Article

Not peer-reviewed version

# Enhancing Bioactive Compound Classification through the Synergy of Fourier-Transform Infrared Spectroscopy and Advanced Machine Learning Methods

Pedro Sampaio * and Cecília Calado

Posted Date: 2 April 2024

doi: 10.20944/preprints202404.0195.v1

Keywords: Antimicrobial; Cynara cardunculus; Machine learning; MIR-Spectroscopy; PCA; PLS-DA; SVM; KNN; BPN

 doi:10.20944/preprints202404.0195.v1



*Article*

# Enhancing Bioactive Compound Classification through the Synergy of Fourier-Transform Infrared Spectroscopy and Advanced Machine Learning Methods

**Pedro Sampaio [1,2,*] and Cecília Calado [3,4]**

[1] COPELABS - Computação e Cognição Centrada nas Pessoas, Faculty of Engineering, Lusófona University, Campo Grande, 376, 1749-024 Lisbon, Portugal
[2] GREEN-IT - BioResources for Sustainability Unit Institute of Chemical and Biological Technology António Xavier, ITQB NOVA, Av. da República, 2780-157 Oeiras, Portugal
[3] CIMOSM – Centro de Investigação em Modelação e Optimização de Sistemas Multifuncionais, ISEL, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisbon, Portugal
[4] ISEL-Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Rua Conselheiro Emídio Navarro, 1 1959-007 Lisbon, Portugal
* Correspondence: pedro.sampaio@ulusofona.pt; Tel.: 00 351 21 750 55 00

**Abstract:** Bacterial infections and resistance to antibiotic drugs represent the highest challenges to public health. The search for new and promising compounds with anti-bacterial activity is a very urgent matter. In order to promote the development of platforms enabling to the discovery of compounds with anti-bacterial activity, Fourier Transformed Mid-Infrared (FT-MIR) spectroscopy associated with the machine learning algorithms were used to predict the impact of compounds extracted from *Cynara cardunculus* against *Escherichia coli* cells. According to the plant tissue (seeds, dry and fresh leaves, and flowers) and the solvents (ethanol, methanol, acetone, ethyl acetate, and water), compounds with different compositions concerning, phenol content, antioxidant, and antimicrobial activity were obtained. Principal component analysis of spectra, allowed to discriminate compounds that inhibit the *E. coli* growth, according to the conventional assay. The supervised classification models enabled the prediction of the compound's impact on the *E. coli* growth, with the accuracy for the test data: Partial least squares-discriminant analysis of 94%; Support vector machines of 89%; k-Nearest neighbor of 72%; and a Backpropagation network of 100%. According to the promising results, the integration of FT-MIR spectroscopy with machine learning presents a high potential to promote the discovery of new compounds with antibacterial activity, thereby streamlining the drug exploratory process.

**Keywords:** antimicrobial; *Cynara cardunculus*; machine learning; MIR-spectroscopy; PCA; PLS-DA; SVM; KNN; BPN

## 1. Introduction

Due to the recent increasing resistance to antibiotics, it is urgent to develop fast and cheap systems for the discovery of bioactive compounds with antimicrobial activity and without cytotoxicity for the human host [1]. Antibiotic-resistant bacteria represent a challenging and concerning aspect of modern medicine, with factors such as the decreased development of new antibiotics and the spread of multi-drug-resistant determinants aggravating the problem. Conventional drug discovery methods are characterized by significant expenses, lengthy synthesis and testing processes, the requirement for costly equipment, and substantial demand for human resources, rendering them challenging to access and sustain [2,3]. However, the continuous development of artificial intelligence brings a new perspective to the field of antibiotic discovery [4].

 doi:10.20944/preprints202404.0195.v1



Machine learning and neural networks are revolutionizing the conventional experimental approaches to discovering new antibiotics or enhancing existing ones. These algorithms enable extensive in silico exploration and analysis, transforming the landscape of antibiotic research [4]. The extracts of certain plants are important sources of biologically active substances, being indispensable for the development and synthesis of many drugs, including antibiotics. *Cynara cardunculus* (Cardoon) has been used in traditional medicine [5], namely, extracts of rhizomes present antioxidant and antimicrobial properties [6], while the leaves have shown diuretic, choleretic, and hepatoprotective properties [7], in addition to its in vitro anti-proliferative potential against breast cancer [8,9], and cervical cancer uterus [10].

The antimicrobial activity is evaluated by expensive and time-consuming techniques that require significant amounts of samples. To overcome these limitations, spectroscopic techniques, such as Mid-Infrared (MIR) spectroscopy, present a distinctive sensitivity towards biological features, which enables the characterization of metabolic fingerprints of biological samples with high sensitivity and specificity [11]. Fourier Transform Infrared Spectroscopy (FTIR) evaluates, in detail, the different vibrational modes of the main biomolecules such as proteins, carbohydrates, lipids, and nucleic acids [12], being used, for example, in the classification of bacteria [13], evaluation of the cell cycle, discrimination of cell death mechanisms between apoptosis and necrosis [14], studies and diagnoses related to cancer [15], monitoring the effect of drugs on *Helicobacter pylori* [16,17], in addition to the characterization of cell metabolism in bioreactor cultures [18,19]. In untargeted analyses, whole FTIR spectra have been associated with specific phenotypes, e.g., to elucidate the antimicrobial effect of novel extracts [20], to distinguish the effect of various surfactants on *E. coli* cells [21], to compare the global responses of *E. coli* to diverse stress conditions with transcriptomics and FTIR [22], to develop a bioassay for toxicity testing in yeasts [23], among others. In comparison with the different omics techniques, FTIR is not as data-intensive and provides valuable information without considerable detriment to metabolic sensitivity, as seen by previous examples [24]. Given the significant amount of information present in the spectra, their analysis must be carried out using multivariate statistical methods in which the classification and discrimination methods allow the development of mathematical models capable of discriminating the samples that present the antimicrobial activity through the specific molecular profile, proving that this technique is an asset concerning the search for new pharmacological solutions from plants with unique therapeutic properties. Data mining and machine learning techniques are widely used in data analysis for various objectives such as pattern identification, and decision-making with minimal human intervention.

Machine learning (ML), considered a subgroup of Artificial Intelligence, is based on mathematical models of data that allow the computer to learn without direct instructions and identify patterns or structures in structured and unstructured data [25]. In ML two main types of techniques are used: supervised and unsupervised learning. Supervised learning methods are employed to construct training models that predict future values of data categories or continuous variables. On the other hand, unsupervised methods serve exploratory purposes, aiding in the development of models that facilitate data clustering in a manner not explicitly specified by the user. The unsupervised learning technique identifies hidden patterns or intrinsic structures in the input data and uses these to cluster data in meaningful ways [26]. A wide variety of approaches of pattern recognition techniques have been taken towards this task in the food quality evaluation such as partial least squares discriminant analysis (PLS-DA), the k-nearest neighbors (kNN), support vector machine (SVM), and artificial neural networks (ANN) such as the backpropagation network (BPN) and the counter-propagation networks, a multilayer network based on the combinations of the input, output, and clustering layers [27].

The main objective of this study was to show the advantage of the FT-MIR spectroscopy and machine learning methods to screen and, consequently, to discover the promissory bioactive molecules as antimicrobial agents. The extracts obtained from seeds, dry and fresh leaves, and flowers, extracted using different solvents were evaluated. The spectra of *E. coli* cells following rapid exposure to various compounds could predict its impact on cell growth as an alternative to conventional methods. Specific spectral bands of *E. coli* cells were analyzed using both unsupervised

 doi:10.20944/preprints202404.0195.v1



and supervised machine-learning algorithms. This approach serves as an invaluable tool in guiding the design of synthesis procedures and subsequent biological experiments, particularly in the targeted exploration for novel compounds with significant potential as antimicrobial activity.

## 2. Results and Discussion

### *2.1. Analysis of C. cardunculus Extracts*

Samples exhibiting antimicrobial activity against *E. coli* cells were extracted from various parts of the plant. Seeds were subjected to extraction using both water and ethanol, leaves were processed with methanol, dry leaves were treated with water, while flowers were extracted using ethanol and methanol (Figure 1). The extracts obtained from seeds showed a significant antioxidant activity (AA): 75% AA for methanol and 60% AA for water. Fresh leaves also demonstrated significant antioxidant activity (76% AA for methanol and 86% AA for water). The antioxidant activity then registered may be related to the specific nature of the tissues and their respective functions in terms of defense. Furthermore, extracts obtained from fresh leaves and seeds using a methanol and water mixture revealed substantial antioxidant activity.

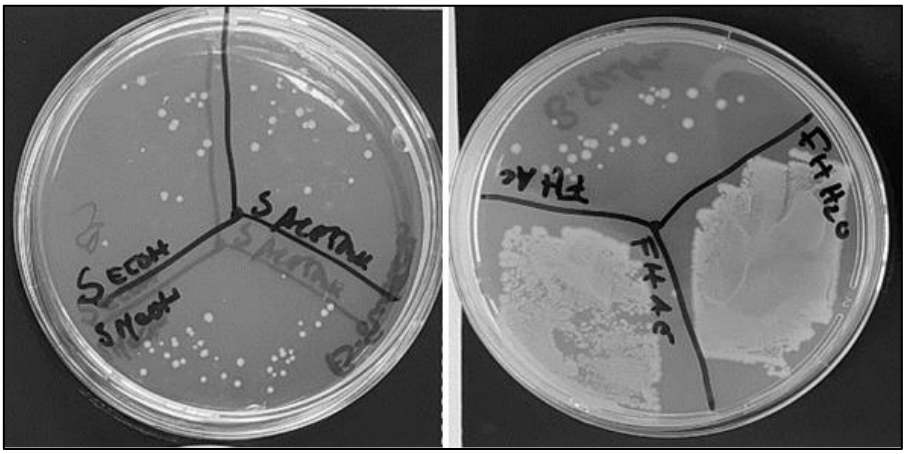

**Figure 1.** Antimicrobial activity of different assays using the *C. cardunculus* extracts in *E. coli* cells in petri dishes: seeds/Ethanol; Seeds/Methanol; Seeds/Acetone; Fresh leaves/Acetone; Fresh leaves/Ethyl acetate; and Fresh leaves/Water.

The highest phenol concentrations were registered, respectively, in the aqueous extracts from flowers and seeds (8.03 µg gallic acid equivalent (GAE)/g; 3.93 µg GAE/g) as a consequence of the extraction conditions, such as the temperature (~100 °C), following the extracts obtained by methanol and ethanol due to the affinity that exists among phenolic compounds and organic solvents. Meanwhile, the extracts obtained with ethyl acetate (0.80 µg GAE/g) and acetone (1.22 µg GAE/g) presented low values, suggesting that these solvents present low affinity for the extraction of these compounds. According to previous studies, extracts from *C. cardunculus* L. presented polyphenolic compounds, specifically flavonoids and phenolic acids. The pronounced lipophilic nature of these phenolic compounds enhances their antimicrobial activity, potentially through interactions with the cell membrane [6,8,28,29]. The polyphenols act as antioxidants due to their hydrogen-donor, and metal-chelating capacities. The antimicrobial activity observed depends on several factors. Firstly, it is influenced by the variation in the permeability of cell membranes, which can be affected by the extraction process and the specific compounds present in the samples. Additionally, changes in various intracellular functions can be induced by hydrogen bonding of phenolic compounds to enzymes, altering their activity and impacting microbial growth. Moreover, interactions with the cell membrane can lead to changes in cell wall rigidity, resulting in the loss of integrity and ultimately inhibiting microbial proliferation [30]. Phenolic compounds are the main plant antioxidants that have been associated with beneficial effects on health, which include anti-allergic, anti-atherogenic, anti-inflammatory, and antimicrobial effects, antithrombotic, cardio protectors, and vasodilators [31].

 doi:10.20944/preprints202404.0195.v1



Regarding antioxidant activity, the solvent types influence the extraction yield of different active compounds. Methanol, ethanol, and water were the most suitable solvents to extract compounds characterized by significant antioxidant activity. According to Ibrahim et al. (2015), seed extracts showed antimicrobial activity against *Escherichia coli*, *Staphylococcus aureus*, *Staphylococcus sprophyticus,* and *Klebsiella pneumoniae* species [32]. Based on the phenol content and antioxidant activity, the ethanol (0.63) and methanol (0.96) extracts, showed a significant correlation, representing a strong relationship between the phenolic compounds and the antioxidant activity, which assumes that alcoholic solvents favour the extraction of compounds with antioxidant activity. The methanolic extract exhibits significant free radical scavenging activity, indicating that it contains bioactive compounds with antioxidant properties derived from various tissues of the *C. cardunculus* species. Studies developed by Fu et al. (2011) showed that the phenolic content presented a positive correlation with its antioxidant activity [33].

### *2.2. PCA of FT-MIR Spectra*

The FT-MIR spectra of *E. coli* cells tested with different extracts were processed by different algorithms, such as the multiplicative scatter correction (MSC), straight normal variate (SNV), 1st derivative, and 2nd derivative, and its impact evaluated by principal component analysis (PCA). The most effective pre-processing methods for clustering scores in the PCA score plot according to the antimicrobial activity of extracts were found to be MSC (Multiplicative Scatter Correction) and the 2nd derivative algorithms. The 2nd derivative allowed to highlight the relevance of increasing band resolution. Interestingly, the scores were clustered according to the extract antimicrobial activity as observed by the conventional assay (based on counting colonies on agar plates) (Figure 2a,b). The clusters related to the extracts with antimicrobial activity were from seeds (extracted with water and ethanol), leaves (extracted with methanol), dry leaves (extracted with water), and flowers (extracted with ethanol and methanol). These extracts also presented a significant antioxidant activity, which may be related to the specific nature of the tissues and their respective functions in terms of defense. According to the results obtained, 14 samples tested showed a residual antimicrobial activity, while 6 samples showed significant antimicrobial activity. Meanwhile, the PCA loadings highlight spectral regions that contributed to the resolution of scores between the two clusters, characterized by high and low antimicrobial activity, especially the PC2 loading, which is defined by the following regions: 1000 to 1800 $cm^{-1}$, and 2700 to 3000 $cm^{-1}$ (Figure 2b). The spectral regions between 1500 and 1800 $cm^{-1}$ are mainly due to amide I and amide II adsorptions related to the proteins and, usually, represent the most prominent bands in a bacterial infrared spectrum since proteins tend to be, on average, half of bacterial cell composition. The peak at 1740 $cm^{-1}$ is assigned to the ester C=O stretching of the phospholipids. In general, at 1080 $cm^{-1}$ and 1240 $cm^{-1}$ there is a strong contribution of the P=O bond present in phosphate groups of molecules as nucleotides as ATP and nucleic acids, phosphorylated proteins, and lipids. Between 1000 and 1100 $cm^{-1}$ a general contribution from C–O, C–C, C–OH, C–O–C bonds present in carbohydrates occurs. The region between 2700 and 3000 $cm^{-1}$ presents a high contribution of stretching vibrations of the C-H bonds from lipids, especially from the bacteria membrane.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2024 doi:10.20944/preprints202404.0195.v1



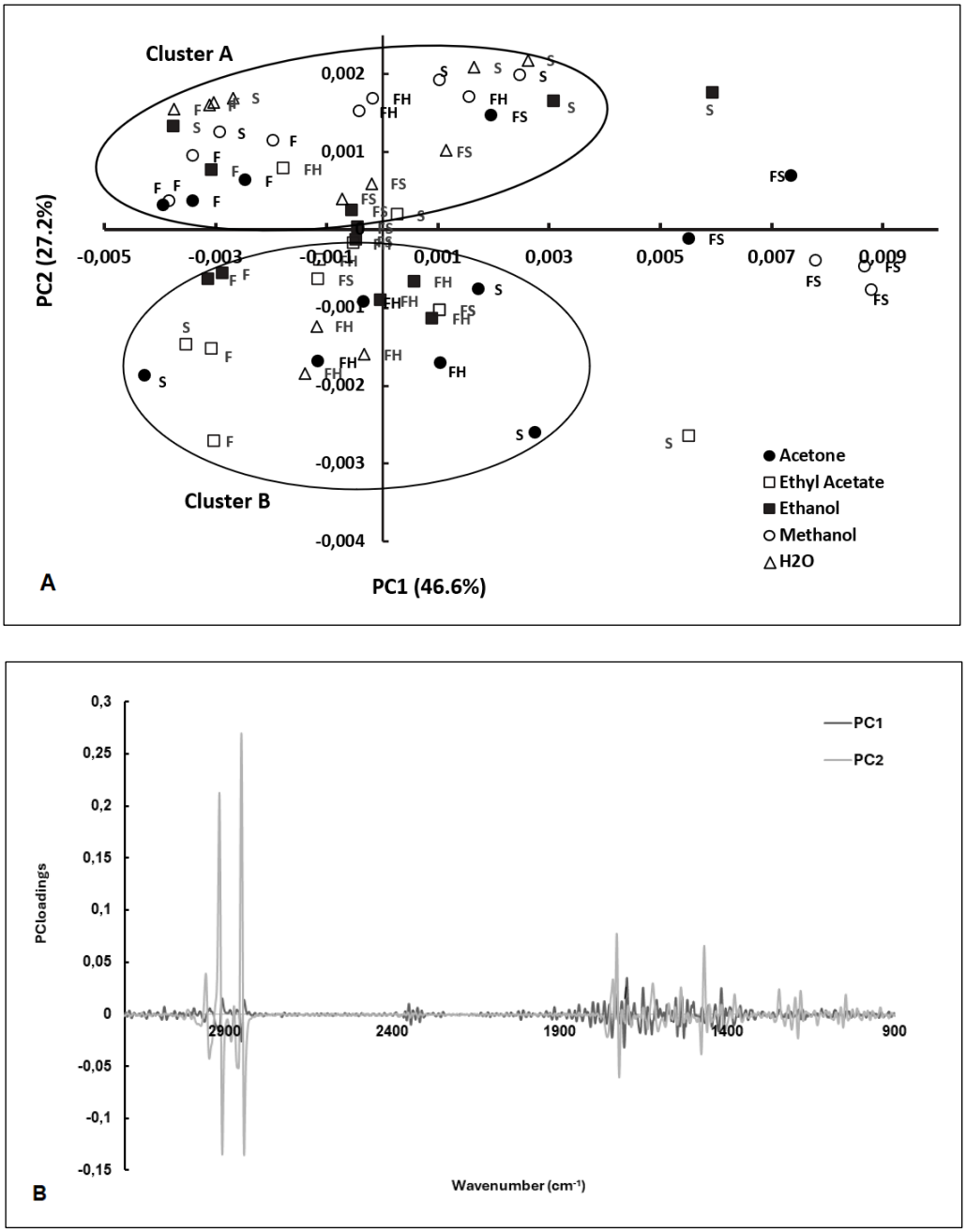


**Figure 2.** PCA of *E. coli* FT-MIR spectra tested with extracts from *C. cardunculus* tissues (S-Seeds; F-Flowers; FH-Fresh leaves; FS; Dry leaves) obtained by different solvents (AE-Ethyl acetate, EtOH-Ethanol; Ac – Acetone; MeOH - Methanol, and $H_2O$ - Water) (A); and the corresponding loading vectors (B). Spectra were pre-processed by MSC and 2nd derivative. Cluster A is defined by samples characterized by antimicrobial activity; Cluster B is constituted by samples characterized by low antimicrobial activity.

### *2.3. Analysis of Spectral Bands*

The second derivative of the spectra reveals distinct bands, seemingly indicative of variations between *E. coli* subjected to antimicrobial and low antimicrobial compound treatments (Figure 3a,b). To further identify spectral bands that exhibit significant differences between bacteria exposed to inhibitory compounds and those left untreated, several ratios between the bands, emphasized by the PCA loading vector, were examined. Twelve ratios between spectral bands were analyzed, where "A" followed by the number represents the absorbance at that wavenumber (Table 1). From the twelve ratios evaluated, the following nine ratios were statistically different at 5% significance between the two clusters: A2856/A1705, A1740/A1656, A1740/1545, A2847/1545, A1617/A1545,

 doi:10.20944/preprints202404.0195.v1



A1476/1545, 1215/1179, A1215/1545 and A1244/1230. These ratios emphasize the influence of the compounds on the cellular metabolism of the *E. coli* strain, as evidenced by changes in protein synthesis (as indicated by the ratio between amide I and II, i.e., A1617/A1545, and various ratios involving the amide I peak). Additionally, changes in the composition of the cell membrane are highlighted by the $CH_3/CH_2$ ratio (i.e., A2912/A2856), as well as ratios incorporating peaks at 1740 $cm^{-1}$ and 1179 $cm^{-1}$ corresponding to carbonyl vibrations of esters of phospholipids, and peaks at 2856 $cm^{-1}$ and 1476 $cm^{-1}$ originating from methyl and methylene groups. Figure 4 represents the box-plots for the ratios of spectral bands that were statistically different. These findings underscore the remarkable sensitivity of FT-MIR spectra in capturing the influence of compounds on *E. coli* metabolism.

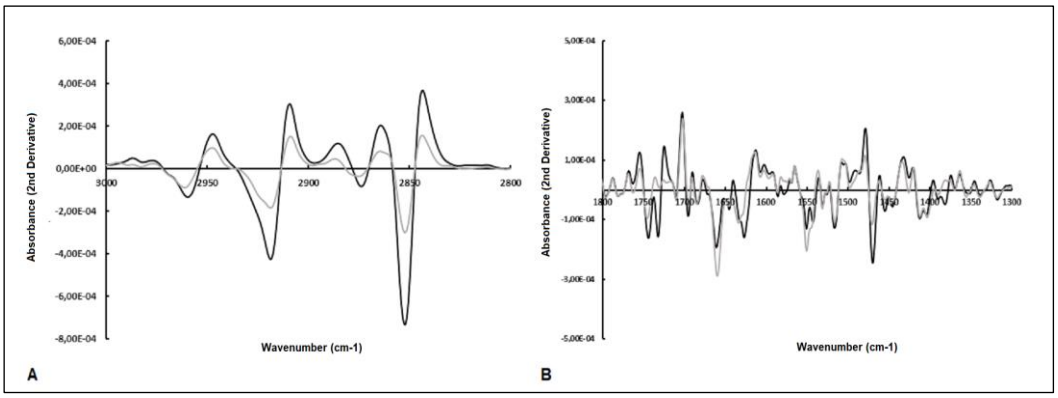


**Figure 3.** Average of second derivative spectra of *E. coli* submitted to antimicrobial (gray line) and non-antimicrobial (black line) compounds. Arrows highlight spectral bands that are different between the two bacteria populations (A and B).

**Table 1.** Ratios between spectral bands of *E. coli* spectra, submitted or not to antimicrobial compounds.

| Ratio bands ($cm^{-1}$) | Average | | Standard deviation | | *p-value* |
|---|---|---|---|---|---|
| | Antimicrobial | No-Antimicrobial | Antimicrobial | No-Antimicrobial | |
| A2912/2856 | 1,359 | 1,356 | 0,114 | 0,162 | 0,472 |
| A2912/1740 | 3,921 | 4,121 | 1,086 | 1,114 | 0,263 |
| A2856/1705 | 2,601 | 1,334 | 0,753 | 0,248 | **0,000** |
| A1740/1656 | 0,261 | 0,114 | 0,089 | 0,038 | **0,000** |
| A1740/1545 | 0,434 | 0,218 | 0,151 | 0,082 | **0,000** |
| A2847/1545 | 0,977 | 0,537 | 0,253 | 0,124 | **0,000** |
| A2847/1740 | 2,395 | 2,720 | 0,657 | 0,966 | 0,102 |
| A1617/1545 | 1,129 | 1,035 | 0,124 | 0,131 | **0,008** |
| A1476/1545 | 0,304 | 0,146 | 0,167 | 0,081 | **0,000** |
| A1215/1179 | 1,269 | 2,950 | 0,605 | 2,314 | **0,004** |
| A1215/1545 | 0,942 | 0,708 | 0,174 | 0,082 | **0,000** |
| A1244/1230 | 1,312 | 1,158 | 0,180 | 0,027 | **0,000** |

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2024 doi:10.20944/preprints202404.0195.v1



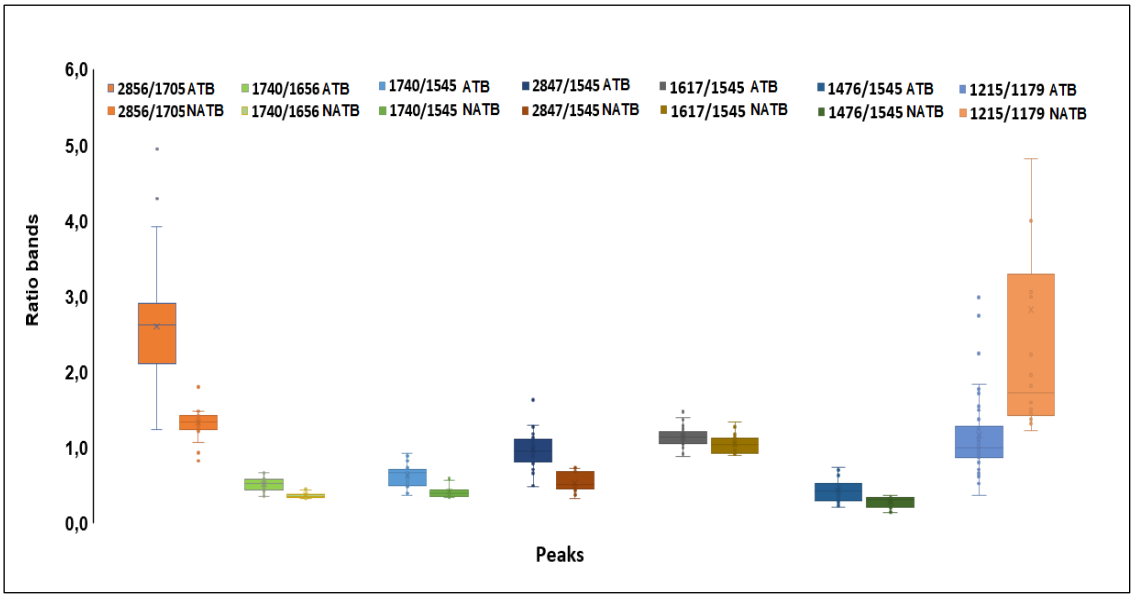


**Figure 4.** Box-plots of ratios of spectral bands of *E. coli* cells submitted to compounds with and without inhibitory activities.

*2.4. Classification Models*

To assess the feasibility of predicting the impact of compounds on *E. coli* inhibition, the study evaluated several supervised classification machine learning algorithms, including PLS-DA, k-NN, SVM, and BPN. For all models, 70% and 30% of spectra were used as the training data (n=42) and for test data (n=18), respectively (Table 2). The calibration, cross-validation performances of the tested models were compared in terms of classification parameters, such as Non-Error Rate (NER) and accuracy with the analysis of confusion matrices, providing a comprehensive overview and a rational means of selecting the approach for the analysis of MIR data for antimicrobial effect.

**Table 2.** Performance of prediction models of the impact of compounds on inhibiting *E. coli*, based on the FT-MIR spectra of *E. coli*, after diverse pre-processing methods. Data are shown to the training and the cross-validation data set.

| Model | Calibration | | | Cross-Validation | | | Test | | |
|---|---|---|---|---|---|---|---|---|---|
| | NER | ER | Accuracy | NER | ER | Accuracy | NER | ER | Accuracy |
| PLS-DA | 100 | 0 | 100 | 88 | 22 | 86 | 92 | 8 | 89 |
| PLS-DA msc | 100 | 0 | 100 | 89 | 11 | 88 | 96 | 4 | 94 |
| PLS-DA snv | 97 | 3 | 95 | 89 | 11 | 88 | 92 | 8 | 89 |
| PLS-DA 1st | 100 | 0 | 100 | 78 | 22 | 81 | 50 | 50 | 72 |
| PLS-DA 2nd | 100 | 0 | 100 | 50 | 50 | 100 | 0 | 100 | 0 |
| kNN | 70 | 30 | 76 | 69 | 31 | 69 | 52 | 48 | 67 |
| kNN msc | 82 | 18 | 83 | 93 | 7 | 90 | 65 | 35 | 67 |
| kNN snv | 79 | 21 | 83 | 85 | 15 | 86 | 65 | 35 | 67 |
| kNN 1st | 87 | 13 | 88 | 91 | 9 | 90 | 50 | 50 | 72 |
| kNN 2nd | 74 | 26 | 76 | 69 | 31 | 71 | 50 | 50 | 28 |
| SVM | 77 | 23 | 83 | 58 | 42 | 55 | 72 | 28 | 78 |
| SVM msc | 91 | 9 | 93 | 80 | 20 | 81 | 92 | 8 | 89 |
| SVM snv | 87 | 13 | 90 | 80 | 20 | 81 | 92 | 8 | 89 |

 doi:10.20944/preprints202404.0195.v1



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SVM 1st | 92 | 8 | 95 | 79 | 21 | 83 | 50 | 50 | 28 |
| SVM 2nd | 100 | 0 | 100 | 68 | 32 | 79 | 50 | 50 | 28 |
| BPN:1:10 | 100 | 0 | 100 | 60 | 40 | 62 | 68 | 32 | 72 |
| BPN:1:10 msc | 96 | 4 | 95 | 90 | 10 | 90 | 81 | 19 | 71 |
| BPN:1:10 snv | 98 | 2 | 98 | 77 | 23 | 79 | 68 | 32 | 72 |
| BPN:1:10 1st | 100 | 0 | 100 | 73 | 27 | 80 | 62 | 38 | 72 |
| BPN:1:10 2nd | 85 | 15 | 87 | 51 | 49 | 52 | 47 | 53 | 50 |
| BPN:1:20 | 100 | 0 | 100 | 68 | 32 | 71 | 100 | 0 | 100 |
| BPN:1:20 msc | 100 | 0 | 100 | 78 | 22 | 81 | 86 | 14 | 89 |
| BPN:1:20 snv | 100 | 0 | 100 | 86 | 14 | 86 | 81 | 19 | 87 |
| BPN:1:20 1st | 100 | 0 | 100 | 60 | 40 | 74 | 67 | 33 | 88 |
| BPN:1:20 2nd | 96 | 4 | 94 | 63 | 37 | 69 | 43 | 57 | 53 |
| BPN:2:10 | 98 | 2 | 97 | 74 | 26 | 79 | 82 | 18 | 83 |
| BPN:2:10 msc | 98 | 2 | 98 | 86 | 14 | 87 | 78 | 22 | 76 |
| BPN:2:10 snv | 100 | 0 | 100 | 77 | 23 | 80 | 72 | 28 | 78 |
| BPN:2:10 1st | 84 | 16 | 85 | 72 | 25 | 82 | 75 | 25 | 80 |
| BPN:2:10 2nd | 100 | 0 | 100 | 62 | 38 | 60 | 66 | 34 | 78 |
| BPN:2:20 | 100 | 0 | 100 | 60 | 40 | 64 | 50 | 50 | 72 |
| BPN:2:20 msc | 100 | 0 | 87 | 87 | 13 | 88 | 82 | 18 | 83 |
| BPN:2:20 snv | 100 | 0 | 100 | 72 | 28 | 78 | 68 | 32 | 72 |
| BPN:2:20 1st | 100 | 0 | 100 | 53 | 47 | 58 | 58 | 42 | 75 |
| BPN:2:20 2nd | 100 | 0 | 100 | 58 | 42 | 68 | 53 | 47 | 50 |

*NER – No error rate; ER – error rate.*

### 2.4.1. PLS-DA Model

The most appropriate PLS-DA model was developed through the implementation of MSC spectral preprocessing, with an explaining total variance of 99%, for 11 LV. The error for the validation procedure was 11%, with an accuracy of 88%. The fitting value of the model for antimicrobial samples presented a sensibility, specificity, precision, and accuracy of 86%, 92%, 96%, and 88%, respectively. The PLS-DA model performance can be affected by outliers, being predicted by several parameters such as leverages, Q residuals, and Hotelling's T2. The Hotelling T2 and Q residuals analysis pointed to the outliers on the training and test data. The samples are characterized by low Q residual and Hotelling T2 value, once the higher Q residual means that there is a significant difference between the original data and those reconstructed from the PLS-DA model, characterized by huge out-of-model residuals [34]. The results registered in Figure 5a depict the calculated response for cluster A (Antimicrobial activity) for the training samples. As represented, the classification threshold in the established PLS-DA model for differentiating cluster A (Antimicrobial activity) from cluster B (Low antimicrobial activity) was 0.210 (dotted line). The threshold value is used to classify into one group or the other. Based on these values, the samples that present a calculated response higher than 0.210 were classified as belonging to cluster A, while the samples that present lower values correspond to cluster B.

Meanwhile, there are some misclassified samples such as cluster A samples recognized as cluster B (i.e., false positives), and samples of group A not included in the cluster A class (*i.e.*, false negatives) (Figure 5a). Based on the score plot, an interesting pattern in the samples' distribution was registered with a certain degree of overlap. The first latent variable explains 38.92% of the variance, while the second latent variable explains just 12.41% (Figure 5b). As expected by the classification performance,

 doi:10.20944/preprints202404.0195.v1



the samples of cluster A characterized by antimicrobial action present higher scores on the first LV comparatively with samples of cluster B (Low antimicrobial activity) (Figure 5b). Relatively to the second LV, the samples of the cluster B are represented in the negative score's values, while the samples of cluster A present positive scores, which shows a significant resolution among both plant extracts. Based on the classification results, the PLS-DA classification model can be reliable and stable, as it is not influenced by samples taken out from the calibration set during the cross-validation procedure. In the validation step, it was observed that the samples utilized in the test phase for both groups were placed close to the training set. This finding indicates that the model exhibits high accuracy, thereby adding value to the classification process facilitated by the PLS-DA technique. Studies developed by Pellegrini et al. (2017) showed the advantages in terms of the PLS-DA models for a classification model for different essential oils from of *B. latifolia*, *Buddleja globosa*, *Solidago chilensis*, and *Aloysia polystachia* against *Paenibacillus larvae* that showed antimicrobial activity [35].

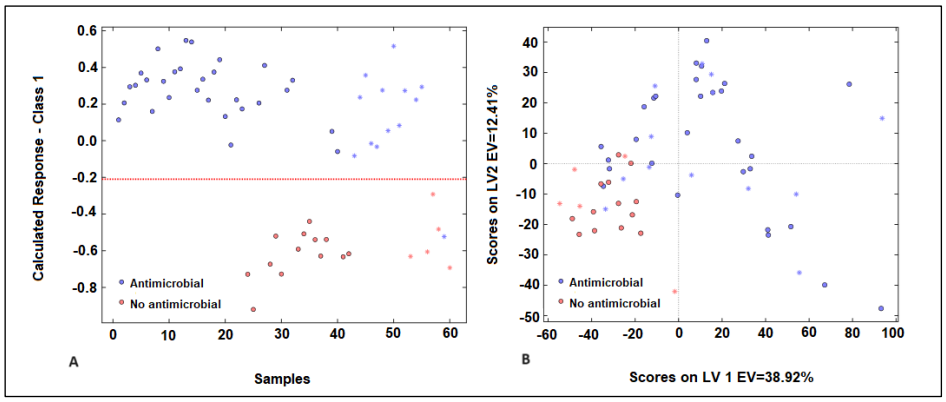


**Figure 5.** Representation of samples regarding group (Antimicrobial activity) and group (No antimicrobial activity) of plant extracts (A) for the PLS-DA model developed using the MSC preprocessing of the spectra related to the *E. coli* assay; Score plot obtained during the PLS-DA model development (B). Blue marks - predicted compounds characterized by antimicrobial activity (blue) and without antimicrobial activity (red).

2.4.2. K-Nearest Neighbor (KNN)

The KNN-optimized models, after the MSC processing, showed a significant fitting accuracy (90%), a classification error of 12%, and a cross-validation error rate of 20% (Table 2). This can be elucidated by the nature of KNN as a simple linear and non-parametric tool that classifies groups based on the classes predominantly represented among its K nearest neighbors. The training model is characterized by an accuracy of 88% and a calibration error 13%. The class A (antimicrobial activity) presented a sensibility of 90% and a specificity of 85%. In terms of cross-validation, an accuracy value of 90% and a validation error of 10% were registered, being characterized by a significant specificity/sensitivity of 90%/92% for cluster A, and 92%/90% for cluster B (Table 2). In terms of the prediction step, the external samples were used, obtaining a value of 72% for accuracy (0%) for the specificity, and 100% for sensitivity in terms of cluster A, while cluster B, was characterized by a specificity of 100% and sensitivity of 0% (Table 2). According to the classification parameters, the KNN model was not significant for the class model. The performance of KNN revealed that for the classification of samples characterized by antimicrobial activity, the classification rate was 88% and 90% for the training and validation process. KNN is a supervised learning method that can be applied for classification and regression tasks and is effectively applied in medicinal chemistry for novel antibacterial drug design. Karakoc et al. (2007) introduced a classification procedure employing KNN method. This approach was designed for the classification of small molecules, focusing on the identification of the most pertinent chemical descriptors. These descriptors are crucial for effectively distinguishing between active and inactive compounds across diverse biological systems [36]. He et al. (2021) provided an extensive compilation of applications of the KNN method in both classification and regression tasks. These applications were specifically directed towards drug delivery for the

 doi:10.20944/preprints202404.0195.v1



treatment of infectious diseases. They encompassed areas such as treatment regimen optimization, drug delivery system design, administration route optimization, and prediction of drug delivery outcomes [37].

2.4.3. Support Vector Machines (SVM)

The SVM model, based on the MSC spectral processing, was developed after the fine-tuning of several parameters, characterized by a radial basis function SVM-kernel, C=100, showed a significant fitting accuracy (90%), being characterized by a classification error (12%), and a cross-validation error rate (20%) (Table 2). The training model was characterized by a significative accuracy (93%), and a calibration error (9%). The sensibility and the specificity for cluster A (antimicrobial activity) were defined by 97% and 85%, respectively, while for cluster B (no antimicrobial activity), the sensitivity and specificity achieved were 88% and 97%, respectively. In terms of the calibration model, the specificity and sensitivity for cluster A were 80% and 85%, respectively, while for cluster B, the specificity and sensitivity were 85% and 80%, respectively. In terms of cross-validation, the accuracy and the validation error were, respectively, 81% and 20%, being characterized by a significant specificity/sensitivity (92%/93%) for cluster A, and 93%/92% for cluster B (Table 2). In terms of the prediction, the model developed presented an accuracy value of 89%, being characterized by 80% and 85% for specificity and sensitivity, respectively, for cluster A, and specificity of 85%, and sensitivity of 80% for cluster B. The SVM tools combined with MIR spectra data, had therefore been considered the most robust model, allowing to develop a reliable method for discrimination of the extracts characterized by antimicrobial activity. SVM supervised learning models are also widely applied for classification, regression, and ranking/virtual screening tasks in medicinal chemistry in a range of fields such as novel anticancer research, design of antivirals, protein–protein interaction research, among others [38]. Focusing on antibacterial drug design, Li et al. reported SVM model development from the fingerprint-featurized ChEMBL database to identify novel antibacterial compounds [39].

2.4.4. Backpropagation Network (BPN)

The success of an artificial neural network depends on the size and quality of the training data of the network, as well as its structure and the learning algorithms used. Samples were assigned randomly to a training set (65%) and the test set (35%). To optimize the network, an appropriate learning rate and different numbers of input nodes (10 and 20) and hidden layer neurons (1 and 2) were evaluated. According to the test set values obtained for the different models, the BPN algorithms (1:10; 1:20; 2:10 and 2:20) presented the best results compared to the other algorithms (Table 2). Based on the MSE value, the model developed with 1 layer and 20 nodes was more suitable for the classification process. For each condition, 20 neurons per layer were considered, with a learning rate (learning rate = 0.01), 1000 iterations, and a momentum term of 0.5, which allows for faster convergence with the use of smaller learning coefficients. Two-layer BP model was achieved at last, with the sigmoid transfer function. The scores of the first 15 PCs were applied as the input variables in the BP-ANN models. The epoch number was set at 1000 and after the multi-training step, the optimal number of nodes in the hidden layer and the learning rate were defined as 20 and 0.01, respectively. The optimized BPN model, developed based on the MSC processing, showed a significant fitting accuracy (90%), a classification error of 12%, and a cross-validation error rate of 20% (Table 2). The training model was characterized by an accuracy of 100% and a calibration error (22%). The sensibility was 100% and a specificity of 100% for class A (antimicrobial effect), while the sensitivity and specificity were 100% for class B (no-antimicrobial activity) (Table 2). In terms of the prediction step, the error rate was 14% and the accuracy value was 89%, being characterized by a specificity of 80%, sensitivity of 92%, and precision of 92% for cluster A, while the cluster B is characterized by specificity (92%), sensitivity (80%) and (80%) precision.

The BPN 1:20 model was the ones with the best classification results in terms of the test condition, followed by the BPN model (2 layers and 10 nodes) and by the BPN models defined by 2 layers and 20 nodes. Based on these results, the BPN algorithms provide the best guarantees for the

 doi:10.20944/preprints202404.0195.v1



development of classification models to predict extracts with antimicrobial activity (Figure 6). The neural networks have an advantage over the conventional mathematical methods for modeling complex biological systems, because (1) they are based only on an actual measured set of input and output variables, without requiring prior information about the interrelationship between the variables, and (2) possess strong generalization and prediction ability. The main advantages of ANN techniques include the learning and generalization ability of data, fault tolerance, and inherent contextual information processing in addition to fast computation capacity [40].

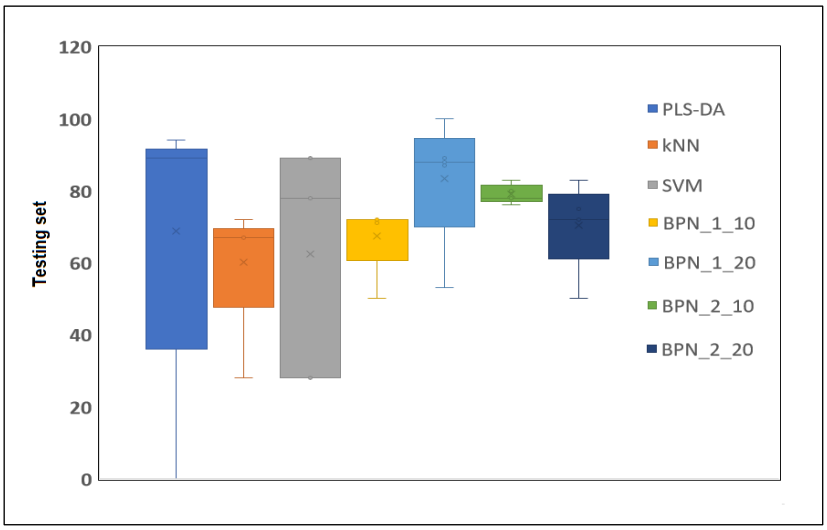


**Figure 6.** Comparison of test parameters related to different classification methods studied (PLS-DA, kNN, SVM, BPN:1:10; BPN:1:20; BPN:2:10; and BPN:2:20 models).

According to Badura et al. (2021) the ANNs represent an excellent tool supporting the work of a researcher in a laboratory, once allowed to estimate whether the tested compound has the desired antimicrobial activity. It should also be emphasized that the greatest advantage of the ANN is to facilitate the work of a researcher who needs to focus on detecting complex relationships in the physicochemical properties of a compound. This approach allows the researcher to target the chemical synthesis of compounds with the desired activity based on their theoretical models obtained by computational chemistry [41]. According to studies developed by Cabrera et al. (2010), artificial neural networks are reliable, fast, and cheap tools and open the way for predicting the antioxidant activity of essential oils [42].

## 3. Materials and Methods

### *3.1. Extraction Process*

The plant material (flowers, seeds, fresh and dried leaves from *C. cardunculus*), was macerated using a mortar and pestle, and subsequently transferred to flask, containing the respective solvent (water, ethanol, acetone, methanol, and ethyl acetate), in the proportion of 5% (w/v), that was kept stirring at room temperature, for 16 hours. The extracts were subsequently filtered under vacuum and the solvents were evaporated using a rotavapor at 60 °C. The dry extract, after being duly weighted, was recovered with a suitable volume of $H_2O$ for a final concentration of 10 mg/ml. Regarding aqueous extraction, the material was boiled (~100 °C) for 10 minutes.

### *3.2. Antioxidant Activity*

The antioxidant activity was determined according to the procedure developed by Mensor et al. (2001) [43], with slight adaptations as follows: samples were diluted in methanol at the following concentrations: 10, 50, and 200 µg/ml. To 2.5 ml of sample was added 1 ml of alcoholic solution of 2,2-diphenyl-1-picryl-hydrazyl (DPPH) (50 µg/ml), remaining in the dark for 30 minutes. The absorbance was evaluated using a spectrophotometer at 518 nm (Shimadzu, Japan). The percentage

 doi:10.20944/preprints202404.0195.v1



of free radical scavenging (%FRS) was evaluated according to the equation [1]: Ac – Control absorbance; As - Sample absorbance; sample, while the blank was prepared by replacing DPPH with ethanol in the reaction mixture. The negative control was prepared by adding 1 ml of DPPH in 2.5 ml of ethanol.

$$FRS\ (\%) = \frac{(Ac-As)}{Ac} \times 100 \quad [1]$$

*3.3. Total Phenols*

The spectrophotometric determination of phenolic compounds was conducted according to the method developed by Slinkard and Singleton (1977) [44]. The calibration curve was developed using different dilutions of gallic acid (0-1000 mg/l). The sample (1 ml) was transferred to a 25 ml flask, while the blank was prepared with distilled $H_2O$. Subsequently, 1 ml of Folin-Cicalteau reagent was added, then stirred, and left to rest for 5 minutes. To the previous mixture, 7 ml of (0.2 g/l) $Na_2CO_3$ solution was added and kept in the dark for 2 hours. The absorbance of the samples was determined at 765 nm, and the values were consequently converted into gallic acid equivalent (GAE) per gram of sample.

*3.4. Antimicrobial Activity*

The extracts obtained from different tissues of the *C. cardunculus,* obtained using diverse solvents, were tested previously in agar plates against *E. coli* cells. The bacteria were grown at 37° C with a stirrer (200 rpm) for 16 hours. Then, 100 µl of extracts were incubated with a similar volume of culture medium containing a specific concentration of cells. After an incubation period (2 hours, at 37 °C), 100 µl of this mixture was plated in a solid culture medium, containing 2% (w/v) of agar. The plates were incubated at 37 °C, for 16 h, before counting the number of colonies.

*3.5. FTIR Spectral Analysis*

3.5.1. Acquisition of Spectra

The samples related to *E. coli* cells were subsequently diluted to obtain a similar optical density (OD) for each assay. All assays were performed in triplicate. The plant extracts (100 µl) were incubated with the cells in the culture medium for 2 hours, at 37° C. After that, the microtubes were centrifuged at 10,000 rpm, for 3 minutes. In order to maintain the same OD value, after the supernatant was discarded, 100 µl of 0.9% (w/v) NaCl was added. Then, 30 µl of each assay was pipetted, in triplicate, in Zn-Se plates and, consequently, dried using a vacuum system for 3.5 h, at room temperature. Spectra were registered with an HTS-XT system (Vertex, Bruker Optics) associated with a FTIR spectrometer (Bruker Optics, Germany), between 4000-400 $cm^{-1}$ with a resolution of 4 $cm^{-1}$.

3.5.2. Spectral Pre-Processing

Spectra were pre-processed by baseline correction, multiplicative scatter correction (MSC), the Standard Normal Variate (SNV), and derivatives (1st and 2nd derivatives).

*3.6. Chemometric Methods*

3.6.1. Principal Component Analysis

The compression and classification of data can be performed using the principal component analysis (PCA) technique. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variables that nonetheless retains most of the sample's information. The method groups the samples according to the existing information so that the individuals in a group are as like each other as possible and as different from the remaining groups as possible.

 doi:10.20944/preprints202404.0195.v1



*3.7. Classification Methods*

3.7.1. Partial Least Squares-Discriminant Analysis (PLS-DA)

Partial least squares-discriminant analysis (PLS-DA) is a linear classification tool that allows to calculate the predictive models based on partial least squares regression algorithm that searches for latent variables with maximum covariance [45,46]. The PLS-DA models were calibrated using a total of 42 samples selected randomly: 29 samples for antimicrobial effect samples and 13 no-antimicrobial effect samples. The cross-validation was performed employing randomly selected training samples with 5 data splits (20% of the calibration samples each) and 20 iterations. The test samples were constituted by 18 samples selected randomly: 13 samples for antimicrobial effect and 5 for no-antimicrobial effect samples. The prediction was performed with the samples of the testing set. The evaluation of the prediction ability of the models was done through the sensitivity (Sn), specificity (Sp), and accuracy, as defined as follows considering true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

Sensitivity (Sn): (Equation (2)):

$$Sn = \frac{tp}{tp+fp} \times 100\% \quad [2]$$

Specificity (Sp): (Equation (3)):

$$Sp = \frac{tn}{tn+fp} \times 100\% \quad [3]$$

Accuracy (Acc): (Equation (4)):

$$Acc = \frac{tp+tn}{tp+fn+tn+fp} \times 100\% \quad [4]$$

3.7.2. K-Nearest Neighbours

K-Nearest Neighbours (KNN) is a supervised machine learning algorithm where classification rules are based on the neighborhood of training set objects, where *k* represents the number of neighbors around the new data point. The distance functions can be used to calculate the distance between the new data point and its k neighbors [47]. The neighborhoods were calculated using Euclidian distances or correlation coefficients between the unknown object and the training set objects. For a training set with *n* samples, *n* distances or correlations are calculated. The unknown object is attributed to the class to which most of the *k* objects with the smallest Euclidean distances or the highest correlations, of the training samples belong.

3.7.3. Support Vector Machine

Support Vector Machine (SVM) is based on finding a hyperplane that best separates a dataset into two groups. However, if the data is non-linear, the SVM algorithm has little ability to separate the hyperplanes. There are other methods available to classify non-linear data, which are known as kernel functions. Kernel functions are mathematical functions that take data as input and transform it into the required form [48]. Kernel functions mapped data to a higher dimension. The parameters and kernels were optimized for the best performance of SVM classifiers. The regularisation parameter C, which controls the trade-off between the minimum training error and minimum model complexity, along with the kernel parameter $\gamma$, represents the width of the kernel function and the degree of generalization, being determined by a grid-search procedure in SVM. The optimal conditions in terms of the C (100) and the linear kernel function were tested in this study.

3.7.4. Back Propagation Network (BPN)

Back-propagation network (BPN) is one of the most popular neural networks, which is a viable, reliable, and attractive approach for data processing because (1) BPNs are capable of modeling non-linear processes; (2) the data-driven features of BPNs make them powerful in parallel computing and capable of handling large amounts of data, and (3) BPNs have good fault tolerance and adaptability [49]. A BPN uses a gradient descent-based delta learning rule (known as backpropagation) for training the artificial neuronal network [50]. This systematic method is computationally efficient in changing the weight in the network with function units to study a set of input-output patterns. This

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 April 2024 doi:10.20944/preprints202404.0195.v1



method aims to minimize the total squared error of the output. The trained supervised learning network achieved through this approach can effectively balance its ability to accurately respond to input patterns. The BPN algorithm minimizes the error in predictions and produces satisfactory results by adjusting each weight of the networks, which was utilized in the establishment of the ANN model [51]. BPN consisted of 10 or 20 neurons in one or two hidden layers according to the optimization study. In modeling, 70% of the total samples were used and treated as the training data set for building the model (42 samples). To prevent overfitting, 30% of the samples were used as the validation data set for early stopping methodology (18 samples). Based on the preliminary experiment, the divider and function were chosen to randomly divide the sample data in this work. The Mean Square Error (MSE) is a commonly used metric in the context of machine learning and neural networks, including BPN. MSE represents a measure of the average squared difference between the actual and predicted values. In the context of a BPN model during the training process, MSE is used to quantify the extent to which the model's predictions deviate from the actual target values. The formula for MSE is typically expressed as (Equation (5)):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad [5]$$

where: $N$ is the number of data points in the dataset; $y_i$ represents the actual target value for the ith data point; $\hat{y}_i$. The MSE is calculated by taking the average of the squared difference between the actual and predicted values for all data points. The use of the squared helps to penalize larger errors more significantly than smaller errors. The BPN neural network is a kind of multilayer feed-forward neural network. The specific topological structure of the BP neural network is shown in Figure 7. The BP-ANN model was developed using ANN tool of MATLAB software.

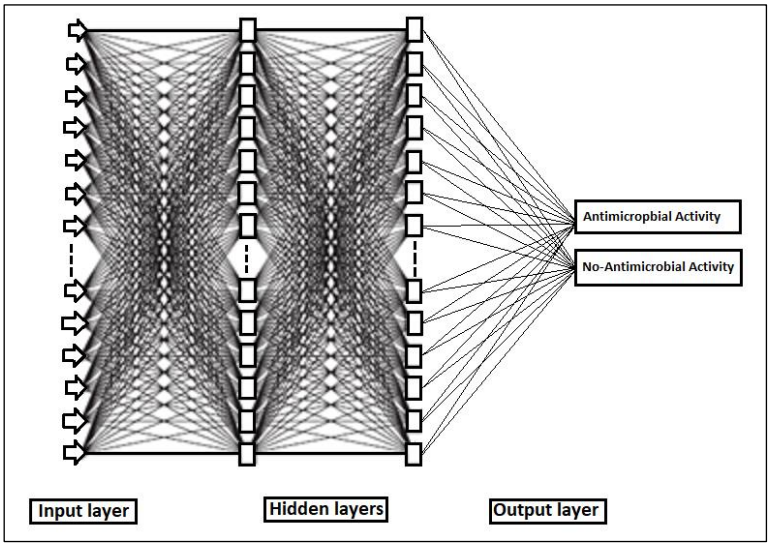


**Figure 7.** Topological structure of neural network used for the identification of antimicrobial activity.

Neurons are arranged in each layer. The layers between the input signal applied to the input and output layer that contribute to the output signal are called the hidden layer.

In supervised learning the network creates a model that relates a training input data (i.e., spectral data) with the desired response or output as defined by the corresponding training target row (i.e., the antimicrobial classes). The network learns every time that a training pattern (input and output data) is applied to the network, by modifying the connection weights, and this is performed several times (iterations or epochs) until a suitable level of error is achieved. The MLP parameters were trained through back-propagation by minimizing the MSE, hence optimizing the weights and bias for each neuron. Finally, the output layer provided the result of the hidden layers as the antimicrobial activity. The training process was performed with a learning rate of 0.001 for 500 epochs of 1000 batches, where each batch was composed of 1154 absorbance spectra. At each epoch, the MLP was

 doi:10.20944/preprints202404.0195.v1



evaluated with a validation set of the same size as the training set. In turn, MLP parameters corresponding to the epoch.

### *3.8. Other Statistical Analysis*

The Student's t-test was performed using the data analysis tool package (Microsoft Excel®) to evaluate the difference between the ratio bands between samples with high antimicrobial effect and samples with low antimicrobial effect. A 5% significance was considered in this work. Spectral pre-processing and processing were conducted with a classification toolbox (PLS-DA, SVM, k-NN, and BPN) developed by Milano Chemometrics and QSAR Research Group (http://michem.disat.unimib.it/chm), using MATLAB® 2023a (The MathWorks, Natick, MA, USA).

## 4. Conclusions

According to the promising results, the study suggests that the integration of FT-MIR spectroscopy and machine learning is a promising strategy to identify compounds from plant extracts with antimicrobial properties. This holistic approach holds the potential for efficiently classifying and discovering new bioactive molecules, thereby streamlining the drug exploratory process. It was possible to develop a model based on the back-propagation network for plant extracts classification, containing active compounds based on the biomolecular characteristics recorded in the FT-MIR spectra of the bacteria. The FTIR spectroscopy and machine learning techniques create the opportunity for preliminary research of promising antimicrobial compounds, saving time and resources compared to tests that require greater logistics in terms of consumables and laboratory techniques. Data analysis employing artificial neural networks offers a powerful method to optimize and minimize labor costs by streamlining the synthesis process, focusing only on compounds with anticipated desired properties. This approach serves as an invaluable tool in guiding the design of synthesis procedures and subsequent biological experiments, particularly in the targeted exploration for novel compounds with significant potential as antimicrobial activity.

**Acknowledgments:** We thank Dr. Paulo Barracosa from the Instituto Politécnico de Viseu, and Dr. César Garcia from Botanical Garden – Faculty of Sciences of the University of Lisbon. PLABIA - Research Platform for Bioactive Compounds through Artificial Intelligence. Project funding by ILIND – Lusófona University.

**Authors Contributions:** Conceptualization, Pedro Sampaio and Cecília Calado; Methodology, Pedro Sampaio and Cecília Calado; Software, Pedro Sampaio; Validation, Pedro Sampaio; Formal Analysis, Pedro Sampaio and Cecília Calado; Investigation, Pedro Sampaio and Cecília Calado; Writing – Pedro Sampaio; Writing – Review & Editing, Pedro Sampaio and Cecília Calado; Visualization, Pedro Sampaio and Cecília Calado; Supervision, Cecília Calado; Project Administration, Pedro Sampaio and Cecília Calado; Funding Acquisition, Pedro Sampaio and Cecília Calado.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Chinemerem, D.N.; Ugwu, M.C.; Oliseloke, A.C. Antibiotic resistance: The challenges and some emerging strategies for tackling a global menace. *J Clin Lab Anal*. **2022**, *36(9), 9*, e24655. DOI: 10.1002/jcla.24655
2. O'Neill, J. *Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations*. **2014**.
3. Mohr, K.I. History of Antibiotics Research. *Curr Top Microbiol Immunol*. **2016**, *398*, 237-272.
4. David, L.; Brata, A.M.; Mogosan, C.; Pop, C.; Czako, Z.; Muresan, L.; Ismaiel, A.; Dumitrascu, D.I.; Leucuta, D.C.; Stanculete, M.F.; Iaru, I.; Popa, S.L. Artificial Intelligence and Antibiotic Discovery. *Antibiotics (Basel)* **2021**, *10*, 1376. DOI: 10.3390/antibiotics10111376.
5. Petropoulos, S.; Fernandes, Â.; Pereira, C.; Tzortzakis, N.; Vaz, J.; Soković, M.; Ferreira, I.C.R. Bioactivities, chemical composition and nutritional value of *Cynara cardunculus* L. seeds. *Food chem* **2019**, *289*, 404-412. DOI: 10.1016/j.foodchem.2019.03.066
6. Falleh, H.; Ksouri, R.; Chaieb, K.; Karray-Bouraoui, N.; Trabelsi, N.; Boulaaba, M.; Abdelly, C. Phenolic composition of *Cynara cardunculus* L. organs, and their biological activities. *C R Biol* **2008**, *331*, 372-9.
7. Paris, R.R. ; Moyse, H. Médicale Massons & Cie. Paris. Masson, **1971**.
8. Velez, Z.; Campinho, M.A.; Guerra, Â.R.; García, L.; Ramos, P.; Guerreiro, O.; Felício, L.; Schmitt, F.; Duarte, M. Biological characterization of *Cynara cardunculus* L. methanolic extracts: antioxidant, anti-proliferative,

anti-migratory and anti-angiogenic activities. *Agriculture* **2012**, *2*, 472-492. https://doi.org/10.3390/agriculture204047

9. Mileo, A. M.; Di Venere, D.; Linsalata, V.; Fraioli, R.; Miccadei, S. Artichoke polyphenols induce apoptosis and decrease the invasive potential of the human breast cancer cell line MDA-MB231. *J Cell Physiol*, **2012,** *227(9),* 3301-3309.
10. Pais, M.S.; Sampaio, P.; Soares, R. Pharmaceutical composition containing the enzyme cyprosin B an aspartic peptidase from *Cynara cardunculus* and its inclusion in anti-tumorals formulations. WO 2009/040778 A2 (PT nº 103839) **2012**.
11. O'Connell, K.M.; Hodgkinson, J.T.; Sore, H.F.; Welch, M.; Salmond, G.P.; Spring, D.R. Combating multidrug-resistant bacteria: Current strategies for the discovery of novel antibacterials. Angew. *Chem Int Ed Engl* **2013**, *20120, (52)*, 10706–10733.
12. Rosa, F.; Sales, K.C.; Cunha, B.R.; Calado, C.R.C. A comprehensive high-throughput FTIR spectroscopy-based method for evaluating the transfection event: estimating the transfection efficiency and extracting associated metabolic responses. *Anal Bioanal Chem* **2015**, *407*, 8097–8108.
13. Baldauf, N.A.; Rodriguez-Romo, L.A.; Männig, A.; Yousef, A.E.; Rodriguez-Saona, L.E. Effect of selective growth media on the differentiation of *Salmonella enterica* serovars by Fourier-transform mid-infrared spectroscopy. *J. Microbiol. Methods.* **2007**, *68*,106–114.
14. Jamin, N.; Miller, L.; Moncuit, J.; Fridman, W.H.; Dumas, P.; Teillaud, J.L. Chemical heterogeneity in cell death: combined synchrotron IR and fluorescence microscopy studies of single apoptotic and necrotic cells. *Biopolymers*. **2003**, *72*, 366–373.
15. Hughes, C.; Liew, M.; Sachdeva, A.; Bassan, P.; Dumas, P.; Hart, C.A.; Gardner, P. SR-FTIR spectroscopy of renal epithelial carcinoma side population cells displaying stem cell-like characteristics. *Analyst*. **2010**, *135*, 3133–3144.
16. Scholz, T.; Lopes, V.V.; Calado, CRC. High-throughput analysis of the plasmid bioproduction process in *Escherichia coli* by FTIR spectroscopy. *Biotechnol Bioeng* **2012**, *109*, 2279-2285.
17. Sampaio, P.N.; Calado, C.R.C. Potential of FTIR-spectroscopy for drug screening against *Helicobacter pylori*. *Antibiotics*. **2020**, *9*, (12), 897.
18. Sales, K.K.; Rosa, F.; Cunha, B.R.; Sampaio, P.N.; Lopes, M.B.; Calado, C.R.C. Metabolic profiling of recombinant *Escherichia coli* cultivations based on high-throughput FT-MIR spectroscopic analysis. *Biotechnol Progr.* **2017**, *33*, 285-298.
19. Sampaio, P.N.; Sales, K.K.; Rosa, F.; Lopes, M.B.; Calado, C.R.C. High-throughput FTIR-based bioprocess analysis of recombinant cyprosin production. J *Ind Microbiol Biotechnol*. **2018**, *44*, 49-61.
20. Alvarez-Ordonez, A.; Carvajal, A.; Arguello, H.; Martínez-Lobo, F.J.; Naharro, G.; Rubio, P. Antibacterial activity and mode of action of a commercial citrus fruit extract. *J Appl Microbiol.* **2013**, *115,* 50-60.
21. Corte, L.; Tiecco, M.; Roscini, L.; De Vincenzi, S.; Colabella, C.; Germani, R.; Tascini, C.; Cardinali, G. FTIR metabolomic fingerprint reveals different modes of action exerted by structural variants of N-alkyltropinium bromide surfactants on *Escherichia coli* and *Listeria innocua* cells. *PLoS ONE*. **2015**, *10, 1*–15.
22. Moen, B.; Janbu, A.O.; Langsrud, S.; Langsrud, Ø.; Hobman, J.L.; Constantinidou, C.; Kohler, A.; Rudi, K. Global responses of *Escherichia coli* to adverse conditions determined by microarrays and FT-IR spectroscopy. *Can J Microbiol*. **2009**, *55*, 714–728.
23. Corte, L.; Rellini, P.; Roscini, L.; Fatichenti, F.; Cardinali, G. Development of a novel, FTIR (Fourier transform infrared spectroscopy) based, yeast bioassay for toxicity testing and stress response study. *Anal Chim Acta*. **2010**, *659*, 258–265.
24. Ribeiro da Cunha, B.; Fonseca, L.P; Calado, C.R.C. A phenotypic screening bioassay for Escherichia coli stress and antibiotic responses based on Fourier-transform infrared (FTIR) spectroscopy and multivariate analysis. *J Appl Microbiol*. **2019**, *127, (6)*, 1776-1789.
25. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *Sn Comput Sci.* **2021**, *2*, 160.
26. Vamathevan, J.; Clark, D.; Czodrowski, P. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*. **2019**, *18*, 463–477.
27. Vora, N.; Tambe, S.S.; Kulkarn, B.D. Counter-propagation neural networks for fault detection and diagnosis. *Comput Chem Eng* **1997**, *21*, 2.
28. Valentão, P.; Fernandes, E.; Carvalho, F.; Andrade, P.B.; Seabra, R.M.; Bastos, M.L. Antioxidative properties of cardoon (*Cynara cardunculus* L.) infusion against superoxide radical, hydroxyl radical, and hypochlorous acid. *J Agric Food Chem* **2002**, *50*, 498. DOI: 10.1021/jf020225o.
29. Pauli, A.; Knobloch, K. Inhibitory effects of essential oil components on growth of food-contaminating fungi. Z. *Lebensm Unters Forsch.* **1987**, *185*, 10-3.
30. Ikigai, H.; Nakae, T.; Hara, Y.; Shimamura, T. Bactericidal catechins damage the lipid bilayer. *Biochim Biophys Acta.* **1993**,*1147*, 132-136. DOI: 10.1016/0005-2736(93)90323-r.

31. Correia Da Silva, T.B.; Souza, V.K.; Lyra, L. Determination of the phenolic content and antioxidant potential of crude extracts and isolated compounds from leaves of *Cordia multispicata* and *Tournefortia bicolor*. *Pharm Biol.* **2010**, *48, 1*, 63-9.
32. Ibrahim, D.; Lee, C.C.; Yenn, T.W.; Zakaria, L.; Sheh-Hong, L. Effect of the extract of endophytic fungus, *Nigrospora sphaerica* CL-OP 30, against the growth of methicillin-resistant *Staphylococcus aureus* (MRSA) and *Klebsiella pneumonia* cells. *Trop J Pharm Res.* **2015**, *14, 11*, 2091-2097.
33. Fu, L.; Xu, B.T.; Xu, X.R.; Gan, R.Y.; Zhang, Y.; Xia, E.Q.; Li, H.B. Antioxidant capacities and total phenolic contents of 62 fruits. *Food Chem*. **2011**, *129, 2*, 345-350. DOI: 10.1016/j.foodchem.2011.04.079.
34. Ballabio, D.; Consonni, V. Classification tools in chemistry. Part 1: linear models PLS-DA. *Anal Methods* **2013**, *5*, 3790–3798.
35. Pellegrini, M.C.; Alonso-Salces, R.M.; Umpierrez, M.L.; Rossini, C.; Fuselli, S.R. Chemical Composition, Antimicrobial Activity, and Mode of Action of Essential Oils against *Paenibacillus larvae*, Etiological Agent of American Foulbrood on *Apis mellifera*. *Chem Biodiversity* **2017**, 14: e1600382. doi.org/10.1002/cbdv.201600382
36. Karakoc, E.; Cherkasov, A.; Sahinalp, S.C. (**2007**). Novel Approaches for Small Biomolecule Classification and Structural Similarity Search. *SIGKDD Explor Newsl 9(1)*, 14–21. DOI:10.1145/1294301.1294307
37. He, S.; Leanse, L.G.; Feng, Y. Artificial Intelligence and Machine Learning Assisted Drug Delivery for Effective Treatment of Infectious Diseases. *Adv Drug Deliv Rev.* **2021,** *178*, 113922. DOI:10.1016/j.addr.2021.113922
38. Romero-Molina, S.; Ruiz-Blanco, Y.B.; Harms, M.; Münch, J.; Sanchez-Garcia, E. PPI-detect: A Support Vector Machine Model for Sequence-based Prediction of Protein–Protein Interactions. *J Comput Chem.* **2019**, *40 (11)*, 1233–1242.
39. Li, W.X.; Tong, X.; Yang, P.P.; Zheng, Y.; Liang, J. H.; Li, G.H.; Liu, D.; Guan, D.G.; Dai, S.X. Screening of antibacterial compounds with novel structure from the FDA approved drugs using machine learning methods. *Aging* (Albany NY). **2022**, 12;14(3):1448-1472. DOI: 10.18632/aging.203887.
40. Jain, A.K.; Mao, J.; Mohiuddin, K.M. Artificial Neural Networks: A Tutorial. *IEEE Computer* **1996**, *29*, 31-44.
41. Badura, A.; Krysiński, J.; Nowaczyk, A.; Buciński, A. Prediction of the antimicrobial activity of quaternary ammonium salts against *Staphylococcus aureus* using artificial neural networks. *Arab J Chem* **2021,** 14, 7, 103233. doi.org/10.1016/j.arabjc.2021.103233.
42. Cabrera, A.C.; Prieto, J.M. Application of artificial neural networks to the prediction of the antioxidant activity of essential oils in two experimental in vitro models. *Food Chem* **2010**, *118*, 1, 141-146. https://doi.org/10.1016/j.foodchem.2009.04.070.
43. Mensor, L.L.; Menezes, F.S.; Leitão, G.G.; Reis, A.; dos Santos, T.C.; Coube, C.S.; Leitão, S.G. Screening of Brazilian plant extracts for antioxidant activity by the use of DPPH free radical method. *Phytother Res* **2001,** 15(2):127-30. doi: 10.1002/ptr.687.
44. Slinkard, K.; Singleton, V.L. Total Phenol Analysis: automation and comparison with manual methods. *Am J Enol Vitic* **1977**, *28,* 49-55.
45. Ballabio, D.; Consonni, V. Classification tools in chemistry. Part 1: linear models PLS-DA. *Anal Methods* **2013**, *5*, 3790–3798.
46. Barker, M.; Rayens, W. Partial least squares for discrimination. *J Chemom* **2003**, *17*, 166–173.
47. Deconinck, E.; Sokeng Djiogo, C.A.; Courselle, P. Chemometrics and chromatographic fingerprints to classify plant food supplements according to the content of regulated plants. *J Pharm Biomed Anal* **2017**, *143*, 48-55.
48. Westerhuis, J. A.; Hoefsloot, H.C.J.; Smit, S.; Vis, D.J.; Smilde, A.K.; van Velzen, E.J.; van Duijnhoven, J.P.M.; van Dorsten, F.A. Assessment of PLS-DA cross validation. *Metabolomics* **2008**, *1*, 81–89.
49. Berry, M.J.; Linoff, G. Data mining techniques: for marking, sales and customer support. *Eds*. **2001**.
50. Russell, S.; Norvig, P. Artificial Intelligence: A Modern Approach. **2003**.
51. Tu, J.V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* **1996**, *49*, 1225–1231.