Short communication

Applicability of computer vision in seed identification: deep learning, random forest, and support vector machine classification algorithms

Francielli Bao[1]* and Deborah Bambil[2]

ABSTRACT

The use of computer image analysis can assist the extraction of morphological information from seeds, potentially serving as a resource for solving taxonomic problems that require extensive training by specialists whose primary method of examination is visual identification. We propose to test the ability of deep learning, SVM and random forest algorithms to classify seeds from twelve species of aquatic plants as an alternative to traditional classification methods. A total of 150 seeds of the species were collected. The attributes of colour, shape, and texture were analysed through the machine learning algorithms of deep learning, random forest, and support vector machine (SVM). Computer vision proved to be efficient at classifying species using all three algorithms, with an accuracy rate for SVM of 97.91 %, random forest 97.08 % and deep learning 92.5 %. We believe that the method performed well in our experiment and improved seed classification accuracy. As a result, the algorithms SVM and random forest were found to be enough at aquatic plant seed recognition.

Keywords: aquatic macrophyte seeds, colour, machine learning, shape, texture

Seed identification poses a challenge to researchers worldwide, for a variety of different reasons. Studies with wetland seed banks require accuracy for identifying seeds of amphibious and aquatic plants that may be less than 1mm in size (Tirintan *et al*. 2018). The difficulty of accessing certain areas during seasonal floods emphasises the necessity of seed banks for the examination of all species present in vegetation (Souza *et al*. 2016). For that purpose, the direct seed count assessment method is ideal because it reflects the entire plant community (Bonis *et al*. 1995; Bao *et al*. 2021). However, to accurately identify seeds it is necessary to know their morphological characteristics.

1 Departamento de Biodiversidade, Instituto de Biociências, Universidade Estadual Paulista, 13506-752, Rio Claro, SP, Brazil
2 Departamento de Biologia Celular, Universidade de Brasília, 709010-900, Brasília, DF, Brazil

* Corresponding author: franbao@yahoo.com.br

Morphological studies with herbaceous seeds from wetlands are rare, sometimes restricted to a single genus or family (*e.g.*, Groth 1983; Kaul 1978; 1985; Gil & Bove 2006; Souza & Giulietti 2014) or floristic inventories and general catalogues (Tirintan *et al.* 2018); alongside this, specialists are rare. Therefore, the use of computer vision technology is an alternative that aims to facilitate and accelerate these processes using algorithmic bases (Wäldchen & Mäder 2018a). However, to create a software capable of providing accurate identification, it is necessary to verify the performance of different classification algorithms using image databases (Wäldchen & Mäder 2018b).

The algorithm classification allows one to verify which has the best performance of classification and, thus, to analyse and evaluate the best method for solving problems (Bambil *et al.* 2020). The use of different algorithms is necessary for information extraction and seed classification, as these algorithms address morphological aspects such as shape (Granitto *et al.* 2005), colour and texture (Granitto *et al.* 2002), as well as size (Granitto *et al.* 2003), aspects that are adequate for the classification of different types of seeds (Wäldchen *et al.* 2018). The most common algorithms of this type of extraction are the Naïve Bayes algorithm, the neural network classification algorithm, and the boosting algorithm in weed seeds (Granitto *et al.* 2002; 2003; 2005). However, the resources extracted by these algorithms can miss important information for the identification of noise and blockages (Xinshao & Cheng 2015). Thus, deep learning has presented the best results for extracting high-level information from the image resource in weed seeds (Xinshao & Cheng 2015), as well as the SVM and random forest algorithms that have only been tested in studies with rice seeds (Hong *et al.* 2015) and also showed positive results in studies with pollens (Allen *et al.* 2008), which make these algorithms better alternatives for studies with aquatic plant seeds, due to their small size and the high presence of detailed structures that mark the tegument.

Many significant studies have been done on the automatic classification of seeds, however, these have all been focused on agricultural seeds or grasses used in grasslands, for an industrial of economic use (Granitto *et al.* 2002; 2003; 2005; Xinshao & Cheng 2015; Wäldchen *et al.* 2018). Aquatic plants have a major role in maintaining plant diversity globally, and the seed bank assessment method has been instrumental in capturing and developing a database of seed images (Bao *et al.* 2021). Deepening the knowledge in species classification helps not only taxonomic and systematic studies but also in studies of the ecological and evolutionary processes of different ecosystems. In this article, we propose to test deep learning, SVM, and random forest algorithms to solve issues in classification regarding twelve species of aquatic plants that were acquired in a long-term seed bank study.

Seeds were collected from aquatic macrophytes occupying a seasonally flooded grassland, Fazenda São Bento

(19°29'27.3" S; 57°01'55.9" W), in the Aboral sub-region, Pantanal, Mato Grosso do Sul (Central-West Brazil) (Bao *et al.* 2014). These fields have an annual summer flooding period (between February and May), resulting from rainfall and the overflowing of nearby rivers (Silva & Abdon 1998), whereby extensive seasonal ponds are formed due to a slight unevenness in the topography of the area which is characterized by a seed bank rich in native species of different growth forms (Bao *et al.* 2018).

We chose twelve aquatic macrophytes with high abundance in the soil seed bank (Bao *et al.* 2018): *Bacopa australis* V.C. Souza, *B. salzmannii* (Benth.) Wettst. ex Edwall, *B. stricta* (Schrad.) Edwall (Plantaginaceae), *Eleocharis acutangula* (Roxb.) Schult. (Cyperaceae), *Hydrocleys parviflora* Seub. (Alismataceae), *Helanthium tenellum* (Mart. ex Schult. & Schult. f.) Britton (Alismataceae), *Heteranthera limosa* (Sw.) Vahl (Pontederiaceae), *Limnocharis flava* (L.) Buchenau (Pontederiaceae), *Ludwigia leptocarpa* (Nutt.) H. Hara, *L. octovalvis* (Jacq.) P.H. Raven (Onagraceae), *Rotala ramosior* (L.) Koehne (Lythraceae) and *Scirpus supinus* L. (Cyperaceae). *H. parviflora, H. limosa* and *L. flava* are strictly aquatic species, while the other species can germinate only in moist soil and are found among the vegetation during the flood and drought seasons (Bao *et al.* 2014; 2018). All chosen species are found in the seed bank throughout the year (Bao *et al.* 2014; Souza *et al.* 2016). The seeds were collected from a long-term experiment with seed banks (cf. Bao *et al.* 2014; 2018). All seeds were separated from the soil using the direct counting method, with the aid of a microscope (Bonis *et al.* 1995). The seeds were collected from ca. 80 matrices, where 20 seeds of each species were separated to capture images with the aid of a sterile microscope encoded model Leica M125.

The Inovtaxon software was used to extract 226 attributes ([Tab. S1 in supplementary material](#)), which were divided by colour, shape, and texture. The RBG colour model was used to identify red, green, and blue colours, and HSV to distinguish blue, red, yellow, green and purple. The saturation was classified by patterns: a) the highest saturation was the most vivid colour and b) the lowest saturation was the lightest colour (*e.g.*, Amma *et al.* 2013; Danelljan *et al.* 2014), while CIELab patterns indicated luminosity and chromatic coordinates between red and green, and yellow and blue (Kruse *et al.* 2014). To extract shape features, 7 Hu moments were used to identify the invariant region in scale, translation, and rotation (Jia *et al.* 2014). In addition, the orientation was assessed using a gradient histogram that looks for colour variations and identifies the object by examining shape segmentation (Xiao *et al.* 2010). The extracted texture attributes were analysed with "co-occurrence matrix", with the aim of evaluating the co-occurrence values based on a grayscale in the entire image (Harralick 1979; Jafarpour 2012). Thus, the binary local pattern was used in pixel intensity comparisons between

grayscale neighbours, filtering the edges into a uniform pattern (Zhou *et al.* 2012).

The deep learning algorithm is then executed from a convolutional instance with spatial interpretation of various attributes. To use this model several GPUs are necessary (however if there is one GPU or no GPU infrastructure available this can be ignored) (Lang *et al.* 2019). Random forest divides class data into sets of trees that form forests from which the classification is made, working efficiently with large data sets (Rajagopal *et al.* 2013). The SVM algorithm classifies using machine learning techniques, plotting a hyperplane dividing the classes, and the data of these classes that are closest to the margin of the hyperplane are the support vectors (He *et al.* 2016; Kremic & Subasi 2016). For further information, features extracted are available (see Video S1 in supplementary material) and a user guide that explains step-by-step how to use the software (Inovtaxon) is also provided. This guide helps the implementation and execution of the software. The Inovtaxon tool and user guide are available for download and are open source (https://github.com/DeborahBambil/Inovtaxon).

The percentage of correct classification (PCC) was used to analyse the results (*e.g.*, Bambil *et al.* 2020). The results of the general percentage of the algorithm's classification of all trained classes and ROC area are characterized by the weighted average of precision and recall (Amarnath *et al.* 2018). The analysis of variance (One-Way ANOVA) was done to compare the ROC area between SVM, random-forest and deep learning followed by a Tukey post hoc test (HSD = 0.05) using the packages vegan (Oksanen *et al.* 2017), permute (Simpson 2016) and lattice (Deepayan 2008) in the software R (R Development Core Team 2020). To analyse the individual performance of each seed species the confusion matrix was used, which allows us to see the number of samples classified of each species (cf. Arafat *et al.* 2016).

The percentage of correct classification was 97.91 to SVM, 97.08 to random forest, and 92.5 to deep learning. The average of the ROC curve was 0.996 to deep learning, 1,000 for random forest, and 0.998 to SVM (Fig. 1). The classification algorithms showed differences between random forest and deep learning (ANOVA: $F_{1.05}$=4.045, P<0.05, Fig. 1). However, SVM did not show differences with deep learning (ANOVA: $F_{1.05}$=3.145, P=0.262, Fig.1) and random forest (ANOVA: $F_{1.05}$=4.241, P=0.178, Fig.1).

In the heatmap of the confusion matrix, it was possible to analyse the classification of the samples individually and verify the performance of the algorithm for each species (class). It is possible to see, in the individual performance, the discrepancy between the results of SVM and deep learning. The correct classification scale of the SVM stands out between 19 and 20 samples, from a total of 20, meaning that 100 % of the samples were correctly classified. The SVM achieved total success with 6 species (classes): *B. salzmannii*
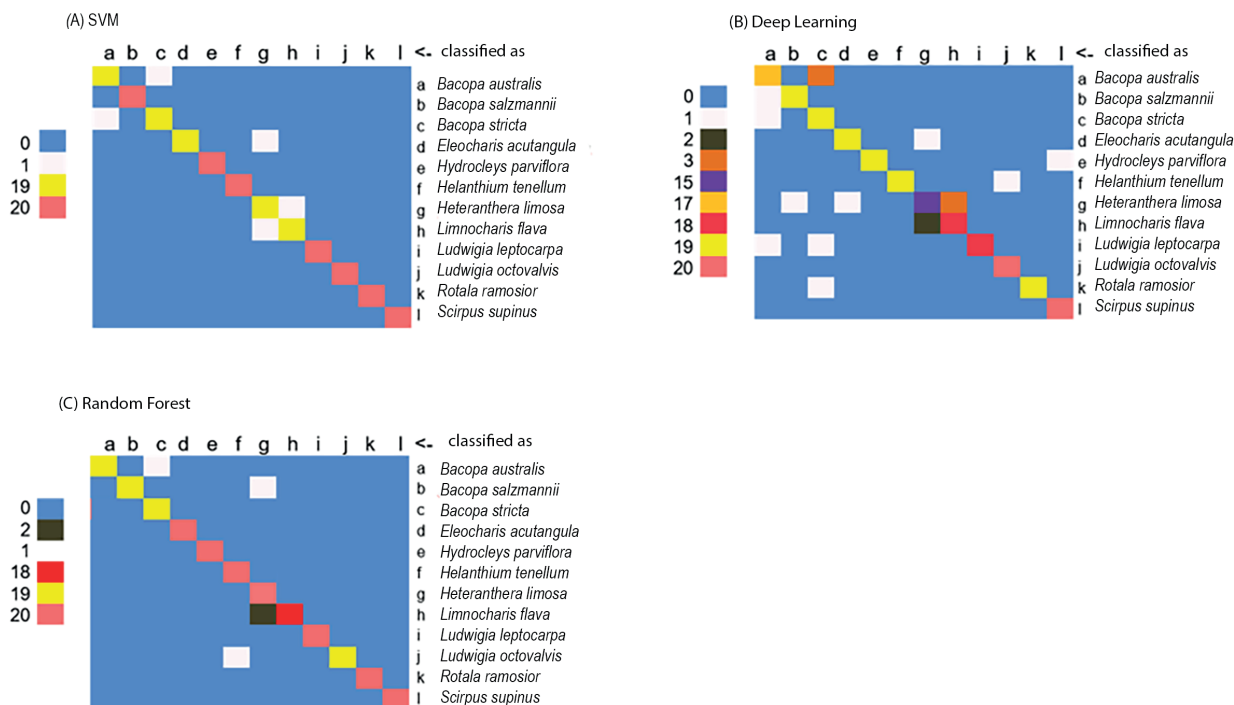
(b), *E. acuntangula* (d), *H. parviflora* (e), *L. leptocarpa* (i), *L. octovalvis* (j), *R. romosior* (k) (Fig. 2A). The deep learning matrix also had a higher margin of error compared to SVM, as the number of samples which were classified correctly range from 15 to 20; the network could only get 100 % (20) of the samples correctly in *L. octovalvis* (j) (Fig. 2B). Finally, for random forest the species *H. parviflora* (e) and *L. flava* (h) showed the same classification error (Fig.2C).
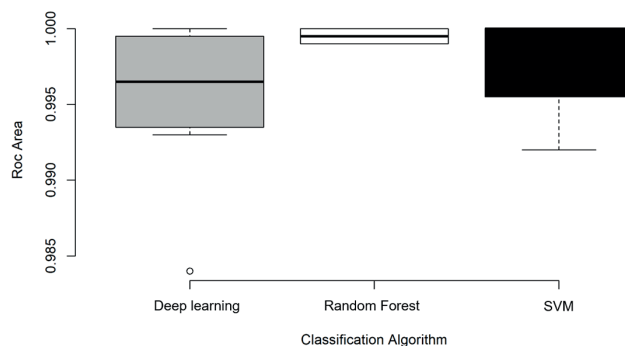
Figure 1. Analysis of variance (ANOVA) of the general percentage of the ROC area from twelve macrophyte aquatic seeds between Deep learning, Random Forest and SVM algorithms.

The analysis of colour, shape, and texture in the classification of seeds generated excellent results for our study, as well as in other studies that described the same efficiency of SVM, random forest and deep learning algorithms using the features of colour, shape, and texture (Patil & Kumar 2017). We used a method that has been successful in classifying biological data, performed to classify leaves and pollens. In this experiment, the deep learning network obtained a result very close to SVM, in which both performed better (Bambil *et al.* 2020). Comparing different classification techniques makes it possible to reduce the issues surrounding algorithm choice in the classification of features (Wen *et al.* 2015).

The best results with random forest emphasize that this is a promising set learning algorithm which has been increasingly used for image classification (Shi & Yang 2016). The random forest deals with many input variables and selects those considered significant and assigns importance to each, thus better balancing the errors in the data (Wen *et al.* 2015). The same can be said for SVM, which obtained excellent results in seed segmentation as it is trained in the initial super pixels of an input image (Park *et al.* 2016). This demonstrates the ability of SVM to classify complex elements, dealing mainly with morphological variables (Kremic & Subasi 2016). Despite deep learning being the most popular method contemporaneously we found better results with SVM, and random forest, unlike studies with the use of the deep learning network in recent times, such as the record obtained in the correct classification of FaceNet, which is used for accurate facial recognition (Schroff *et al.* 2015).

Figure 2. Co-occurrence matrix used to describe the performance of each algorithm: (**A**) the correct classification scale of the SVM stands out between 19 and 20 samples, the total of 20 means that 100 % of the samples were correctly classified. The SVM achieved complete success with the species: *B. salzmannii* (b), *H. parviflora* (e), *H. tenellum* (f), *L. leptocarpa* (i), *L. octovalvis* (j), *R. romosior* (k) and *S. supinus* (l). However, the other species did not reach a total of 100 % correct classification. In matrix (**B**) the deep learning matrix had a higher margin of error compared to SVM, as the number of samples which were classified correctly ranges from 15 to 20, the network could only classify 100 % (20) of the samples correctly in *L. octovalvis* (j) and *S.supinus* (l). However, the other species were shown not to have reached a total of 100% correct classification. In matrix (**C**) Finally, for random forest the species matrix was similarly correct when compared to SVM, considering the samples which were correctly classified 100 %: *E. acutangula* (d), *H. paviflora* (e), *H. tenellum* (f), *H. limosa* (g), *L. leptocarpa* (i), *L. octavalvis* (j) and *S. spinus* (l). However, the other species did not reach a total of 100 % correct classification.

In this article we presented the results of three algorithms that were used to classify aquatic plant seeds. We built linear classifiers of many classes between seeds of aquatic plants by extracting 226 features. We argue that this method performed well in our experiment and improved the accuracy rate of seed classification. At the same time, we found that the SVM and random forest algorithms are significantly robust in their use for aquatic plant seeds recognition. Our algorithms maintained a high recognition rate even with deformation of seed images, such as translation and rotation. Likewise, there was a high rate of precision in terms of shape, texture and color. However, we see that the application of deep learning in the classification of aquatic plant seeds is not satisfactory and, therefore, a new approach is needed to improve the classification.

The use of machines could be an exceptional tool to assist researchers, considering the accuracy of computers in completing complex tasks such as plant identification, however it is not sufficient to replace the work of human botanists, rather to help save time and resources. These results pave the way for taxonomy and conservation work via evaluation of seed banks in wetlands that present high diversity. The analysis of these results will take us to the second phase of seed classification with a larger database of seeds images that are already catalogued and will be used in any future software development which aims to identify native and exotic seeds from wetlands.

References

Amma KY, Yaguchi Y, Niitsuma T, Matsuzaki K, Oka K. 2013. A comparative study of gesture recognition between RGB and HSV colours using time-space continuous dynamic programming. International Joint Conference on Awareness Science and Technology and Ubi-Media Computing: Can We Realize Awareness via Ubi-Media? iCAST 2013 UMEDIA. https://s3-us-west-2.amazonaws.com/ieeeshutpages/xplore/xplore-shut-page.html. 01 Jul. 2020.

Allen GP, Hodgson RM, Marsland SR, Flenley JR. 2008. Machine vision for automated optical recognition and classification of pollen grains or other singulated microscopic objects. 15th International Conference on Mechatronics and Machine Vision in Practice. https://s3-us-west-2.amazonaws.com/ieeeshutpages/xplore/xplore-shut-page.html. 01 Jul. 2020.

Amarnath JJ, Shwetha P, Rajeswari P, Sahoo PK. 2018. Plant Identification Using Leaves with Particle Swarm Optimization and Firefly Algorithms. Indian Journal of Public Health Research & Development 9: 366-371.

Arafat SY, Saghir MI, Ishtiaq M, Bashir U. 2016. Comparison of Techniques for Leaf Classification. 3-8. 2016 Sixth International Conference on Digital Information and Communication Technology and its Applications (DICTAP). IEEE. https://s3-us-west-2.amazonaws.com/ieeeshutpages/xplore/xplore-shut-page.html. 01 Jul. 2020.

Bambil D, Pistori H, Bao F, *et al.* 2020. Plant species identification using colour learning resources, shape, texture, through machine learning and artificial neural networks. Environment Systems and Decisions 40: 480-484.

Bao F, Pott A, Ferreira FA, Arruda R. 2014. Soil seed bank of floodable native and cultivated grassland in the Pantanal wetland: effects of flood gradient, season, and species invasion. Brazilian Journal of Botany 37: 239-250.

Bao F, Elsey-Quirk T, Assis MA, Pott A. 2018. Seed bank of seasonally flooded grassland: experimental simulation of flood and post-flood. Aquatic Ecology 50: 1-13.

Bao F, Assis MA, Pott A. 2021. Applicability of seed bank assessment methods in wetlands: advantages and disadvantages. Oecologia Australis 25: 22-33.

Bonis A, Lepart J, Grillas P. 1995. Seed bank dynamics and coexistence of annual macrophytes in temporary and variable habitat. Oikos 74: 81-92.

Danelljan M, Khan FS, Felsberg M, Van De Weijer J. 2014. Adaptive colour attributes for real-time visual tracking. Proceedings IEEE Computer Society Conference Computer Vision Pattern Recognition. https://openaccess.thecvf.com/content_cvpr_2014/html/Danelljan_Adaptive_Color_Attributes_2014_CVPR_paper.html. 01 Jul. 2020.

Deepayan S. 2008. Lattice: Multivariate Data Visualization with R. New York, Springer Science & Business Media.

Gil ASB, Bove CP. 2006. Eleocharis R.Br. (Cyperaceae) no estado do Rio de Janeiro, Brasil. Biota Neotropica 7: 163-193.

Granitto PM, Verdes PF, Ceccatto HA. 2005. Large-scale investigation of weed seed identification by machine vision. Computers and Electronics in Agriculture 47: 15-24.

Granitto PM, Navone HD, Verdes PF, Ceccatto HA. 2002. Weed seeds identification by machine vision. Computers and Electronics in Agriculture 33: 91-103.

Granitto PM, Garralda PA, Verdes PF, Ceccatto HA. 2003. Boosting Classifiers for Weed Seeds Identification. Journal of Computer Science & Technology 3: 34-39.

Groth D. 1983. Estudo morfológico das unidades de dispersão e respectivas plantas de seis espécies invasoras da família Cyperaceae. Planta Daninha 5: 25-38.

Harralick RM. 1979. Statistical and structural approach to texture. Proceedings IEEE Computer Society Conference Computer Vision Pattern Recognition 67: 786-804.

He C, Zhu Q, Huang M, Mendoza F. 2016. Model Updating of Hyperspectral Imaging Data for Variety Discrimination of Maize Seeds Harvested in Different Years by Clustering Algorithm. Transactions of the ASABE 59: 1529-1537.

Hong PTT, Hai TTT, Hoang VT, Hai V, Nguyen TT. 2015. Comparative study on vision-based rice seed varieties identification. In Seventh International Conference on Knowledge and Systems Engineering (KSE). Proceedings IEEE Computer Society Conference Computer Vision Pattern Recognition. https://s3-us-west-2.amazonaws.com/ieeeshutpages/xplore/xplore-shut-page.html. 01 Jul. 2020.

Jafarpour S. 2012. A Robust Brain MRI Classification with GLCM Features. International Journal of Computer Applications 37: 1-5.

Jia S, Zhao X, Li Y, Wang K. 2014. A particle filter human tracking method based on HOG and Hu moment. ICMA International Conference on Mechatronics and Automation. https://s3-us-west-2.amazonaws.com/ieeeshutpages/xplore/xplore-shut-page.html. 01 Jul. 2020.

Kaul RB. 1978. Morphology of germination and establishment of aquatic seedlings in Alismataceae and Hydrocharitaceae. Aquatic Botany 5: 139-147.

Kaul RB. 1985. Reproductive phenology and biology in annual and perennial Alismataceae. Aquatic Botany 22: 153-164.

Kremic E, Subasi A. 2016. Performance of Random Forest and SVM in Face Recognition. The International Arab Journal of Information Technology 13: 287-293.

Kruse OMO, Prats-Montalbán JM, Indahl UG, Kvaal K, Ferrer A, Futsaether CM. 2014. Pixel classification methods for identifying and quantifying leaf surface injury from digital images. Computers and Electronics in Agriculture 108: 155-165.

Lang S, Bravo-Marquez F, Beckham C, Hall M, Frank E. 2019. WekaDeeplearning4j: a deep learning package for Weka based on Deeplearning. Knowledge Based Systems 178: 48-50.

Oksanen JF, Blanchet G, Friendly M, *et al.* 2017. Vegan: Community Ecology Package. R package version 2: 4-3. https://CRAN.R-project.org/package=vegan. 01 Jul. 2020.

Park S, Lee HS, Kim J. 2016. Seed growing for interactive image segmentation using SVM classification with geodesic distance. Electronics Letters 53: 22-24.

Patil JK, Kumar R. 2017. Analysis of content-based image retrieval for plant leaf diseases using colour, shape and texture features. Engineering in Agriculture, Environment and Food 10: 69-78.

R Development Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org/. 01 Jul. 2020.

Rajagopal N, Xie W, Li Y, *et al.* 2013. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. PLOS Computational Biology 9: e1002968. doi: 10.1371/journal.pcbi.1002968

Schroff F, Kalenichenko D, Philbin J. 2015. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html. 01 Jul. 2020.

Shi D, Yang X. 2016. An assessment of algorithmic parameters affecting image classification accuracy by random forests. Photogrammetric Engineering & Remote Sensing 82: 407-417.

Silva JSV, Abdon MM. 1998. Delimitação do Pantanal Brasileiro e suas sub regiões. Pesquisa Agropecuária Brasileira 33: 1703-1711.

Simpson GL. 2016. Permute: Functions for Generating Restricted Permutations of Data. R package version 0.9-4. https://cran.r-project.org/web/packages/permute/vignettes/permutations.pdf.

Souza EB, Ferreira FA, Pott A. 2016. Effects of flooding and its temporal variation on seedling recruitment from the soil seed bank of a Neotropical floodplain. Acta Botanica Brasilica 31: 64-75.

Souza DKL, Giulietti AM. 2014. Flora da Bahia: Pontederiaceae. Sitientibus-série Ciências Biológicas 14: 1-10.

Tirintan G, Catian G, Da Luz GP, Manvailer V, Scremin-Dias E. 2018. Plântulas e sementes de macrófitas aquáticas de lagoas do Pantanal Sul-Mato-Grossense. Iheringia, Série Botânica 73: 69-87.

Wäldchen J, Mäder P. 2018a. Plant species identification using computer vision techniques: A systematic literature review. Archives of Computational Methods in Engineering 25: 507-543.

Wäldchen J, Mäder P. 2018b. Machine learning for image-based species identification. Methods in Ecology and Evolution 9: 2216-2225.

Wäldchen J, Rzanny M, Seeland M, Mäder P. 2018. Automated plant species identification-Trends and future directions. PLOS Computational Biology 14: e1005993. doi: 10.1371/journal.pcbi.1005993

Wen X, Shao L, Xue Y, Fang W. 2015. A rapid learning algorithm for vehicle classification. Information Science 295: 395-406.

Xiao XY, Hu R, Zhang SW, Wang XF. 2010. HOG-Based Approach for Leaf Classification. In: Huang DS, Zhang X, Reyes García CA, Zhang L. (eds.) Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. ICIC 2010. Lecture Notes in Computer Science. Vol. 6216. Berlin, Heidelberg, Springer Science & Business Media.

Xinshao W, Cheng C. 2015. Weed seeds classification based on PCANet deep learning baseline. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). https://s3-us-west-2.amazonaws.com/ieeeshutpages/xplore/xplore-shut-page.html. 01 Jul. 2020.

Zhou S, Liu Q, Guo J, Jiang Y. 2012. ROI-HOG and LBP Based Human Detection via Shape Part-Templates Matching. In: Huang T, Zeng Z, Li C, Leung CS. (eds.) Neural Information Processing. ICONIP 2012. Lecture Notes in Computer Science. Vol. 7667. Berlin, Heidelberg, Springer Science & Business Media. p. 109-115.