Classificação de sementes aquáticas do Pantanal Brasileiro utilizando árvores de decisão

André Núncio de Oliveira Sol, Edilaine Gonçalves Costa de Faria, Francielli Bao, Deborah Bambil
Daniel Chaves Café e Francisco Assis de Oliveira Nascimento

***Resumo*— Este artigo compara métodos de classificação de sementes aquáticas. Estas sementes foram adquiridas de um banco de dados de sementes do Pantanal Brasiliense. Ao total, 12 espécies foram selecionadas, cada uma com 20 amostras. O banco de dados tem 240 imagens. Estas imagens são processadas num vetor de 226 *features*. Os métodos propostos foram *Naive-Bayes, Logistic Regression, Decision Tree, Extra-Trees, Multi-class AdaBoost, Histogram-based Gradient Boosting and Ensable Extra-Trees*. Dentre esses algoritmos, o que mais se destacou foi o Ensable Extra-Trees, atingindo uma acurácia de 98,12%.**

***Palavras-Chave*— Aprendizado de Máquina, Banco de sementes, Pantanal, Classificação.**

***Abstract*— This article compares methods of classifying aquatic seeds. These seeds are taken from seed banks of the Brazilian Pantanal. Twelve species were selected, each with 20 samples. The data bank has 240 images. Those images are processed into a vector with 226 attributes. The proposed methods are Naive-Bayes, Logistic Regression, Decision Tree, Extra-Trees, Multi-class AdaBoost, Histogram-based Gradient Boosting and Ensable Extra-Trees. Among these algorithms, the Ensable Extra-Trees achieved the higher accuracy score, reaching 98.12%.**

***Keywords*— Machine Learning, Seed bank, Pantanal, Classification.**

I. Introdução

Classificação de sementes é uma técnica muito utilizada na agricultura para a escolha das sementes que geraram melhores colheitas [1]. A automação dessa atividade inspirou a integração de técnicas computacionais para compensar a baixa precisão e baixa velocidade dos seres humanos para a execução dessa tarefa ao longo dos anos. Com a evolução das técnicas de visão computacional e aprendizagem de máquina novas plataformas foram desenvolvidas [2]. Verifica-se forma, cor e textura da sementes para determinar qual é a melhor candidata. Outras características das plantas também são utilizadas, como a morfologia das folhas [3] na integração da computação à botânica.

Outra abordagem de interesse para o País é o estudo do processo de invasão de biomas por espécies exóticas, como acontece no Pantanal brasileiro. Esse bioma possui uma característica peculiar: durante o período de inundação, algumas plantas morrem e deixam suas sementes para germinarem durante o período de estiagem. Assim, forma-se no solo um banco de sementes. Este banco é peça fundamental para a regeneração do bioma.

Realizou-se o estudo sobre a invasão de Urochloa humidicola [4] no Pantanal. Colheu-se substrato do solo em caixotes, e em seguida retirou-se as sementes presentes em cada amostra de solo coletada. Com essas amostras fotografou-se uma a uma as sementes presentes. Isso resultou em um banco de fotos de sementes contendo 12 espécies e 240 amostras. As amostras são separadas manualmente por um profissional e catalogadas, tarefa que demanda muito tempo, pois as sementes são diminutas. A proposta desse trabalho é conceber uma solução capaz de classificar a espécie correta a partir destas fotos utilizando algoritmos de inteligencia artificial.

Para tal, a seguinte abordagem é proposta: primeiramente as imagens são segmentadas para extrair atributos, do inglês *features*. Em seguida, os algoritmos utilizados para classificação foram: *Logistic Regression, Multi-class AdaBoost, Decision Tree, Naive-Bayes Classifier, Extra-Trees Classifier, Ensemble Extra-Tree Classifier, Histogram-based Gradient Boosting Classification Tree*. Por fim, mede-se o desempenho de cada classificador.

II. Revisão Bibliográfica

Diversos trabalhos apresentam resultados obtidos da classificação de sementes. Um estudo, por exemplo, consiste em classificação de sementes de abóboras. Neste trabalho foi utilizada a técnica de validação cruzada denominada k-fold com k igual a 10. E obteve-se como resultado de acurácia máxima 88,64% utilizando Máquinas de Vetores de Suporte - do inglês *Support Vector Machine (SVM)* [5].

Outro estudo obteve resultado de acurácia de até 99,9% para a classificação de sementes de milho. Este utilizando também 10 k-fold [6].

Além disso, as Redes Neurais Convolucionais, do inglês *Convolutional Neural Network (CNN)*, que estão presentes no estado da arte nos estudos de aprendizagem de máquina, foram utilizadas para um sistema de classificação de sementes atingindo resultado de 99% de acurácia [7] e 99,42% de acurácia em outro estudo semelhante [8].

Ainda neste contexto, a identificação automática de sementes em um artigo obteve resultados de acurácia de 97,91%, 97,08% e 92,50%, utilizando SVM, *random forest* e *deep learning*, respectivamente [9].

André Núncio de Oliveira Sol, Departamento de Engenharia Elétrica, Universidade de Brasília (UnB), Brasília-DF, email: andrenuncio@gmail.com; Edilaine Gonçalves Costa de Faria, Departamento de Engenharia Elétrica, Universidade de Brasília (UnB), Brasília-DF, email: edilaine@ieee.org; Francielli Bao, Universidade Estadual Paulista Júlio de Mesquita Filho, Campus Experimental de Rio Claro - SP. UNESP - Universidade do Estado de São Paulo, Rio Claro, SP, e-mail: franbao@yahoo.com.br; Deborah Bambil, Departamento de Sistemas e Computação, Universidade Regional de Blumenal (FURB), Blumenau-SC, e-mail: deborahbambil@gmail.com; Daniel Chaves Café, Departamento de Engenharia Elétrica, Universidade de Brasília (UnB), Brasília-DF, email: dccafe@gmail.com; Francisco Assis de Oliveira Nascimento, Departamento de Engenharia Elétrica, Universidade de Brasília (UnB), Brasília-DF, email: assis@unb.br.

Em outro trabalho, sementes de berinjelas foram utilizadas para a classificação utilizando SVM, rede neural convolucional unidimensional (1D-CNN) e rede neural convolucional bidimensional (2D-CNN). Nestes, as acurácias obtidas foram de 90,12%, 94,80% e 90,67% respectivamente [10].

A classificação automática de sementes também pode ser realizada em sementes de arroz. Em que a melhor acurácia obtida utilizando SVM foi 90,61% e utilizando técnicas de aprendizado profundo foi de 95,15% [11].

Percebe-se, a partir dos trabalhos citados, que o uso de classificadores em variantes de uma mesma espécie geralmente possuem índice elevado de acurácia. No presente trabalho, entretanto, estamos interessados em verificar a performance de diferentes algoritmos classificadores para um conjunto de diferentes espécies.

III. Métodos e Materiais

Nessa seção, são descritos os algoritmos empregados no processamento e o banco de dados utilizado para treinamento. As simulações foram implementadas em linguagem Python utilizando a ferramenta Scikit-learn e utilizando a configuração paramétrica *"default"* da biblioteca.

A. Banco de Dados

Bancos de sementes no solo são a diversidade de sementes encontradas no solo de uma determinada região, assim ao se retirar um parcela do terreno pode-se verificar a diversidade de sementes presentes, contando a quantidade. Também verifica-se a presença de espécies exóticas.

Realizou-se [12] um estudo de captação e classificação das espécies presentes no banco de sementes de terrenos no Pantanal brasileiro. Coletou-se dezenas de espécies utilizando o método da contagem. Segundo o estudo, a coleta foi realizada na Fazenda São Bento (19°29'27.3"S; 57°01'55.9"W), para a proposta deste trabalho utilizou-se somente as sementes advindas de espécies aquáticas, resultando em 12 espécies diferentes. Que são:

1) ***Bacopa Australis***: Flores brancas azuladas, fica submersa durante a cheia, erva aquática emergente
2) ***Bacopa Salzmanii***: Erva rastejante com corola alvoazulada
3) ***Bacopa Stricta***: Erva de folhas e caule verdes, corola lilás claro, conhecida pelo bom cheiro
4) ***Echinodorus Tenellus***: Erva encontrada em margens de lagoas, possui pétalas brancas
5) ***Eleocharis Acuntansula***: Planta de folhas anguladas e duras
6) ***Hetheranthera Limosa***: Erva sem odor com flores azuis e amarelas, planta emergente
7) ***Hydrocleys Paruiflora***: Planta flutuante fixa, com flores amarelas e fruto marrom
8) ***Liminocharis Flava***: Planta com parte interna do cálice amarela e flores brancas
9) ***Ludwigia Leptocarpha***: Possui flores amarelas e frutos verdes-avermelhados
10) ***Ludwigia Octovalus***: Planta com característica de arbusto, flores amarelas
11) ***Rotala Romosior***: Planta com folhas membranáceas e flores rosas, tem frutos esverdeados
12) ***Scirpus Supinos***: Planta com folhas simples que crescem até 15 centímetros

Fig. 1. Exemplos de fotografias presentes no banco de dados.

O estudo de Bao gerou um banco de imagens formado por sementes fotografadas com um auxílio de um microscópio estéril codificado Leica modelo M125 [12]. Exemplos de imagens de algumas sementes utilizadas neste trabalho são ilustradas na Figura 1. Foram obtidas 20 amostras diferentes para cada espécie, totalizando 240 amostras no banco de fotos utilizado neste trabalho.

B. Extrator de Features

Foi construído um *framework* para a extração de *features* para o banco de dados de sementes digitalizadas. A Figura 2 mostra um diagrama em blocos para o *framework* construído. O banco de imagens processadas produz como saída, um arquivo com grandezas numéricas no formato *CSV (Comma-Separated Values)*. Cada imagem contribui com uma linha desta matriz. A matriz (o arquivo CSV) possui a quantidade de linhas delimitada pela quantidade de imagens presentes no banco de dados e as suas colunas correspondes a quantidade de *features* computados.

A Tabela I mostra todos os algoritmos utilizados para construção da matriz de *features*. A grande maioria são algoritmos clássicos que se acham implementados em linguagem Python. Alguns precisaram de um pouco mais de esforço para a sua implementação, como é o caso dos Padrões Direcionais Locais (*LDP – Local Directional Patters*).

Para se construir os *features* de distribuição da energia no domínio das frequências utilizou-se uma FFT (*Fast Fourier Transform*) bidimensional e segmentamos o espectro da magnitude (elevado ao quadrado) em um conjunto de discos circulares. Cada par de discos circulares é delimitado por uma

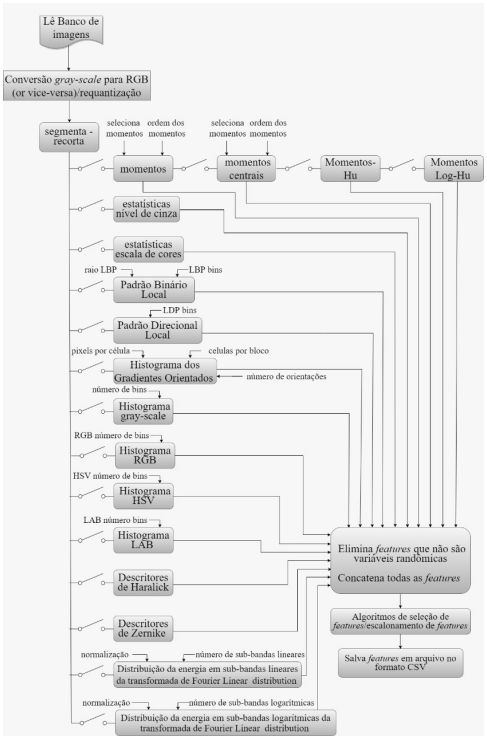

Fig. 2. Diagrama de blocos do extrato de *features*

TABELA I
ESPECIFICAÇÃO E QUANTIDADE DE *FEATURES* CALCULADOS.

Atributo	Quantidade
Momentos cartesianos	16
Momentos centrais	16
Momentos Hu	7
Momentos Log-Hu	7
Histograma de níveis de cinza	32
Histograma RGB	32
Histograma HSV	32
Histograma CIE-LAB	32
Estatísticas de mapas Gray/RGB/HSV/CIE Lab	210
Matrizes de co-ocorrência de níveis de cinza	48
Haralick descriptors	13
Momentos de Zernike	25
Histogramas de Gradientes Orientados	128
Padrões Binários Locais	32
Padrões Locais Direcionais	32
Distribuição da Energia no domínio das frequências	8

frequência inferior e outra superior, correspondendo a uma sub-banda no domínio da transformada discreta de *Fourier*.

Para a obtenção dos resultados apresentados neste trabalho foram extraídos inicialmente 670 *features*. Depois de um processo de eliminação das grandezas que não são variáveis randômicas, a matriz de *features* se reduziu a 458 colunas.

C. Algoritmos de Classificação

Os algoritmos foram implementados com a linguagem de programação Python e as bibliotecas de aprendizado de maquina: scikit-learn, TensorFlow e SciPY.

1) ***Logistic Regression*** [13]: Modelo linear de classificação, também conhecido como classificador de máxima entropia, aplica-se a função logística para modelar uma variável binária, esse modelo pode ser estendido para múltiplas classes.
2) ***Naive-Bayes Classifier*** [14]: Modelo probabilístico que usa o teorema de Bayes para classificar entre as classes, considera as variáveis fortemente independentes.
3) ***Decision Tree*** [15]: Modelo de árvore de decisão utilizam a configuração a partir de observações sobre um item tirar conclusões. As folhas representam as classes e as ramificações indicam o processo de predição.
4) ***Multi-class AdaBoost*** [16]: Modelo que cria diversos classificadores utilizando um outro paradigma para a classificação, neste caso árvores de decisão. Os classificadores gerados primeiro são responsáveis pelo problema geral e os subsequentes são ajustados para casos mais difíceis e específicos.
5) ***Extra-Trees Classifier*** [17]: Modelo de árvore de decisão que verifica cada *feature* antes de separar os nós. Utiliza-se um subconjunto do espaço de características, assim geram-se diversas árvores de decisão.
6) ***Ensemble Extra-Tree Classifier*** [17]: Modelo que emprega diversos modelos de *Extra Trees* sobre diversos subconjuntos do espaço de características.
7) ***Histogram-based Gradient Boosting Classification Tree***: Modelo com suporte para valores *NaN (Not a Number)* e inspirado em *LightGBM* [18]. Como alguns algoritmos as vezes não retornam valores numéricos.

D. Medida estatística utilizada

Foi utilizada como medida estatística a acurácia, que é determinada pela Equação 1. Ela pode ser utilizada em um algoritmo de mais de duas classes.

$$A_{cc} = \frac{\sum_{c=1}^{N} T_P[c]}{\sum_{c=1}^{N} T_P[c] + F_P[c]}, \quad (1)$$

sendo que A_{cc} é acurácia, do inglês *accuracy*; T_P é verdadeiro positivo, do inglês *true positive*; F_P é falso positivo, do inglês *false positive*; c consiste em cada classe e N é o número máximo de classes.

IV. RESULTADOS

O processo de treinamento utilizou 80% das amostras para treinamento e 20% para teste, sendo assim foram 192 amostras para o treinamento e 48 para o teste. Repetiu-se a etapa de treinamento e validação 10 vezes com seleção randômica para os conjuntos de dados de treinamento e de teste. Calculou-se a média da acurácia obtida pelos modelos propostos.

A Tabela II resume os resultados obtidos de acurácia através da Equação 1, entretanto, neste caso, a equação é relacionada ao total de instâncias.

A Figura 3 ilustra a matriz de confusão média após 10 iterações, sendo que representa o resultado do modelo de *Ensemble Extra-Tree Classifier*, que obteve o melhor resultado de acurácia. A cor azul denomina a baixa escolha por parte do modelo, quanto mais escolhido mais vermelhado, os valores

TABELA II
ACURÁCIA.

Algoritmo	**Acurácia (%)**
Logistic Regression	$94,79$
Naive-Bayes Classifier	$85,42$
Decision Tree	$95,00$
Extra-Trees Classifier	$87,50$
Multi-class AdaBoost	$90,21$
Histogram-based Gradient Boosting Classification Tree	$96,67$
Ensemble Extra-Tree Classifier	$98,12$

estão em percentual em relação ao total de instâncias no momento de teste. A diagonal principal denota os casos de verdadeiros positivos, quanto mais concentrado os valores nela, melhor o desempenho do modelo.

Calculou-se a acurácia por classe dos modelos propostos, sendo esta a taxa de verdadeiros positivos em relação ao total de casos da classe no grupo de teste, como ilustra a Equação 1. Estes resultados estão compilados na Tabela III.

V. DISCUSSÃO

O classificador probabilístico *Naive-Bayes* obteve a menor acurácia média, com $85,42\%$ de acurácia geral e atingindo valores menores que 65% para algumas classes. O modelo de regressão logística, que é um modelo linear, obteve um resultado superior ao algoritmo *Naive-Bayes*. Demonstra-se que há uma separação linear entre as amostras no espaço do banco de dados. Duas classes causaram a maior discrepância para este modelo durante o teste.

O modelo simples de árvore de decisão obteve 95% de acurácia, valor superior aos modelos de *Naive-Bayes* e regressão logística. Como utiliza aferições nas *features* para determinar a divisão dos nós de seu modelo de predição, um banco de dados composto por um vetor extenso de características facilita a utilização deste método. Aplicou-se *Multi-Class AdaBoost*. O primeiro resolve o problema criando diversos classificadores de árvore de decisão, cada um com o objetivo de resolver parte do problema. O primeiro classificador gerado tenta resolver o problema geral enquanto os seguintes procuram resolver casos mais difíceis em que o primeiro classificador errou. Após todos os classificadores serem gerados são atribuídos pesos a cada um, assim classificadores mais acurados tem um impacto maior na decisão final do modelo. Contudo este classificador se mostrou menos acurado que os anteriores, um sinal de sobreajuste, pois seu paradigma permite que gere mais árvores a fim de melhor classificar o banco de treino.

Extra-trees, que utiliza subconjuntos de amostras para realizar a classificação gerando múltiplos modelos de árvore de decisão. Ele obteve taxas baixas nas medidas realizadas, demonstrando novamente o sobreajuste, pois esse modelo possui mais graus de liberdade que a proposta de árvore de decisão. Superando somente o classificador probabilistico *Naive-Bayes*. Dado que esse modelo ataca subconjuntos do problema por vez, isso auxilia na segmentação entre classes.

O modelo *Histogram-based Gradient Boosting* consegue interpretar instâncias de $NaNs$ nos vetores de *feature*. Outros algoritmos somente substituem por um valor numérico, como 0 ou -1. Assim obteve-se um resultado superior nas taxas medidas, superando o modelo simples de árvore de decisão. Como alguns extratores de *features* retornam valores nulos, acredita-se que essa característica do modelo favoreceu seu desempenho.

Por fim, o modelo *Ensable Extra-tree*, cujo paradigma é a aplicação de diversos modelos de *Extra-trees* sobre variados subconjuntos de amostras. Utilizando esta estratégia obteve-se a maior acurácia média dentre os modelos aferidos.

A árdua tarefa de classificar sementes pode ser auxiliada por um computador e uma câmera, como demonstram os resultados obtidos dos modelos propostos. O melhor modelo obteve uma acurácia de $98,12\%$, próximo aos valores encontrados em outros trabalhos, mencionados na Seção II, que atingiram valores de acurácia de aproximadamente 99%.

VI. CONCLUSÃO

A abordagem adotada se mostrou eficiente na tarefa de discriminar entre as sementes, podendo assim ser utilizada para auxiliar profissionais no trabalho de coleta em campo. Nesse estudo verificou-se a classificação de imagens baseando-se em características retiradas a partir de dados estatísticos e qualitativos de fotografias.

Avaliou-se os modelos *Naive-Bayes Classifier, Logistic Regression, Decision Tree, Multi-class AdaBoost, Extra-Trees Classifier, Ensemble Extra-Tree Classifter, Histogram-based Gradient Boosting Classification Tree* detertminado que o melhor desempenhou foi *Ensemble Extra-Tree Classifier*. Mostrando assim que as técnicas se adaptaram bem ao problema proposto.

Espera-se no futuro poder determinar mais *features* das imagens e a aplicação de modelos inteligentes baseados em sistemas biológicos, como as redes neurais. Além disso, incluir a classificação de outros tipos de sementes.

REFERÊNCIAS

[1] A. Gudipalli, N. Prabha, and P. R. Ch, "A review on analysis and grading of rice using image processing," *ARPN Journal of Enginecring and Applied Sciences*, vol. 11, no. 23, 2016.

[2] G. L. Grinblat, L. C. Uzal, M. G. Larese, and P. M. Granitto, "Deep learning for plant identification using vein morphological patterns," *Compurers and Electronics in Agriculrare*, vol. 127, pp. 418–424, 2016.

[3] D. Bambil, H. Pistori, F. Bao, V. Weber, F. M. Alves, E. G. Goncalves, L. F. de Alencar Figueiredo, G. P. A. U., and R. Arroda, "and l," *M. Bortolotto, "Plant species identification using color learning resources, shape, texture, through machine leaming and artificial neural networks,"*, vol. 40, no. 4, pp. 10 669–020.

[4] F. Bao, T. Elsey-Quirk, M. A. Assis, R. Arruda, and A. Pott, "Seasonal flooding, topography, and organic debris interact to influence the emergence and distribution of seedlings in a tropical grassland," *Biotropica*, vol. 50, no. 4, pp. 616–624, 2018.

[5] M. Koklu, S. Sarigil, and O. Ozbek, "The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.)," *Genetic Resources and Crop Evolution*, vol. 68, no. 7, pp. 2713–2726, oct 2021. [Online]. Available: https://doi.org/10.1007/s10722-021-01226-0 https://link.springer.com/10.1007/s10722-021-01226-0

[6] A. Ali, S. Qadri, W. K. Mashwani, S. Brahim Belhaouari, S. Naeem, S. Rafique, F. Jamal, C. Chesneau, and S. Anam, "Machine learning approach for the classification of corn seed using hybrid features," *International Journal of Food Properties*, vol. 23, no. 1, pp. 1097–1111, 2020. [Online]. Available: https://doi.org/10.1080/10942912.2020.1778724

[7] Y. Gulzar, Y. Hamid, A. B. Soomro, A. A. Alwan, and L. Journaux, "A convolution neural network-based seed classification system," *Symmetry*, vol. 12, no. 12, pp. 1–18, 2020.

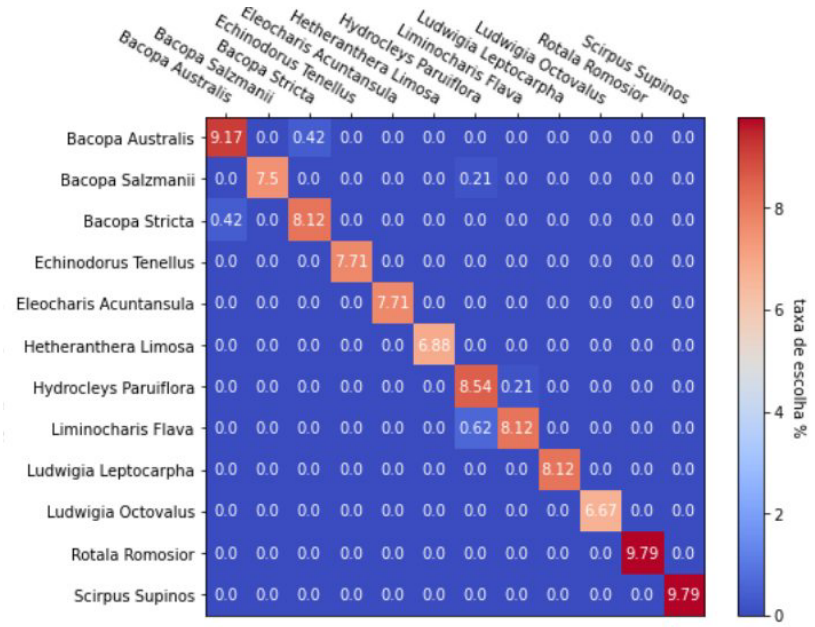

Fig. 3. Matriz de Confusão do modelo *Ensable Extra-tree*

TABELA III
ACURÁCIA POR CLASSE (%).

	Logistic Regression	Multi-class AdaBoost	Decison Tree	Naive-Bayes	Extra-trees	Ensemble Extra-Tree	HGBCT
Bacopa Australis	93, 88	93,88	91,30	86,96	84,78	95,65	95,65
Bacopa Salzmanii	100,00	97,22	91,89	81,08	91,89	97,30	94,59
Bacopa Stricta	97, 67	100,00	100,00	87,80	87,80	95,12	90,24
Echinodorus Tenellus	95, 00	100,00	97,30	91,89	94,59	100,00	100,00
Eleocharis Acuntansula	100,00	98,19	94,59	97,30	100,00	100,00	100,00
Hetheranthera Limosa	100,00	87,80	96,97	84,85	87,88	100,00	93,94
Hydrocleys Paruiflora	80, 95	95,24	88,10	95,24	61,90	97,62	90,48
Liminocharis Flava	81, 08	56,76	95,24	69,05	83,33	92,86	97,62
Ludwigia Leptocarpha	100,00	97,62	94,87	79,49	89,74	100,00	100,00
Ludwigia Octovalus	100,00	87,50	100,00	62,50	96,88	100,00	100,00
Rotala Romosior	91, 11	80,00	93,62	100,00	76,60	100,00	97,87
Scirpus Supinos	100, 00	96,43	97,87	82,98	100,00	100,00	100,00

[8] R. Eryigit and B. Tugrul, "Performance of Various Deep-Learning Networks in the Seed Classification Problem," *Symmetry*, vol. 13, no. 10, p. 1892, oct 2021. [Online]. Available: https://www.mdpi.com/2073-8994/13/10/1892

[9] F. Bao and D. Bambil, *Applicability of computer vision in seeds identification: deep learning, random forest, and support vector machine classification algorithms*. ACTA Botânica Brasilica.

[10] L. Sun, X. Fan, S. Huang, S. Luo, L. Zhao, X. Chen, Y. He, and X. Suo, "Research on Classification Method of Eggplant Seeds Based on Machine Learning and Multispectral Imaging Classification Eggplant Seeds," *Journal of Sensors*, vol. 2021, pp. 1–9, sep 2021. [Online]. Available: https://www.hindawi.com/journals/js/2021/8857931/

[11] K. Kiratiratanapruk, P. Temniranrat, W. Sinthupinyo, P. Prempree, K. Chaitavon, S. Porntheeraphat, and A. Prasertsak, "Development of Paddy Rice Seed Classification Process using Machine Learning Techniques for Automatic Grading Machine," *Journal of Sensors*, vol. 2020, pp. 1–14, jul 2020. [Online]. Available: https://www.hindawi.com/journals/js/2020/7041310/

[12] F. Bao, "A vegetação campestre em gradientes inundáveis: composiçâo floristica, dinâmica do banco de sementes e de plântulas.","Ph.D. dissertation, RIO CLARO SP, 2017.

[13] C. M. Bishop, *Pantem recognition and machine leaming*. New York: Springer, 2006. [Online]. Available: https://search.library.wisc.edu/catalog/9910032530902121

[14] H. Zhang, "The optimality of naive bayes," in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, V. Barr and Z. Markov, Eds., 2004.

[15] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," *Monterey, CA: Wadsworth and Brooks*, 1984.

[16] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," *Journal of Contpuer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002200009791504X

[17] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006. [Online]. Available: https://doi.org/10.1007/s 10994-006-

[18] G. Ke, Q. Meng, T. Finley, T. Wang, C. W., M. W., Q. Ye, , and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advaces in Neamal lnformation Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, and H. Wallach, Eds., 2017.