Journal of Food Science and Technology
https://doi.org/10.1007/s13197-025-06433-6

ORIGINAL ARTICLE

Optimization of extraction of bioactive compounds from *Piper corcovadensis* C.DC leaves using a generalized linear model

Bruno Henrique Fontoura[2] · Luciano de Souza Ramos[1] · Lucas Vinícius Dallacorte[2] · Michelle Fernanda Faita Rodrigues[2] · José Abramo Marchese[2] · Tiago Adriano Fernandes[3,4] · Mário Antônio Alves da Cunha[1] · Vanderlei Aparecido de Lima[1] · Solange Teresinha Carpes[1]

Received: 29 January 2025 / Revised: 5 July 2025 / Accepted: 26 July 2025

Abstract

This concerns *P. corcovadensis*, an endemic plant of Brazil commonly used by the population due to its therapeutic properties. Optimizing chemical extraction conditions is critical for increasing the availability of bioactive compounds from plants. These compounds have antioxidant potential derived from a plant's specialized metabolism and can exhibit a variety of biological actions. Therefore, statistical tools such as the Random Forest and Lazy KStar machine learning algorithms were used to determine the optimal condition for the extraction of phenolic compounds from *P. corcovadensis* leaves, with model evaluated by coefficient of determination (R^2), mean square root of calibration error (RMSEC), and residual predictive deviation (RPD). The optimal extraction condition was obtained using a mixture of 80/20% (ethanol/water) at 70 °C for 120 min. For those extracts, there were 11.64 ± 0.04 mg GAE g^{-1} and antioxidant activity of 21.27 ± 0.53 mmol Trolox g^{-1}, 33.15 ± 11.66 mmol Trolox g^{-1}, and 13.47 ± 1.37 mmol Fe^{2+} by DPPH, ABTS and FRAP tests. With this study, we have shown that mathematical modelling can also be helpful in experimental sciences and can be used to develop predictive models. It was possible to develop predictive models for total phenolic compounds determination using the Random Forest and Lazy KStar machine learning algorithms. The Random Forest algorithm performed very well for DPPH modelling, giving us the confidence to use it to prediction antioxidant activity.

✉ Solange Teresinha Carpes
carpes@utfpr.edu.br

1 Chemistry Department, Universidade Tecnológica Federal do Paraná, Campus Pato Branco, PO Box 571, Pato Branco, PR CEP 85503-390, Brazil

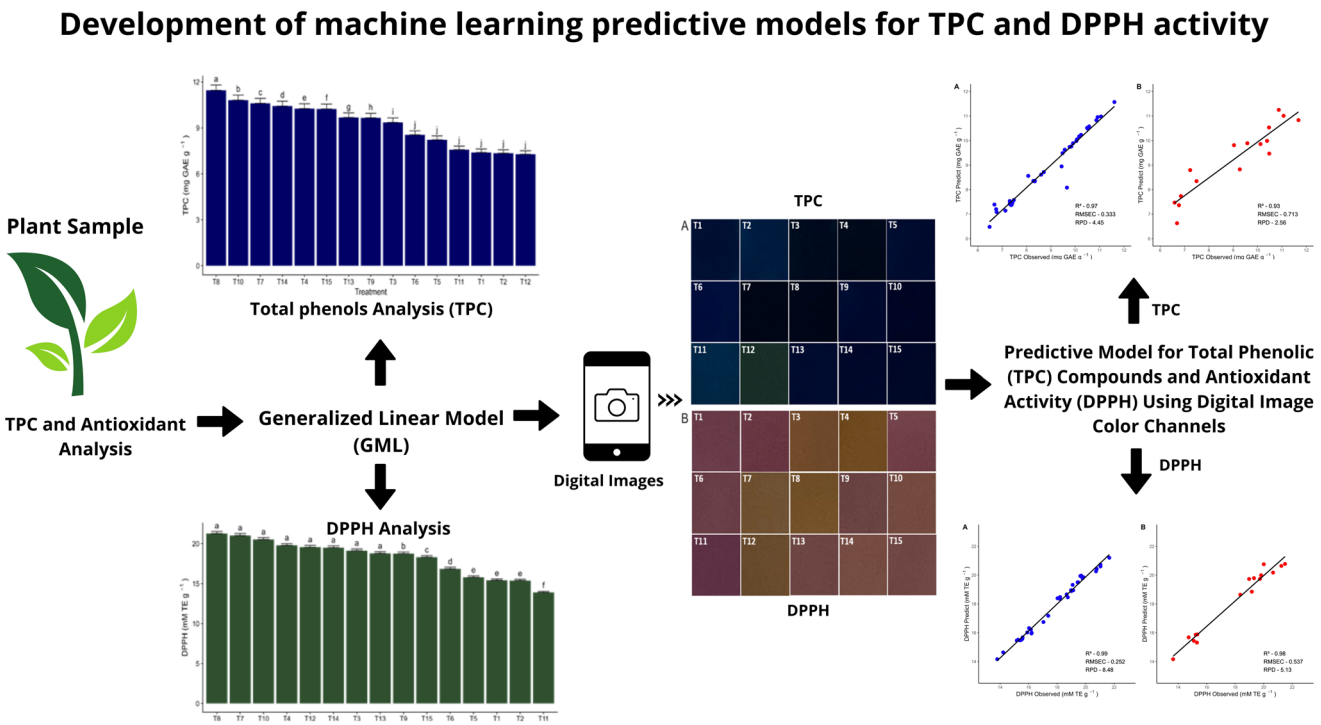
2 Agronomy Department, Universidade Tecnológica Federal do Paraná, Campus Pato Branco, PO Box 571, Pato Branco, PR CEP 85503-390, Brazil

3 MINDlab: Molecular Design & Innovation Laboratory, Centro de Química Estrutural, Institute of Molecular Sciences, Departamento de Engenharia Química, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisbon 1049-001, Portugal

4 Departamento de Ciências e Tecnologia, Universidade Aberta, Rua da Escola Politécnica, 141-147, Lisboa 1269-001, Portugal

Published online: 12 August 2025

Graphical Abstract



Keywords Specialized metabolism · Phenolic compounds · Antioxidant activity · Generalized linear model · Machine learning

Introduction

Natural products are compounds synthesized by the specialized metabolism of plants for defense and adaptation within their biological environment (Yao et al. 2023). Phenolic compounds characterized by aromatic rings attached to one or more hydroxyl groups are the most abundant type of chemical in this metabolic process (Yusoff et al. 2022). In addition to their primary function for plants, these compounds may have various biological activities for humans and animals, including antioxidant, antimicrobial, antiparasitic, anti-inflammatory, photoprotective, and antitumor activities (Mehdizadeh et al. 2024; Sulejmanović et al. 2024; Zeng et al. 2024).

Piper corcovadensis C.DC is a Brazilian natural plant that is widely found in regions covered by the Atlantic and Amazon rainforests. It is commonly used empirically to treat colds, coughs, rheumatism, and general pain (Herrera et al. 2022), and its properties are reported as antioxidant, antimicrobial, and antitumor activities derived from its chemical components (Fontoura et al. 2024). However, extraction strategies must be studied to improve the quantity of potential extracted, bioavailability, and stability of compounds obtained from specific metabolism in this plant (de Marsiglia et al. 2023).

The use of solvents with a higher chemical affinity for target compounds, e.q. different polarities, as well as the use of appropriate time and temperature to avoid degrading of the chemical structures, can be cited as important factors for maximizing natural antioxidant extraction (Fontoura et al. 2024). Furthermore, statistical approaches can be used to assess how different variables, such as solvent, temperature, time, concentration, pressure, etc., can affect the extraction process and identify optimal factors and levels to ensure that the process takes place under conditions that guarantee the highest availability of compounds (Virág et al. 2024). When the data obtained cannot be analyzed using analysis of variance (ANOVA) due to its non-normal distribution, data normalization can solve this problem by changing the data distribution and making it parametric (Hamasha et al. 2022).

Another possibility is to use non-parametric approaches that use different probability distributions. One such approach is the Generalized Linear Model (GLM), in which the data can have an exponential (Adesina et al. 2021). This approach is widely used in classifying of hydrocarbons (Wang et al. 2021), epidemiological risk predictions

(Mudele et al. 2021), and weather forecasting (Chandler 2020), but it is seldom used in natural products.

In this context, extraction models can be validated using machine learning algorithms, prediction and validation models are created using decision tree algorithms, displayed as a response to the fit of the experimental data with the models developed by the algorithm (Sircar et al. 2021). In addition, prediction models can be developed using digital images obtained by commonly and daily used objects, such as smartphones or tablets, making data predictability faster and with lower operational costs than traditional experimental data acquisition approaches (Fontoura et al. 2023).

The application of data analysis models alternative to ANOVA, such as Generalized Linear Models (GLM), and machine learning (ML) algorithms in the extraction of natural products represents a methodological advancement by enabling the modeling of complex and non-normally distributed data. This approach improves the prediction of optimal extraction factor levels, contributing to the reduction of reagent use, experimental time, and operational costs (Shi et al. 2024).

In this study, we aimed to optimize the extraction of phenolic compounds with antioxidant potential from *P. corcovadensis* leaves by adjusting the concentration, temperature, and extraction time of the solvent. Additionally, the data was treated using a generalized linear model, and a chemometric prediction model of the data was constructed through digital images obtained via smartphone from the total phenolic compounds and antioxidant activity assays.

Materials and methods

Chemicals and reagents

Ethanol and Folin-Ciocalteu phenol reagent were obtained from Êxodo Cientifica LTDA (Sumaré, SP, Brazil). Gallic acid, 2,2-diphenyl-1-picrylhydrayl (DPPH), 2,4,6-tris(2-pyridyl)-s-triazine (TPTZ), 2–2'-azino-di-(3-ethylbenzthiazoline sulfonic acid) (ABTS) and Trolox were purchased from Sigma-Aldrich Chemical Co. (St. Louis, MO, USA).

Piper corcovadensis leaves

Samples were collected in the rural area of the municipality of Alto Paraíso (Rondônia state – Brazil) (9 ° 44'43' S; 63 ° 17'0' W, at 143 m altitude) in October 2021 and gently provided by the Fonte Clara Indústria e Comércio de Produtos Naturais LTDA company based in the municipality of Bom Sucesso do Sul (Paraná state – Brazil). *Piper corcovadensis* was identified and deposited in the herbarium of the Universidade Estadual do Centro Oeste – UNICENTRO – Campus Guarapuava (Paraná state – Brazil) under voucher # ARAUCA 1154.

The plant material was dried in an oven at 37 ° C until a constant mass. Subsequently, it was ground (30 mesh) using a knife mill (Tecnal R-TE-650/1 model, Piracicaba, SP, Brazil). The grounded samples were placed in plastic bags and stored in a freezer at -12 °C until extraction.

Factorial planning and extraction of phenolic compounds

For the extraction of phenolic compounds, a factorial experiment design with three factors was performed: (1) solvent mixture (% ethanol/water), (2) extraction temperature (°C), (3) and extraction time (minutes) in a water bath. Factors were evaluated at five levels, including a central level (0), two higher levels (1; 1.68), and two lower levels (-1; -1.68) from the central level (Table 1).

For all levels studied, the extract was prepared with 0.1 g of dried *P. corcovadensis leaves mass* in 10 mL of solvent. The extracts obtained were stored in a freezer at -12 °C until analysis was performed.

Total phenolic compounds (TPC) and antioxidant activity (AA)

The TPC analysis was performed using the Folin-Ciocalteu spectrophotometric method at 740 nm (UV-Vis, KASVI K37), and the results were expressed as mg GAE g^{-1} (GAE: Gallic acid equivalent) (Singleton et al. 1999).

The antioxidant activity was carried out using the DPPH method and evaluated at a wavelength of 517 nm. The results were expressed as mmol TE g^{-1} (TE: Trolox equivalent). The extract obtained under the best conditions was also determined using the ABTS radical scavenging method and Ferric Reductor Activity Power (FRAP) methods. Both were evaluated by UV-Vis spectrophotometer at wavelengths 734 and 595 nm, respectively. The result of the ABTS test was expressed as mmol TE g^{-1} (TE: Trolox equivalent). The result of the FRAP test was expressed as mmol de Fe^{2+} g^{-1}(Re et al. 1999).

Statistical analysis of the data set using a generalized linear model (GLM)

Statistical analysis was performed using GLM, with a gamma probability distribution, using R Studio software. The Bonferroni test was conducted to compare the estimated marginal means of each treatment.

The odds ratio between treatments was calculated using Eq. (1).

Table 1 Values encoded and decoded from the three-factor experiment and the results of total phenolic compounds (TPC) and antioxidant activity (AA) using the DPPH method

Treatment	Factor 1 %Ethanol/Water	Factor 2 Temperature (°C)	Factor 3 Time (min)	TPC (mg GAE g^{-1})		DPPH (mmol TE g^{-1})	
T1	-1 (20/80)	-1 (50)	-1 (50)	7.48±0.00*	j**	15.42±0.733	e
T2	-1 (20/80)	-1 (50)	1 (120)	7.34±0.06	j	15.37±0.19	e
T3	1 (80/20)	-1 (50)	-1 (50)	9.52±0.06	i	19.13±0.17	a
T4	1 (80/20)	-1 (50)	1 (120)	10.25±0.30	e	19.79±0.19	a
T5	-1 (20/80)	1 (70)	-1 (50)	8.30±0.04	j	15.80±0.28	e
T6	-1 (20/80)	1 (70)	1 (120)	8.65±0.07	j	17.20±0.24	d
T7	1 (80/20)	1 (70)	-1 (50)	10.86±0.00	c	21.01±0.53	a
T8	1 (80/20)	1 (70)	1 (120)	11.64±0.04	a	21.27±0.53	a
T9	0 (50/50)	-1.68 (40)	0 (80)	9.42±0.20	h	18.76±0.40	b
T10	0 (50/50)	1.68 (80)	0 (80)	10.98±0.09	b	20.52±0.63	a
T11	-1.68 (0/100)	0 (60)	0 (80)	6.53±0.07	j	13.89±0.29	f
T12	1.68 (100/100)	0 (60)	0 (80)	7.29±0.14	j	19.56±0.18	a
T13	0 (50/50)	0 (60)	-1.68 (30)	9.80±0.05	g	18.77±0.492	a
T14	0 (50)	0 (60)	1.68 (150)	10.52±0.05	d	19.50±0.44	a
T15 (C)	0 (50)	0 (60)	0 (80)	10.24±0.24	f	18.27±0.34	c

Values followed by different letters in the same column are significantly different ($p<0.05$). GAE: Gallic acid equivalent, TE: Trolox equivalent, C: Central point. The results are expressed as mean ± standard error ($n=3$)

$$logit\,(\widehat{\pi}\,) = \beta\ (Intercept) + \beta\ 1xT1 + \beta\ 2xT2 \ldots + \beta\ 15xT15 \quad (1)$$

The overall response to optimize extraction factors was calculated using Eq. (2).

$$RG = \left[\begin{matrix} R\,(x_1)\,/MR\,(x_1) + R\,(x2)\,/MR\,(x2) \\ +\ldots + R\,(x_n)\,/MR\,(x_n) \end{matrix}\right] \quad (2)$$

where $R(x_n)$ is the response for an element in a specific experiment and $MR(x_n)$ is the maximum response in the set for element n.

Chemometric modeling using digital images and machine learning (ML) algorithms

The models were constructed using images obtained from the TPC and DPPH optimization assays. Images were captured in a mini photographic studio (56 LED-Lampe Evodobox 60 cm plus – Evodobox plus) using the camera of a smartphone (Apple iPhone®, model SE, 12 MP camera), under standardized lighting and distance conditions (Perin et al. 2020).

The digital images (8 bits) were cropped to dimensions of 400 × 400 dpi using open-source software, GIMP (version 2.10.18). RGB colour channels (red, blue, and green), HSL standards (hue, saturation, and luminosity), and V and I patterns (chromatic vortex, intensity, and brightness) were extracted using ChemoStat® software.

The variables were subjected to the VIF (Variance Inflation Factor). This test identifies multicollinearity variables and can inflate the model accepted values of VIF ≤ 15 (Cheng et al. 2022). Regression between models was performed using WEKA® 3.8.5 software, and predicted vs. observed plots were developed using R Studio software. Evaluation of the precision and performance of the models was obtained through the root mean square error of calibration (RMSEC), coefficient of determination (R^2), and residual predictive deviation (RPD).

Results and discussion

Optimization of phenolic compound extraction with data processing using GLM

The results obtained by estimating the marginal means (Table 1S for TPC and Table 2 S for DPPH) on a 95% confidence interval indicated that treatment T8 achieved the best response for the variables TPC and DPPH for *P. corcovadensis* leaves (Table 1). In this extract, 11.64 ± 0.04 mg GAE g^{-1} and 21.27 ± 0.53 mmol TE g^{-1} were achieved with 80/20% ethanol/water, 70 °C temperature during 120 min of extraction (Table 1). Furthermore, according to the Bonferroni test (Table 3 S), for TPC, T8 compared to all other runs ($p<0.05$) showed a statistically significant difference.

Treatment T12 showed the lowest levels for TPC values (7.29 ± 0.14 mg GAE g^{-1}). In this condition, ethanol (100%) at 60 °C for 80 min of extraction. It did not present a significant difference ($p<0.05$) compared to treatments T1, T2, T5, T6 and T11 (Fig. 1A).

However, for antioxidant activity (AA) using the DPPH method, T8 did not show a significant difference ($p \geq 0.05$) according to the Bonferroni test (Table 4 S) compared to

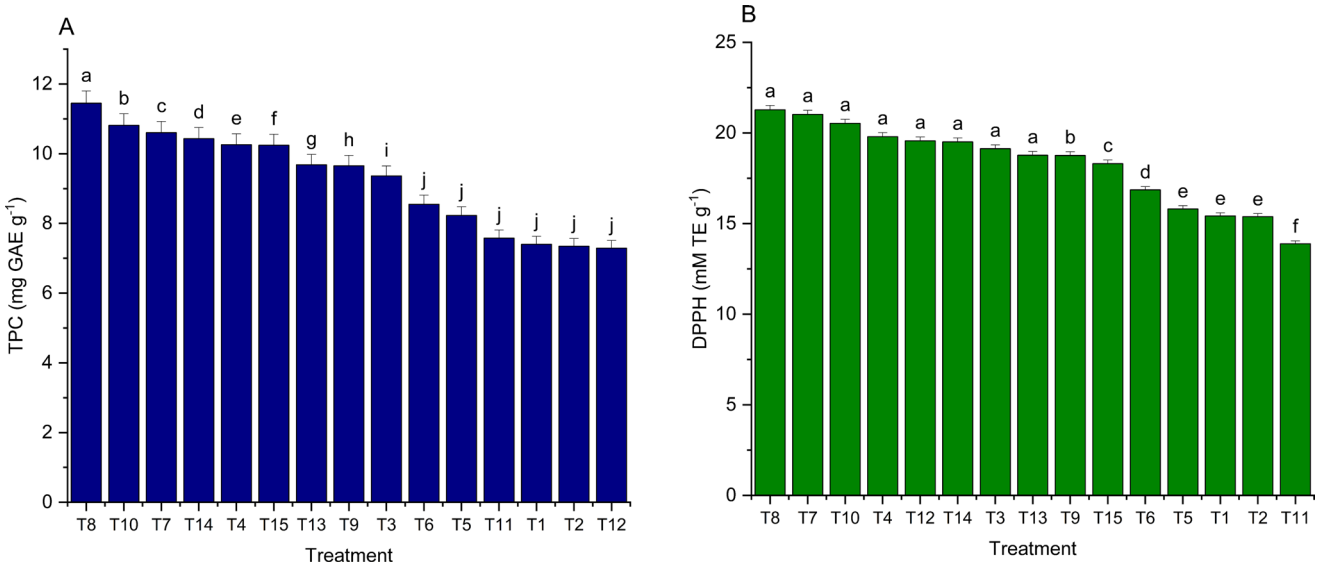


Fig. 1 **A**: Represent the Total Phenolic Compounds (TPC) and **B**: Represent the antioxidant activity by the DPPH method means of the treatments (T1-T15). Different letters indicate a significant difference according to the Bonferroni test ($p \leq 0.05$). GAE: Gallic acid equivalent

treatments T3, T4, T7, T10, T12 to T14 (Fig. 1B). For DPPH, T11 exhibited the lowest antioxidant potential obtained from water (100%), a temperature of 60 °C and an extraction time of 80 min. It showed a significant difference ($p \leq 0.05$) compared to the all other treatments (Fig. 1B).

The solvent concentration factor proved to be crucial in the extraction of phenolic compounds, as well as in the interaction between ethanol and water (Table 1). Treatments T11 and T12, with 0 and 100% ethanol, respectively, showed lower extraction efficiency than treatments with solvent interaction.

Treatment T11 with 0% ethanol exhibited the lowest antioxidant potential, and in comparison, treatment T12 with 100% ethanol showed a higher antioxidant potential and did not show a significant difference from T8 ($p < 0.05$), which obtained the highest antioxidant potential. Treatments with an ethanol concentration in the range of 20 to 80% showed higher levels of TPC and antioxidant activity, indicating that lower concentrations of ethanol and higher concentrations of water may not be efficient. However, there was better extraction efficiency in the 50 to 80% concentration range.

According to Galanakis et al. (2013), the use of solvents with high and intermediate polarities, such as water and ethanol, respectively, are effective in the extraction of phenolic compounds due to the intermolecular interactions (dipole-dipole, London dispersion forces and hydrogen bonds) of the solvent with the chemical structures of the phenolic compounds (phenolic acids, hydrolyzable tannins and flavonoids). In addition to the temperature factor, the agitation rate of the plant material, along with the solvent, helps to disrupt cell membranes and tissues, increasing cellular diffusion. Higher temperature promotes greater outstanding mass transfer between the plant material and the solvent, increasing thus the availability of phenolic compounds in the extracts (Song et al. 2019).

Although the T1, T5 and T6 treatments have been extracted at different temperatures, 50 and 70 °C, according to the Bonferroni test, there is no significant difference between these treatments ($p < 0.05$), indicating that the temperature effect was not sufficient to increase the phenolic content in the extracts. On the other hand, for treatments T3 and T4 (50 °C) and T7 and T8 (70 °C), with ethanol/water with a rate of 80/20%, significant differences were observed (Table 1). It should be noted that as the temperature increases, so does the concentration of phenolic compound content in the extract. However, treatments with the same temperature and solvent ratio also diverged, showing a time-dependent factor.

Treatments T11 and T12, with a temperature of 50 °C and ethanol/water ratios of 0 and 100%, respectively, did not differ from each other ($p < 0.05$), revealing that the temperature factor did not have an effect on extraction. Meanwhile, the treatments T13, 14, and T15, all with an extraction temperature of 60 °C and ethanol concentration of 50%, differed significantly ($p < 0.05$), suggesting that the difference in phenolic compound content may be due to extraction time. However, T10 treatment with 50% ethanol at 80 °C resulted in an increase in the phenolic content, indicating that temperature influences extraction (Table 1).

When analyzing antioxidant activity responses (DPPH), temperature is not a determining factor for T1, T5, and T6. There are no significant differences in the treatments with the same solvent ratio and at different temperatures, 50 °C and 70 °C ($p < 0.05$), respectively. Furthermore, there were

no significant differences between treatments T3 and T4 ($p \geq 0.05$) with an extraction temperature of 50 °C, and T7 and T8 with an extraction temperature of 70 °C and solvent ratio (80/20% ethanol/water) (Table 1).

Treatments T11 and T12 with 0 and 100% ethanol and the same extraction temperature (50 °C) did not diverge ($p \leq 0.05$), indicating that the temperature was not influenced, with only the solvent having an effect. Similarly, treatments T13, T14 and T15, all with 50% ethanol and an extraction temperature of 60 °C, did not show a significant difference between T13 and T14. Only T15 diverged from the other samples, and this disparity could be explained by the extraction time (80 min), different from T13 and T14 (30 and 150 min, respectively) (Table 1).

Treatment T10 with 50% ethanol and 80 °C (the highest temperature tested) was not observed to show a significant difference with T3, T4 and T12 (all with a temperature of 50 °C), T7 and T8 (with 70 °C) and T13 and T14 (60 °C), again suggesting that the temperature factor does not determine the antioxidant potential of the extract.

In this study, the extraction time was a determinant and influenced the content of phenolic compounds of treatments with the same solvent ratio and temperature, such as T7 and T8 (80% ethanol and 70 °C). It was observed that T7, which had an extraction time of 50 min, had a lower phenolic content compared to T8, which had an extraction time of 120 min. Both samples exhibited a significant difference ($p \leq 0.05$). However, for DPPH, there was no significant difference between the treatments.

The samples from treatments T13, T14, and T15 with the same solvent mixture ratio and temperature (50% ethanol and 60 °C) showed a significant difference between them for TPC ($p \leq 0.05$). It could be seen that the extraction times were 30, 150, and 80 min, respectively. It was noted that T14, which had a 150 min extraction time, had a higher content of phenolic compounds in the extract.

The extraction time is crucial in the extraction process as it determines the duration of contact between the plant material and the solvent (Chamali et al. 2023). In cases where all other factors are the same, increasing the extraction time can significantly enhance the content of phenolic compounds and compounds with antioxidant potential. From an analytical perspective, it is evident that the extraction processes are influenced by several key factors; these factors include the solvent mixture composition and its interactions with the compounds present in the plant material, the extraction temperature, and the duration of exposure of the plant material to the solvent and heating. These factors are of utmost importance in determining the success of the extraction process. Changes in any of the studied factors led to variations in the levels of bioactive compounds, as observed in treatments T3 and T4 for the extraction of total phenolic compounds (TPC). In these treatments, the solvent and temperature were kept constant, while only the extraction time was modified, resulting in a significant difference in the phenolic compound content.

Equations for generalized linear model (GLM)

The *odds ratio* is a statistical measure that is used to quantify the strength of association between two categorical variables. In other words, it is a ratio of the odds that an event will occur in one group compared to another group (Chen et al. 2011). The odds ratio associated with these explanatory variables can be calculated using Eq. (3).

$$\mathrm{OR} = e^{\widehat{\beta}_1} \tag{3}$$

Where $\boldsymbol{e}$ is the base of the natural logarithm (approximately 2.71828), and $\boldsymbol{\beta}$ are the model coefficients.

Through Eq. 3 and the significant terms ($p < 0.05$) in the estimation of the parameters (Table 5 S for TPC and Table 6 S for DPPH), the odds ratio of one treatment being greater than the other was calculated for the analyses TPC (Eq. 4) and DPPH (Eq. 5), respectively.

$$\begin{aligned} logit\,(\hat{\pi}) =& 1.986 + 0.395xT10 + 0.284xT13 \\ &+ 0.358xT14 + 0.340xT15 + 0.250xT3 \\ &+ 0.342xT4 + 0.121xT5 + 0.160xT6 \\ &+ 0.375xT7 + 0.452xT8 + 0.281xT9 \end{aligned} \tag{4}$$

$$\begin{aligned} logit\,(\hat{\pi}) =& 2.631 + 0.105xT1 + 0.391xT10 \\ &+ 0.343xT12 + 0.301xT13 + 0.340xT14 \\ &+ 0.276xT15 + 0.102xT2 + 0.320xT3 \\ &+ 0.354xT4 + \ 0.129xT5 + 0.194xT6 \\ &+ 0.414xT7 + 0.426xT8 + 0.301xT9 \end{aligned} \tag{5}$$

In this order, the *odds ratio* was calculated and compared with the treatment that showed the lowest levels of TPC and DPPH, T12, and T11, respectively. In the TPC assay, it was revealed that treatment T8, which exhibits higher levels of TPC, has an odds ratio of 57.1% higher probability of having higher levels compared to samples of treatment T12. The nature of the solvent, the adequate extraction temperature, and the extraction time are essential for the interaction with the compounds of the plant matrix.

Regarding the other treatments, T10, T7, T14, and T4 showed odds ratios of 48.40, 45.50, 43.10, and 40.70%, respectively, to be higher compared to T12. However, they all show significant differences ($p \leq 0.05$) when compared to T8 treatment samples.

Table 1 shows that all the specified treatments had solvent mixtures ranging from 50 to 80% ethanol/water. The enhancement in extraction observed in samples from treatments T10 and T14, which used a mixture of 50% ethanol/

water, can be attributed to the influence of temperature and time factors. These factors compensate for the lower affinity of the solvent towards the phenolic compounds.

For the DPPH antioxidant activity assay, treatment T8, which has an odds ratio of 53.2% higher than T11 (0% ethanol), exhibited reduced antioxidant potential. Therefore, it is evident that the antioxidant potential of the extract increases when using a higher ethanol content in the solvent mixture, longer time, and higher temperature of the extraction process.

Other treatments that did not show significant differences compared to T8 ($p \geq 0.05$) were T7, T10, T4, T12, T14, T3, and T13, which had odds ratios of 51.3, 47.8, 42.5, 40.8, 40.4, 37.7, and 35.2%, respectively.

Treatments with a solvent concentration between 21 and 100% had higher odds ratios of greater than T11. It is evident that treatment T10 (50% (ethanol / water) showed a better antioxidant potential compared to T4 (80%/20% (ethanol/water)). This can be attributed to the combined influence of solvent concentration and temperature, where the temperature factor likely leads to a higher content of compounds with antioxidant potential in the extract.

Global response (GR) and antioxidant activity by ABTS and FRAP methods

Using GR analysis, it was determined that the treatment with the best performance in the extraction of phenolic compounds and antioxidant potential was T8 (80/20% (ethanol/water), 70 °C and 120 min). Therefore, according to the information reported in the previous sections and subjected to statistical analysis, the antioxidant activity of the extract T8 was further tested using the ABTS and FRAP procedures. In fact, the determination of antioxidant activity should be carried out using different approaches that complement each other, considering that each methodology has a specific and distinct reaction mechanism capable of accurately assessing the antioxidant activity of certain groups of compounds with different polarities, varied functional groups, and different antioxidant mechanisms (Munteanu and Apetrei 2021; Oliveira-Alves et al. 2022).

The extract T8 exhibited a significant high antioxidant potential of 33.15 ± 11.66 mmol Trolox g^{-1} and 13.47 ± 1.37 mmol Fe^{+2} g^{-1} as determined by the ABTS and FRAP assays, respectively. These results provide additional support for the findings of the present study.

The antioxidant activity values of *P. corcovadensis* are scarce in the literature, for comparative purposes. However, in a study conducted by Lizcano et al. (2010), infusions of the leaves of plants of the *Piper genus* were made by agitating them in water for 15 min and allowing them to rest for 10 min. The plants used were *Piper glandulosissimum*, *Piper krukoffii*, and *Piper putumayonse*, which had TPC levels of 10.62 ± 0.02, 16.84 ± 0.07, and 22.20 ± 0.11 mg GAE g^{-1}, respectively. These values are consistent with those obtained in the current study.

Plants of the genus *Piper* are rich in lignans, which are specialized metabolites (polyphenols) characteristic of this genus and possess antioxidant and pharmacological activities (Oliveira-Alves et al. 2022).(Parmar et al. 1997).

In a study conducted by Fan et al. 2023; the antioxidant activity of various lignans derived from plants of the *Piper genus* was investigated. The study demonstrated antioxidant activity using the ABTS and DPPH methods. These results validate the hypothesis that these compounds found in Piperaceae plants may be responsible for the observed antioxidant activity.

It is important to take into account that, in addition to the factors examined in this study (solvent mixture and polarity, temperature and extraction time), environmental factors such as location, seasonality, stress levels, and the species under study can have an impact on the amounts of total phenolic compounds (TPC).

Predictive chemometric modeling using digital images and machine learning algorithms

For the construction of predictive models, digital images (Fig. 2A for TPC and 2B for DPPH) were used.

Information obtained through the color channels was subjected to the VIF test. The following color attributes were retained for TPC (Fig. 3A): V (VIF: 6.56), G (VIF: 6.46) and R (VIF: 0.97). Similarly, for DPPH, attributes H (VIF: 13.12), V (VIF: 8.21), and S (VIF: 4.04) were retained (Fig. 3B).

A higher coefficient of determination (R^2) value, close to 1, indicates a good prediction, according to the study conducted by Andrés et al. 2007; values of $R^2 \geq 0.80$ can already be considered adequate in predictive modeling.

According to Grooten et al. (2023) the closer the RMSEC parameter is to 0, the less adjustment the model needs. For the RPD parameter, it can be interpreted in three ways: weak prediction ($RPD < 1.4$), reasonable prediction ($1.4 \geq RPD \leq 2.0$), and strong prediction ($RPD > 2.0$) (Lu et al. 2023).

For the construction of the predictive model for TPC, the Random Forest (RF) algorithm (Fig. 4A-B) and the Lazy Kstar (LKS) algorithm (Fig. 4C-D) were used.

Figure 4A depicts the model produced by the algorithm, while Fig. 4B illustrates the cross-validation process. The R^2 value was 0.99 and 0.95 for the model and validation, respectively. This suggests that the predicted and observed values are quite similar.

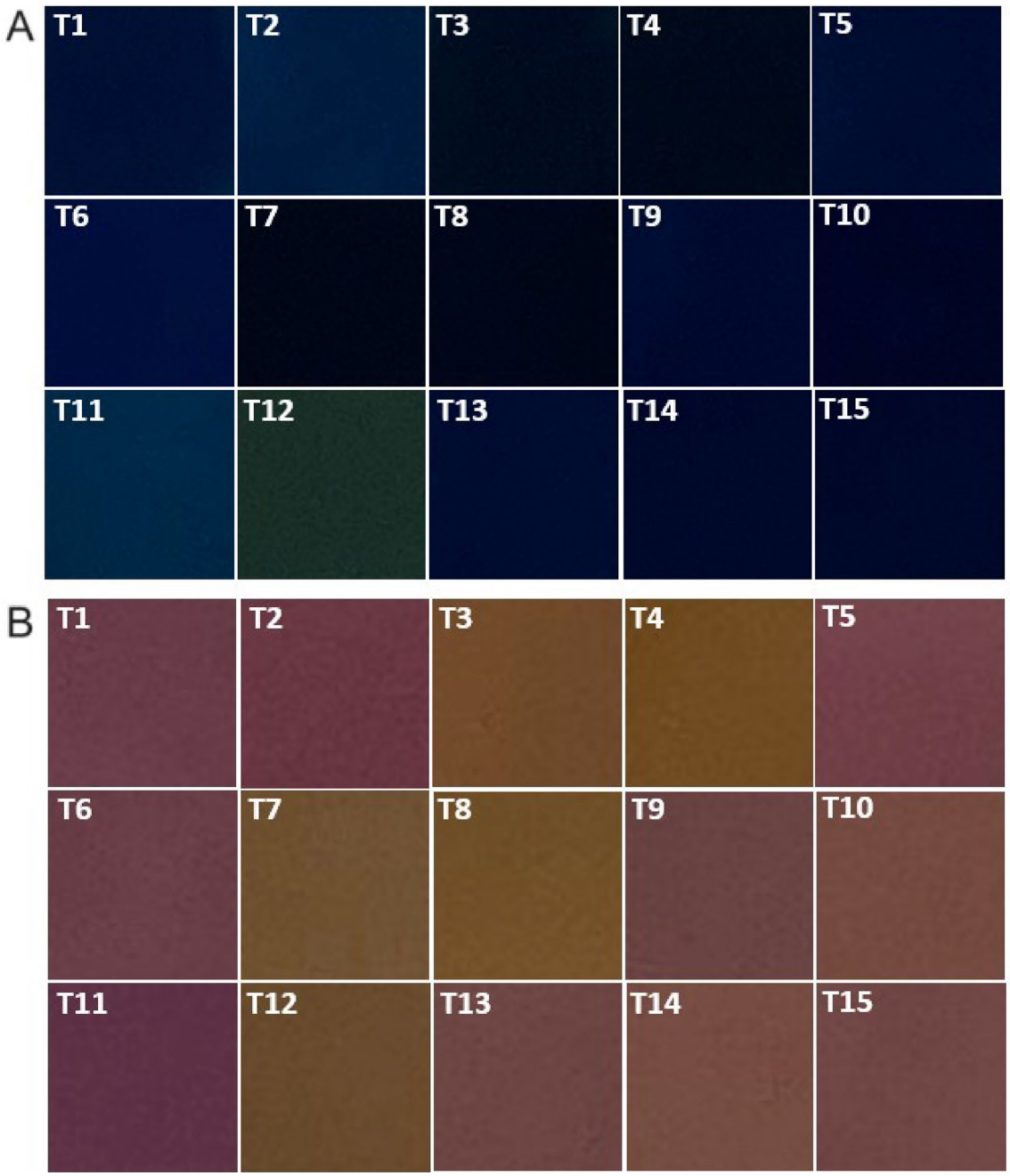


Fig. 2 Images used in the construction of the TPC (**A**) and DPPH (**B**) predictive model

Regarding the RMSEC and RPD values, the model showed a value of 0.276 for RMSEC and 5.39 for RPD, indicating that the model does not require adjustment and that the prediction can be considered adequate for this type of model.

The modelling for TPC using the LKS algorithm obtained an R^2 value of 0.97 for the model (Fig. 4C) and 0.93 for validation (Fig. 4D), values considered adequate for regression.

Regarding the RMSEC and RPD values, the model demonstrated an RMSEC value of 0.333 and an RPD value of 4.45, indicating accurate prediction with minimal adjustment requirements. Regarding validation, it was observed that the model had a lower R^2 value and a higher RMSEC value (RMSEC = 0.713), suggesting a slight need for model data adjustment. However, the RPD score of 2.56, suggests that the prediction is sufficiently accurate.

Predictive modeling of antioxidant activity using the DPPH method was constructed using the RF algorithm. For the model (Fig. 4E), the R^2 value was 0.99, the RMSEC was 0.252, and the RPD was 8.48. It can be considered a good

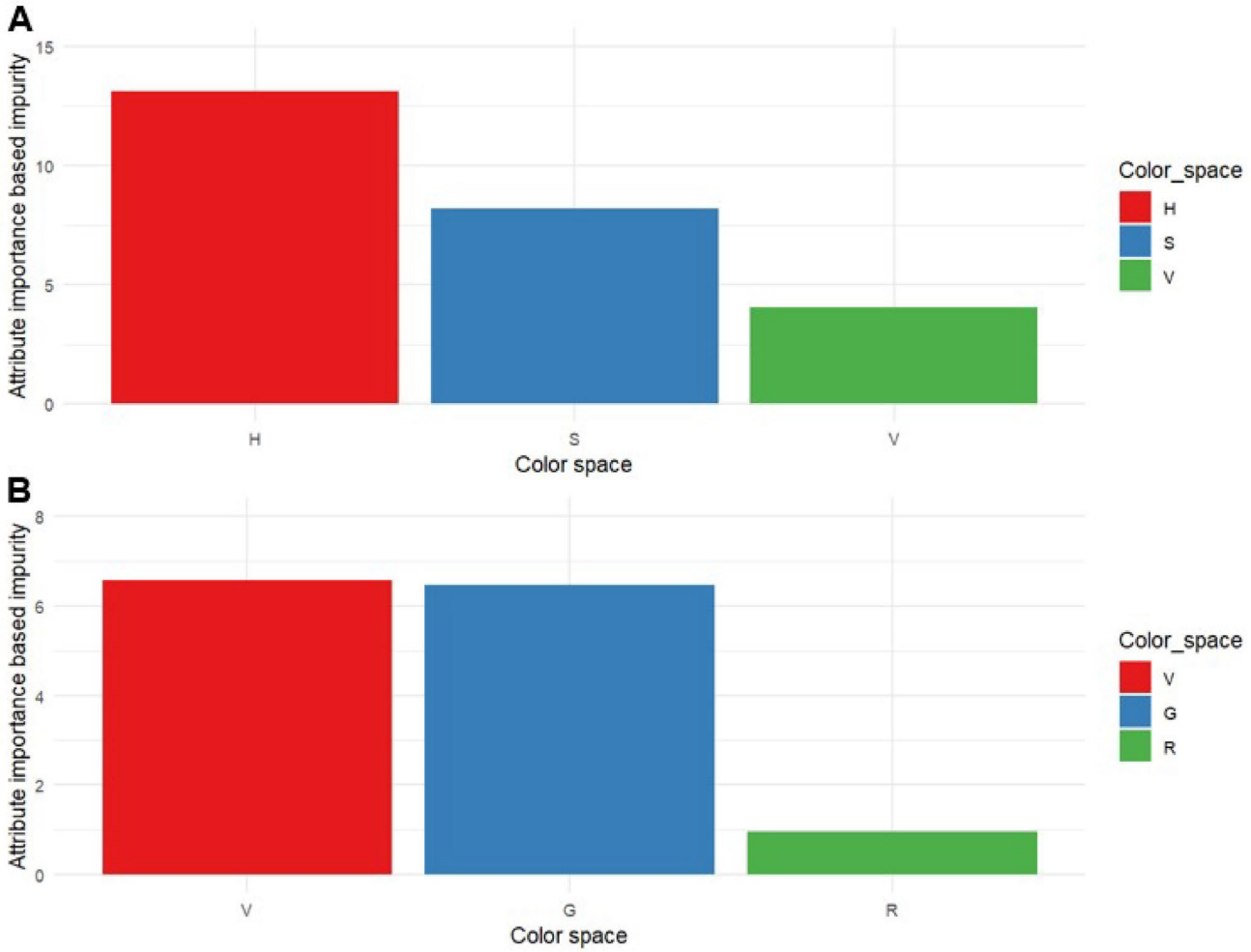


Fig. 3 **A**: Variance Inflation Factor (VIF) for TPC; **B**: Predictive Model for DPPH Test

prediction by analyzing the parameters and evaluating the quality of the model.

For the cross-validation analysis (Fig. 4F), the value of R^2 was found to be 0.98. Furthermore, the mean square error of calibration (RMSEC) and the ratio of performance to deviation (RPD) were 0.537 and 5.13, respectively. These values indicate a higher level of precision in the prediction.

Finally, it is essential to emphasize that when building predictive models, it is important to optimize the extraction process and eliminate experimental errors. It is crucial to carefully consider suitable statistical variables and use machine learning (ML) algorithms that best suit the model to ensure accurate predictions. The use of GLM and ML in data analysis and prediction of bioactive compounds is limited to specific plant species and the predetermined extraction factors and their levels. Changes in extraction factors or the use of a new plant species can invalidate the constructed models, requiring a new optimization and adjustment of parameters to ensure accurate predictions when dealing with new species.

Conclusion

The optimization of the extraction conditions of compounds with antioxidant activity from *P. corcovadensis* yielded optimal results for total phenolic compounds (TPC) and antioxidant activity using an 80/20% ethanol-water mixture, an extraction temperature of 70 °C, and an extraction time of 120 min, as determined through readily available machine learning and statistical tools.

Statistical tools employed in this study, such as GLM, have been demonstrated to be efficient in data analysis based on antioxidant activity data. The TPC levels obtained were consistent with those reported in the literature for the *Piper* genus. The current study successfully developed machine learning predictive models for TPC and DPPH,

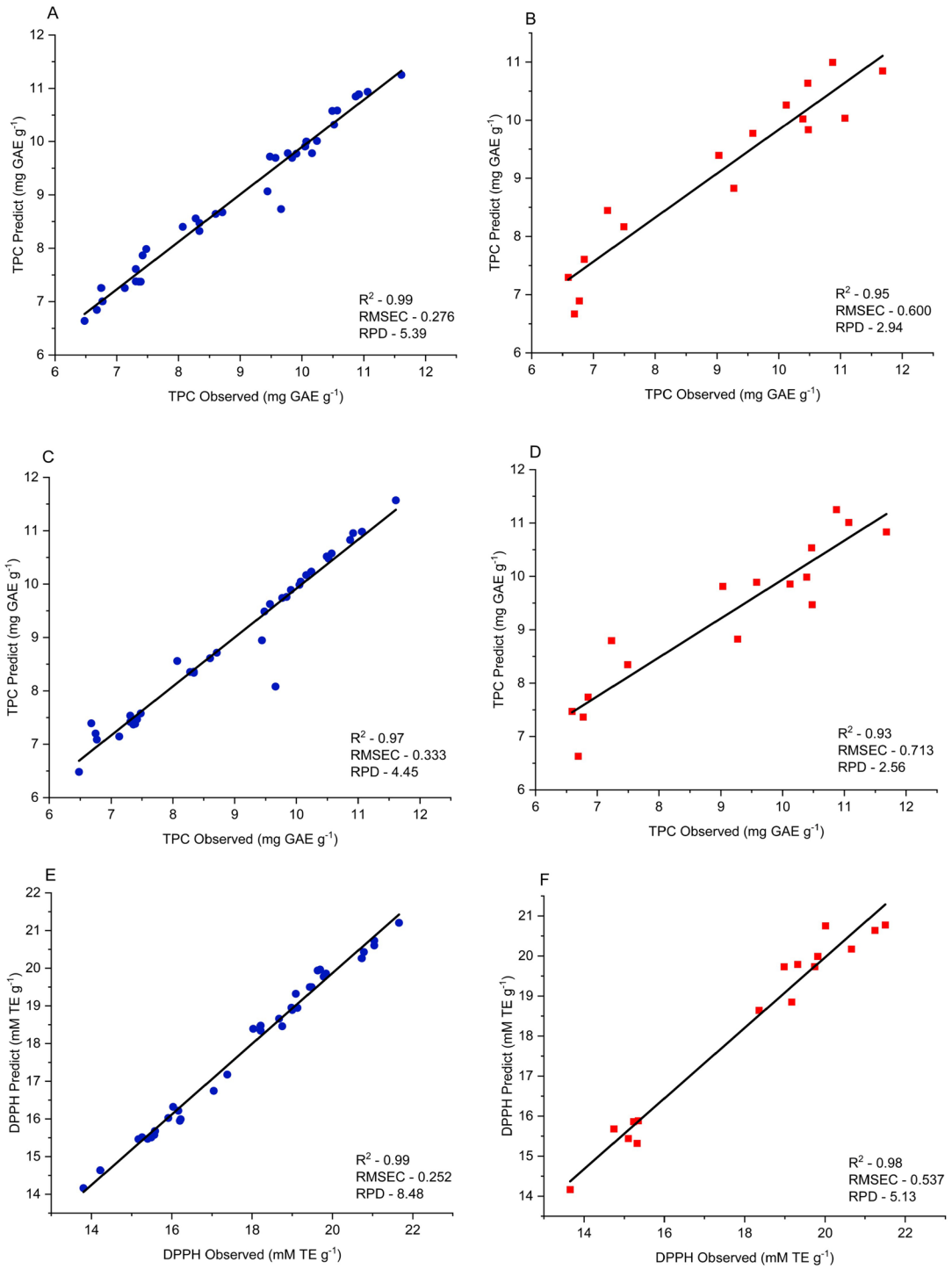


Fig. 4 Model predictive regression (**A**) and validation for TPC (**B**) using the RF algorithm (Random Forest), predictive regression (**C**) and validation (**D**) models for TPC using the LKS algorithm (Lazy Kstar) and model predictive regression (**E**) and validation (**F**) for DPPH using the RF algorithm (Random Forest)

demonstrating a strong predictive capability. Using photography to create predictive models to assess the total phenolic compounds (TPC) and the antioxidant activity (AA) content efficiently reduces both time and operational costs. This approach could emerge as a future strategy for optimizing and accelerating the determination of plant biological activities.

Furthermore, the reduction of operational costs and extraction time, combined with machine learning tools to predict compound levels, can be applied to industrial

processes, making them more accessible and sustainable. Additionally, photography as a method for analyzing and predicting phenolic compounds may represent an innovative and low-cost strategy for optimizing large-scale processes, contributing to the economic and environmental viability of natural product extractions for various industries, such as cosmetics, pharmaceuticals, and food.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13197-025-06433-6.

Acknowledgements The authors gratefully acknowledge scholarship from the Brazilian National Research Council (CNPq), the Coordination for the Improvement of Higher-Level Personnel (CAPES), and Fundação Araucária. The authors also gratefully acknowledge the Foundation for Science and Technology (FCT) through the projects UIDB/00239/2020 [CEF], UIDP/00100/2020 [CQE] (https://doi.org/10.54499/UIDP/00100/2020), UIDB/00100/2020 [CQE] (https://doi.org/10.54499/UIDB/00100/2020), LA/P/0056/2020 (https://doi.org/10.54499/LA/P/0056/2020), contract CEECIND/02725/2018.

Author contributions Bruno Henrique Fontoura: Investigation, Writing – original draft, formal analysis. Luciano de Souza Ramos: Formal analysis. Lucas Vinícius Dallacorte: Formal analysis, data curation. Michelle Fernanda Faita Rodrigues: Formal analysis. José Abramo Marchese: Supervision, resoucer, data curation. Tiago Adriano Fernandes: Supervision, resoucer, data curation. Mario Antonio Alves da Cunha: Supervision, resoucer, data curation. Vanderlei Aparecido de Lima: Supervision, resoucer, data curation. Solange Teresinha Carpes: Project administration, review and editing, supervision, resoucer, data curation.

Funding Not applicable.

Data availability All data generated or analysed during this study are included in this published article and its supplementary information files. Additional data and clarifications can be obtained upon request to the corresponding author.

Code availability Not applicable.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests Not applicable.

References

Adesina OS, Agunbiade DA, Oguntunde PE (2021) Flexible bayesian dirichlet mixtures of generalized linear mixed models for count data. Sci Afr 13:e00963. https://doi.org/10.1016/j.sciaf.2021.e00963

Alasalvar H, Yildirim Z (2021) Ultrasound-assisted extraction of antioxidant phenolic compounds from Lavandula angustifolia flowers using natural deep eutectic solvents: an experimental design approach. Sustain Chem Pharm 22:100492. https://doi.org/10.1016/j.scp.2021.100492

Andrés S, Murray I, Navajas EA, Fisher AV, Lambe NR, Bünger L (2007) Prediction of sensory characteristics of lamb meat samples by near infrared reflectance spectroscopy. Meat Sci 76:509–516. https://doi.org/10.1016/j.meatsci.2007.01.011

Carrizosa E, Galvis Restrepo M, Romero Morales D (2021) On clustering categories of categorical predictors in generalized linear models. Expert Syst Appl 182:115245. https://doi.org/10.1016/j.eswa.2021.115245

Chamali S, Bendaoud H, Bouajila J, Camy S, Saadaoui E, Condoret J-S, Romdhane M (2023) Optimization of accelerated solvent extraction of bioactive compounds from Eucalyptus intertexta using response surface methodology and evaluation of its phenolic composition and biological activities. J Appl Res Med Aromat Plants 35:100464. https://doi.org/10.1016/j.jarmap.2023.100464

Chandler RE (2020) Multisite, multivariate weather generation based on generalised linear models. Environ Model Softw 134:104867. https://doi.org/10.1016/j.envsoft.2020.104867

Chen Z, Shi NZ, Gao W (2011) Nonparametric Estimation of the log odds ratio for sparse data by kernel smoothing. Stat Probab Lett 81:1802–1807. https://doi.org/10.1016/j.spl.2011.06.017

Cheng J, Sun J, Yao K, Xu M, Cao Y (2022) A variable selection method based on mutual information and variance inflation factor. Spectrochim Acta - Part Mol Biomol Spectrosc 268:120652. https://doi.org/10.1016/j.saa.2021.120652

de Marsiglia L, Oliveira WIM, de Lucas Jacinto Almeida L, Santos R, da Silva Neto NC, Santiago JM, de Melo ÂM, Honorato BCA, da Silva FL (2023) Thermal stability of total phenolic compounds and antioxidant activities of Jaboticaba peel: effect of solvents and extraction methods. J Indian Chem Soc 100. https://doi.org/10.1016/j.jics.2023.100995

El Bouchefry K, de Souza RS (2020) Learning in Big Data: Introduction to Machine Learning, in: Škoda, P., Adam, F.B.T.-K.D. in B.D. from A. and E.O. (Eds.), Knowledge Discovery in Big Data from Astronomy and Earth Observation. Elsevier, pp. 225–249. https://doi.org/10.1016/B978-0-12-819154-5.00023-0

Fan D, Zhou C, Chen C, Li X, Ma J, Hu Y, Li G, Ruan J, Wu A, Li L, Gong X (2023) Lignans from the genus Piper L. and their Pharmacological activities: an updated review. Fitoterapia 165:105403. https://doi.org/10.1016/j.fitote.2022.105403

Fontoura BH, Perin EC, Teixeira SD, de Lima VA, Carpes ST (2023) Multivariate and machine learning approaches for prediction of antioxidant potential in Bertholletia excelsa barks. J King Saud Univ - Sci 35:102792. https://doi.org/10.1016/j.jksus.2023.102792

Fontoura BH, Perin EC, Simon AP, Bett CF, Lustosa PR, Oldoni TLC, de Lima VA, Marchese JA, Carpes ST (2024) Chemometric tools to characterize phenolic compounds with antioxidant activity of Melipona quadrifasciata propolis from Brazil. Food Anal Methods. https://doi.org/10.1007/s12161-024-02611-y

Galanakis CM, Goulas V, Tsakona S, Manganaris GA, Gekas V (2013) A knowledge base for the recovery of naturalphenols with different solvents. Int J Food Prop 16:382–396. https://doi.org/10.1080/10942912.2010.522750

Grooten Y, Mangelings D, Vander Heyden Y (2023) Comparison of supercritical fluid chromatographic methods to predict the skin permeability of pharmaceutical and cosmetic compounds. J Chromatogr A 1692:463855. https://doi.org/10.1016/j.chroma.2023.463855

Haida Z, Ab Ghani S, Juju Nakasha J, Hakiman M (2022) Determination of experimental domain factors of polyphenols, phenolic acids and flavonoids of lemon (Citrus limon) Peel using two-level

factorial design: determination of experimental domain factors. Saudi J Biol Sci 29:574–582. https://doi.org/10.1016/j.sjbs.2021.09.022

Hamasha MM, Ali H, Hamasha S, Ahmed A (2022) Ultra-fine transformation of data for normality. Heliyon 8:e09370. https://doi.org/10.1016/j.heliyon.2022.e09370

Henrique Fontoura B, Perin C, Paula Buratto E, Francisco Schreiner A, Menezes Cavalcante J, Dias Teixeira K, Manica S, Antônio D, Narzetti R, da Silva B, Dulce Bagatini G, Oldoni MLC, Carpes TT, S (2024) Chemical profile and biological properties of the Piper corcovadense C.DC. Essential oil. Saudi Pharm J 32:101993. https://doi.org/10.1016/j.jsps.2024.101993

Herrera JG, Ramos MP, de Lima Albuquerque BN, de Oliveira Farias de Aguiar JCR, Agra Neto AC, Guedes Paiva PM, do Amaral Ferraz Navarro DM, Pinto L (2022) Multivariate evaluation of process parameters to obtain essential oil of Piper corcovadensis using supercritical fluid extraction. Microchem J 181. https://doi.org/10.1016/j.microc.2022.107747

Imran I, Bin, Engström MT, Karonen M, Williams AR, Salminen JP (2023) Alkaline oxidization can increase the in vitro antiparasitic activity of proanthocyanidin-rich plant extracts against Ascaris suum. Exp Parasitol 248. https://doi.org/10.1016/j.exppara.2023.108493

Lizcano LJ, Bakkali F, Begoña Ruiz-Larrea M, Ruiz-Sanz IJ (2010) Antioxidant activity and polyphenol content of aqueous extracts from Colombian Amazonian plants with medicinal use. Food Chem 119:1566–1570. https://doi.org/10.1016/j.foodchem.2009.09.043

Lu Q, Tian S, Wei L (2023) Digital mapping of soil pH and carbonates at the European scale using environmental variables and machine learning. Sci Total Environ 856:159171. https://doi.org/10.1016/j.scitotenv.2022.159171

Mehdizadeh L, Moghaddam M, Ganjeali A, Rahimmalek M (2024) Phenolic compounds, enzymatic and non-enzymatic antioxidant activities of mentha Piperita modified by zinc and Methyl jasmonate concentrations. Sci Hortic (Amsterdam) 329:112980. https://doi.org/10.1016/j.scienta.2024.112980

Mudele O, Frery AC, Zanandrez LFR, Eiras AE, Gamba P (2021) Modeling dengue vector population with Earth observation data and a generalized linear model. Acta Trop 215:105809. https://doi.org/10.1016/j.actatropica.2020.105809

Munteanu IG, Apetrei C (2021) Analytical methods used in determining antioxidant activity: A review. Int J Mol Sci 22:3380. https://doi.org/10.3390/ijms22073380

Oliveira-Alves S, Lourenço S, Anjos O, Fernandes TA, Caldeira I, Catarino S, Canas S (2022) Influence of the storage in bottle on the antioxidant activities and related chemical characteristics of wine spirits aged with chestnut staves and micro-oxygenation. Molecules 27. https://doi.org/10.3390/molecules27010106

Parmar VS, Jain SC, Bisht KS, Jain R, Taneja P, Jha A, Tyagi OD, Prasad AK, Wengel J, Olsen CE, Boll PM (1997) Phytochemistry of the genus Piper. Phytochemistry 46:597–673. https://doi.org/10.1016/S0031-9422(97)00328-2

Perin EC, Fontoura BH, Lima VA, Carpes ST (2020) RGB pattern of images allows rapid and efficient prediction of antioxidant potential in Calycophyllum Spruceanum barks. Arab J Chem 13:7104–7114. https://doi.org/10.1016/j.arabjc.2020.07.015

Re R, Pellegrini N, Proteggente A, Pannala A, Yang M, Rice-Evans C (1999) Antioxidant activity applying an improved ABTS radical cation decolorization assay. Free Radic E Med 26:51. https://doi.org/10.1016/S0891-5849(98)00315-3

Salim A, Deiana P, Fancello F, Molinu MG, Santona M, Zara S (2023) Antimicrobial and antibiofilm activities of pomegranate Peel phenolic compounds: varietal screening through a multivariate approach. J Bioresour Bioprod 8:146–161. https://doi.org/10.1016/j.jobab.2023.01.006

Shi S, Huang Z, Gu X, Lin X, Zhong C, Hang J, Lin J, Zhong CC, Zhang L, Li Y, Huang J (2024) Aids Nat Prod Anal 505–522. https://doi.org/10.1007/s42250-024-01154-3. From 2015 to 2023: How Machine Learning

Singleton VL, Orthofer R, Lamuela-Raventós RM (1999) Analysis of total phenols and other oxidation substrates and antioxidants by means of folin-ciocalteu reagent, in: Lipids. pp. 152–178. https://doi.org/10.1016/S0076-6879(99)99017-1

Sircar A, Yadav K, Rayavarapu K, Bist N, Oza H (2021) Application of machine learning and artificial intelligence in oil and gas industry. Pet Res 6:379–391. https://doi.org/10.1016/j.ptlrs.2021.05.009

Song X, Zhang R, Xie T, Wang S, Cao J (2019) Deep eutectic solvent micro-Functionalized graphene assisted dispersive micro Solid-Phase extraction of pyrethroid insecticides in natural products. Front Chem 7:1–10. https://doi.org/10.3389/fchem.2019.00594

Stankevičius M, Maruška A, Jakobsone I, Akuneca I (2010) Analysis of phenolic compounds and radical scavenging activities of spice plants extracts 85–91

Sulejmanović M, Milić N, Mourtzinos I, Nastić N, Kyriakoudi A, Drljača J, Vidović S (2024) Ultrasound-assisted and subcritical water extraction techniques for maximal recovery of phenolic compounds from raw ginger herbal dust toward in vitro biological activity investigation. Food Chem 437. https://doi.org/10.1016/j.foodchem.2023.137774

Vergara-Salinas JR, Pérez-Jiménez J, Torres JL, Agosin E, Pérez-Correa JR (2012) Effects of temperature and time on polyphenolic content and antioxidant activity in the pressurized hot water extraction of deodorized thyme (Thymus vulgaris). J Agric Food Chem 60:10920–10929. https://doi.org/10.1021/jf3027759

Virág L, Egedy A, Varga C, Erdős G, Berezvai S, Kovács L, Ulbert Z (2024) Determination of the most significant rubber components influencing the hardness of natural rubber (NR) using various statistical methods. Heliyon 10. https://doi.org/10.1016/j.heliyon.2024.e25170

Wang Y, Wei W, Zhang Y, Hanson RK (2021) A new strategy of characterizing hydrocarbon fuels using FTIR spectra and generalized linear model with grouped-lasso regularization. Fuel 287:119419. https://doi.org/10.1016/j.fuel.2020.119419

Yao L, Wu X, Jiang X, Shan M, Zhang Z, Li, Yiting, Yang A, Li Y, Yang C (2023) Subcellular compartmentalization in the biosynthesis and engineering of plant natural products. Biotechnol Adv 69. https://doi.org/10.1016/j.biotechadv.2023.108258

Yusoff IM, Mat Taher Z, Rahmat Z, Chua LS (2022) A review of ultrasound-assisted extraction for plant bioactive compounds: phenolics, flavonoids, thymols, saponins and proteins. Food Res Int 157:111268. https://doi.org/10.1016/j.foodres.2022.111268

Zadra M, de Menezes BB, Frescura LM, Essi L, Amaro de Carvalho C, Barcellos da Rosa M (2023) Ruellia angustiflora (Nees) Lindau ex rambo: extraction and characterization of phenolic compounds and evaluation of antiradical, photoprotective and antimicrobial activities. Nat Prod Res 0:1–9. https://doi.org/10.1080/14786419.2023.2244124

Zeng F, Chen M, Yang S, Li R, Lu X, Zhang L, Chen T, Peng S, Zhou W, Li J (2024) Distribution profiles of phenolic compounds in a cultivar of Wampee (Clausena lansium (Lour.) Skeels) fruits and in vitro anti-inflammatory activity. J Ethnopharmacol 319:117168. https://doi.org/10.1016/j.jep.2023.117168