COMMUNICATION

Prediction of antioxidant activity and classification of teas using Artificial Intelligence methods.

Caroline SOLON [a] , Marjolaine FAUVRE[a]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Abstract

Green, black, and white teas all originate from the same leaf; however, the level of oxidation that occurs as the leaf dries varies, giving each type of tea its unique characteristics. For the first time, we have set up a continuous HPLC-DPPH system with DADs that will measure at 230nm and 517 nm the areas before and after reaction with DPPH of 61 samples. We will classify the teas using neural networks and predict antioxidant activity using linear regression of standards and Trolox. We obtained a classification accuracy of 0.92 and correctly predicted 10 out of 11 values on the test set. In addition, we obtained very diverse ECG concentration values within all the teas, ranging from around 0 to 700 ppm and Trolox equivalents from around 1000 to 6000 ppm.

1. Introduction

The aim of this project is to use and compare several artificial intelligence methods to predict the nature (green, black, white, etc.), antioxidant power and origin of unknown teas. Plusieurs molécules contribuent au pouvoir antioxydant du thé. Several molecules contribute to the antioxidant power of tea. In order to collect sufficient data quickly and autonomously, we set up an on-line set-up to first separate the molecules of interest and then react them with DPPH.

Antioxidants in tea come mainly from phenolic compounds, such as catechins (-)-epigallocatechin (EGC), epigallocatechin gallate (EGCG) and (-)-epicatechin gallate (ECG), (-)-gallocatechin (GC), gallocatechingallate (GCG); gallotannins such as 3-galloyquinic acid and 1;2;6 trigalloylglucose; xanthines such as caffeine and theobromine and other polyphenolics such as Procyanidin Dimer **[1] [3]**. They are all originally derived from theaflavins, which are natural by-products of the metabolism of the tea plant (Camellia sinensis), playing several key roles, both for the plant itself and for human health **[4][6][7]**.

DPPH (2,2-diphenyl-1-picrylhydrazyl) is a chemical compound known to be a stable free radical at room temperature **[1][6].** The DPPH test is a rapid, simple and effective method for assessing the antioxidant activity of substances **[1][3][6][9].** The principle of the test is based on the reduction of the violet-colored DPPH radical to a pale yellow non-radical form (DPPH-H) when it reacts with an antioxidant capable of donating a hydrogen atom. This reaction leads to a color change that can be observed visually and quantified by spectrophotometry at a wavelength of 517 nm.

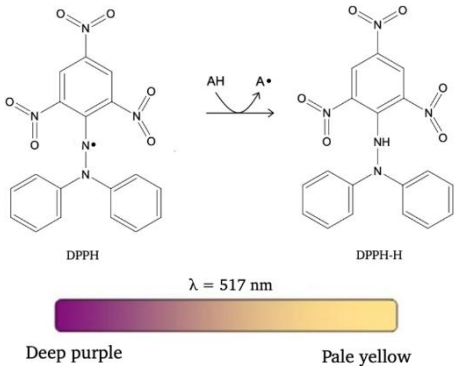

Figure 1: DPPH test reaction

What makes DPPH particularly useful is its stability as a free radical, allowing repeatable and reliable measurements without the need to generate free radicals during the experiment.

[a] *Ecole Européenne de Chimie, Polymères et Matériaux de Strasbourg, Laboratoire de chimie analytique 25 rue Becquerel, F-67076 Strasbourg, France*

Then, Trolox, also known as 6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid, is a water-soluble derivative of vitamin E that occurs naturally in the body and is commonly used as a reference standard in antioxidant activity tests **[7]**. It shares similar antioxidant properties to vitamin E, but its solubility in water makes it the tool of choice for assessing the antioxidant capacity of hydrophilic systems.

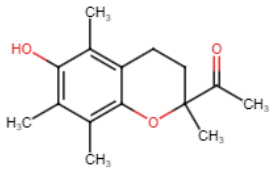

Figure 2: Trolox molecule

Its usefulness lies in the fact that it acts as an antioxidant by neutralizing free radicals, which prevents the oxidation of molecules susceptible to damage, such as lipids, proteins, and DNA. This action protects cells against oxidative stress, causing an imbalance between the production of free radicals and the body's ability to neutralize them, which is associated with various chronic diseases and ageing **[7]**.

It is therefore used as a standard in various methods for measuring antioxidant activity, such as the ORAC (Oxygen Radical Absorbance Capacity) test and the DPPH (2,2-diphenyl-1-picrylhydrazyl) test, which assess the capacity of antioxidants to absorb or neutralize free radicals **[3]**. This standard allows us to obtain a global reference value that can be compared with any sample thanks to the fact that a 'Trolox equivalents' value is associated with each antioxidant activity measurement, thus allowing a standardized and uniform assessment of the antioxidant capacities of the different compounds and extracts tested.

We carried out off-line tests before being able to implement our protocol enabling us to carry out the analyses and the reaction with DPPH automatically.

The scavenging capacity is defined as as the concentration of Trolox (mM) having the same activity as 1 mM of the test compound **[1]** and can be calculated as:

$$\boldsymbol{TEAC} = (\mathbf{1} - \boldsymbol{As}/\boldsymbol{Ac}) \times \mathbf{100}$$

where Ac is the absorbance of the control and As
is the absorbance of the tested sample after to have passed into the capillary.

We could further calculate the theoretical Trolox equivalent concentration **[1]** :

$$\boldsymbol{Ctrolox}(\boldsymbol{\mu M}) = \boldsymbol{content}(\boldsymbol{\mu M}) \times \boldsymbol{TEAC}$$

It is also possible to estimate the C50 values (concentration of samples required to scavenge 50% of DPPH radicals).

2.Materials and methods

2.1 References of products

The Trolox used came from the supplier TCI with a purity greater than 98%. We used DPPH from the same supplier with a purity greater than 97%. Also from TCI, we used epicatechin with a purity greater than 97%. From the same supplier with a purity of over 98%, we used epicatechin gallate and EGCG. We took theobromine from the supplier Fluka with a purity greater than 98%. The caffeine came from Sigma-Aldrich, whose purity was not specified on the packaging.

2.2. Preparation of tea samples:

We were able to collect various tea samples thanks to a collection as well as the remaining samples recovered from the experiments carried out last year at the ECPM on a similar subject **[2]**.

On this project, we wanted to innovate by collecting samples of white tea and infusions to see if the algorithm could classify them in a category of their own. We collected 29 green teas, 22 black teas, 10 white teas.

We began by weighing out 0,5g of each sample before grinding it thoroughly using a mortar and pestle. We infused the tea for 30 minutes at 90°C, while maintaining constant stirring. We used a syringe, we passed the tea through an Agilent cartridge retaining particles larger than 0.45µm. The tea samples were then stored in amber vials. We chose to use amber vials because, when using certain antioxidant molecule standards contained in the tea, we had noticed degradation following exposure to light and especially with epicatechin. In order not to take any risks, even though the tea did not seem to degrade, we still opted for amber vials.

2.3. Preparation of standard solutions:

Our aim was to determine the antioxidant power of the teas studied. To do this, we prepared standard solutions of several compounds such as theobromine, ECG, EGCG, caffeine, epicatechine **(figure 3)** and Trolox that had already been identified in the study of Yuanting Zhang et al **[1]**. We had to prepare standard ranges of the following products: caffeine, theobromine, Trolox, ECG and epicatechin. We prepared standard with concentrations between 160 ppm and 1000 ppm dissolved in distilled water, as this is the solvent in which the teas were infused.

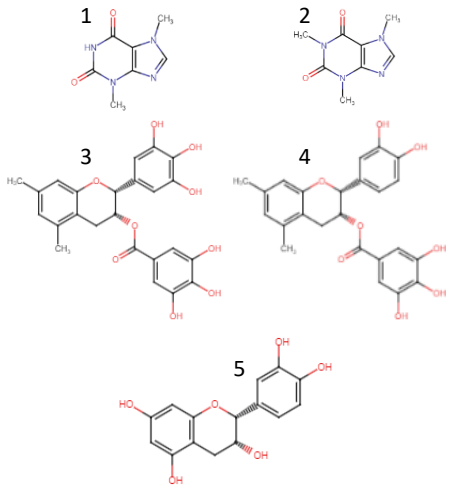

Figure 3 - Molecule of the standards studied (1 : theobromine ; 2 : caffeine ; 3 : epigallocatechin gallate ; 4 : epicatechin gallate ; 5 : epicatechin)

2.4. Preparation of DPPH solution:

The 40 ppm for the offline one and 24 ppm for in-line DPPH solutions were prepared in methanol as this is one of the two solvents used with HPLC. This concentration were chosen in agreement with the publication [1].

2.5. Conditioning and degradation of reagents

DPPH should be kept refrigerated for reasons of free radical stability. The free radical is stable at room temperature, but its reactivity can be affected by temperature, light and exposure to air. In addition, refrigeration minimizes these effects by slowing down oxidation reactions, which is crucial for guaranteeing the reliability and reproducibility of antioxidant activity measurements; subsequently, refrigeration slows down the chemical degradation of DPPH, which can be accelerated by high temperatures; and finally, at lower temperatures, undesirable parasitic reactions, which could otherwise influence the reactivity of DPPH with the antioxidants to be tested, are minimized.

We tested the degradability of the antioxidants by comparing their intensity initially, 2 days later and 7 days later. Generally speaking, we observed no difference in intensity, as there was no exchange with the ambient air, which could react with them, given that they were in vials. However, some antioxidants such as epicatechin reacted with light when we used transparent vials. That's why, we need to pay attention to their packaging, which should be kept away from light and in an airtight container.

As far as tea is concerned, its intensity value does not deteriorate, but it does deteriorate biologically with the appearance of particles. So we don't prepare all the teas in advance.

2.6. Off-line HPLC analysis:

Time (min)	Water (%)	MeOH (%)
0	95	5
5	60	40
10	60	40
12	50	50
15	50	50
20	30	70
22	30	70
23	5	95
29	5	95
30	95	5
35	95	5

Figure 4 : Elution gradient used for tea analysis.

Then, in order to be able to compare with the trolox equivalents, we carried out an off-line calibration of the Trolox by passing it through the UV-visible spectrometer.

We initially chose a 1:2 ratio of tea diluted with DPPH as suggested in publication x, in the dark for 30 min. However, we saw visually that after 30 min, the solution was not completely pale yellow, meaning that not all the DPPH had had time to react. We therefore opted for a 1:1 ratio, which allowed all the DPPH to react without altering the retention time in the dark, which was already long enough by analysis.

For the dilution, we tested different dilutions and finally opted for a dilution by 400 which would allow us to get closer to the points included in the calibration line of the trolox during its passage through the UV-visible spectrometer.

2.7. HPLC-DPPH on-line analysis:

In order to be able to determine the antioxidant power and to avoid having to carry out the reaction with DPPH ourselves, we set up a set-up shown in **Figure 5**, allowing us to automate the analyses. We had to use solvents that did not contain TFA. After a quick reaction test between TFA and DPPH, we noticed that the solution was turning yellow, which would have interfered with the reaction with the tea antioxidants. We therefore used solvent A, which was water, and solvent B, which was methanol. We obtained satisfactory results with the gradient shown in **Figure 4**. First, the tea sample was separated on the same column as above. After passing through the first DAD detecting at 230nm, the sample was passed through an Agilent capillary measuring 0.25x5000mm. Thanks to a T-junction, this capillary was also fed by a SpectraSYSTEM P1000XR pump continuously dispensing DPPH at a concentration of 6.10^{-5} mol/L at a flow rate of 0.800mL/min. After passing through the capillary, where the antioxidant compounds in the tea had time to react, a second DAD detecting at 517nm.

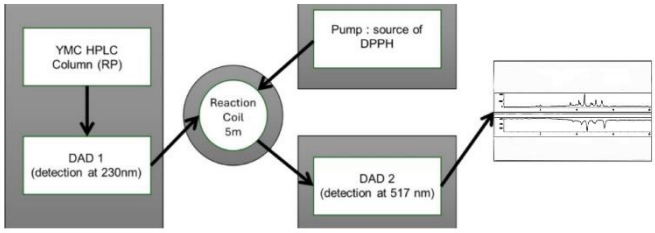

Figure 5 : On-line assembly.

2.8. Method for calculating antioxidant activity

After collecting the data at 230nm before reaction, i.e. after DAD1, at 517nm after DAD2 and at 230nm after DAD2, we will first have to identify each standard among the many peaks.

Then, we select only the areas under the curve of these areas at 230 nm in order to quantify them and at 517 nm in order to be able to know which are really antioxidants and which will allow us to calculate the antioxidant activity. Hypothetically, the peaks corresponding to the antioxidants would not appear in the chromatogram at 230 nm in DAD2, having normally reacted completely with DPPH. If this were not the case, the peaks at 517nm and 230nm after DAD2 would have to be summed to find the full peak of those at 230nm after DAD1.

Next, we will need to quantify the antioxidants in the tea at 230 using the calibration lines for the standards, with concentrations in ppm and the areas under the curve of the chromatograms at 230 nm after DAD1 and that of the chromatogram at 517nm after DAD2.

Next, we need to make an online calibration line for Trolox at 517nm or at 230 nm of the concentration in ppm as a function of the area.

Once we have quantified each sample present in the tea, we can use this line to determine the equivalent concentration in Trolox.

2.9. Visualisation and processing of HPLC results

We used the Agilent ChemStation B.04.03-SP1 [87] for LC & LC/MS Systems software developed by Agilent Technologies.

We exported the data in CSV and JPG format, at 230nm with DAD1 before reaction, at 517nm after reaction and at 230nm after reaction in order to see what did not react.

2.10. Libraries and packages Python

In order to predict the nature and antioxidant power of teas, we produced and compared several predictive models with different approaches and input variables. We carried out our work on Google Colab in order to facilitate collaborative work. The language used was Python version 3.10. We used a number of libraries that enabled us to develop models and visualize them in a user-friendly way. To process the data, we used pandas version 2.0.3, numpy version 1.25.2 and pillow (PIL) version 9.4.0. For visualization, we used version 3.7.1 of the matplotlib library. For prediction models, we used tensorflow version 2.15.0 and scikit-learn version 1.2.2.

Initially we tried to test several models using chromatogram images as input. We used a random forest model, an SVM and a neural network. We then tested a new neural network by giving it as input the csv files containing each apparent point on the chromatograms.

2.11. Artificial Intelligence methods

For classification and prediction, Machine learning methods such as random Forrest and neural networks can be used to address this subject.

Whatever the machine learning method, the data is generally split into a training set and a test set using. We then create the model on the training set with the highest ratio by making it fit. We then apply this model to our test set and perform a cross-validation, which is a technique designed to assess a model's ability to generalise to a dataset by testing it on several subsets of a dataset. Finally, we can calculate metrics such as the Root Mean Square Values Error to check the reliability and quality of our model.

For classification, we can also use the PCA (Principal component analysis) method to reduce the dimensions and SVC (Support Vector Classifier) which is supervised learning in which the labels or classes of the training data are known, unlike clustering. The objective is to find a hyperplane (or several in the case of multiclass classifications) in a multidimensional space which separates the different classes as well as possible.

2.11.1. Images as Input

For the first method with the images, we had to start by importing them in .jpg format into our environment. We resized each image so that they all had the same format and converted them into a numpy table. The data was labelled 'green', 'black', 'white' and 'infusions'. We trained and tested a random forest model, an SVM model, a model using K-nearest-neighbors and a neural network.

For the first method with the images, we had to start by importing them in .jpg format into our environment. We resized each image so that they all had the same format and converted them into a numpy table. The data was labelled 'green', 'black', 'white' and 'infusions'. We trained and tested a random forest model, an SVM model, a model using K-nearest-neighbors and a neural network.

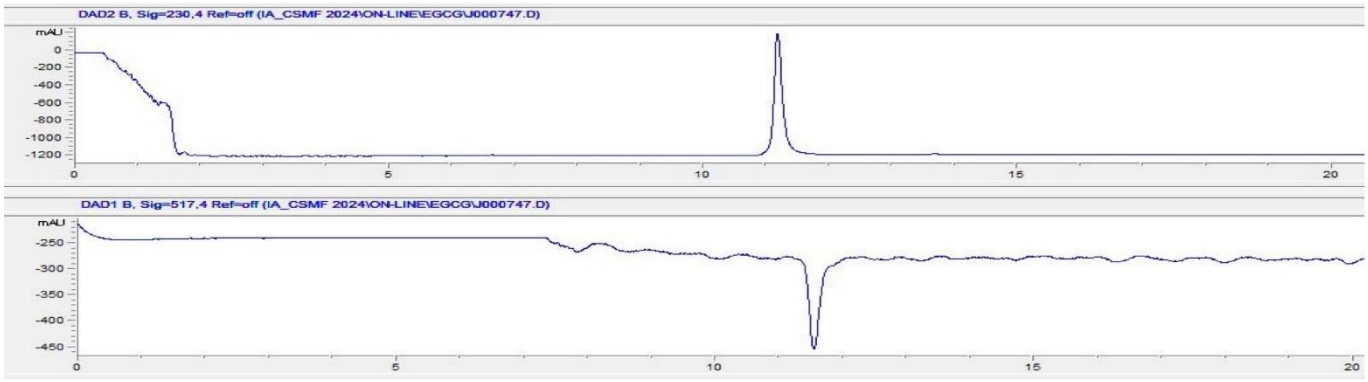

Figure 6: Chromatogram of EGCG at 230nm and 517 nm

2.11.2. Numerical values as Input

For the second method, we created a dictionary containing the 61 dataframes containing the chromatogram point data. Here we only looked at green, black and white tea, leaving out infusions to obtain a more powerful model. We assigned each dataframe a label of "0", "1" or "2", each number being associated with a green, black or white color respectively. We focused here on the use of a neural network.

3.Results and discussion

3.1. Method validation

The calibration curve parameters, encompassing both regression equations and correlation coefficients (R), exhibit excellent linearity with values exceeding 0.999 across the designated analytical ranges. The detection limits, quantified as the limits of detection (LODs) and limits of quantification (LOQs), were determined to lie within the range of 0.12 to 0.97 µg/mL and 0.27 to 3,02 µg/mL, respectively. Furthermore, the methodological precision, as demonstrated by the relative standard deviations (RSDs) for precision, repeatability, and stability assessments, was consistently maintained below 3.8%. These findings substantiate the analytical system's robustness and its suitability for the quantitative analysis of tea samples.

3.2. Processing of tea standards and samples

We decided to test our set-up with EGCG in order to focus on coordinating the reaction times of a single compound.

We observed only a very slight time lag, demonstrating a good distribution of flow rates between the HPLC outlet and the pump.

So, we put green, black and white teas and infusions through the line, supplying DPPH when it was needed so as not to leave the vacuum pump running, which could damage it, as well as A and B solvents.

Chromatograms generally have the appearance shown in **figure 7**:

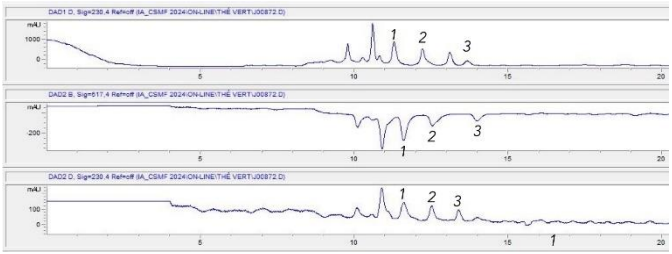

Figure 7 : Example of a chromatogram (1 : EGCG ; 2 : epicatechine ; 3 : ECG).

The main difference between the different types of samples will either be in the diversity and intensity of their compounds.

3.3. Calibration line for Trolox at 517 nm online

The Trolox calibration line was well usable at 517 nm as shown below but we have drawn it for 230 nm too:

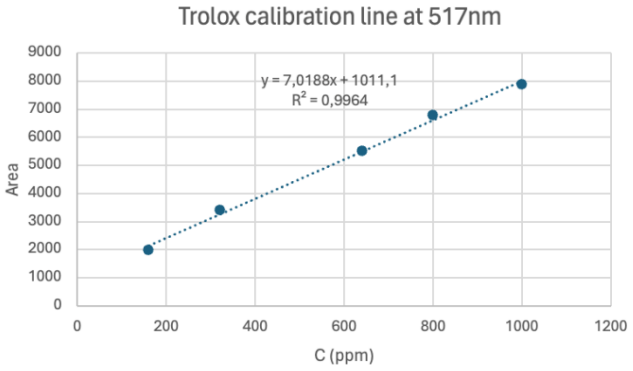

Figure 8 : Trolox calibration line

3.4. Tea classification

3.4.1. Neural Networks

3.4.1.1. Images as Input

The only model giving a potentially usable result was the neural network model. We obtained an accuracy greater than 0.90. However, we will not use this method because the images used contained noise and the integration values were present, which can lead the model into error.

3.4.1.2. Numerical values as Input

The method using csv files containing all the points present on the chromatograms is quite efficient and much more reliable than the method using images. We used neural networks build as follow: 3 Dense layers with 64 neurons and an activation function “relu” ; 1 Dense layer with an activation function “softmax”.

The weight allocated to each neuron is adjusted in the different layers of the network. The layer performs a linear combination of the inputs with the associated weights and then applies a non-linear activation function to the result. The purpose of an activation function is to introduce non-linearity into the network. Here the ReLu layer is used for deep learning because it is simple and efficient. We ran the training over 20 epochs using the "adam" optimizer. These parameters enabled us to achieve an accuracy of 0.92. The data set was split randomly into a training set (80%) and a test set (20%). We can see the predictions and mistakes of the model on the confusion matrix in ***Figure 9***.

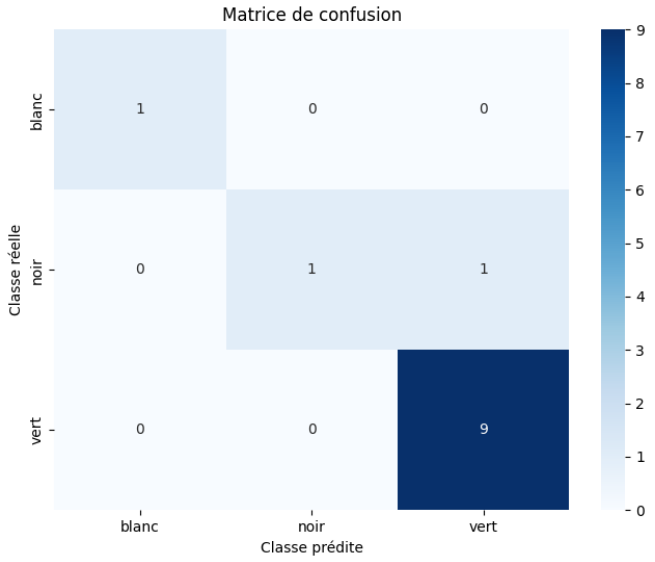

Figure 9 - Confusion matrix

We found that the only errors made were in predicting green tea as white and green tea as black. The performance of this model is very satisfactory.

3.5. Prediction of antioxidant activity

As explained above, we create a csv with the integrations at 230 nm or 517 nm (here 230nm because the results are more interpretable) of the peaks of interest, i.e. EGCG, Epicatechine and and ECG. Then, using the regression equations for the standards and Trolox, we can calculate the concentration of each standard and its Trolox equivalent.

We cannot generalise about the ECG content in the different types of tea, which could differentiate them, as the sample is very diverse. Nonetheless, we can assume that there are lower levels in the green samples, but these are the ones with the lowest concentrations at 0. There are therefore more green teas with a detectable concentration but lower levels on average than black and white teas.

3.6. Analyzing Experimental Limitations and Paths to Enhancement

From a purely analytical chemistry point of view, on some of our chromatograms we obtained noise after DAD1 which may be due to a difference in concentration in the HPLC solvent bottles or to a faulty DAD lamp. In addition, we observed a lot of noise on the chromatograms with DPPH at 517 nm, probably due to irregular pump flow rates. This is due to wear and tear on the pump and pump uncertainty. In addition, we did not observe complete consumption of the DPPH, perhaps due to the capillary being too short or too large in diameter. Then there is the uncertainty of the DPPH concentration, given that we had to prepare 500mL of solution many

	V5	V10	V14	V22	V25	V27	V30
ECG (ppm)	0	17.01	51.58	248.0	274,0	1.711	29.47
Eq. Trolox (ppm)	1011	1130	1373	2752	2934	1023	1218

N5	N10	N14	N22	B2	B4	B5	B8
0	0	0	607.1	142.3	23.30	708.4	244.9
1011	1011	1011	5272	2010	1175	5983	2730

Figure 10 - Calculation of the trolox equivalent for EGCG

times in order to keep the pump continuously fed. As a result, not all the analyses were carried out at the same DPPH concentration. Finally, the greatest uncertainty lay in the fact that negative peaks had to be manually integrated into the software.

From the point of view of artificial intelligence methods, for the classification part, the "Dense" layers of the neural network are very commonly used. Here, each layer has 64 neurons; we didn't have the opportunity to compare a large number of different networks with more complex layers because we were quickly limited by the RAM available on google Colab. Then, for the prediction part, The real difficulty was the imprecision of all the values, which explained the fact that the prediction was made with the values at 230 nm, as the peaks are automatically integrated by the software, and the lack of data on the standards.

Conclusions

The aims of this subject were to find out how DPPH reacts with antioxidants, and to be able to set up an HPLC-DPPH set-up. Then we had to classify the teas according to their nature using artificial intelligence methods such as random Forrest, neural networks, SVC and PCA according to the signal at each wavelength. Next, we were able to discover the ability to calculate antioxidant activity using Trolox and the use of regression lines on the antioxidant standards (EGCG, caffeine, theobromine, ECG, epicatechin) to calculate the concentrations of each in the tea and, using the Trolox calibration line, to find equivalents.

In the end, our classification model of neuronal network achieved an accuracy of 92%, successfully predicting 10 out of 11 values within the test dataset. Furthermore, we observed a wide range of ECG concentration values across different teas, varying from approximately 0 to 700 ppm, with the Trolox equivalents spanning from about 1000 to 6000 ppm.

Analytically, we encountered chromatographic noise post-DAD1, potentially caused by HPLC solvent concentration variances or a defective DAD lamp. Moreover, significant noise was noted in DPPH chromatograms at 517 nm, likely due to inconsistent pump flow, attributed to pump wear or uncertainty. Incomplete DPPH consumption could stem from inadequate capillary length or diameter. Repeated preparation of 500mL DPPH solutions to maintain pump supply introduced variability in DPPH concentration across analyses. Additionally, the necessity for manual integration of negative peaks in the software introduced substantial uncertainty.

In the realm of artificial intelligence, we utilized "Dense" neural network layers with 64 neurons each for classification, without exploring more complex networks due to Google Colab's RAM limitations. The prediction challenge was compounded by the imprecision of values, leading to reliance on 230 nm values for automatic software integration and a scarcity of standard data.

Conflicts of interest

There are no conflicts to declare.

Notes and references

[1]: Yuanting Zhang et al., “Evaluation of Antioxidant Activity of Ten Compounds in Different Tea Samples by Means of an On-Line HPLC–DPPH Assay,” Food Research International 53, no. 2 (2013): 847–856.
[2]: Marine Michel, Hannah Kaczorowski, “Tea recognition by HPLC and machine learning”, ECPM, (2023)
[3]: Molay K Roy et al., “ORAC and DPPH Assay Comparison to Assess Antioxidant Capacity of Tea Infusions: Relationship between Total Polyphenol and Individual Catechin Content,” International Journal of Food Sciences and Nutrition 61, no. 2 (2010): 109–124.
[4]:Zhiyong Zhang et al., “On-Line Screening of Natural Antioxidants and the Antioxidant Activity Prediction for the Extracts from Flowers of Chrysanthemum Morifolium Ramat,” Journal of Ethnopharmacology 294 (2022): 115336.
[5]:J. Ricardo Lucio-Gutiérrez et al., “Multi-Wavelength High-Performance Liquid Chromatographic Fingerprints and Chemometrics to Predict the Antioxidant Activity of Turnera Diffusa as Part of Its Quality Control,” Journal of Chromatography A 1235 (2012): 68–76.
[6]:Jyh-Horng Wu et al., “Online RP-HPLC-DPPH Screening Method for Detection of Radical-Scavenging Phytochemicals from Flowers of Acacia Confusa,” Journal of Agricultural and Food Chemistry 56, no. 2 (January 1, 2008): 328–332
[7]:Q Wei et al., “Synergistic Effect of Green Tea Polyphenols with Trolox on Free Radical-Induced Oxidative DNA Damage,” Food Chemistry 96, no. 1 (2006): 90–95
[8]:Irina I. Koleva, Harm A. G. Niederländer, and Teris A. Van Beek, “An On-Line HPLC Method for Detection of Radical Scavenging Compounds in Complex Mixtures,” Analytical Chemistry 72, no. 10 (May 1, 2000): 2323–2328
[9]:Martina Bancirova, “Comparison of the Antioxidant Capacity and the Antimicrobial Activity of Black and Green Tea,” Food Research International 43, no. 5 (2010): 1379–1382