ABRATES

Journal of Seed Science
ISSN 2317-1545
www.abrates.org.br/revista

ARTICLE

Machine learning in the identification of native species from seed image analysis

Journal of Seed Science, v.46, e202446002, 2024

http://dx.doi.org/10.1590/2317-1545v46277554

Francival Cardoso Felix[1]*, Dagma Kratz[2], Richardson Ribeiro[3], Antônio Carlos Nogueira[2]

ABSTRACT: The identification of seeds from native species is a complex assessment due to the high Brazilian biodiversity and varied characteristics between species. The objective was to apply different machine learning classifiers associated with image analysis to identify seeds of forest species. In total, 155 native species belonging to 42 botanical families were analyzed. In addition, to determine the appropriate machine learning classifier, five supervised learning classification techniques were implemented: decision trees (DT), artificial neural networks (ANN), k-nearest neighbors (k-NN), Naive-Bayes classifier (NBC) and support vector machine (SVM), which had their performance evaluated. For modeling, 66% of the seeds' morphobiometric data were used to train the classifiers, while 34% were reserved for validation. The classifiers are promising tools for identifying species from seed images. The decision tree (DT) classifier showed greater accuracy for correct species identification (82.8%), followed by ANN (81.7%), k-NN (81.7%), NBC (81.1%) and SVM (78.7%). Therefore, it is possible to identify seeds of native species from images and machine learning with a satisfactory accuracy rate. Finally, the decision tree classifier is recommended.

Index terms: artificial intelligence, forest seeds, image processing, morphobiometry, seed identification.

RESUMO: A identificação de sementes de espécies nativas é uma avaliação complexa devido a elevada biodiversidade brasileira e características variadas entre as espécies. Objetivou-se aplicar diferentes classificadores de aprendizado de máquina associado à análise de imagens para identificar sementes de espécies florestais. Foram analisadas 155 espécies nativas pertencentes a 42 famílias botânicas. Para determinar o classificador de aprendizado de máquina adequado, cinco técnicas de classificação por aprendizado supervisionado foram implementadas: árvores de decisão (DT), redes neurais artificiais (ANN), k-vizinhos mais próximos (k-NN), classificador Naive-Bayes (NBC) e máquina de vetores de suporte (SVM), os quais tiveram seu desempenho avaliados. Para modelagem, 66% dos dados morfobiométricos das sementes foram usados para treinamento dos classificadores, enquanto 34% foram reservados para validação. Os classificadores são ferramentas promissoras para a identificação de espécies a partir das imagens de sementes. O classificador por árvores de decisão (DT) apresentou maior acurácia para identificação correta das espécies (82,8%), seguido dos classificadores ANN (81,7%), k-NN (81,7%), NBC (81,1%) e SVM (78,7%). Portanto, é possível realizar a identificação de sementes de espécies nativas a partir de imagens e aprendizado de máquina com taxa satisfatória de acurácia. Recomenda-se o classificador por árvores de decisão.

Termos para indexação: inteligência artificial, sementes florestais, processamento de imagens, morfobiometria, identificação de sementes.

***Corresponding author**
francival.felix@ufrpe.br

Received: 08/13/2023.
Accepted: 11/03/2023.

[1]Universidade Federal Rural de Pernambuco (UFRPE), 52171-900, Recife, PE, Brazil.

[2]Universidade Federal do Paraná (UFPR), 80210-170, Curitiba, PR, Brazil.

[3]Universidade Tecnológica Federal do Paraná (UTFPR), 85503-390, Pato Branco, PR, Brazil.

INTRODUCTION

Seed identification represents a global challenge for researchers for different reasons (Bao and Bambil, 2021), especially for native species due to high biodiversity, similarity between seeds of the same genus, and variations in color, sizes, and shapes. In Brazil, 35,653 plants with seeds have been recognized, of which 8,320 are tree species distributed in 138 families and 938 botanical genera (Reflora, 2020). Generally, the identification and differentiation of forest species is carried out by means of botanical and morphological descriptors based on vegetative and reproductive structures of the plants (Urbanetz et al., 2010; Costa et al., 2016; Ferreira et al., 2020), such as leaves, flowers, fruits, and seeds. However, identifying or differentiating hundreds or thousands of species from seeds is an unfeasible task for the analyst or professional in the forestry area.

In recent years, there have been advances in the identification of species from images of seeds and fruits (Farris et al., 2020). However, the use of digital images for the differentiation of native seeds is a method that has been little explored and has not been validated due to the absence of a dataset. In the forestry sector, artificial intelligence and associated technologies have significant potential for allowing faster and greater data processing (Franklin and Ahmed, 2017; Cao et al., 2018; Xi et al., 2020). For example, with machine learning it is possible to identify complex patterns and correlations at different levels of detail, which can be explored in seed studies.

Machine learning techniques from seed images have been successfully explored in studies with agricultural species. Examples are the use of a high-resolution scanner to assess the texture of tomato seeds and artificial neural networks for the classification of cultivars (Ropelewska and Piecko, 2022), or the classification of barley varieties based on the shape, color, and texture of the seeds (Shi et al., 2022), as well as the use of hyperspectral imaging and machine learning for detection of varieties of soybean seeds (Tan et al., 2019; Zhu et al., 2019; Zhu et al., 2020) and maize seeds (Bao et al., 2019). However, there is a lack of studies with native species of Brazil presenting accessible alternatives of equipment and computational resources for the identification of seeds.

The application of freely accessible and easy-to-use tools for image processing is an option to the use of less accessible equipment. Thus, tools such as ImageJ® software (Ferreira and Rasband, 2012), which allows the extraction of data from an image, can be applied to seeds (Noronha et al., 2018; Felix et al., 2020; Medeiros et al., 2020). In addition, employing artificial intelligence with the use of Weka® software, which contains a collection of machine learning algorithms for data mining tasks, tools for data preprocessing, classification, regression, clustering, association rules, and visualization for further analysis, is a viable option (Weka®, 2018). Weka® software was developed at the University of Waikato, New Zealand, with the aim of identifying information of data obtained from agricultural domains due to its usability; however, its use has been extended to other fields (Škrubej et al., 2015).

Therefore, it is necessary to prove that the associated use of different descriptors in each seed is capable of reducing identification errors due to greater measurement accuracy and low human interference, as well as establishing methodologies for capturing and processing seed images in an accessible, efficient and reproducible way. The objective of this study was to apply different machine learning classifiers associated with image analysis to identify seeds of native forest species in Brazil based on morphobiometric attributes.

MATERIAL AND METHODS

The methodology followed the following work flowchart (Figure 1): (i) acquisition and analysis of images to obtain morphobiometric attributes of the seeds, (ii) data processing and application of different machine learning classifiers, and (iii) selection of the classifier with superior performance in the task of species identification.

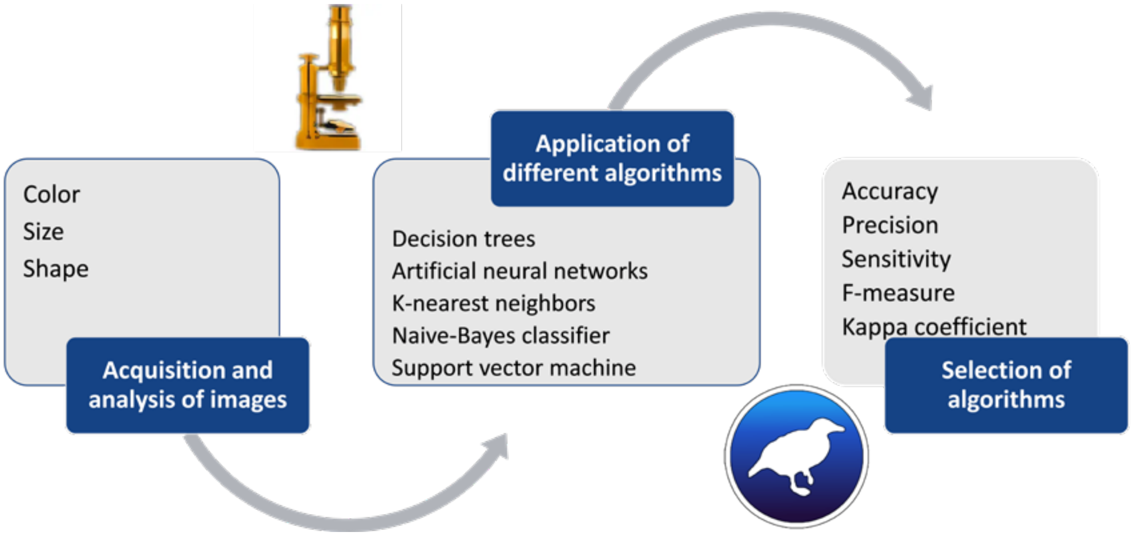

Figure 1. Workflow for acquisition and analysis of images and application and selection of machine learning classifiers.

Acquisition and processing of seed images

Samples of native seeds were photographed and analyzed for size, shape and color. A total of 155 native Brazilian species belonging to 42 botanical families were evaluated. The seeds of each species were spatially arranged on an ethylene-vinyl acetate (EVA) sheet. Then, the images were captured with a Canon PowerShot SX500 IS (f/4) camera with 12 MP lens at a distance of 50 cm from the seeds with millimeter reference template. The images were acquired using a mini photographic studio (50 x 50 x 50 cm) with white artificial light (LED) to standardize the lighting conditions. The methodological detailing for capturing the images was based on the methodology proposed by Felix et al. (2023) for the characterization of forest species. For each photograph, new seed samples were taken by species, totaling 465 images that made up the seed image base used in this study.

Images in their original format (.JPEG) were transferred to a microcomputer, and the reference scale in millimeters was determined in ImageJ® software, version 1.53 (https://imagej.nih.gov/ij/index.html). Next, the Threshold mask was applied to contrast the components of the image. In summary, eight attributes were used for size, three for shape, and six for color, totaling 17 attributes analyzed (Ferreira and Rasband, 2012) (Table 1).

A set of 1.827 million morphobiometric data were obtained from the processing of images of 101,521 seeds for 155 species. Files in *Comma-separated values* format (.CSV) containing the morphobiometric data of the seeds were used for processing of the machine learning models in Weka® software, version 3.8.3 (Weka®, 2018).

Machine Learning & Classification

To determine the appropriate machine learning classifier for species identification, five supervised learning classification techniques were implemented and compared in this study: decision trees (DT), artificial neural networks (ANN), k-nearest neighbors (k-NN), Naive-Bayes classifier (NBC) and support vector machine (SVM). The classifier was selected based on the superior performance in accuracy, precision, sensitivity and F-measure (Witten and Frank, 2005) derived from the confusion matrix for the classification of species from the validation sets, and which also has the highest Kappa correlation coefficient (0.0-1.0). For modeling, 66% of the morphobiometric data of the seeds were used for training the classifiers, while 34% of the data not seen by the classifiers were left for validation.

Decision trees: organizes the knowledge extracted from the dataset in a hierarchical structure similar to a tree, composed of nodes and branches; each internal node represents an attribute and is associated with a test for data classification, while the nodes and leaves of the tree correspond to the classes and the branches represent each of the possible results of the tests applied (Quinlan, 1996). A new example can be classified by following the nodes and

Table 1. Morphobiometric attributes used for analysis of images of native Brazilian forest seeds.

Seed size	Description of attributes
Area	selection of the seed surface (mm²), calculated from the limits defined by the perimeter.
Perimeter	outer limit of seed selection (mm), calculated from the centers of the limit pixels.
Width	width measurement (mm) defined by the smallest bounding rectangle that surrounds the seed selection.
Height	height measurement (mm) defined by the smallest bounding rectangle that surrounds seed selection.
Major	major axis (mm) fitted to an ellipse that surrounds seed selection.
Minor	minor axis (mm) fitted to an ellipse that surrounds seed selection.
Feret	greatest distance (mm) between two points along the seed selection boundary set at an angle of up to 180°.
MinFeret	shortest distance (mm) between two points along the seed selection boundary set at an angle of up to 180°.
Seed shape	**Description of attributes**
Circularity	scalar value (0.0 to 1.0), indicating a perfect circle when close to 1.0 for the shape of the seed relative to its perimeter and an elongated shape when close to zero.
Proportion	relationship between the major and minor axes from an ellipse fitted to the seed image.
Solidity	scalar value (0.0 to 1.0), indicating the relationship between the area of the seed captured in the image and the convex area of each seed.
Seed color	**Description of attributes**
Color	mean gray value (0 to 255) resulting from the sum of pixel values from the seed image surface selection divided by the number of pixels.
Standard deviation (StdDev)	calculation of the standard deviation of the gray color values of the seed image surface used to generate the mean gray value.
Modal gray value (Modal)	gray value (0 to 255) for the color of the most frequently occurring seed surface selection, corresponding to the highest peak in the histogram.
Minimum gray level	minimum values (0 to 255) for the gray color of the seed surface selection.
Maximum gray level	maximum values (0 to 255) for the gray color of the seed surface selection.
Median	median value (0 to 255) of pixels for the color of the seed surface selection.

branches until a leaf is reached. The decision tree modeling process aims to maximize the correct classification of all training data. The J48 algorithm (C4.5) was adopted for this procedure because it is the learning algorithm with this approach most used to generate a decision tree.

Artificial neural networks: simulates the behavior of the human brain, composed of a large number of highly interconnected processing elements similar to the functioning of biological neurons, linked with weighted connections corresponding to brain synapses (McCulloch and Walter, 1943). Multilayer perceptron (MLP) is a common type of artificial neural network that is widely used for classification purposes. For neural network architecture, the learning rate value was set at 0.3 and the impulse rate at 0.2 (Škrubej et al., 2015). The number of neurons in the input and output layers was defined as 17 and 155, respectively, since the number of attributes evaluated was 17 and the total number of species analyzed was 155. The middle layer was constructed with 240 neurons and the training time was set at 500 epochs.

K-nearest neighbors: learns based on instances, analyzing the instances or examples around a specific case. This model calculates the distance between each training sample and the test case based on the Euclidean distance. After classifying all distances, the model selects the nearest *k* from those that are considered to be the nearest *k* neighbors (Aha et al., 1991). If the algorithm returns more than one *k* neighbor, these are voted to form the final ranking.

Naive-Bayes classifier: predicts the class for which the a posteriori probability is higher, given the predictor variables of the case to be classified, based on probability theory using Thomas Bayes' theorem (Shannon, 1948). The Naive-Bayes classifier is one of the Bayesian learning methods.

Support vector machine: constructs a hyperplane with a decision line for classification of instances widely used in various applications (Cortes and Vapnik, 1995; Vapnik, 1995).

Performance Evaluation

Accuracy: percentage of correct predictions made from the tested model when compared with the actual classification of the validation dataset, calculated as a function of the number of correctly classified seeds divided by the total number of seeds, according to the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where: TP refers to the true positive; TN refers to the true negative; FP refers to the false positive; and FN refers to the false negative. Thus, TP + TN + FP + FN is the total number of seeds in the validation set and TP + TN is the number of seeds correctly identified.

Precision: proportion of predicted positive cases that are actually the real ones among all seeds that were classified for each species, calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

Where: TP refers to the true positive and FP refers to the false positive. A false positive occurs when the seed is incorrectly predicted to be positive when in fact it is negative.

Sensitivity: defined as the ratio of the true positive to the sum of the true positive and false negative, calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

Where: TP refers to the true positive and FN refers to the false negative. False negative occurs when the seed predicted as negative is actually positive.

F-measure: defined as a harmonic mean calculated based on precision and sensitivity, calculated as follows:

$$F\ measure = 2 \times \frac{precision \times sensitivity}{precision + sensitivity}$$

Kappa coefficient: defined as a metric that evaluates the level of agreement of a classification task between two or more datasets, indicating which of those classified data have greater cohesion, calculated as follows:

$$Kappa = 2 \times \frac{TP \times TN - FP \times FN}{TP \times FN + TP \times FP + 2 \times TP \times TN + FN^2 + FN \times TN + FP^2 + FP \times TN}$$

Where: TP refers to the true positive; TN refers to the true negative; FP refers to the false positive; and FN refers to the false negative.

RESULTS AND DISCUSSION

The machine learning classifiers tested proved to be promising for the identification of native species from image processing and obtaining of morphobiometric attributes of forest seeds. The decision tree model showed greater accuracy for correct identification of seeds from the validation dataset (82.8%), whose Kappa coefficient was 0.780 (Table 2), followed by classifiers using artificial neural networks (81.7%; 0.763), k-nearest neighbors (81.7%; 0.760), Naive-Bayes classifier (81.1%; 0.749) and support vector machine (78.7%; 0.699). The other performance parameters evaluated were also superior for the decision tree compared to the other classifiers.

The decision tree model proved to be effective for the correct identification of species based on attributes related to seed color, size and shape, as evidenced by the higher precision of seed identification for the validation dataset (0.782) (Table 2). The use of machine learning to solve some question requires some prerequisites, since not all algorithms solve all types of problems, requiring robust dataset and examples, as well as their construction and constant updating (Mitchell, 1997). In addition, it is necessary to select the appropriate sets of classifiers for the problem to be solved. After training, the validation of the classifiers needs to be measured at a level of precision for the problem being solved.

In a study with machine learning and computer vision techniques to classify watermelon seeds, images were captured through a camera, resulting in a classification precision that ranged from 69.5% to 84.3% with the use of the support vector machine (Mukasa et al., 2022). On the other hand, using artificial neural networks to detect the authenticity of maize seeds, a recognition precision of 98.0% was achieved (Tu et al., 2021). A classification system based on alternate circumrotating mechanisms to expose the external characteristics of soybean seeds achieved 97.8% precision through deep learning (Zhao et al., 2021). These different results demonstrate the need for validation of the techniques, since they may show varying levels of precision according to the problem analyzed.

The Kappa coefficient found for the decision tree classifier is considered high (0.780) (Table 2), on a scale from 0.0 to 1.0 (Almeida et al., 2022). However, it should be noted that the classifier revealed differences of precision in the correct identification of the species evaluated from the analysis of seed images, ranging from low to very high precision. The decision tree classifier correctly considered six species with low precision (0.20 to 0.39), 40 species with moderate precision (0.40 to 0.69), 72 species with high precision (0.70 to 0.89) and 37 species with very high precision (0.90 to 1.00) (Table 3).

Unlike agricultural crops that have mostly homogeneous seeds due to genetic improvement (Duan et al., 2022), forest species exhibit significant variation in seed characteristics for the same species. In addition, the proximity of botanical genera represents a confounding factor for the correct identification of seeds, by the method proposed using a mini studio and a camera.

Within the genera evaluated, the similarity or proximity of seed characteristics was one of the reasons for confusion by the decision tree classifier for the correct identification of seeds of species of the genera *Annona* (*A. cacans*, *A. emarginata*, *A. mucosa* and *A. sylvatica*), *Butia* (*B. capitata* and *B. eriospatha*), *Cassia* (*C. grandis* and *C. leptophylla*), *Cecropia* (*C. glaziovii*, *C. hololeuca*, *C. pachystachya* and *C. sciadophylla*), *Enterolobium* (*E. contortisiliquum* and *E. timbouva*) and *Senna* (*S. macranthera*, *S. pendula* and *S. spectabilis*), which were correctly classified with moderate precision (Table 3). Seeds of *Annona* (Annonaceae), *Cassia* (Fabaceae), *Enterolobium* (Fabaceae) and *Senna* (Fabaceae) are larger and very similar to each other within the genus, while those of *Cecropia* (Urticaceae) are also similar and very small, which may have contributed to a lower effectiveness of correct classification.

Table 2. Performance of machine learning classifiers tested for identification of native forest seeds based on morphobiometric attributes.

Classifiers tested	Accuracy (%)	Precision	Sensitivity	F-measure	Kappa coefficient
Decision trees	82.8	0.782	0.783	0.782	0.780
Artificial neural networks	81.7	0.767	0.766	0.767	0.763
K-nearest neighbors	81.7	0.762	0.763	0.762	0.760
Naive-Bayes classifier	81.1	0.754	0.753	0.750	0.749
Support vector machine	78.7	0.703	0.703	0.704	0.699

Table 3. Level of precision in the correct identification of native forest seeds from image analysis associated with machine learning using the decision tree classifier.

Precision for correct identification	*n*	Species
Very low 0.00 to 0.19	0	None
Low 0.20 to 0.39	6	*Cenostigma pyramidale* (0.283), *Mimosa ophthalmocentra* (0.340), *Croton blanchetianus* (0.361), *Manihot carthagenensis* (0.364), *Albizia niopoides* (0.374) and *Genipa americana* (0.398).
Moderate 0.40 to 0.69	40	*Vitex megapotamica* (0.402), *Bixa orellana* (0.408), *Annona cacans* (0.418), *Annona emarginata* (0.428), *Jatropha molissima* (0.429), *Mimosa bimucronata* (0.431), *Cassia grandis* (0.446), *Cecropia hololeuca* (0.451), *Enterolobium contortisiliquum* (0.465), *Cnidoscolus quercifolius* (0.473), *Butia eriospatha* (0.489), *Cenostigma microphyllum* (0.500), *Annona sylvatica* (0.517), *Butia capitata* (0.529), *Lophanthera lactescens* (0.540), *Senna macranthera* (0.541), *Croton floribundus* (0.542), *Senna pendula* (0.556), *Cecropia pachystachya* (0.559), *Senna spectabilis* (0.570), *Cecropia glaziovii* (0.575), *Cecropia sciadophylla* (0.581), *Dalbergia ecastaphyllum* (0.585), *Mimosa tenuiflora* (0.599), *Handroanthus heptaphyllus* (0.613), *Enterolobium timbouva* (0.614), *Euphorbia heterophylla* (0.615), *Euterpe edulis* (0.620), *Lonchocarpus cultratus* (0.626), *Alchornea glandulosa* (0.626), *Schinus terebinthifolia* (0.629), *Piptadenia gonoacantha* (0.633), *Annona mucosa* (0.638), *Lafoensia glyptocarpa* (0.638), *Ceiba speciosa* (0.641), *Leptolobium dasycarpum* (0.643), *Poecilanthe parviflora* (0.647), *Cordia superba* (0.655), *Copernicia prunifera* (0.676) and *Cassia leptophylla* (0.680).
High 0.70 to 0.89	72	*Inga lentiscifolia* (0.703), *Erythroxylum argentinum* (0.704), *Mabea fistulifera* (0.705), *Allophylus guaraniticus* (0.709), *Astronium urundeuva* (0.712), *Gymnanthes klotzschiana* (0.716), *Libidibia ferrea* (0.717), *Platypodium elegans* (0.721), *Diospyros inconstans* (0.727), *Solanum lycocarpum* (0.73), *Manilkara elata* (0.731), *Handroanthus chrysotrichus* (0.738), *Guazuma ulmifolia* (0.738), *Neocalyptrocalyx longifolium* (0.740), *Clitoria fairchildiana* (0.740), *Handroanthus impetiginosus* (0.743), *Psidium myrtoides* (0.743), *Pachira glabra* (0.743), *Erythrina speciosa* (0.747), *Sesbania punicea* (0.752), *Terminalia mameluco* (0.753), *Ilex paraguariensis* (0.755), *Senegalia bonariensis* (0.755), *Allophylus edulis* (0.757), *Helietta apiculata* (0.760), *Dimorphandra mollis* (0.763), *Hymenaea courbaril* (0.763), *Psidium guajava* (0.767), *Mimosa flocculosa* (0.769), *Cassia ferruginea* (0.770), *Mimosa scabrella* (0.772), *Bauhinia forficata* (0.774), *Calopogonium mucunoides* (0.774), *Anadenanthera colubrina* (0.775), *Leucochloron incuriale* (0.776), *Nectandra lanceolata* (0.778), *Sarcomphalus joazeiro* (0.780), *Dalbergia frutescens* (0.781), *Mucuna pruriens* (0.790), *Pityrocarpa moniliformis* (0.792), *Solanum viarum* (0.795), *Lithraea molleoides* (0.798), *Stryphnodendron adstringens* (0.800), *Jatropha curcas* (0.801), *Balfourodendron riedelianum* (0.803), *Matayba elaeagnoides* (0.803), *Monteverdia ilicifolia* (0.804), *Sesbania virgata* (0.818), *Campomanesia xanthocarpa* (0.819), *Gallesia integrifolia* (0.824), *Cryptocarya aschersoniana* (0.833), *Aegiphila integrifolia* (0.833), *Drimys brasiliensis* (0.838), *Podocarpus lambertii* (0.840), *Phytolacca dioica* (0.843), *Pleroma raddianum* (0.843*), Hymenaea altissima* (0.847), *Parapiptadenia rigida* (0.851), *Berberis laurina* (0.851), *Cupania vernalis* (0.855), *Abrus precatorius* (0.857), *Cecropia peltata* (0.858), *Xiquexique gounellei* (0.860), *Handroanthus ochraceus* (0.870), *Zanthoxylum rhoifolium* (0.876), *Machaerium acutifolium* (0.882), *Calliandra brevipes* (0.884), *Cereus jamacaru* (0.885), *Encholirium spectabile* (0.886), *Peltophorum dubium* (0.889), *Ormosia arborea* (0.89) and *Syagrus romanzoffiana* (0.892).
Very high 0.90 to 1.00	37	*Psidium cattleyanum* (0.900), *Prunus brasiliensis* (0.900), *Gaylussacia brasiliensis* (0.900), *Machaerium villosum* (0.901), *Aspidosperma parvifolium* (0.910), *Spondias tuberosa* (0.910), *Mimosa pigra* (0.911), *Araucaria angustifolia* (0.921), *Commiphora leptophloeos* (0.921), *Luehea divaricata* (0.922), *Sapindus saponaria* (0.924), *Pleroma sellowianum* (0.925), *Schizolobium parahyba* (0.933), *Terminalia glabrescens* (0.937), *Pterogyne nitens* (0.939), *Ateleia glazioveana* (0.942), *Handroanthus albus* (0.942), *Hevea brasiliensis* (0.947), *Lafoensia pacari* (0.949), *Solanum granulosoleprosum* (0.951), *Machaerium stipitatum* (0.957), *Vassobia breviflora* (0.958), *Dalbergia brasiliensis* (0.958), *Aspidosperma subincanum* (0.960), *Tabebuia aurea* (0.961), *Schinopsis brasiliensis* (0.963), *Cochlospermum orinocense* (0.965), *Aspidosperma pyrifolium* (0.970), *Amburana cearenses* (0.973), *Cedrela fissilis* (0.983), *Dictyoloma vandellianum* (0.990), *Styrax leprosus* (0.990), *Vernonanthura discolor* (0.991), *Miconia theaezans* (0.992), *Senna multijuga* (0.994), *Myrocarpus frondosus* (0.999) and *Pterocarpus rohrii* (0.999).
Total	155	

n: number of species identified using the decision tree classifier.

On the other hand, seeds of the genera *Allophylus* (*A. edulis* and *A. guaraniticus*), *Aspidosperma* (*A. parvifolium*, *A. pyrifolium* and *A. subincanum*), *Dalbergia* (*D. brasiliensis* and *D. frutescens*), *Handroanthus* (*H. albus*, *H. chrysotrichus*, *H. impetiginosus* and *H. ochraceus*), *Hymenaea* (*H. altissima* and *H. courbaril*), *Machaerium* (*M. stipitatum* and *M. villosum*), *Pleroma* (*P. raddianum* and *P. sellowianum*), *Psidium* (*P. cattleyanum*, *P. guajava* and *P. myrtoides*), *Sesbania* (*S. punicea* and *S. virgata*), *Solanum* (*S. granulosoleprosum*, *S. lycocarpum* and *S. viarum*) and *Terminalia* (*T. glabrescens* and *T. mameluco*) were correctly identified by the decision tree classifier with high or very high precision (Table 3).

It is worth pointing out that *Aspidosperma* (Apocynaceae), *Handroanthus* (Bignoniaceae), *Machaerium* (Fabaceae) and *Terminalia* (Combretaceae) have seeds with dispersal structures (anemochory), while *Pleroma* (Melastomataceae) has tiny seeds. Therefore, these results denote the potential in the identification of forest seeds of the same genus, based on morphobiometric attributes obtained with image analysis and machine learning. However, it is not possible to make generalizations about the effectiveness of correct classification within the botanical genera.

The genera *Cenostigma* and *Croton* were classified with low (*Cenostigma pyramidale* and *Croton blanchetianus*) and moderate (*Cenostigma microphyllum* and *Croton floribundus*) precision by the decision tree classifier for seed identification (Table 3). There was greater confusion of the classifier for these species, due to the similar morphobiometric characteristics of the seeds captured by image. In turn, seeds of the genus *Mimosa* were classified with different levels of precision, being low precision for *Mimosa ophthalmocentra*, moderate precision for *M. bimucronata* and *M. tenuiflora*, and high precision for *M. flocculosa*, *M. pigra* and *M. scabrella*. It is worth noting that the classifier may not have correctly identified some seeds of the species mentioned, but considered them within the genus *Mimosa*.

The decision tree classifier confused seeds of *Cenostigma pyramidale*, for which it showed low precision (0.283), mainly with *Cassia leptophylla*, and to a lesser extent for *Annona cacans*, *Dalbergia ecastaphyllum*, *Piptadenia gonoacantha*, *Anadenanthera colubrina*, *Poecilanthe parviflora* and *Annona sylvatica*. Except for the genus *Annona*, the other species belong to the Fabaceae family, which may have contributed to the lower precision of correct identification of *Cenostigma pyramidale* seeds, as verified by the confusion matrix. A point to be considered is the fact that the extraction of characteristics analyzed for the species in question was not enough to correctly differentiate it by images of seeds.

For *Mimosa ophthalmocentra* (0.340), greater confusion occurred with *Mimosa bimucronata*, a species of the same genus and with similar morphobiometric characteristics of seeds. The genus *Mimosa* is known to be rich in species diversity, with about 378 native species occurring in Brazil, 42 of which are arboreal (Reflora, 2020). The differentiation of species of this genus by seed images is a challenge and should be explored in future studies covering more species. High precision of correct identification from seed images was achieved for differentiation of *M. bimucronata*, *M. tenuiflora*, *M. flocculosa*, *M. pigra* and *M. scabrella*.

Croton blanchetianus (0.361) showed greater confusion with *Croton floribundus*, *Mimosa bimucronata*, *Gymnanthes klotzschiana*, *Euphorbia heterophylla*, *Senna spectabilis,* *Mimosa ophthalmocentra*, *Schinus terebinthifolia* and *Monteverdia ilicifolia*. Seeds of the Euphorbiaceae family stand out, except for *Mimosa bimucronata* (Fabaceae). The higher rate of confusion for *C. blanchetianus*, *C. floribundus*, *G. klotzschiana* and *E. heterophylla* can be attributed to the fact that the seeds have a caruncle, a type of aril that persists after seed maturation, remains adhered and confers distinct characteristics of color and shape to the seeds, also influencing their position at the time of image acquisition.

On the other hand, winged seeds or seeds with dispersal structures that confer different shapes due to the conformation of the wings were classified with high rates of correct identification of the species, as in the case of the genera *Aspidosperma*, *Handroanthus*, *Machaerium*, *Tabebuia* and *Terminalia* (Table 3). The differentiation of species of the Bignoniaceae family *(Handroanthus* and *Tabebuia*) may be more complex compared to other botanical families, due to the similar morphology of their seeds, and because they have 29 genera and 414 native species (Reflora, 2020). For *Manihot carthagenensis* (0.364), greater confusion occurred with *Dalbergia ecastaphyllum* and *Cenostigma pyramidale*, while *Albizia niopoides* (0.374) showed greater confusion with *Euphorbia heterophylla*, and *Genipa americana* (0.398) with *Vitex megapotamica*, *Cenostigma microphyllum* and *Cordia superba*.

Genipa americana seeds do not have a standard shape and size, so seeds with different characteristics can occur. Thus, the skewness coefficient of this species indicates that smaller seeds predominate in the sample, while kurtosis (k<3) indicates that there is a greater amplitude of distribution of the frequency of their biometric characteristics relative to a normal curve (Sobrinho et al., 2017). This fact justifies the low capacity of the decision tree classifier to recognize this species only with the morphobiometry of the seed.

It is worth noting that the use of the proposed methodology for capturing images with a camera associated with a mini studio, image processing in open access software and the use of machine learning proved to be adequate for more than 70% of the species studied. Therefore, it is relevant to consider that image processing and machine learning to identify native forest seeds involve additional challenges and complexities.

Among the challenges encountered, we highlight the variety of seed sizes of the same species, similar characteristics for some botanical genera, minute seeds that cannot be photographed by common means, and the presence of diaspores or dispersal structures linked to the seeds that give them distinct characteristics, such as wings, arils, caruncles, and persistent pericarps. In addition, it is important to perform the extraction of relevant attributes and the proper selection of these to feed the classifier, and the availability of a representative and high-quality training dataset is critical to obtain reliable results.

Among the possibilities explored, the implementation of applications with machine learning techniques for seed identification stands out, which can bring significant benefits in terms of automation, inspection, and taxonomic identification. Therefore, it is important to consider conducting future studies using advanced deep learning techniques in native seeds, with a view to the creation of applications or integrated tools for capturing and processing images aimed at identification.

Deep learning can play a prominent role in identifying seeds in future studies, through complex algorithms and artificial neural networks. The capacity to handle variations in lighting, viewing angles, and noise in images makes it extremely efficient in recognition tasks. However, it is important to highlight the requirement of advanced programming knowledge, robust computational resources, and interpretation skills of the researcher.

Finally, the proposed method, which involves the use of a digital camera and a mini studio, is applicable for the identification of forest seeds. However, it should be noted that in cases of species with a low rate of recognition by the classifier or a higher degree of confusion, the proposed method may not be adequate. In these cases, further research and improvements in the methodology are needed to meet the demands of the native forest species in question, such as exploring alternatives for image capture, larger sample sizes, and deep learning techniques.

CONCLUSIONS

Image processing and the use of machine learning techniques make it possible to identify native forest seeds with a satisfactory accuracy rate. Classifiers based on decision trees are recommended.

ACKNOWLEDGEMENTS

To CNPq (National Council for Scientific and Technological Development) for granting a scholarship and financial support to the first author, and to the national and regional research, production and teaching institutions that provided seeds used in this study: (i) Embrapa Forests, (ii) Water and Land Institute (IAT/PR), (iii) Chauá Society, (iv) Forest Seed Laboratory (UFPR), (v) *Bolsa de Sementes* Program (AFUBRA/UFSM), (vi) Biodiversity Conservation Center of the São Paulo State Environmental Research Institute, and (vii) Ecology and Environmental Monitoring Center (NEMA/UNIVASF).

REFERENCES

AHA, D.W.; KIBLER, D.; ALBERT, M.K. Instance-based learning algorithms. *Machine Learning*, v.6, n.1, p.37-66, 1991. https://link.springer.com/article/10.1007/BF00153759

ALMEIDA, F.A.; ROMÃO, E.L.; GOMES, G.F.; GOMES, J.H.F.; PAIVA, A.P.; FILHO, J.M.; BALESTRASSI, P.P. Combining machine learning techniques with Kappa–Kendall indexes for robust hard-cluster assessment in substation pattern recognition. *Electric Power Systems Research*, v.206, e107778, 2022. https://doi.org/10.1016/j.epsr.2022.107778

BAO, F.; BAMBIL, D. Applicability of computer vision in seed identification: deep learning, random forest, and support vector machine classification algorithms. *Acta Botanica Brasilica*, v.35, n.1, p.17–21, 2021. https://doi.org/10.1590/0102-33062020abb0361

BAO, Y.; MI, C.; WU, N.; LIU, F.; HE, Y. "Rapid classification of wheat grain varieties using hyperspectral imaging and chemometrics." *Applied Sciences*, v.9, n.19, e4119, 2019. https://doi.org/10.3390/app9194119

CAO, J.; LIU, K.; LIU, L.; ZHU, Y.; LI, J.; HE, Z. Identifying mangrove species using field close-range snapshot hyperspectral imaging and machine-learning techniques. *Remote Sensing*, v.10, n.12, e2047, 2018. https://doi.org/10.3390/rs10122047

CORTES, C.; VAPNIK, V. Support-vector network. *Machine Learning*, v.20, n.3, p.273–297, 1995. http://dx.doi.org/10.1007/BF00994018.

COSTA, M.F.; LOPES, A.C.A.; GOMES, R.L.F.; ARAÚJO, A.S.F.; ZUCCHI, M.I.; PINHEIRO, J.B.; VALENTE, S.E.S. Characterization and genetic divergence of *Casearia grandiflora* populations in the Cerrado of Piaui State, Brazil. *Floresta e Ambiente*, v.23, n.3, p.387-396, 2016. https://doi.org/10.1590/2179-8087.007115

DUAN, Z.; MIN, Z.; ZHIFANG, Z.; SHAN, L.; LEI, F.; XIA, Y.; YAQIN, Y.; YI, P.; GUOAN, Z.; SHULIN, L.; ZHIXI, T. Natural allelic variation of GmST05 controlling seed size and quality in soybean. *Plant Biotechnology Journal*, v.20, n.9, p.1807-1818, 2022. http://dx.doi.org/10.1111/pbi.13865

FARRIS, E.; ORRÙ, M.; UCCHESU, M.; AMADORI, A.; PORCEDDU, M.; BACCHETTA, G. Morpho-colorimetric characterization of the Sardinian endemic taxa of the genus *Anchusa* L. by seed image analysis. *Plants*, v.9, n.10, p.1–14, 2020. https://doi.org/10.3390/plants9101321

FELIX, F.C.; MEDEIROS, J.A.D.; FERRARI, C.S.; VIEIRA, F.A.; PACHECO, M.V. Biometry of *Pityrocarpa moniliformis* seeds using digital imaging: implications for studies of genetic divergence. *Brazilian Journal of Agricultural Sciences*, v.15, n.1, e6128, 2020. https://doi.org/10.5039/agraria.v15i1a6128

FELIX, F.C.; KRATZ, D.; RIBEIRO, R.; NOGUEIRA, A.C. Characterization and differentiation of forest species by seed image analysis: a new methodological approach. *Ciência Florestal*, v.33, n.3, e73427, 2023. https://doi.org/10.5902/1980509873427

FERREIRA, R.L.A.; CERQUEIRA, R.M.; CARDOSO-JUNIOR, R.C. Analysis of botanical identification in forest inventories of sustainable management plans on wester Pará state, Brazil. *Nature and Conservation*, v.13, n.3, p.136-145, 2020. https://doi.org/10.6008/CBPC2318-2881.2020.003.0014

FERREIRA, T.; RASBAND, W**.** *ImageJ: user guide (IJ 1.46r)*, 2012. 198p.

FRANKLIN, S.E.; AHMED, O.S. Deciduous tree species classification using object-based analysis and machine learning with unmanned aerial vehicle multispectral data. *International Journal of Remote Sensing*, v.39, p.5236-5245, 2017. https://doi.org/10.1080/01431161.2017.1363442

McCULLOCH, W.S.; WALTER, P. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, v.5, n.4, p.115-133, 1943.

MEDEIROS, A.D.; PINHEIRO, D.T.; XAVIER, W.A.; SILVA, L.J.; DIAS, D.C.F.S. Quality classification of *Jatropha curcas* seeds using radiographic images and machine learning. *Industrial Crops and Products*, v.146, p.112-162, 2020. https://doi.org/10.1016/j.indcrop.2020.112162

MITCHELL, T. M. *Machine Learning*. McGraw–Hill Science/Engineering/Math, 1997. 421p.

MUKASA, P.; WAKHOLI, C.; FAQEERZADA, M.A.; AMANAH, H.Z.; KIM, H.; JOSHI, R.; SUH, H.K.; KIM, G.; LEE, H.; KIM, M.S.; BAEK, I.; CHO, B.K. Nondestructive discrimination of seedless from seeded watermelon seeds by using multivariate and deep learning image analysis. *Computers and Electronics in Agriculture*, v.194, e106799, 2022. https://doi.org/10.1016/j.compag.2022.106799

NORONHA, B.G.; PEREIRA, M.D.; FLORES, A.V.; DEMARTELAERE, A.C.F.; MEDEIROS, A.D. Morphometry and physiological quality of *Moringa oleifera* seeds in the function of their fruit position. *Journal of Experimental Agriculture International*, v.25, n.6, p.1-10, 2018. https://doi.org/10.9734/JEAI/2018/43375

QUINLAN, J.R. Learning decision tree classifiers. *ACM Computing Surveys*, v.28, n.1, p.71-72, 1996. https://doi.org/10.1145/234313.234346

REFLORA. *Flora do Brasil 2020*. Jardim Botânico do Rio de Janeiro. http://floradobrasil.jbrj.gov.br/

ROPELEWSKA, E.; PIECKO, J. Discrimination of tomato seeds belonging to different cultivars using machine learning. *European Food Research and Technology*, v.248, n.3, p.685–705, 2022. https://doi.org/10.1007/s00217-021-03920-w

SHANNON, C.E. A mathematical theory of communication. *Bell System Technical Journal*, v.27, n.4, p.623-656, 1948.

SHI, Y.; PATEL, Y.; ROSTAMI, B.; CHEN, H.; WU, L.; YU, Z.; LI, Y. Barley variety identification by IPhone images and deep learning. *Journal of the American Society of Brewing Chemists*, v.80, n.3, p.215-224, 2022. https://doi.org/10.1080/03610470.2021.1958602

ŠKRUBEJ, U.; ROZMAN, C.; STAJNKO, D. Assessment of germination rate of the tomato seeds using image processing and machine learning. *European Journal of Horticultural Science*, v.80, n.2, p.68-75, 2015. http://dx.doi.org/10.17660/eJHS.2015/80.2.4

SOBRINHO, S.P.; ALBUQUERQUE, M.C.F.; LUZ, P.B.; CAMILI, E.C. Physical characterization of fruits and seeds of *Lafoensia pacari*, *Alibertia edulis* and *Genipa americana*. *Revista de Ciências Agrárias*, v.40, n.2, p.382-389, 2017. https://doi.org/10.19084/RCA16034

TAN, K.; RUNTAO W.; MINGYING L.; ZHENPING G. Discriminating soybean seed varieties using hyperspectral imaging and machine learning. *Journal of Computational Methods in Sciences and Engineering*, v.19, n.4, p.1001-1015, 2019. http://dx.doi.org/10.3233/JCM-193562

TU, K.; WEN, S.; CHENG, Y.; ZHANG, T.; PAN, T.; WANG, J.; WANG, J.; SUN, Q. A non-destructive and highly efficient model for detecting the genuineness of maize variety 'JINGKE 968' using machine vision combined with deep learning. *Computers and Electronics in Agriculture*, v.182, e106002, 2021. https://doi.org/10.1016/j.compag.2021.106002

URBANETZ, C.; TAMASHIRO, J.Y.; KINOSHITA, L.S. Chave de identificação de espécies lenhosas de um trecho de floresta ombrófila densa atlântica, no sudeste do Brasil, baseada em caracteres vegetativos. *Biota Neotropica*, v.10, n.2, p.350-388, 2010. https://doi.org/10.1590/S1676-06032010000200036

VAPNIK, V.N. *The nature of statistical learning theory*. Springer, New York. 1995. 188p. http://dx.doi.org/10.1007/978-1-4757-2440-0

WEKA, *Waikato Environment for Knowledge Analysis*, versão 3.8.3. Universidade de Waikato, Nova Zelândia. 2018. https://weka.softonic.com.br/

WITTEN, I.; FRANK, E. *Data mining: practical machine learning tools and techniques*, Morgan Kaufmann, 2005. 629p.

XI, Z.; HOPKINSON, C.; ROOD, S.B.; PEDDLE, D.R. See the forest and the trees: effective machine and deep learning algorithms for wood filtering and tree species classification from terrestrial laser scanning. *ISPRS Journal of Photogrammetry and Remote Sensing*, v.168, p.1-16, 2020. https://doi.org/10.1016/j.isprsjprs.2020.08.001

ZHAO, G.; QUAN, L.; LI, H.; FENG, H.; LI, S.; ZHANG, S.; LIU, R. Real-time recognition system of soybean seed full-surface defects based on deep learning. *Computers and Electronics in Agriculture*, v.187, e106230, 2021. https://doi.org/10.1016/j.compag.2021.106230

ZHU, S.; CHAO, M.; ZHANG, J.; XU, X.; SONG, P.; ZHANG, J.; HUANG, Z. Identification of soybean seed varieties based on hyperspectral imaging technology. *Sensors*, v.19, n.23, e5225, 2019. http://dx.doi.org/10.3390/s19235225

ZHU, S.; ZHANG, J.; CHAO, M.; XU, X.; SONG, P.; ZHANG, J.; HUANG, Z. A rapid and highly efficient method for the identification of soybean seed varieties: hyperspectral images combined with transfer learning. *Molecules*, v.25, n.1, e152, 2020. http://dx.doi.org/10.3390/molecules25010152