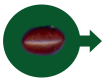
In the context of technological advances and searching for automatization of production processes, including for the agricultural sector, neural networks and image analysis became alternatives for fast obtaining results of seed physiological quality (VASCONCELOS et al., 2018).

The use of these technologies can provide morphological information of seeds focused on breeding programs and exploration of the production potential of crops with specific market niches (ADVÍNCULA et al., 2015), such as fava crops. In this context, the objective of this work was to evaluate different machine learning models to classify fava genotypes, using data obtained from image analysis.

MATERIAL AND METHODS

The fava seeds used were obtained from family farmers in the municipality of Redenção, CE, Brazil. The seeds were from the varieties Orelha-de-Vó, Fava-Rajada-Preta, Espírito-Santo-Vermelho, Espírito-Santo-Marrom, Fava-Rajada, Fava-Amarela, Mulatinha, Fígado-de-Frango, Fava-Manteiga and Fava-S (Table 1). The study was conducted at the Laboratory of Seed Analysis of the Department of Plant Production of the Federal University of Ceará (UFC), Brazil.

Table 1. Description of seeds of the fava genotypes evaluated.

Genotypes		Description of seeds
	Orelha-de-Vó	Beige bottom with variegated brown; large, with elliptical and flat shape.
	Fava-Rajada-Preta	Gray bottom with variegated black; large, with elliptical and flat shape, and round ends.
	Espírito-Santo-Vermelho	Gray bottom with streaked red dark and dotted colors; large, with kidney-like shape; slightly bold profile.
	Espírito-Santo-Marrom	White bottom with streaked clear brown and dotted colors; large, slightly flat, with to kidney-like shape.
	Fava-Rajada	white bottom with dotted dark brown in the upper part; large, with a kidney-like shape and flat profile.
	Fava-Amarela	Brown bottom with a streaked slightly visible dark brown color; large, with elliptical, slightly bold shape.
	Mulatinha	Beige bottom; small, with elliptical, slightly bold shape.
	Fígado-de-Frango	Beige bottom; large, with elliptical, slightly bold shape.
	Fava-Manteiga	Brown bottom, medium size, with elliptical, slightly bold shape.
	Fava-S	Purple red bottom; small, with spherical shape and bold profile.

Acquisition of images

Images of fava seeds were captured using a table scanner (HP Scanjet 2004), set to true color mode, arranged upside down inside of an aluminum box fully closed during the capture of the images for an adequate illumination and prevention of environmental noises. The process was carried out using a resolution of 300 dots per inch (DPI), and the resulting image was saved as a Joint Photographic Experts Group (JPEG) file with resolution of 2550 × 3510 pixels (Figure 1).

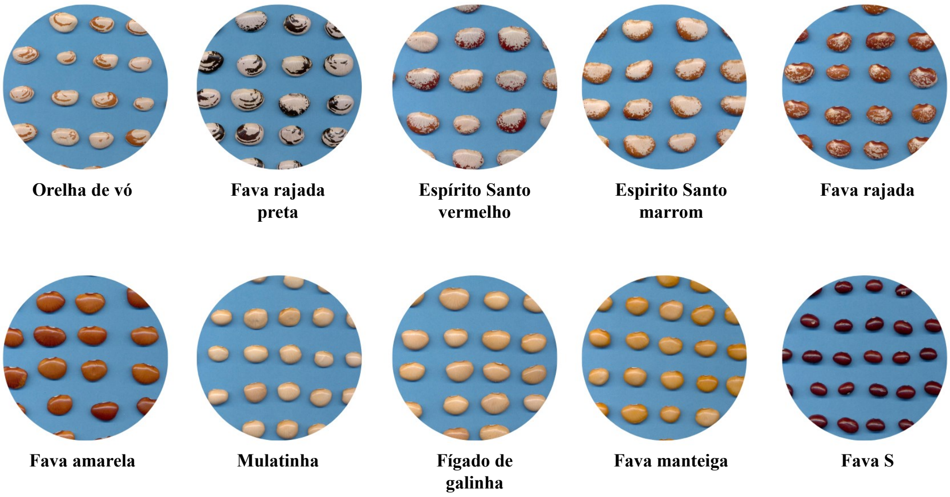


Figure 1. Images of seeds of fava genotypes.

One-hundred seeds of each genotype were randomly selected and arranged on blue paper, with a basis weight of 120 g m^{-2}. This color was selected in initial tests that showed a higher contrast between seeds of the 10 genotypes and the image background, facilitating the segmentation process in the evaluation step.

Segmentation and extraction of data

After the capture and storage, the images were analyzed through the following steps: the first step for the analysis was the plotting of spectral data: red, green, and blue (RGB); and RGB indexes (BI, SCI, BGI, HUE, VARI, SI, NGRDI, GLI, BIM, and HI) to choose the index that better represented the objects in the images (seeds and background) for later segmentation. The second step of analysis consisted of choosing the background removal index, evaluating which histogram had the best index (SI), and defining the cut value to separate the background from the seeds.

The third and last step of analysis was carried out after the removal of the background and isolation of seeds; the RGB spectral data and biometric data of seeds were extracted (Table 2).

The RGB color model uses the primary colors red, green, and blue, with each component varying in an interval between 0 and 255. The combination of these spectra forms secondary colors, such as cyan (green + blue), magenta (red + blue), and yellow (red + green) (WANG et al., 2020). When the 3 channels (RGB) have maximum value (255), the color is white, and when they have the minimum (0), the color is black (WANG et al., 2016).

Table 2. List of colors and biometric variables obtained through image analysis of fava (*Phaseolus lunatus* L.) seeds.

Variable	Description
Red (R)	Spectral values of channel R
Green (G)	Spectral values of channel G
Blue (B)	Spectral values of channel B
Area (pixels)	Seed area and number of pixels within their limits
Width (pixels)	Longest row perpendicular to the main axis
Length (pixels)	Longest row between the seed ends
Perimeter (pixels)	Total number of pixels in the external outline of each seed

Exploratory analysis of the data

The exploratory analysis of the data obtained was used to subsidize the understanding of the numeric and statistical nature of the variables used in the machine learning models. Summarized statistics and violin graphs were developed to evaluate the RGB spectral variables and biometric variables for each genotype.

The violin graph was used to show the data variation, together with their distribution, denoting the dynamics of the genotypes, and dotplot was used to show the distribution and characteristics of the machine learning models used in the work.

Machine learning models developed

Classification models are important for automatic systems of decision making. There are many algorithms that make this classification process. They can provide different results for different datasets. Thus, using a more adequate classifier, according to the data obtained, is important for the decision making. The following models were used in the present work: K-Nearest Neighbor (KNN), Naive Bayes (NB), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Gradient Boosting (GBM), Bootstrap Aggregating (BAGGING), Classification and Regression Trees (CART), Random Forest (RF), and C50.

Validation of the model

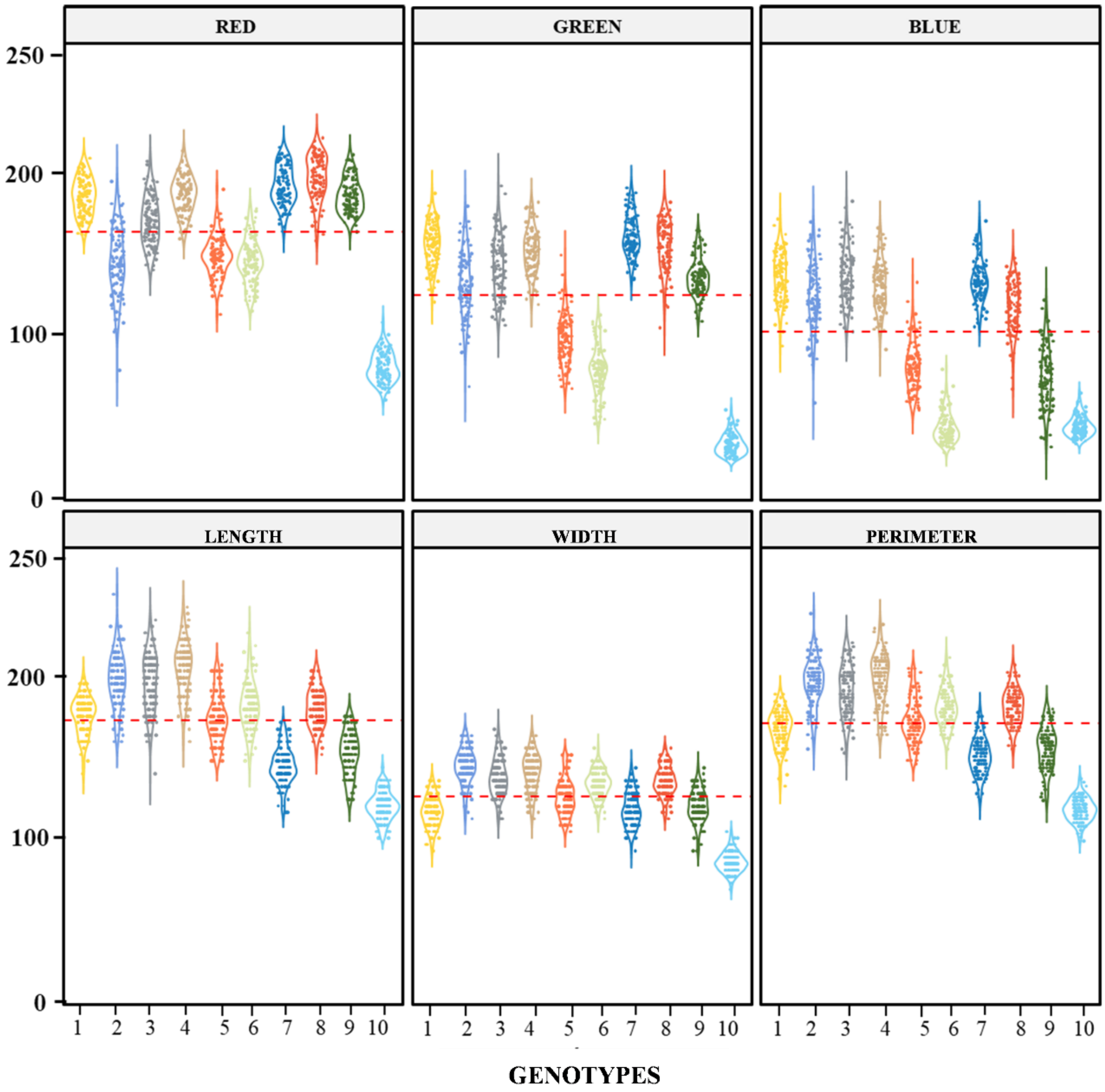
The validation of the model, i.e., estimation of the accuracy of the models tested, was carried out using a 10-fold cross validation, in which the dataset was divided into 10 parts (9 for training and 1 for testing) after the release of all combinations of divisions of the training test.

The process was repeated 3 times for each algorithm, with different divisions of the data into 10 parts to obtain a more accurate estimate. They were evaluated based on True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) results.

The quality of the classification by the machine learning models tested were evaluated by Accuracy and Kappa Indexes (CUNHA; COSTA, 2020).

RESULTS AND DISCUSSION

The different values of RGB of each genotype are shown in Figure 2. The lowest RGB values was found for genotype 10, denoting that it had darker colors, which is consistent with the visual characterization of the genotype (Table 1). Similar dynamics was found for the other genotypes, with RGB numerical data consistent with the morphological descriptors of integument color.



1 = Orelha-de-Vó; 2 = Fava-Rajada-Preta; 3 = Espírito-Santo-Vermelho; 4 = Espírito-Santo-Marrom; 5 = Fava-Rajada; 6 = Fava-Amarela; 7 = Mulatinha; 8 = Fígado-de-Frango; 9 = Fava-Manteiga; 10 = Fava-S.

Figure 2. Violin graphs for color (RGB) and biometric variables.

The biometric data shown through the violin graphs (Figure 2) denoted that the image analysis was able to distinguish the different genotypes, since genotypes 2, 3, and 4 stood out, presenting the highest values. Ponce et al. (2019) showed that seed size can affect germination and vigor of seed lots. In addition, these data are important for conducting characterization activities for germplasm banks and breeding programs, considering that they enable to differentiate genotypes according to their size.

The seed market in Brazil, as well as international markets, searches for a better characterization of seeds quality and time saving practices (MEDEIROS; PEREIRA; SILVA, 2018). The use of digital images stands out for providing a faster and more precise analysis, when compared to traditional seed analysis techniques, which are subjective and time consuming (REGO et al., 2020). Therefore, computational analyses are good tools for seed characterization, and present high capacity to optimize steps in plant breeding programs.

The results of the validation of the model were used to identify the best fit of the data for machine learning, which enables a better classification of the different fava genotypes. Considering the Accuracy and Kappa indexes, the best model for the seed classification (Tables 03 and 04) was the LDA. Medeiros et al. (2020a) used the same metrics for choosing the best model, and recommended LDA, RF, and SVM for quality classification of soybean seeds and seedlings; they concluded that machine learning for identification of seeds and seedlings through images has a high precision.

Table 3. Accuracy Index of the models used with lower limit, first quartile, median, mean, third quartile, and upper limit.

	Accuracy					
	Lower limit	First quartile	Median	Mean	Third quartile	Upper limit
LDA	0.85	0.90	0.91	0.91	0.93	0.96
SVM	0.78	0.85	0.87	0.87	0.90	0.93
KNN	0.75	0.79	0.81	0.82	0.86	0.89
NB	0.66	0.70	0.73	0.73	0.75	0.80
CART	0.30	0.39	0.43	0.42	0.46	0.49
C50	0.79	0.83	0.85	0.85	0.87	0.92
BAGGING	0.79	0.81	0.85	0.85	0.87	0.92
RF	0.78	0.83	0.85	0.85	0.87	0.90
GBM	0.80	0.83	0.85	0.86	0.89	0.93

LDA = Linear Discriminant Analysis; SVM = Support Vector Machine; KNN = K-Nearest Neighbor; NB = Naïve Bayes; CART = Classification and Regression Trees; BAGGING = Bootstrap Aggregating, RF = Random Forest and GBM = Gradient Boosting (GBM).

Table 4. Kappa Index used for testing the efficiency of the models used with with lower limit, first quartile, median, mean, third quartile, and upper limit.

	Kappa					
	Lower limit	First quartile	Median	Mean	Third quartile	Upper limit
LDA	0.83	0.89	0.90	0.90	0.92	0.96
SVM	0.76	0.83	0.86	0.85	0.89	0.92
KNN	0.72	0.77	0.79	0.80	0.84	0.88
NB	0.62	0.66	0.69	0.70	0.72	0.78
CART	0.22	0.32	0.37	0.36	0.40	0.43
C50	0.77	0.81	0.83	0.83	0.86	0.91
BAGGING	0.77	0.79	0.83	0.83	0.85	0.91
RF	0.76	0.81	0.83	0.83	0.86	0.89
GBM	0.78	0.81	0.84	0.84	0.88	0.92

LDA = Linear Discriminant Analysis; SVM = Support Vector Machine; KNN = K-Nearest Neighbor; NB = Naïve Bayes; CART = Classification and Regression Trees; BAGGING = Bootstrap Aggregating, RF = Random Forest and GBM = Gradient Boosting (GBM).

Soyeurt et al. (2020) compared different machine learning models to predict lactoferrin contents in cow milk using infrared spectra and found that two of the four models tested presented better results when combined. It denotes the importance of studies to identify the best models for classification of data or improvement of models.

According to the accuracy indexes of the models for seed classification shown in Table 3, LDA was the model that presented the best performance for classifying the genotypes, presenting an accuracy of 0.83 in the lower limit, 0.90 in the first quartile, 0.91 in the median, 0.91 in the mean, 0.93 in the third quartile, and 0.96 in the upper limit. It shows that the model LDA had a high accuracy for identifying the genotypes; it is important that machine learning models present accuracies close to 1 (HOLANDA et al., 2021).

Kappa index is another efficiency indicator for machine learning models and a metric that denotes, statistically, the consistency between the reference data and the classified data. The mean Kappa Index of the LDA model was 0.90 (Table 4). The Kappa Index varies from 0 to 1; the closest to 1, the better the classification of the data by the models (CUNHA; COSTA, 2020). Thus, LDA is the model that better classified the fava genotypes.

The other models tested also reached satisfactory results. The models SVM, GBM, RF, BAGGING, C50, KNN, and NB reached means of 0.85, 0.84, 0.83, 0.83, 0.83, 0.80, and 0.70, respectively, which are considered high and can be used for differentiation of fava genotypes. Only the model CART presented a low mean (0.36), but it is still considered as a significant value (CUNHA; COSTA, 2020).

Figure 3 shows the values of Tables 4 and 5 in a boxplot, with boxes representing the medians and encompassing the percentiles between 25 and 75, lines at the ends representing the minimum and maximum values, and circles representing the mean. These results reinforce that LDA was the model that had the best values for the classification of genotypes, since it reached a mean value of 0.90, considered excellent (CUNHA; COSTA, 2020).

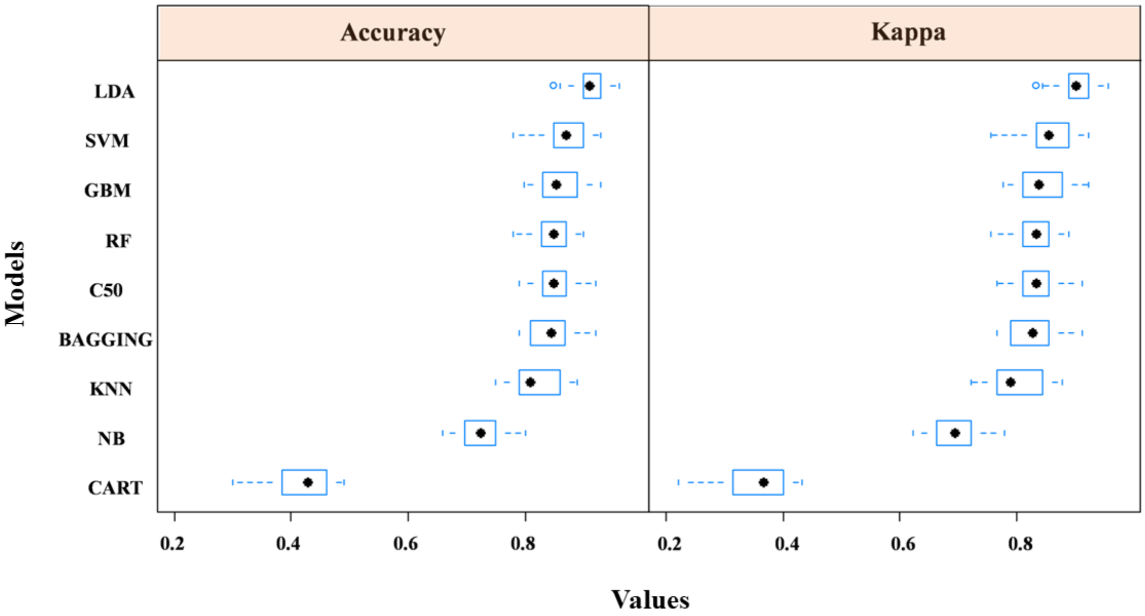


Figure 3. Boxplot for the Accuracy and Kappa Indexes of the models used.

The confusion matrix presented correct and incorrect classifications of the model used, comparing the obtained and expected results. It was used to predict the performance of the machine learning models used. The correct classes (true positive), are distributed in the main diagonal and the incorrect (false positive) in the other elements of the matrix (RAMOS et al., 2018). Thus, the genotypes Fava-Rajada, Fava-Amarela, Fava-Manteiga, and Fava-S showed the highest efficiency in the classification of genotypes by the LDA model, since all individuals were correctly classified, i.e., the accuracy was 100% for these genotypes. Only 17 out of the 19 seeds classified for the genotype Orelha-de-Vó were correctly classified; one was classified as Espírito-Santo-Marrom and one as Mulatinha. The accuracy for this genotype was 93%. Eighteen out of the 20 seeds classified for genotype Fava-Rajada-Preta were correctly classified; two were classified as Espírito-Santo-Vermelho. The accuracy was 95%. Other variations in the classification of genotypes were found, with accuracies above 90% (Table 5).

Table 5. Matrix of confusion generated using data of validation of the LDA model.

	Reference									
Prediction	1	2	3	4	5	6	7	8	9	10
1	17	0	1	3	0	0	1	1	0	0
2	0	18	0	0	0	0	0	0	0	0
3	0	2	17	0	0	0	0	0	0	0
4	1	0	1	17	0	0	0	1	0	0
5	0	0	0	0	20	0	0	0	0	0
6	0	0	0	0	0	20	0	0	0	0
7	1	0	1	0	0	0	19	1	0	0
8	0	0	0	0	0	0	0	17	0	0
9	0	0	0	0	0	0	0	0	20	0
10	0	0	0	0	0	0	0	0	0	20
Total	19	20	20	20	20	20	20	20	20	20

1 = Orelha-de-Vó; 2 = Fava-Rajada-Preta; 3 = Espírito-Santo-Vermelho; 4 = Espírito-Santo-Marrom; 5 = Fava-Rajada; 6 = Fava-Amarela; 7 = Mulatinha; 8 = Fígado-de-Frango; 9 = Fava-Manteiga; 10 = Fava-S.

Koklu, Sarigil and Ozbek (2021) used a confusion matrix to show the success of a model used for classification of pumpkin (*Cucurbita pepo* L.) seeds. In addition, Altuntas, Comert and Kocamaz (2019) used the same classification to predict the performance of CNN models on the identification of haploid and diploid maize seeds. Several metrics for evaluating performance, such as accuracy, sensitivity, and specificity can be derived from confusion matrices (ALTUNTAS; COMERT; KOCAMAZ, 2019). It denotes the importance of these matrices for the validation of the model used for the analysis, based on performance.

Nine machine learning models were tested and compared to classify the different fava genotypes based on morphological, color, and biometric aspects obtained through image analysis. The results showed a high performance in the classification of genotypes for the models tested (Figure 3). The models based on machine learning presented a general accuracy mean of 0.81, which were low due to the model CART, that reached a mean of only 0.42, whereas the other models reached values higher than 0.73 (Table 3).

The LDA model stood out, reaching a mean accuracy of 0.91 and a mean Kappa coefficient of 0.9 (Tables 4 and 5), denoting that image analysis combined with discriminant analysis is an excellent classification tool. This is because the LDA model minimizes the distance between individuals within each genotype and maximizes the distance between different genotypes, using linear algorithms (REGO et al., 2020). The seeds were sampled randomly for training (70% of the samples) and validation (30% of the samples). All genotypes presented a high classification success rate, with sensitivity higher than 0.85 (Table 6). The LDA model presented an accuracy of 100% for the classification of genotypes 5, 6, 9, and 10. Considering the violin graphs (Figure 2), the genotypes that presented higher classification success rates were those that presented lower variation for the parameters evaluated. It indicates that these genotypes are homogeneous and their external characteristics are well defined, which can be important for the characterization of germplasm of the species and, consequently, for the conservation of this genetic resource (LIMA et al., 2018).

Table 6. General statistics for the classification of genotypes by LDA model.

Statistics by Class										
Class	1	2	3	4	5	6	7	8	9	10
Sensitivity	0.89	0.90	0.85	0.85	1.00	1.00	0.95	0.85	1.00	1.00
Specificity	0.97	1.00	0.99	0.98	1.00	1.00	0.98	1.00	1.00	1.00
Positive prediction value	0.74	1.00	0.89	0.85	1.00	1.00	0.86	1.00	1.00	1.00
Negative prediction value	0.99	0.99	0.98	0.98	1.00	1.00	0.99	0.98	1.00	1.00

In addition, genotypes that present integument with only one color were correctly classified, except for genotype 5; it can be one of the factors that favored their identification. However, although genotype 5 shows main and secondary colors, it shows a predominance of secondary color (Figure 1), forming a well-defined pattern in the seed, which may have favored its classification, making it different from the other genotypes with two colors, which present spots or streaks that are visually similar to other genotypes. Similar results were found by Medeiros et al. (2020b) using the LDA model, which enabled to distinguish *Jatropha curcas* seeds with an accuracy of 93.93%.

The genotypes were grouped based on the two first discriminant factors (LD1 and LD2) (Figure 4), which explained 49% and 35% of the total variance, respectively. However, most of those located in the negative quadrant of LD2 were separated and differentiated from the others. Genotypes 1, 2, 3, 4, 7, and 8 were closely grouped, with the centroids closer to each other (Figure 4), contributing to decrease the accuracy of the model; nevertheless, the accuracy rate remained high. The findings were also shown in the confusion matrix, since it is based on the selection of random individuals and the probability of errors during the classification increases as the unevenness is increased.

Thus, genotypes closer to each other and closer to their centroid presented higher classification success rates in the confusion matrix (Table 5). This interaction, which can be seen in the LDA graph, can be attributed to the large variation in physical attributes of individuals. Seeds of traditional varieties have high genetic variability due to natural crossings, exchange of seeds between farmers, and mixtures of varieties during harvest or storage (SOARES, 2018), which may explain these results.

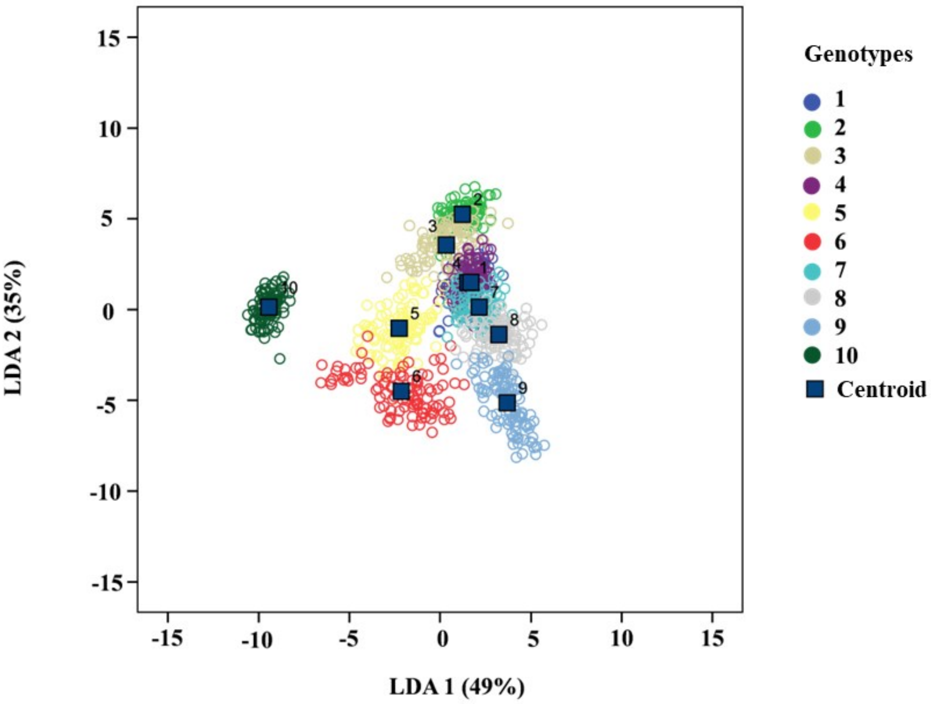


Figure 4. Graph of individuals in the dimensions of linear discriminant analysis (LDA).

Seed and fruit classification using digital image, combined with the LDA model, through colorimetric and biometric data, had been used for other species and shown positive results, denoting a high potential for the use of these tools in agriculture, mainly, for seed classification (ELMASRY et al., 2019). Thus, machine learning techniques are efficient tools for optimization of activities in germplasm banks, in the introduction, characterization, and species conservation phases (ALMEIDA et al., 2021). Thus, machine learning models, combined with data obtained through image analysis are an efficient tool for the classification of fava genotypes.

These results can be used to generate useful information for pre-improvement of genotypes (BAEK et al., 2020) and for evaluation phases of germplasm banks (LODDO; LODDO; DI RUBERTO, 2021). In addition, it can be used for other important evaluations for the market, such as seed analysis (REGO et al., 2020).

CONCLUSION

Variables related to color and size characteristics of genotypes of fava (*Phaseolus lunatus* L.), assessed through two-dimension images, successfully discriminated the different fava genotypes. The results of Accuracy and Kappa Indexes showed that the models tested were able to classify the fava genotypes, except for CART. LDA was the model that presented the highest efficiency for classifying the different fava genotypes, presenting an accuracy of 90%.

REFERENCES

ABUD, H. F. et al. Image analysis of the seeds and seedlings of *Vigna radiata* L. **Revista Ciência Agronômica**, 53: 1-9, 2022.

ADVÍNCULA, T. L. et al. Qualidade física e fisiológica de sementes de *Phaseolus lunatus* L. **Revista Brasileira de Ciências Agrárias**, 10: 341-346, 2015.

ALMEIDA, R. C. et al. Árvore de decisão como ferramenta na classificação de acessos de feijão-fava. **Revista Caatinga**, 34: 471-478, 2021.

ALTUNTAS, Y; COMERT, Z; KOCAMAZ, A. F. Identification of haploid and diploid maize seeds using convolutional neural networks and a transfer learning approach. **Computers and Electronics in Agriculture**, 163: 1-11, 2019.

BAEK, J. et al. High throughput phenotyping for various traits on soybean seeds using image analysis. **Sensors**, 20: 1-9, 2020.

CARLEO, G. et al. Machine learning and the physical sciences. **Reviews of Modern Physics**, 91: 1-47, 2019.

CARMO, M. S. et al. Avaliação de acessos de feijão-fava, para resistência a *Colletotrichum truncatum*, em condições de folhas destacadas e campo. **Summa Phytopathologica**, 41: 292-297, 2015.

CUNHA, M. A; COSTA, S. M. F. Mapeamento da palmeira de açaí (*Euterpe oleracea* Mart.) na floresta Amazônica utilizando imagem de satélite de alta resolução espacial. **Revista Espinhaço**, 9: 40-49, 2020.

ELMASRY, G. et al. Utilization of computer vision and multispectral imaging techniques for classification of cowpea (*Vigna unguiculata*) seeds. **Plant Methods**, 15: 1-16, 2019.

HOLANDA, M. E. S. et al. Aplicação de aprendizado de máquina profundo para detecção por imagens de doenças em frutos do cacaueiro. **International Journal of Development Research**, 11: 47378-47384, 2021.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Produção de Fava**. Disponível em: <https://www.ibge.gov.br/explica/producao-agropecuaria/fava/br>. Acesso em: 11 jun. 2020.

KAYABASI, A. et al. Automatic classification of agricultural grains: Comparison of neural networks. **Neural Network World**, 28: 213-224, 2018.

KOKLU, M.; SARIGIL, S.; OZBEK, O. The use of machine learning methods in classification of pumpkin seeds (*Cucurbita pepo* L.). **Genetic Resources and Crop Evolution**, 68: 2713-2726, 2021.

LIMA, L. F. et al. Manejo de recursos genéticos vegetais. **Anais da Academia Pernambucana de Ciência Agronômica**, 15: 109-126, 2018.

LODDO, A.; LODDO, M.; DI RUBERTO, C. A novel deep learning based approach for seed image classification and retrieval. **Computers and Electronics in Agriculture**, 187: 1-11, 2021.

MEDEIROS, A. D. et al. Interactive machine learning for soybean seed and seedling quality classification. **Scientific reports**, 10: 1-10, 2020a.

MEDEIROS, A. D. et al. Quality classification of *Jatropha curcas* seeds using radiographic images and machine learning. **Industrial Crops and Products**, 146, 1-7, 2020b.

MEDEIROS, A. D.; PEREIRA, M. D.; SILVA, J. A. Processamento digital de imagens na determinação do vigor de sementes de milho. **Revista Brasileira de Ciências Agrárias**, 13: 1-7, 2018.

MOREIRA, I. B. et al. Separation of coriander seeds by Red, Green and Blue image processing. **Ciência Rural**, 52: 1-7, 2022.

MORETI, M. P. et al. Inteligência Artificial no Agronegócio e os Desafios para a Proteção da Propriedade Intelectual. **Cadernos de Prospecção**, 14: 60-77, 2021.

PERINI, L. J. et al. Diversidade genética entre acessos de soja tipo alimento com base no algoritmo de Gower. **Colloquium Agrariae**, 14: 47-57, 2018.

PONCE, R. M. et al. Tamanho da semente e potencial fisiológico de trigo sarraceno. **Revista Científica Rural**, 21: 259-268, 2019.

RAMOS, J. L. C. et al. Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD. In: VII CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2018, Fortaleza. **Anais...** Porto Alegre: SBIE, 2018. p. 1463-1472.

REGO, C. H. Q. et al. Using multispectral imaging for detecting seed-borne fungi in cowpea. **Agriculture**, 10: 1-12, 2020.

SILVA, W. I.; DULTRA FILHO, J. A. Avaliação de caracteres agronômicos e genéticos em acessos de feijão fava no município de Pombal, no semiárido paraibano. In: XV CONGRESSO DE INICIAÇÃO CIENTÍFICA DA UNIVERSIDADE FEDERAL DE CAMPINA GRANDE, 2018, Campina Grande. **Anais...** Campina Grande: XV CICUFCG, 2018. p. 1 -5.

SOARES, L. A. C. **Conservação on farm e avaliação agronômica de variedades crioulas de feijão-fava**. 2018. 94 f. Dissertação (Mestrado em Agronomia: Área de concentração: Agricultura Tropical) - Universidade Federal do Piauí, Teresina, 2018.

SOUSA, C. A. F. et al. Nova abordagem para a fenotipagem de plantas: conceitos, ferramentas e perspectivas. **Revista Brasileira de Geografia Física**, 8: 660-672, 2015.

SOYEURT, H. et al. A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra. **Journal of dairy science**, 103: 11585-11596, 2020.

TORRES, G. X. **Diversidade genética em população segregante de Passiflora via características de sementes**. 2018. 82 f. Dissertação (Mestrado em Produção vegetal) – Universidade Estadual do Norte Fluminense, Rio de Janeiro, 2018.

VASCONCELOS, M. C. et al. Radiography and biometric analysis of broadleaf vegetable seeds. **Revista de Ciências Agrárias**, 61: 1-9, 2018.

WANG, J. et al. Theta-modulated generation of chromatic orbital angular momentum beams from a white-light source. **Optics Express**, 24: 1-6, 2016.

WANG, T. et al. A feasible image-based colorimetric assay using a smartphone RGB camera for point-of-care monitoring of diabetes. **Talanta**, 206: 1-5, 2020.