Engenharia Agrícola

ISSN: 1809-4430 (on-line)

www.engenhariaagricola.org.br

Scientific Paper

Doi: http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v43n2e20220193/2023

FITTING DATA MINING SETTINGS FOR RANKING SEED LOTS

Ruan Bernardy[1*], Gizele I. Gadotti[1], Rita de C. M. Monteiro[2], Karine Von Ahn Pinto[1], Romário de M. Pinheiro[3]

[1*]Corresponding author. Federal University of Pelotas/Pelotas - RS, Brazil.
E-mail: ruanbernardy@yahoo.com.br | ORCID ID: https://orcid.org/0000-0001-9285-1993

KEYWORDS

artificial intelligence, agriculture, quality control.

ABSTRACT

To enhance speed and agility in interpreting physiological quality tests of seeds, The use of algorithms has emerged. This study aimed to identify suitable machine learning models to assist in the precise management of seed lot quality. Soybean lots from two companies were assessed using the Supplied Test Set, Cross-Validation (with 8, 10, and 12 folds), and Percentage Split (with 66% and 70%) methods. Variables analyzed through Tetrazolium tests included vigor, viability, mechanical damage, moisture damage, bed bug damage, and water content. Method performance was determined by Kappa, Precision, and ROC Area metrics. Classification Via Regression and J48 algorithms were employed. The technique utilizing 66% of data for training achieved 93.55% accuracy, with Precision and ROC Area reaching 94.50% for the J48 algorithm. Applying the cross-validation method with 10 folds resulted in 90.22% of correctly classified instances, with a ROC Area outcome like the previous method. Tetrazolium Vigor was the primary attribute used. However, these results are specific to this study's database, and careful planning is necessary to select the most effective application methods.

INTRODUCTION

When implementing agriculture, an essential input required for crop success is high-quality seeds, along with other components. It is crucial for achieving good productivity in any crop. Over the years, farmers have been encouraged to produce seeds with pure genetics and high germination rates (Elias, 2018). According to these authors, high-quality seeds are essential for sustainable food production and stable profits. Therefore, processes have been organized to certify this genetic material, making the process easier (Medeiros et al., 2020b).

However, the seed sector still faces various problems in certifying these seeds (Gadotti et al., 2022b). Analyzing all tests that determine seed quality generates a vast amount of information that makes it impossible for human intellectual capacity to conduct a rapid and effective analysis within a quality control laboratory in the short term (Pinheiro et al., 2021). Therefore, erroneous results can lead to economic losses for seed companies (Gadotti et al., 2022a).

To meet that demand, research in Seed Technology has focused on identifying various aspects associated with ranking lots based on their seed physiological potential. Thus, information technology emerges as a tool to find solutions that make analytical processes faster and more reliable (Patrício & Rieder, 2018).

According to Gadotti et al. (2022b), Artificial Intelligence, within which machine learning is included, can be applied to promote sustainability in the agricultural sector to facilitate a better understanding of the production chain, optimizing resources at various stages composing it. Some published works demonstrate good results in the use of machine learning in seed classification. However, for the development of further studies, it is important to analyze in more depth the training and testing techniques used in algorithms, with a deeper investigation into how these classifiers are trained before they are tested. In this context, it is essential that, when training an algorithm to perform an activity, the sample is random, that is, the

[1] Universidade Federal de Pelotas/Pelotas - RS, Brasil.
[2] Ministério da Agricultura, Pecuária e Abastecimento/Porto Alegre - RS, Brasil.
[3] Instituto Nacional de Pesquisas da Amazônia/Rio Branco - AC, Brasil.

Area Editor: Paulo Carteri Coradi
Received in: 10-24-2022
Accepted in: 3-18-2023

examples used should not be pre-selected by the technician (Peng & Xu, 2022).

Therefore, the training method used must be appropriate to maximize the model performance. Hence, this study focused on identifying, among the available methods for training, the best fits in algorithms to better classify lots of soybean seeds.

MATERIAL AND METHODS

The study was carried out at the Laboratory of Seed Analysis of Base Assessoria Agronômica Ltda company, located in Silveira Martins – RS (Brazil), from March 2021 to April 2022. Samples of soybeans from different varieties received by the routine laboratory during this period were used, totaling 93 samples.

Besides seed vigor and viability, we analyzed moisture percentage, mechanical damage, moisture content, and bug damage using Tetrazolium (TZ) tests. The TZ tests followed the procedure in the Seed Analysis Rules (Brasil, 2009), in which 100 seeds (two 50-seed subsamples) were preconditioned in germination papers moistened to 2.5 times their dry weight and kept at 25°C for 16 hours.

After preconditioning, seeds were immersed in a 0.075% tetrazolium solution, where they were kept at 35 to 40°C for 150 to 180 minutes in a BOD incubator. After staining, they were longitudinally sectioned, exposing their embryos, and classified as "viable" (1 to 5) or "non-viable" (6 to 8), based on their coloration. Then, types of damage were indicated following guidelines by França-Neto e Krzyzanowski (2022). Ninety-two lines with seven attributes were generated (Table 1), with 49 lots accepted for commercialization and 43 rejected.

TABLE 1. Description of the attributes analyzed by data mining.

Attribute	Description	Value
Humidity	Humidity	{0-100}
Tetrazolium	Vigor	{0-100}
	Viability	
	Humidity Damage	
	Mechanical Damage	
	Bedbug Damage	
Bath Classification	Decision Taken	{accepted, rejected}

Soybean lots were classified during this research since companies did not provide any classes. Therefore, an expert scientist in the field was used to approve and reject lots according to the requirements for commercialization.

For processing and prediction, data were initially preprocessed so that the tool could accurately perform readings and analyses. The data was received in *.xls* (Excel) format, with all attributes in a single line and each value in columns, below its respective attribute. This file was afterward converted to *.csv* format. Furthermore, lines with missing values or considered incorrect were excluded during this process.

The Weka software, version 3.8.5, developed by the University of Waikato (Eibe et al., 2020), was used for data mining. Three available forms were used to train the algorithms, namely: Supplied Test Set, Cross-Validation (with 8, 10, and 12 folds), and Percentage Split (with 66 and 70% of the data for training). The algorithms tested here comprised Classification Via Regression (CVR) and J48, following the works by Gadotti et al. (2022a) and Gadotti et al. (2022b).

Then, k-fold cross-validation was performed. It subdivides datasets into 90% for training and 10% for testing. This process is repeated for the number of times (folds) proposed by the operator, changing the parts used for training and testing (i.e., performing a new data subdivision). This repetition in training reduces the underrating or overrating of an algorithm's performance in a certain setting. An ideal number of repetitions should be identified because, although algorithm training rarely generates unsatisfactory results, it may occur when training is excessive. Therefore, it is fundamental for proper data processing by an algorithm.

The Percentage Split, unlike k-fold, divides data into training and testing once, and Weka can define the percentage to be allocated for training. This often generates doubts, and many attempts are made to choose the best percentage.

Finally, the Supplied Test Set uses a second dataset for training. This set was obtained from a second company so that the seed parameters evaluated were identical, with the same quantity as the set to be tested.

Therefore, a second set of data was used, provided by an internal seed analysis laboratory (LAIS) of a company from Mato Grosso/MT, with 62 classified by the company as Accepted and 30 as Rejected. The varieties are indicated for cultivation in this state, being produced in the 18/19 harvest. All evaluated parameters followed the same process as the first set so that both had values under the same conditions to validate the obtained results.

Seed analysis data received are inherently unbalanced, especially those originating from companies that prioritize high-quality lots. To solve this problem and

not bias the algorithm, the Resample filter was used, an unsupervised instance filter that maintains the class distribution in the subsample, where alternatively it can be configured to bias the class distribution towards a uniform distribution (Gadotti et al., 2022b). This technique performs subsampling and is the best among conventional approaches (Sarada & Devi, 2019). Feature selection is a key step for the classifier to function properly, presenting its best performance (Sarada & Devi, 2019).

According to Mariano (2021), in classification problems, there are two possible solutions: correct or incorrect. However, in this work, there are also positive and negative classes, as it is considered a binary problem (Mariano, 2021). For explanation, each class can be understood as follows:

- **True Positive (TP):** when the classifier places the class as positive and upon verification, identifies that it is indeed positive;
- **True Negative (TN):** when the algorithm understands it as a negative class and then certifies the information;
- **False Positive (FP):** when the computational model concludes that the class is positive, however, this class is negative,
- **False Negative (FN):** when the classifier indicates that the class is negative, however, it is positive.

Thus, the following evaluation metrics were used to calculate the performance of the methods: Correct Instances; Kappa, Precision, and ROC Area according to Lever et al. (2016). True positive (TP), false positive (FP), true negative (TN), and false negative (FN) values extracted from the confusion matrix were used to calculate the Precision metrics, through [eq. (1)], as proposed by Medeiros et al. (2020a). With the results obtained, the best learning technique was determined.

$$Precision = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where:

TP = True Positives;

TN = True Negative;

FP = False Positive,

FN = False Negative.

After these steps, the set is ready to be processed by the main task in the entire process: mining. From this, algorithms are employed, often repetitively, searching for patterns and rules in the data. Finally, the discovered information is interpreted and evaluated, often in the form of graphs or reports, selecting useful knowledge (Vasconcelos & Carvalho, 2018). This process can be understood in Figure 1.

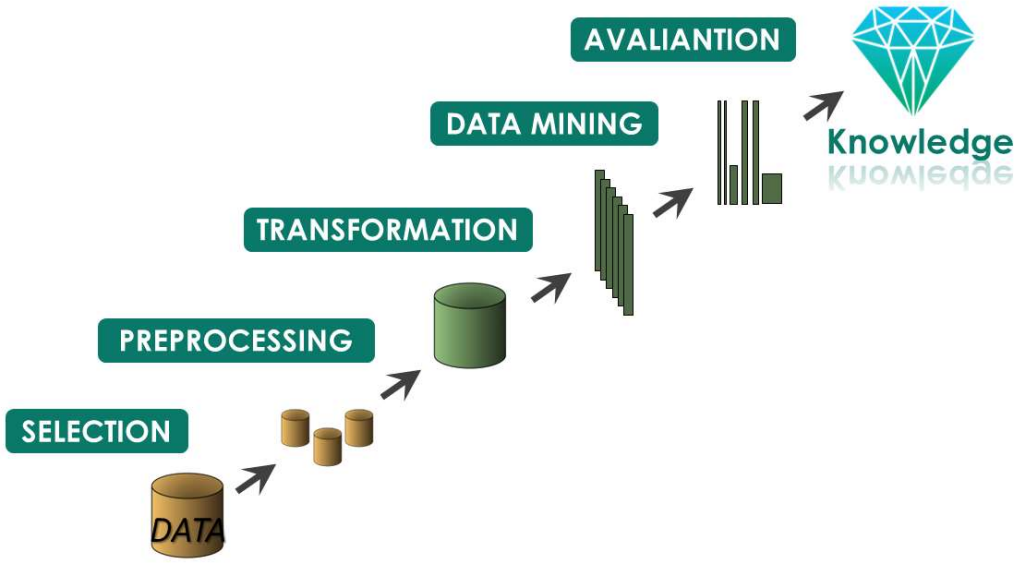

FIGURE 1. Sequence of operations using data mining.

RESULTS AND DISCUSSION

Data mining is a way of finding unknown information in large datasets (Eti & İnel, 2019). With the development of computers and advances in algorithm innovation, as well as greater data availability, mining methods have undergone significant technological advances (Artrith et al., 2021). In this sense, machine learning techniques and statistical analyses have been developed to analyze large data sets. Machine learning can be divided into two categories: supervised and unsupervised (Eti & İnel, 2019).

Seed lots are ranked to provide accurate information on their health aspects to farmers, who seek germination and vigor nearly 100%. This means a good quality genetic material to start new cropping, reducing the use of agrochemicals for crop establishment. According to Costa et al. (2018) and Rocha et al. (2017), genetic material quality is fundamental for crop development, generating healthier plants and highly uniform stands, reducing pathogen attacks, and improving competitiveness against invasive plants. These factors are also decisive for reductions in fertilizers and agrochemicals throughout

crop development, besides reducing negative impacts on the environment.

Furthermore, the choice of method for analyzing a dataset is determined by its characteristics, as well as the expected result. In large datasets, more exotic and deep algorithms, such as neural networks, should be adopted, while in smaller datasets, more classical techniques, such as linear regression and decision trees, are recommended, taking care to adapt each approach to data characteristics (Nichols et al., 2018). Decision tree algorithms create a sequence of rules to increase knowledge gain, enabling classifying data from a model composed of branches and leaves (Eti & İnel, 2019). For this technique, there are some models, such as ID3, C4.5, and J48, which is an evolution of the latter (Joshuva et al., 2020).

Table 2 shows the accuracy parameters for the training methods using the J48 algorithm. In this case, when using 66% of the data for training, results showed better performance (93.55% accuracy), with precision and ROC (Receiver Operator Characteristic) area achieving 94.50% accuracy. Cross-validation with 10 folds, widely used in machine learning research, achieved 90.22% of instances classified correctly, but with a ROC Area result close to the previous method.

TABLE 2. Accuracies of the J48 algorithm for the different training methods.

Training Method	Correct Instances (%)	Kappa (%)	Precision	ROC Area
Supplied Test Set	81,52	63,15	82,00	81,80
8 folds	90,22	80,21	90,60	92,70
10 folds	90,22	80,21	90,60	93,80
12 folds	88,04	75,88	88,20	91,80
Percentage Split 66%	93,55	86,81	94,20	94,50
Percentage Split 70%	92,86	85,49	93,70	93,80

The ROC curve or area demonstrates the relationship between the sensitivity (VP rate) and specificity (FP rate) of the classifier; the higher the value, the better the curve is fitted (Gadotti et al., 2022a). However, the J48 classifier aims to generate a decision tree based on labeled data, involving qualitative variables, and it is the best algorithm to use this technique (Joshuva et al., 2020). To induce a decision tree, it divides a complex problem into simpler ones, applying the same strategy again until getting to a final solution (Costa et al., 2014).

As mentioned above, Gadotti et al. (2022a) and Gadotti et al. (2022b) used the 10-fold cross-validation technique for analyzing data on corn and soybean seeds. But, when these methods are used with a CVR algorithm, accuracy is reduced. Conversely, in our study, the 66% training technique maintained the highest amount of correctly classified instances (Table 3).

TABLE 3. Accuracies of CVR algorithm for the different training methods.

Training Method	Correct Instances (%)	Kappa (%)	Precision	ROC Area
Supplied Test Set	84,78	69,52	84,90	86,10
8 folds	85,87	71,58	85,90	90,00
10 folds	85,87	71,58	85,90	90,06
12 folds	83,69	67,30	83,70	88,70
Percentage Split 66%	90,32	80,34	90,40	89,10
Percentage Split 70%	89,28	78,35	89,40	87,70

However, when the ROC Area results were evaluated, the 8- and 10-fold cross-validation obtained better figures, training a better-applied model. This training was used by Gadotti et al. (2022a; 2022b) to work with a low amount of data, thus being an alternative for these cases.

In addition, the confusion matrix of the J48 algorithm was extracted when it was trained with 66% of the data (Table 4). The rejected class obtained some errors, but the test data used were mostly classified correctly. This shows that the training technique was suitable for this activity.

TABLE 4. Confusion matrix of J48 algorithm with the 66% Percentage Split method.

		Prediction	
		Accepted	**Rejected**
Real Class	**Accepted**	17	0
	Rejected	2	12

Table 5 shows the confusion matrix using the same training method, but with the CVR classifier.

TABLE 5. Confusion matrix of CVR algorithm with the 66% Percentage Split method.

		Prediction	
		Accepted	**Accepted**
Real Class	**Accepted**	16	1
	Rejected	2	12

When compared to the previous model, which presented better accuracy values, the results were similar, and the Rejected class had the same lot separation, while the Accepted class had one incorrectly classified lot.

Cross-validation consists of resampling data to validate the predictive capacity of models, preventing overfitting. The data is divided into equal or double subsets,

one used for testing and the remaining for training (Berrar, 2018). No studies have indicated an ideal number of folds, and it is necessary to perform tests with several possibilities, selecting the best when training a model, which is the objective of this work.

Thus, a confusion matrix evaluates an algorithm's performance based on the classifier's errors and successes (Ribeiro et al., 2016). As recently performed research in the seed sector, a 10-fold cross-validation technique was used here, with a confusion matrix of the same classifier using this training method (Table 6).

TABLE 6. Confusion matrix of the J48 algorithm with the 10-fold cross-validation method.

		Prediction	
		Accepted	**Accepted**
Real Class	**Accepted**	47	2
	Rejected	7	36

In this case, errors were higher; therefore, accuracy and precision for prediction and ranking of seed lots also depend on the computational model applied, not only the training technique used as a benchmark for results. According to Ribeiro et al. (2016), the machine learning process must consider the relationship between the technique used and the goal aimed at. This is because once a classifier is established, its method must be analyzed. Despite using the same training technique in all chosen algorithms, the results will be different since the mathematical model used is the one to define which one relates best to the initially proposed objective.

A CVR classifier uses regression methods to create a binary class by generating a regression model for each class value (Joshuva et al., 2020). Table 7 displays the CVR matrix, which indicates that the Rejected class achieved a comparable outcome, despite using a distinct computational model.

TABLE 7. Confusion matrix of CVR algorithm with the 10-fold cross-validation method.

		Prediction	
		Accepted	**Accepted**
Real Class	**Accepted**	43	6
	Rejected	7	36

Figure 2 presents the decision tree for the J48 algorithm using the 66%- and 10-fold techniques. The highlighted attribute was the Tetrazolium Vigor, which was the first criterion used. According to Gadotti et al. (2022a), this analysis enables a quick determination of seed viability, even for the most dormant seeds, in comparison to the germination test.

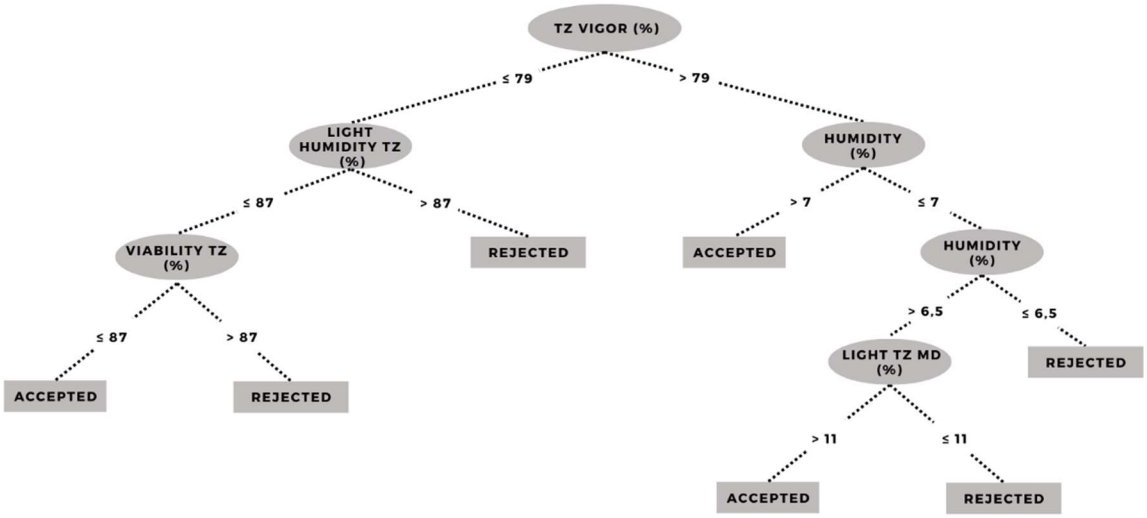

FIGURE 2. Decision tree resulting from J48 algorithm using the Percentage Split 66% and Cross-Validation 10 folds method.

According to Soares et al. (2016), accurate classification through proper algorithm training is essential to provide reliable information about seed viability in a short time, so that decisions can be made quickly and accurately.

CONCLUSIONS

Based on the results obtained in the different algorithm training techniques, dividing the data into 66% for training showed better results when compared to cross-validation, a method commonly used by researchers. However, the training method used does not solely determine the results of each algorithm, as they also depend on the chosen model. Therefore, planning and studying are necessary to choose the best methods to be applied to this type of data.

From our findings, further research can be developed with a stronger scientific basis, understanding how each training method works to improve the results of a chosen algorithm.

ACKNOWLEDGMENTS

We would like to thank the *Fundação Assistencial e Previdenciária da Extensão Rural no Rio Grande do Sul* (FAPERGS), the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), and the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), under process n° 311722/2020-2, for the financial support in providing research scholarships to carry out this study.

REFERENCES

Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, Walsh A (2021) Best practices in machine learning for chemistry. Nature Chemistry 13: pp. 505-508. https://doi.org/10.1038/s41557-021-00716-z

Berrar D (2018) Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology 1: 542-545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Brasil - Ministério da Agricultura, Pecuária e Abastecimento (2009) Regras para análise de sementes. Brasília, DF, Secretaria de Defesa Agropecuária. 399p.

Costa EM, Nunes BM, Ventura MVA, Arantes BHT, Mendes GR (2018) Efeito fisiológico de inseticidas e fungicida sobre a germinação e vigor de sementes de soja (*Glycine max* L.). Científic@ - Multidisciplinary Journal 5(2): 77-84. http://dx.doi.org/10.29247/2358-260x.2018v5i2.p77-84

Costa JJ, Bernardini FC, Viterbo Filho J (2014) A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. Atoz - novas práticas em informação e conhecimento 3(2): 139. http://dx.doi.org/10.5380/atoz.v3i2.41346

Eibe F, Mark AH, Ian HW (2020) The WEKA Workbench. Online appendix for data mining: practical machine learning tools and techniques. Burlington, Morgan Kaufmann.

Elias SG (2018) The importance of using high quality seeds in agriculture systems. Agricultural Research & Technology 15(4): 100. https://doi.org/10.19080/ARTOAJ.2018.15.555961

Eti S, İnel MN (2019) A research on the comparison of classification algorithm in finance. Contaduría y Administración 65(4): 204. http://dx.doi.org/10.22201/fca.24488410e.2020.2497

França-Neto JB, Krzyzanowski FC (2022) Use of the tetrazolium test for estimating the physiological quality of seeds. Seed Science And Technology 50(2): 31-44. http://dx.doi.org/10.15258/sst.2022.50.1.s.03.

Gadotti GI, Ascoli CA, Bernardy R, Monteiro RCM, Pinheiro RM (2022a) Machine learning for soybean seeds lots classification. Engenharia Agrícola 42: spe. http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v42nepe20210101/2022

Gadotti GI, Moraes NAB, Silva JG, Pinheiro RM, Monteiro RCM (2022b) Prediction of ranking of lots of corn seeds by artificial intelligence. Engenharia Agrícola 42(4): e20210005. http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v42n4e20210005/2022

Joshuva A, Arjun M, Murugavel R, Shridhar VA, Gangadhar GSS, Dhanush SS (2020) Predicting wind turbine blade fault condition to enhance wind energy harvest through classification via regression classifier. Lecture Notes in Electrical Engineering 687: 13-20. http://dx.doi.org/10.1007/978-981-15-7245-6_2

Lever J, Krzywinski M, Altman N (2016) Classification evaluation. Nature Methods 13: 603-604. https://doi.org/10.1038/nmeth.3945

Mariano D (2021) Métricas de avaliação em machine learning: acurácia, sensibilidade, precisão, especificidade e f-score. Bioinfo - Revista Brasileira de Bioinformática e Biologia Computacional 1: 1-9. http://dx.doi.org/10.51780/978-6-599-275326-15

Medeiros AD, Capobiango NP, Silva JM da (2020a) Interactive machine learning for soybean seed and seedling quality classification. Scientific Reports 10: 11267. https://doi.org/10.1038/s41598-020-68273-y

Medeiros AD, Silva LJ, Ribeiro JPO, Ferreira KC, Rosas JTF, Santos AA, Silva CB da (2020b) Machine learning for seed quality classification: an advanced approach using merger data from ft-nir spectroscopy and x-ray imaging. Sensors 20(15): 4319. http://dx.doi.org/10.3390/s20154319

Nichols JA, Chan HWH, Baker MAB (2018) Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophysical Reviews 11(1): 111-118. http://dx.doi.org/10.1007/s12551-018-0449-9

Patrício DI, Rieder R (2018) Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review. Computers and Electronics in Agriculture 153: 69-81. https://doi.org/10.1016/j.compag.2018.08.001

Peng J, Xu J (2022) Deep learning analysis on the resulting impacts of weekly load training on students' biological system. Revista Brasileira de Medicina do Esporte 29 (spe1): 1-5. https://doi.org/10.1590/1517-8692202329012022_0197

Pinheiro RM, Gadotti GI, Monteiro RCM, Bernardy R (2021) Inteligência artificial na agricultura com aplicabilidade no setor sementeiro. Diversitas Journal 6: 2984-2995. https://doi.org/10.48017/Diversitas_Journal-v6i3-1857

Ribeiro BG, Abrahão CP, Nery MC, Nascimento RM, Rezende JC, Fialho CMT (2016) Image analysis of coffee seeds submitted to the LERCAFE test. Acta Scientiarum. Agronomy 38(3): 355-361. http://dx.doi.org/10.4025/actasciagron.v38i3.28268

Rocha GC, Rubio Neto A, Cruz SJS, Campos GWB, Castro ACO, Simon GA (2017) Physiological quality of treated and stored soybean seeds. Científic@ - Multidisciplinary Journal 4(1): 50. http://dx.doi.org/10.29247/2358-260x.2017v4i1.p50-65

Sarada C, Devi MS (2019) Imbalanced big data classification using feature selection under-sampling. CVR Journal of Science and Technology 17: 78-82. https://doi.org/10.32377/cvrjst1714

Soares VN, Elias SG, Gadotti GI, Garay AE, Villela FA (2016) Can the tetrazolium test be used as an alternative to the germination test in determining seed viability of grass species? Crop Science 56(2): 707-715. http://dx.doi.org/10.2135/cropsci2015.06.0399

Vasconcelos LMR, Carvalho CL (2018) Aplicação de regras de associação para mineração de dados na Web. Revista Telfract 1: 1-20. Available: https://telematicafractal.com.br/revista/index.php/telfract/article/view/8. Accessed Oct 18, 2022.