Distinguishing Healthy and Diseased Chestnuts via THz Spectroscopy and Unsupervised Learning

Anna Martinez[1,*], *Valentina* Di Sarno[2], *Pasquale* Maddaloni[2], *Vito* Pagliarulo[3], *Domenico* Paparo[3], *Melania* Paturzo[3], *Alessandra* Rocco[2], *Michelina* Ruocco[4]

[1]Scuola Superiore Meridionale, Università di Napoli "Federico II", Napoli, Italy
[2]Istituto Nazionale di Ottica INO-CNR, Consiglio Nazionale delle Ricerche, Pozzuoli, Italy
[3]ISASI, Institute of Applied Sciences and Intelligent Systems, Consiglio Nazionale delle Ricerche, Pozzuoli, Italy
[4]IPSP, Istituto per la Protezione Sostenibile delle Piante, Consiglio Nazionale delle Ricerche, Portici, Italy

Abstract. Classifying chestnuts as healthy or diseased remains a complex challenge in quality assessment. In our study, we use THz imaging to determine accurately the health status of chestnuts. Through innovative spectroscopic analysis, we explore the potential of three distinct unsupervised data analysis techniques: Principal Component Analysis (PCA), K-Means Clustering (KMC), and Agglomerative Clustering (AC). Compared to traditional analysis methods, our findings unveil the remarkable ability of these methods to differentiate between healthy, diseased and in an intermediate state chestnuts, even when concealed beneath the peel. This research not only advances our understanding of quality control in chestnut production but also highlights the potential of THz imaging in agricultural applications.

1 Introduction

Chestnut quality is pivotal in the food industry, impacting consumer satisfaction and marketing. Early detection of fungal diseases is crucial for food safety. Various non-destructive methods, including spectroscopy, assess chestnut quality. For example, Corona et al. [1] integrated sensory and NIR spectroscopy to classify chestnuts by quality and market destination. Moscetti et al. [2] demonstrated the ability of NIR spectroscopy to detect hidden mold infections. Also, THz spectroscopy is a valid non-destructive technique for fungal infection detection in chestnuts, as shown by Di Girolamo et al. [3]. By analysing light attenuation and physical parameters like mass and volume, they distinguished healthy from infected chestnuts. In our study, by employing effective preprocessing and unsupervised data analysis methods, THz imaging accurately identifies chestnut health status. We compared PCA, K-Means, and Agglomerative Clustering results, showing consistent classifications and confirming the robustness of our findings.

2 Experimental Setup

2.1 THz Setup

Room-temperature terahertz transmission measurements were conducted using the Tera ASOPS spectrometer by Menlo Systems. This system employs asynchronous optical sampling (ASOPS) with two femtosecond lasers connected to the transmitting and receiving antennas via optical fibers (Fig. 1(a)). It includes a two-dimensional scanning system with a maximum area of 30 x 30 cm^2 (Fig. 1(b)). The nominal spectral window of the system is 4 THz and the scan interval in the time domain is 10 ns. The signal to noise ratio is > 70 dB (with frequency difference = -10 Hz, sampling rate = 10MHz, gain = 106, bandwidth = 1.8 MHz, 1000 averages).

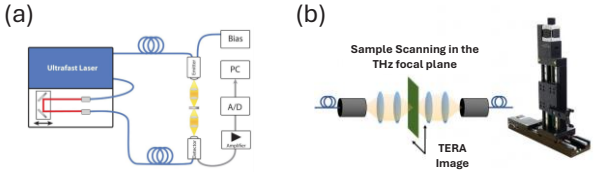

Figure. 1: The basic setup of a Terahertz Time Domain Spectrometer (a), set-up diagram for transmission spectroscopy measurements with a two-dimensional scanning system (b).

The lateral resolution of the measurement is determined by the size of the THz pulse at the focal point, which is approximately 1.5 mm, and by the step of the scanning system, which cannot be less than 0.1 mm. The in-depth resolution depends on the usable bandwidth of the system and is about 0.5 ps, corresponding to 60 μm (in air). The maximum depth of analysis is 7.5 mm (in air).

2.2 Samples Analyzed

Chestnuts of the Palomina cultivar were harvested from the Montella forest in Campania, Avellino, during the

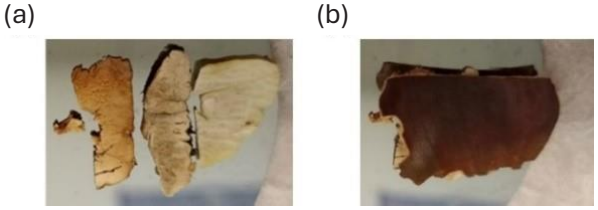

Figure. 2: Three analyzed samples: diseased chestnut (left), intermediate (center), and healthy (right). (b) Same chestnuts with peel placed above.

* Corresponding author: anna.martinez@unina.it

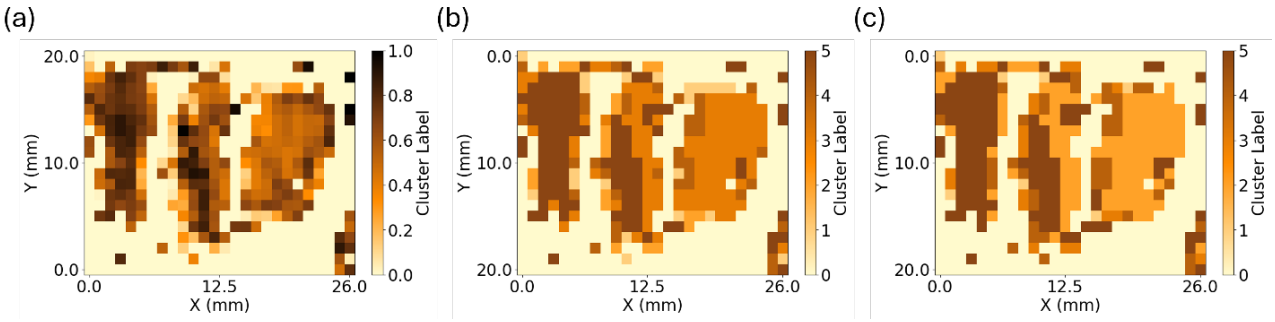

Figure. 3: In (a), (b), and (c), we present the classification results obtained using PCA, k-Means, and Agglomerative Clustering, respectively. Here, label 0 represents the background, while the other labels represent the assigned clusters for each data point.

2023 season. Slices of chestnut fruit, 400 µm thick, were prepared using a rotatory microtome (Microm HM 350s) with a thickness resolution of 50 µm. The samples under analysis are illustrated in Fig. 2(a), comprising three chestnut pieces: one diseased (on the left), one intermediate (in the centre), and one healthy (on the right). The measurement is conducted under the conditions outlined in Fig. 2(b), wherein the chestnut peel is positioned above the samples.

3 Data Analysis Techniques

The data analysis protocol involves importing and normalizing the dataset using the following normalization method:

$$\widehat{\boldsymbol{X}} = \{\bar{X}_i\} = \left\{\frac{X_{ij}}{\max(\bar{X}_i)}\right\}.$$

Here $\bar{X}_i$ represents the spectrum associated with the i-th pixel of the THz imaging. With this procedure, the spectrum associated with each pixel is normalized to 1. This normalization reduces variance in the data caused by intensity differences among pixels, which may arise from factors unrelated to the chemical composition of the sample, such as variations in thickness, color, porosity, roughness, or water content. The use of this specific normalization is essential for accurately determining the health status of the chestnuts. Next, we use KMC to identify background pixels, fixing the number of clusters at two. The identified background pixels are then removed from the dataset. This preprocessing step reduces variance, thereby enhancing the accuracy of chestnut classification. After data preparation, we apply various data analysis techniques, such as PCA, KMC, and AC, to classify the chestnut samples. These techniques are directly applied to the image pixels to cluster them effectively. Finally, we reintegrate the background pixels to visualize the outcome of the classification procedure.

4 Results

In Figure 3, we present the results obtained from the analysis. Particularly in (a), the heatmap is derived from the scores obtained from PCA. These scores are calculated as the sum of the squares of the differences between the normalized data values and the first loading of the PCA, where loadings are defined as $\vec{l}_k = \sqrt{\lambda}_k \vec{v}_k$, with $\vec{v}_k$ e λ_k being the eigenvectors and eigenvalues of the covariance matrix, respectively. In (b)-(c), the heatmaps are derived from the results of the KMC and AC. The determination of the number of clusters in these methods was based on the results provided by the elbow method and the dendrogram method, respectively. Both methods suggest that the optimal number of clusters is equal to 5. From a comparison between the analyzed samples (Fig. 2(a)) and the obtained results (Fig. 3), it becomes evident that employing any of the mentioned analysis methods allows for the accurate classification of healthy chestnuts from diseased ones. Specifically, these analysis methods can also precisely classify chestnuts in an intermediate state, such as the central chestnut, accurately identifying the portion affected by the disease.

5 Conclusion

The combination of THz spectroscopy and unsupervised data analysis techniques offers a robust approach for chestnut classification and quality assessment. Our findings underscore the importance of appropriate data preprocessing and analysis methods in extracting meaningful information from THz spectral data. By leveraging these techniques, researchers and practitioners can enhance the efficiency and accuracy of chestnut quality control and disease management strategies in the agricultural industry.

References

1. Corona, P.; Frangipane, M.T.; Moscetti, R.; Lo Feudo, G.; Castellotti, T.; Massantini, R., Spectral Data. Foods **10**, 11 (2021): 2575.
2. Moscetti, R.; Monarca, D.; Cecchini, M.; Haff, R.P.; Contini, M.; Massantini, R., Postharvest Biology and Technology **93** (2014): 83-90.
3. Di Girolamo, F.V.; Pagano, M.; Tredicucci, A.; Bitossi, M.; Paoletti, R.; Barzanti, G.P.; Benvenuti, C.; Roversi, P.F.; Toncelli, A., Food Control **123** (2021): 107700.