Food Control 168 (2025) 110878

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

Chestnut quality classification by THz Time-Domain Hyperspectral Imaging combined with unsupervised learning analysis

Anna Martinez [a,1], Valentina Di Sarno [b,1], Pasquale Maddaloni [b], Alessandra Rocco [b], Melania Paturzo [c], Michelina Ruocco [d], Domenico Paparo [c,*]

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

ARTICLE INFO

Keywords:
THz-TDS
Machine learning
THz imaging
Principal components analysis
K-means clustering
Agglomerative clustering

ABSTRACT

Chestnut crops are threatened by fungal pathogens such as *Gnomoniopsis castaneae*, which cause significant degradation of quality. Early detection of such infections is crucial to maintain the quality of chestnuts in the food industry. This study explores the application of Terahertz Time-Domain Hyperspectral Imaging (THz-TDHIS) combined with unsupervised learning techniques to identify fungal infections in chestnuts. Unlike conventional methods that rely on light attenuation, this approach leverages the unique spectral signatures of infected tissues. By employing Principal Component Analysis, K-Means Clustering, and Agglomerative Clustering, we effectively differentiate between healthy and infected portions of chestnuts. Our findings indicate that spectral features, rather than just intensity variations, provide more reliable markers for infection. In addition, we demonstrate that these methods enable the quantification of the degree of infection in chestnuts. The robustness of these unsupervised learning methods in handling large and heterogeneous data sets further underscores their potential in agricultural applications. This integrated THz-TDHIS and machine learning approach presents a promising solution to ensure chestnut quality and safety.

1. Introduction

Chestnut fruits, rich in saccharides, attract insects and fungi, posing a significant threat to chestnut crops (Bernárdez et al., 2004). An emerging menace to chestnut cultivation, particularly in European regions like Italy, is *Gnomoniopsis castaneae*, a fungal pathogen causing discoloration of the chestnut endosperm and subsequent mummification and brown rot (Dobry & Campbell, 2023). Additionally, various fungi such as *Penicillium* sp., *Aspergillus* sp., *Fusarium* sp., *Phomopsis castanea*, and *Sclerotinia pseudotuberosa* have been identified in chestnuts globally (Bertuzzi et al., 2015; Donis-Gonzalez et al., 2012; Vettraino et al., 2005; Washington et al., 1997).

Considering the crucial importance of chestnut quality in the food industry, impacting both consumer satisfaction and marketability, the early detection of fungal diseases is essential for ensuring food safety. Non-destructive techniques, such as spectroscopy, are pivotal for assessing chestnut quality (Chen et al., 2013). For example, Corona et al. integrated sensory evaluation with Near Infrared (NIR) spectroscopy to classify chestnuts based on their quality and suitability for the market (Corona et al., 2021). Moscetti et al. demonstrated the efficacy of NIR spectroscopy in detecting concealed mold infections (Moscetti et al., 2014). Additionally, Terahertz (THz) spectroscopy has emerged as a valuable method for non-destructively identifying fungal infections in chestnuts (Di Girolamo et al., 2021). By analyzing light attenuation and physical characteristics like mass and volume, the authors successfully differentiated between healthy and infected chestnuts. Similar results have been obtained for hazelnuts (Gennari et al., 2023). However, these findings heavily rely on the assumption that localized fungal infection alters the water content. This assumption is highly debatable, as numerous other factors could influence the water content in various parts of the fruit (Li et al., 2022; Ruocco et al., 2016). Thus, identifying a spectral feature specifically linked to diseased chestnuts would provide a more reliable marker of their condition, independent of other variables.

In this context, THz spectroscopy has demonstrated its invaluable

* Corresponding author.
E-mail address: domenico.paparo@cnr.it (D. Paparo).
[1] Anna Martinez and Valentina di Sarno contributed equally to this work.

https://doi.org/10.1016/j.foodcont.2024.110878
Received 5 July 2024; Received in revised form 4 September 2024; Accepted 13 September 2024
Available online 19 September 2024

utility for investigating the vibrational characteristics of molecules, encompassing torsional and rotational modes. Various molecules exhibit unique absorption or scattering patterns within the THz range, rendering THz radiation an exceptional non-ionizing alternative to X-rays for generating detailed images of internal structures within objects (Blanchard et al., 2007). THz spectroscopy reveals distinct spectral signatures in numerous soft-matter and bio-systems (George & Markelz, 2013; Mou et al., 2017, 2018; Tielrooij et al., 2009). Recently, it has emerged as a potent tool in agricultural applications too (Afsah-Hejri et al., 2020). However, THz imaging and spectroscopy are often utilized independently, missing the opportunity to augment standard imaging with spectral information. Conversely, in our study, we endeavor to merge THz imaging with Time-Domain Spectroscopy, i.e., to apply THz Time-Domain Hyperspectral Imaging (THz-TDHIS), for more effectively delineating the infected areas in a highly heterogeneous system like chestnuts.

Managing a substantial volume of data is crucial in such methodologies. Moreover, the intrinsic heterogeneity of these systems can obscure spectral characteristics that signify the health or infection status of chestnuts. Therefore, we employ various Unsupervised Learning-based techniques to tackle these obstacles, which are increasingly common in THz imaging applications (Park et al., 2021). Specifically, we assess the results of Principal Component Analysis (PCA) (Greenacre et al., 2022), K-Means Clustering (KMC), (Hartigan & Wong, 1979) and Agglomerative Clustering (AC) (Murtagh & Contreras, 2012). Our findings demonstrate that these approaches can effectively differentiate between healthy and infected chestnuts, a task not achievable solely through signal intensity assessment. Furthermore, all three methods produce consistent classification results, highlighting the reliability of our conclusions. Finally, we will show that these methods are able to quantify the degree of infection of each sample.

2. Materials and methods

2.1. Chestnut samples

Chestnuts *Palomina cultivar* were harvested from chestnuts forest of Montella in Campania region in the province of Avellino during the autumn season 2023. 400 μ m thick slices of chestnut fruit were obtained by means of a rotatory microtome (Microm HM 350s). The thickness resolution of the latter was 50 μ m.

2.2. THz-TDHIS set-up

Room-temperature terahertz transmission measurements were conducted with the TeraASOPS spectrometer by Menlo Systems. This system exploited the principle of Asynchronous Optical Sampling (ASOPS) thanks to the use of two ultra-fast femtosecond laser sources which were connected to the transmitting and receiving antennas via optical fibers (Fig. 1a). This system implemented THz Time Domain Spectroscopy (THz-TDS), thus it generated and detected THz pulses in the time domain, as the one shown in Fig. 1c. The THz pulses transmitted through the sample were attenuated and delayed in time, so they carried information about the refractive index and absorption spectrum at the different frequencies composing the pulse (Lee, 2009). This information could be retrieved in the frequency space by applying a Fast Fourier

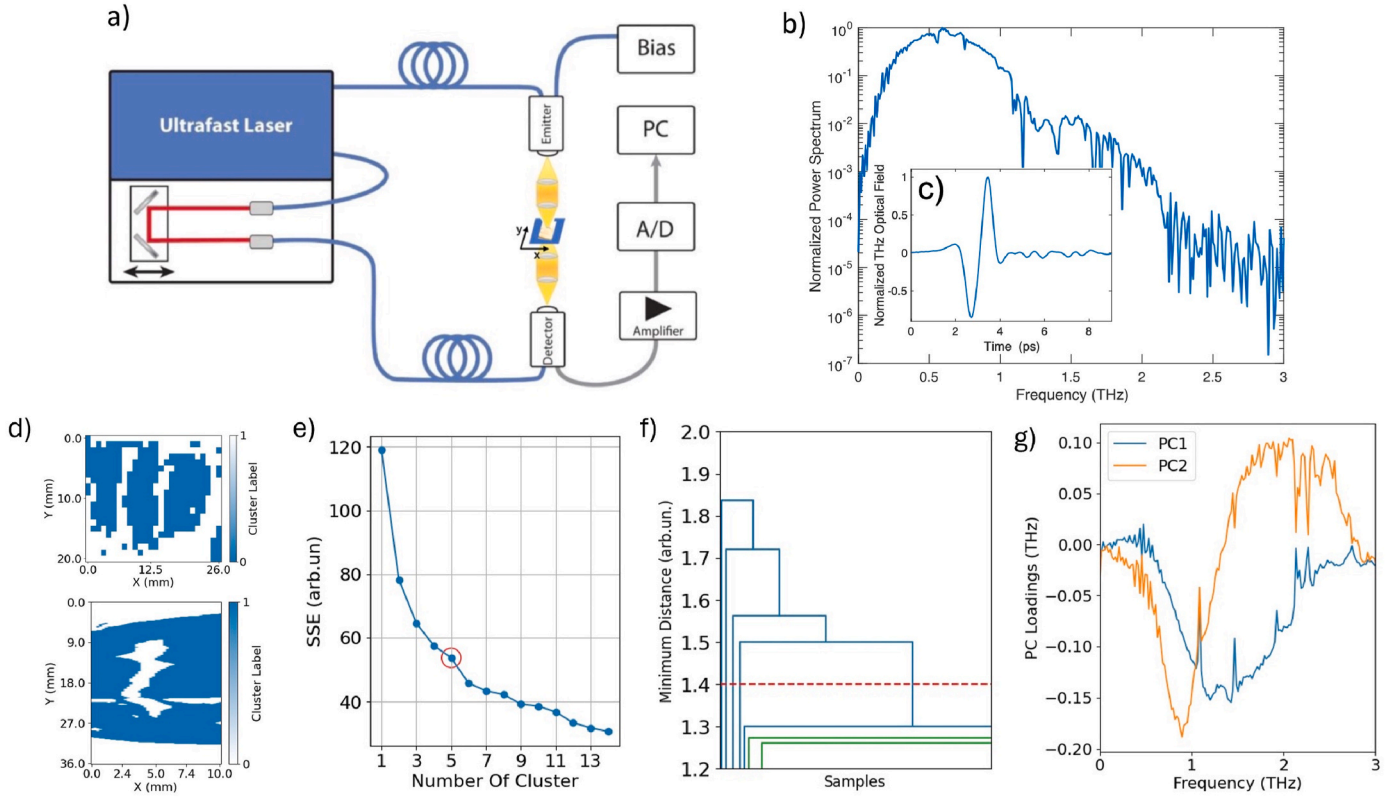

Fig. 1. Methods. a) The THz-TDHIS imaging system by Menlo Systems comprises detection and generation components, consisting of two photo-antennas connected by optical fibers to two high-repetition femtosecond lasers. A portion of the laser beam is deliberately delayed before reaching the detection photo-antenna to perform THz-TDS. b) The power spectrum of the THz pulse transmitted through air. The latter shown in panel c). d) Results of the background removal process on three slices of chestnut (upper panel) and a slice of chestnut (lower panel) covered with a piece of peel in the middle. e) A typical graph generated using the 'elbow' method to determine the optimal number of clusters for the KMC analysis. The red circle highlights the point where a sharp change in the SSE function occurs. f) This graph features a dendrogram for hierarchical clustering. The red line intersects at the point indicating the largest acceptable dissimilarity before merging. g) The 'loadings' from PCA analysis are derived from the eigenvalues and eigenvectors of the covariance matrix.

transform to the THz time waveform.

The Menlo Systems apparatus was associated with a two-dimensional scanning system, whose maximum scanning area was 30 × 30 cm^2. The nominal spectral window of the system was 4 THz. However, as seen in Fig. 1b, where we report the power spectrum of the pulse transmitted in air, the frequency interval with a significant signal was 0.3–1 THz. The scan interval in the time domain was 10 ns. The signal to noise ratio was >70 dB (with frequency difference = −10 Hz, sampling rate = 10 MHz, gain = 106, bandwidth = 1.8 MHz, 1000 averages). The lateral resolution of the measurement was determined by the size of the THz pulse at the focal point, which was approximately 1.5 mm, and by the step of the scanning system, which could not be less than 0.1 mm. The in-depth resolution depended on the useable bandwidth of the system and was about 0.5 ps, corresponding to 60 μm (in air). The maximum depth of analysis was 7.5 mm in air.

2.3. Unsupervised-learning methods

In THz spectral imaging, the transmitted THz optical field are captured. We indicated the latter as $\boldsymbol{E}(t)$, for each pixel. Note that the latter is a discrete array of values, indexed with i. From this, we calculated two important metrics: the total amplitude, $\int |\boldsymbol{E}(t)|dt$, and the spectral amplitude, $\overline{\boldsymbol{E}}(\omega_i) = |\boldsymbol{E}(\omega)|$, using a Fast Fourier Transform. Here and in the following, the bar will indicate the discretized spectral amplitude. We organized all the data from the image into a matrix $\widehat{\mathbf{E}}_0 = \{\overline{E}_{ij}\}$, where i was the index introduced above and j identified each pixel of the image as better shown in Fig. 1 of the Supplementary Material. Therefore, $\overline{\boldsymbol{E}}_j$ indicates the spectrum-array corresponding to pixel j. Here and in the following the hat will indicate the matrices. An essential pre-processing step was then applied to the matrix to optimize its readiness for analysis. This involved normalizing the dataset and removing spectra associated with background pixels from the matrix $\widehat{\mathbf{E}}_0$. The result was a new matrix $\widehat{\mathbf{E}}$, which was now free from background interference and normalized. Below, we outline the pre-processing steps and provide further details about these analytical techniques. We applied three Unsupervised Machine Learning techniques — Principal Component Analysis (PCA), K-Means Clustering (KMC), and Agglomerative Clustering (AC) — each separately on the pre-processed matrix. The goal was to determine if each technique could distinguish the healthy from the unhealthy portions of chestnuts. We then compared the results obtained from all three techniques to evaluate their effectiveness. Below, we describe the pre-processing steps in greater detail and provide insights into the analysis techniques used.

For the analysis of our spectral data, we introduced two types of data normalization. Specifically, with Type 1, we refer to normalization with respect to the global maximum of the dataset:

$$\widehat{\mathbf{E}}^{N1} = \left\{\frac{\overline{E}_{ij}}{\max(\widehat{\mathbf{E}}_0)}\right\} = \left\{\frac{\overline{\boldsymbol{E}}_j}{\max(\widehat{\mathbf{E}}_0)}\right\}, \tag{1}$$

In this normalization, each value $\overline{E}_{ij}$ in the dataset was divided by the maximum value present in the entire dataset $\widehat{\mathbf{E}}_0$. The entire dataset was scaled so that the maximum value in the entire dataset became 1, and all other values were between 0 and 1. The spectrum containing the maximum value in the dataset had a value that reached 1. The other spectra had values less than 1 or equal 1 in case of many global maxima.

In the second type of normalization (Type 2), we normalized each spectrum $\overline{\boldsymbol{E}}_j$ to its own maximum:

$$\widehat{\mathbf{E}}^{N2} = \left\{\frac{\overline{\boldsymbol{E}}_j}{\max(\overline{\boldsymbol{E}}_j)}\right\}. \tag{2}$$

In the first normalization (Type 1), we were more sensitive to the amplitude fluctuations due to the inhomogeneities of the sample, which could be attributed to factors other than changes in the physico-chemical composition of the sample (e.g., varying thickness, porosity, roughness, water content, etc.). The second procedure (Type 2) naturally enhanced the weight of spectral signatures.

Concerning removal of the background pixels, they were discerned using KMC. It was performed using unsupervised binary clustering, where the algorithm divides the image into two classes: 'background' and 'sample'. With this choice, we were assuming that the background spectrum was sufficiently different from the chestnut spectrum to allow for clear distinction between the two through binary classification. This approach, as depicted in Fig. 1d, effectively distinguished the various samples from their backgrounds. However, apart from the vertical cut in the lower panel of Fig. 1d, which corresponded to an actual fracture in the chestnut slice, we also observed regions of the chestnut that were incorrectly classified as background and vice versa (such as the bottom-right corner in the upper panel and the horizontal white lines in the lower panel). We note that this misclassification was already present at the preprocessing stage and hence it affected all the subsequent analysis. The misclassification of these pixels can be attributed to several factors, including measurement disturbances, background impurities, reflections, or instrumental noise, which can strongly distort the spectral response so to confuse it with that of the background. However, it is important to note that these "false" events represent only a small fraction of the pixels.

PCA worked by diagonalizing the covariance matrix to identify the eigenvectors, known as principal components (Greenacre et al., 2022). These principal components were ordered based on the largest eigenvalues of the covariance matrix, so the first component captured the greatest amount of variance. This algorithm allowed to identify the dominant patterns in a data matrix $\widehat{\mathbf{E}}$, which had dimensions $n \times m$ and was formed by measuring m variables across n samples. PCA accomplished this by decomposing the original data matrix into the product of two smaller matrices: the scores matrix and the loadings matrix. In mathematical terms, this decomposition separated the data matrix into a structured component and a noise component, expressed by the equation (Wold et al., 1987):

$$\widehat{\mathbf{E}} = \widehat{\mathbf{T}}\widehat{\mathbf{P}}^{\mathrm{T}} + \widehat{\mathbf{R}}$$

where the structured component $\widehat{\mathbf{T}}\widehat{\mathbf{P}}^{\mathrm{T}}$ captured significant data patterns, and the residual matrix $\widehat{\mathbf{R}}$ represented nonsystematic noise. The scores matrix $\widehat{\mathbf{T}}$ reflected the variance among the samples, assigning each a coordinate in the principal component space. In contrast, the loadings matrix $\widehat{\mathbf{P}}$ conveyed how the original variables related to the principal components, revealing their correlations. Examples of the calculated loadings are shown in Fig. 1g. The residuals matrix $\widehat{\mathbf{R}}$ accounted for the variation that did not align with the principal components. The scores matrix plot distinguished healthy chestnuts from diseased ones. Due to significant differences in their features, healthy and diseased samples cluster separated in the principal component space, allowing for visual identification based on PCA's dominant patterns.

K-Means Clustering and Agglomerative Clustering were powerful clustering algorithms that segmented spectral data into homogeneous groups based on their attributes. KMC worked by initially placing a predefined number of cluster centroids and assigning each data point to the nearest one based on a distance metric, such as Euclidean distance (Murtagh & Contreras, 2012). After assigning the points to clusters, it recalculated the centroids as the mean of all points in the cluster. This iterative process continued until no significant changes occurred, providing stable groupings. The elbow method helped choosing the optimal number of clusters in KMC by plotting the sum of squared distances from each point to its cluster centroid against different values of cluster number K. The optimal K was identified where the plot showed a clear bend or "elbow," indicating diminishing returns in reducing variance with additional clusters. Fig. 1e displays a typical graph generated using the elbow method to determine the optimal number of clusters for

data analysis. In contrast, AC started with each data point as an individual cluster and progressively merged the closest clusters based on their similarity, often using linkage criteria like single, complete, or average linkage (Greenacre et al., 2022). This process continued until only one cluster remained or a specified number was reached. The optimal number of clusters could be determined by analyzing the dendrogram, a tree-like diagram that visually depicts the merging process. To determine the ideal juncture for cutting the dendrogram, it was sought a level where the disparity between consecutive heights was most noticeable. In the dendrogram, each horizontal line represented a level where clusters merged, and the vertical height of the line indicated the distance or dissimilarity between the clusters that were merged. If the difference in height between two consecutive mergers was significant, it meant there was a notable increase in the distance between clusters merged at that level compared to those merged in the previous level. Fig. 1f illustrates a typical dendrogram from our samples.

Together, these three techniques revealed intrinsic groupings within complex spectral datasets, helping to identify subtle differences in composition or structure that were crucial in spectroscopic imaging of our samples.

3. Results

In this section, we present the outcomes derived from employing the PCA, KMC, and AC methodologies on various samples. Initially, we examined a healthy chestnut with a strip of peel positioned at its center, shown in Fig. 2a. Note the uniform white color indicating its healthy status, typically discerned through visual inspection (Ruocco et al., 2016).

The THz spectroscopic imaging outcomes are delineated in Fig. 2 (b–d). The spatial resolution of these images is 0.15 mm. Utilizing the Tera Image software developed by Menlo Systems, the integration of $|E(t)|$ for each pixel yielded the image depicted in Fig. 2b. Additionally, analogous outcomes were attained when visualizing either the maximum or minimum of the THz waveform, a method commonly employed in THz imaging analysis (Catapano et al., 2019; Manca et al., 2023). By employing the PCA method on the spectral amplitudes within the frequency range 0–3 THz for each pixel, with normalization procedures of the spectra delineated above, we could reconstruct the image presented in Fig. 2c-d. In panels (b-d) of Fig. 2, the blue rectangle indicates the chestnut area concealed by the peel. Notice the gray vertical irregular shape in Fig. 2a, which results from a fracture in the slice and which is perfectly reproduced in all the panels (b-d). However, as already noted, there are also limited portions of the chestnut that are wrongly classified as background.

These images demonstrate that in all three cases, THz-TDHIS is capable of detecting the chestnut fruit beneath the peel. However, Fig. 2b reveals that the THz image is highly heterogeneous despite the entire fruit slice being healthy. Consequently, this outcome illustrates that a simple measurement of transmitted intensity is insufficient for unequivocally classifying the chestnut as healthy. This limitation arises because transmitted total amplitude is significantly influenced by several factors of inhomogeneities (e.g., varying thickness, porosity, roughness, water content, etc.). The PCA analysis produced similar outcomes when applied to spectra normalized to the maximum value among all the transmitted amplitude (Fig. 2c). This is because, also in this case, normalizing to the maximum amplitude among all pixels accentuated features associated with inhomogeneities, which strongly impacted the absolute value of transmitted amplitude. Conversely, when employing the second normalization in PCA, thus placing more emphasis on spectral features rather than amplitude variations, distinct results emerged. In this scenario, as depicted in Fig. 2d, the reconstructed image accurately identified the healthy fruit across all sectors of the chestnut slice, except for some small areas which were wrongly classified as background already at the preprocessing stage. In the following, we will omit the images acquired by measuring the total amplitude of the transmitted pulse, as demonstrated in Fig. 2b, and concentrate exclusively on the PCA, KMC, and AC methods.

Both healthy and infected fruit slices were examined to demonstrate the efficacy of PCA with second normalization in accurately classifying them. These findings are presented in Fig. 3(b and c). The spatial resolution of these images is 1 mm. The analyzed slices, depicted in Fig. 3a, are shown without the chestnut husk, which instead covered the fruit slices during the measurements shown in panels (b-c). Similarly to

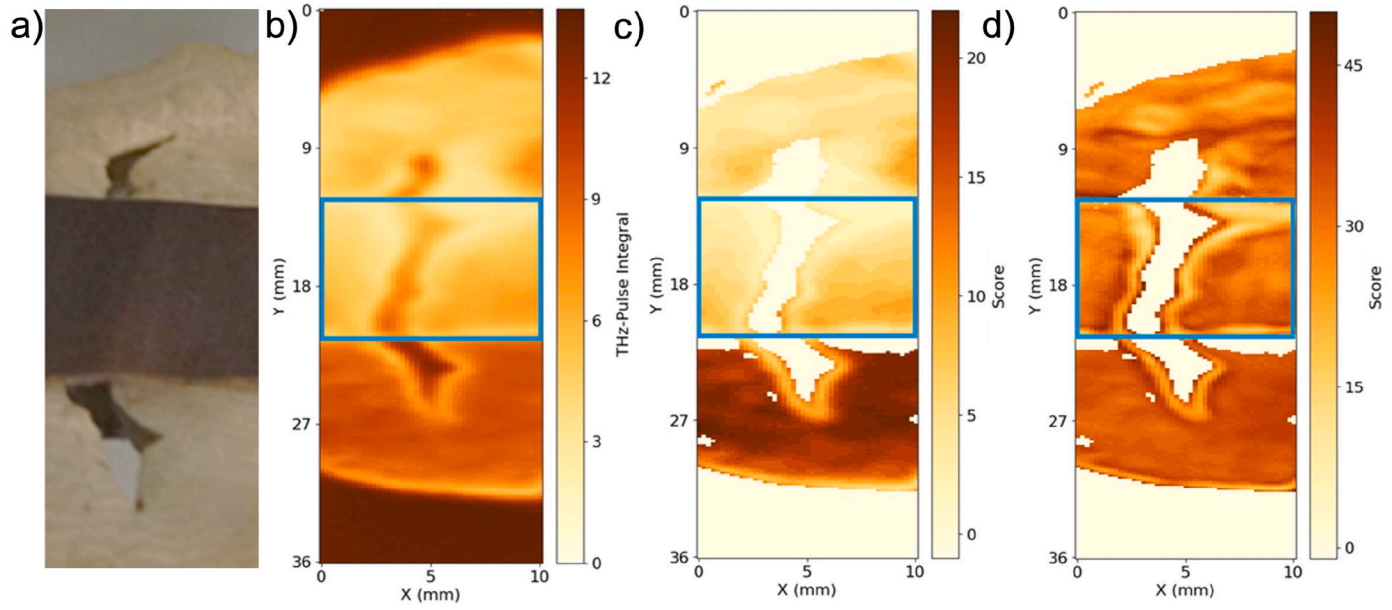

Fig. 2. a) Photograph of the chestnut slice covered by a piece of peel, note the uniformity of the fruit color; b) Image of the chestnut obtained by calculating for each pixel the total amplitude of the transmitted pulse. The spatial resolution of this image is 0.15 mm. Observe the blue square in the center, indicating the area covered by the peel depicted in panel a); c) Score plot obtained by applying the PCA method to the spectral amplitude within the frequency range 0–3 THz of each pixel normalizing all the spectral amplitudes to the maximum of the latter across all pixels; d) The same image as in panel d), but with each spectrum normalized to its maximum, thereby accentuating the spectral distinctions of each pixel.

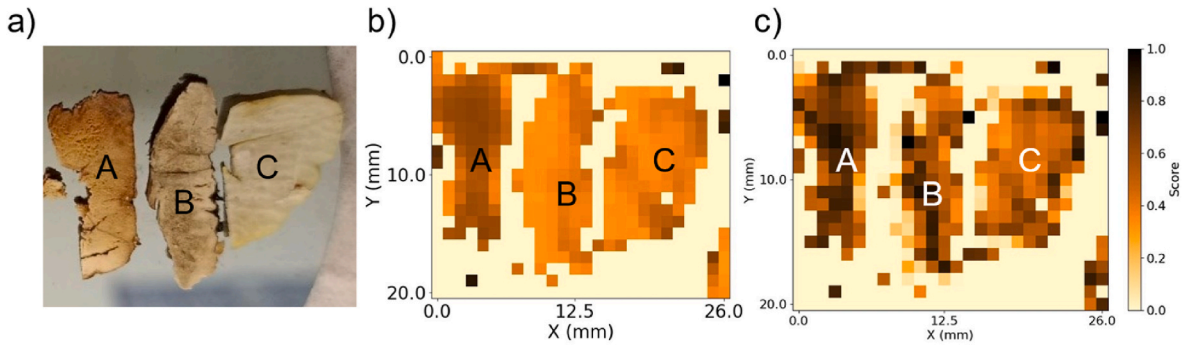

Fig. 3. a) Three slices of chestnut fruit classified by visual inspection: strongly infected (A); partially infected (B); healthy (C). In the remaining panels the images refer to these samples but fully covered with a piece of peel. The spatial resolution is 1 mm. b) Score plot obtained by applying PCA on spectral amplitudes within the 0–3 THz frequency range, normalized to the dataset maximum. c) The same score plot as in panel b), but with each spectrum normalized to its own maximum value, thereby accentuating the spectral distinctions of each pixel.

previous observations, THz-TDHSI successfully reconstructed the fruit image even beneath the peel. However, it struggled to differentiate between the healthy and partially infected slices when applying PCA to spectra normalized to the global maximum (Fig. 3b). Once again, the second normalization procedure proved more effective in discerning the infection gradient, as illustrated in Fig. 3c.

Finally, we proceeded to examine the outcomes derived from the two other methods: K-Means and Agglomerative Clustering. As elucidated in Sec. 2.3, for both techniques, it was crucial to determine the number of clusters for classifying the spectra. We employed the 'elbow' method for K-Means and the 'dendrogram' method for AC to ascertain the appropriate number of clusters. The optimization procedures yielded an optimal value of 2 clusters for the first type of normalization and 5 clusters for the second type. Fig. 4 presents the results obtained using these cluster numbers. It is important to note that the optimal number of clusters depends on the hyperparameters used to implement the elbow method and dendrogram. For this reason, in the Supplementary Material we demonstrate that variations in the cluster numbers within a certain range do not alter the final outcome of chestnut classification, thus confirming the robustness and reliability of the performed analysis.

In panels (a-b) of Fig. 4, we present the results obtained from K-Means applied to spectra normalized to the global maximum and spectra normalized to their own maxima, respectively. These images reveal that, akin to the PCA method, normalizing the spectra to their own maxima was crucial for distinguishing healthy sections of the chestnut from the infected ones, thereby enhancing the spectral disparities among different pixels. Similar results were observed in the case of the AC method, underscoring the robustness of an analysis predicated on

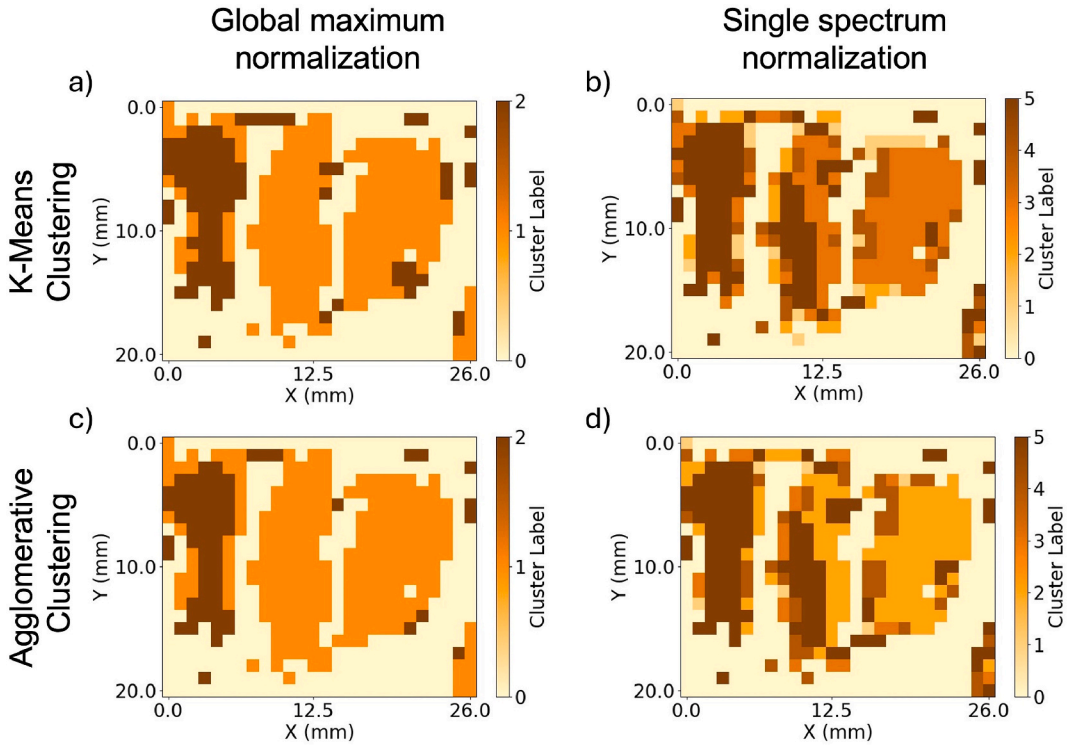

Fig. 4. In panel (a–b) the results of the analysis conducted by means of the K-Means Clustering method are reported: a) spectral amplitude with normalization to the maximum transmitted amplitude among all the pixel spectra; b) the same image with each spectrum normalized to its maximum. In panel (c–d) for the same samples we report the results obtained by means of the Agglomerative Clustering. As in previous sequence panel c) reports the results obtained with normalization to global maximum; d) with normalization to the maximum of each spectrum. Note the different number of clusters for the two normalization procedures as better explained in the main text.

normalizing each spectrum to its own maximum.

4. Discussion

Fig. 5(b–d) and (f-h) display histograms that are used to quantitatively compare the three analytical techniques. We lack independent data on the percentage of healthy areas in the three slices, as the standard methods in agriculture rely mainly on rough visual estimates. To enhance the comparison with our THz spectroscopy findings, we conducted a more detailed visual analysis by converting a photograph, taken prior to the THz measurement, into a binary image. In this process, each pixel was classified as either healthy or diseased (white or black) by isolating the chestnut slice from its background and applying a threshold corresponding to the average RGB value of the original photograph (further details are available in the Supplementary Material). The selection of this threshold is crucial. The error bars shown in Fig. 5a represent an estimate of the variation in the classified portions' percentages when this threshold is adjusted by 10%. Additional information on the statistical analysis is provided in the discussion below.

These histograms distinguish between pixels representing healthy (green bars) and diseased (red bars) chestnuts by establishing an appropriate threshold value. For PCA, the threshold was set as the mean of the scores, whereas scores below the mean are associated with a diseased portion of the chestnut. In the case of KMC and AC, differentiation between Type 1 and Type 2 normalization was necessary. With Type 1 normalization, there were only two clusters, leading to a binary classification where cluster 1 represented healthy pixels and cluster 2 represented unhealthy pixels. For Type 2 normalization, there are five clusters; thus, clusters 4 and 5 were designated as infected pixels, while clusters 1 through 3 were designated as healthy pixels.

Starting from graphs in Fig. 5 we could calculate the percentage of healthy and unhealthy chestnut. The results are summarized in Table 1. Specifically, using Type 1 normalization, samples A, B, and C consisted of 32%, 96%, and 92% healthy portions using PCA; 24%, 92%, and 90% with K-Means; and 35%, 96%, and 96% with Agglomerative Clustering, respectively. Conversely, using Type 2 Normalization, the samples are

Table 1
Percentage of unaffected area as detected by binarization of the photographic image and each unsupervised method and for the two types of normalizations, together with the errors calculated as explained in the main text. The errors for Binary Image correspond to the triangles in Fig. 5, as well as the errors in parenthesis for PCA.

Method	Unhealthy (A)	Partially Healthy (B)	Healthy (C)
Binary Image	22 ± 8%	61 ±10%	76 ±9%
Type 1 norm.			
PCA	32 ±8 (±2)%	96 ± 28 (±9)%	92 ±13 (±9)%
K-Means	24 ±6%	92 ±31%	90 ±14%
AC	35 ±13%	96 ±35%	95 ±19%
Type 2 norm.			
PCA	28 ±7 (±2)%	62 ± 9 (±9)%	74 ±9 (±9)%
K-Means	28 ±6%	61 ± 7%	79 ±7%
AC	25 ±5%	61 ±7%	73 ±7%

28%, 62%, and 74% healthy with PCA; 28%, 61%, and 79% with K-Means; and 25%, 61%, and 73% with Agglomerative Clustering.

To establish the confidence level of these values, we supplemented the results with a statistical analysis. We employed two different approaches. As mentioned earlier, the threshold set for the binarized image and PCA analysis was crucial. Therefore, we varied it by ±10% to assess its impact on the percentage of healthy versus unhealthy areas. The resulting errors are shown in Fig. 5 using triangle symbols. However, this approach is problematic for KMC or AC due to the discrete nature of these methods. Adjusting the threshold by ±10% does not affect the histograms, leading to zero error. Consequently, we adopted a different approach. The circle-marked errors for PCA, KMC, and AC in Fig. 5 were derived by comparing their results to the binarized image, which serves as the expected outcome. Specifically, three datasets of binarized images were created from the original photograph using the threshold, threshold +10%, and threshold −10%. For each dataset, we calculated the maximum error between the dataset and the Unsupervised Learning analysis results. The average of these maximum errors across the three datasets was then determined. All errors are also reported in Table 1.

This analysis shows that almost all methods used are consistent in

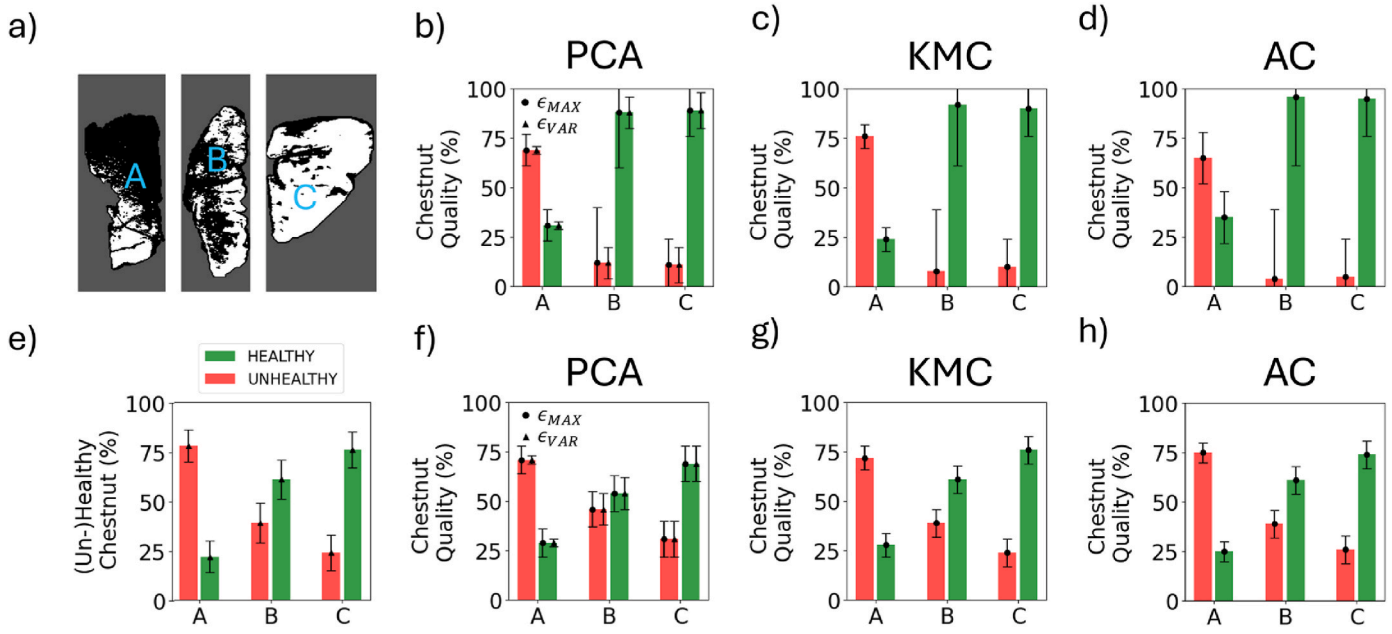

Fig. 5. (a) Shows the binarized image of Fig. 3a, processed according to the method described in the main text. Panels (b–h) display histograms representing the percentage of healthy (green bars) and diseased (red bars) sections of each chestnut, assessed using the binarized image (e), PCA (b, f), K-Means (c, g), and Agglomerative Clustering (d, h). Panels (b–d) illustrate results from spectra normalized to the global maximum (Type 1 procedure), while panels (f–h) show results from spectra normalized to the maximum of each individual spectrum (Type 2 procedure). Each bar includes a maximum error bar (circles, ϵ_{Max}), as described in the main text. Additionally, for the binarized image (panel e) and PCA, a second error bar (triangles, ϵ_{VAR}) is provided, calculated as detailed in the text.

classifying pixels associated with good and bad portions of chestnuts, confirming the robustness and reliability of the proposed analytical protocol. Additionally, the results once again demonstrate that Type 2 Normalization is more effective at distinguishing healthy from infected regions of the chestnut, while Type 1 Normalization fails to classify partially unhealthy chestnuts accurately. Indeed, it is interesting to note that the results obtained from the photographic analysis are entirely consistent with those obtained from the unsupervised analysis using all three methods (PCA, KMC, and AC) with Type 2 Normalization. Finally, this outcome is also evident in the larger relative errors associated with the Type 1 approach compared to the Type 2 approach.

5. Summary

We have investigated the application of Terahertz Time-Domain Hyperspectral Imaging combined with unsupervised learning techniques to detect and quantify fungal infections in chestnuts. Chestnuts, particularly in regions like Italy, are susceptible to infections from fungal pathogens such as *Gnomoniopsis castaneae*, which cause significant damage to the crops. Traditional methods for detecting fungal infections in chestnuts rely on invasive approaches, such as visual inspection after peeling. Recently, optical techniques based on THz attenuation measurements have been employed to detect infections in unpeeled chestnuts. However, this approach may be imprecise since it is based on the hypothesis that fungal infections alter the water content in the fruit—a premise that is debatable given the many other factors that can influence water content.

In this study, we have utilized THz Time-Domain Spectroscopy and Imaging to identify spectral markers that are not influenced by the previously mentioned assumptions or by the heterogeneous properties of the samples, such as varying thickness, porosity, roughness, and water content. To manage the large and complex datasets generated by our approach, we have employed various unsupervised learning techniques: Principal Component Analysis, K-Means Clustering, and Agglomerative Clustering.

We have clearly demonstrated that these three methods help to coherently identify and classify the spectral features of healthy and infected chestnuts. Additionally, by applying these techniques, it is possible to quantify the degree of infection in each sample. This quantification is crucial for assessing the extent of damage and for allowing an early detection of the infection.

In conclusion, the integration of THz-TDHIS with unsupervised learning techniques offers a powerful solution for detecting and quantifying fungal infections in chestnuts. In our study, we used chestnut slices due to the power limitations of the THz source. However, Fig. 2 demonstrates that we were also able to evaluate the quality of chestnut slices hidden by the peel. This serves as a proof-of-concept that our technique can be non-invasive if THz sources with adequate power are utilized, so to enable a measure of the THz spectra without opening the chestnut. Future developments include replacing the emitter with a high-powered one to investigate chestnuts in a completely non-invasive way. Thus, this approach enhances the ability to ensure chestnut quality and safety, providing significant benefits to the food industry and consumers. The study highlights the potential of advanced spectroscopic imaging combined with machine learning to address complex challenges in agricultural quality control.

Supplementary Material

We have included a PDF file as 'Supplementary Material', where we report: (i) further details on the normalization procedure applied to the datasets; (ii) the results of the KMC and AC analysis for different values of the cluster numbers; (iii) details on the binarization procedure applied to the optical images of the chestnut slices. About the second point, the results show the almost independence of the analysis outcome of the KMC and AC methods on the optimal number of clusters provided by the elbow and dendrogram methods.

CRediT authorship contribution statement

Anna Martinez: Writing – review & editing, Visualization, Validation, Software, Formal analysis, Data curation. **Valentina Di Sarno:** Writing – review & editing, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation. **Pasquale Maddaloni:** Writing – review & editing, Resources, Funding acquisition. **Alessandra Rocco:** Writing – review & editing, Project administration, Funding acquisition. **Melania Paturzo:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Michelina Ruocco:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization. **Domenico Paparo:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was funded by the project PSR Campania 2014/2020 Misura 16 –Tipologia di intervento 16.1 –Azione 2 "Sostegno ai Progetti Operativi di Innovazione (POI)"-Progetto 'Migliorcast' (CUP B78H19005230008). The TeraHz ASOPS system was acquired through SHINE project funding (Strengthening the Italian Nodes of E-RIHS, Avviso 424/2018 dell'Azione II.1 PON R&I 2014–2020, DD n. 461 del 14-03-2019, PIR01_00016, CUP B27E19000030007). We also extend our gratitude to Mr. Mario De Angioletti from CNR-IPCB for his assistance in slicing the chestnuts.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodcont.2024.110878.

References

Afsah-Hejri, L., Akbari, E., Toudeshki, A., Homayouni, T., Alizadeh, A., & Ehsani, R. (2020). Terahertz spectroscopy and imaging: A review on agricultural applications. *Computers and Electronics in Agriculture, 177*, Article 105628.

Bernárdez, M. M.g., De la Monta na Miguélez, J., & Queijeiro, J. G. (2004). HPLC determination of sugars in varieties of chestnut fruits from Galicia (Spain). *Journal of Food Composition and Analysis, 17*(1), 63–67.

Bertuzzi, T., Rastelli, S., & Pietri, A. (2015). *Aspergillus* and *Penicillium* toxins in chestnuts and derived products produced in Italy. *Food Control, 50*, 876–880.

Blanchard, F., Razzari, L., Bandulet, H. C., Sharma, G., Morandotti, R., Kieffer, J. C., & Hegmann, F. A. (2007). Generation of 1.5 μJ single-cycle terahertz pulses by optical rectification from a large aperture ZnTe crystal. *Optics Express, 15*(20), 13212–13220.

Catapano, I., & Soldovieri, F. (2019). Chapter 11 - THz imaging and data processing: State of the art and perspective. In R. Persico, S. Piro, & N. Linford (Eds.), *Innovation in near-surface geophysics* (pp. 399–417). Elsevier.

Chen, Q., Zhang, C., Zhao, J., & Ouyang, Q. (2013). Recent advances in emerging imaging techniques for non-destructive detection of food quality and safety. *TrAC, Trends in Analytical Chemistry, 52*, 261–274.

Corona, P., Frangipane, M. T., Moscetti, R., Lo Feudo, G., Castellotti, T., & Massantini, R. (2021). Chestnut cultivar identification through the data fusion of sensory quality and FT-NIR spectral data. *Foods, 10*(11), 2575.

Di Girolamo, F. V., Pagano, M., Tredicucci, A., Bitossi, M., Paoletti, R., Barzanti, G. P., Benvenuti, C., Roversi, P. F., & Toncelli, A. (2021). Detection of fungal infections in chestnuts: A terahertz imaging-based approach. *Food Control, 123*, Article 107700.

Dobry, E., & Campbell, M. (2023). *Gnomoniopsis castaneae*: An emerging plant pathogen and global threat to chestnut systems. *Plant Pathology, 72*, 218–231.

Donis-Gonzalez, I. R., Guyer, D. E., Pease, A., & Fulbright, D. W. (2012). Relation of computerized tomography Hounsfield unit measurements and internal components of fresh chestnuts (*Castanea spp.*). *Postharvest Biology and Technology, 64*(1), 74–82.

Gennari, F., Pagano, M., Toncelli, A., Lisanti, M. T., Paoletti, R., Roversi, P. F., Tredicucci, A., & Giaccone, M. (2023). Terahertz imaging for non-invasive classification of healthy and cimiciato-infected hazelnuts. *Heliyon, 9*, Article e19891.

George, D. K., & Markelz, A. G. (2013). Terahertz spectroscopy of liquids and biomolecules. In , *Springer series in optical sciences: Vol. 171. Terahertz spectroscopy and imaging* (pp. 229–248). Springer.

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nat. Rev. Methods Primers, 2*, 100.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. Series c (applied statistics), 28*(1), 100–108.

Lee, Y.-S. (2009). *Principles of terahertz science and technology*. New York, USA: Springer.

Li, M., Yang, S., Peng, L., Zeng, K., Feng, B., & Jingjing, Y. (2022). Compositional shifts in fungal community of chestnuts during storage and their correlation with fruit quality. *Postharvest Biology and Technology, 191*, Article 111983.

Manca, R., Chiarantini, L., Tartaglia, E., Soldovieri, F., Miliani, C., & Catapano, I. (2023). Non-invasive characterization of maiolica layer structure by terahertz time-domain imaging. *Coatings, 13*, 1268.

Moscetti, R., Monarca, D., Cecchini, M., Haff, R. P., Contini, M., & Massantini, R. (2014). Detection of mold-damaged chestnuts by near-infrared spectroscopy. *Postharvest Biology and Technology, 93*, 83–90.

Mou, S., Rubano, A., & Paparo, D. (2017). Complex permittivity of ionic liquid mixtures investigated by terahertz time-domain spectroscopy. *Journal of Physical Chemistry B, 121*, 7351–7358.

Mou, S., Rubano, A., & Paparo, D. (2018). Broadband terahertz spectroscopy of imidazolium-based ionic liquids. *Journal of Physical Chemistry B, 122*, 3133–3140.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(1), 86–97.

Park, H., & Son, J.-H. (2021). Machine learning techniques for THz imaging and time-domain spectroscopy. *Sensors, 21*, 1186.

Ruocco, M., Lanzuise, S., Lombardi, N., Varlese, R., Aliberti, A., Carpenito, S., Woo, S. L., Scala, F., & Lorito, M. (2016). New tools to improve the shelf life of chestnut fruit during storage. *Acta Horticulturae, 1144*, 309–316.

Tielrooij, K. J., Paparo, D., Piatkowski, L., Bakker, H. J., & Bonn, M. (2009). Dielectric relaxation dynamics of water in model membranes probed by terahertz spectroscopy. *Biophysical Journal, 97*, 2484–2492.

Vettraino, A. M., Paolacci, A., & Vannini, A. (2005). Endophytism of Sclerotinia pseudotuberosa: PCR assay for specific detection in chestnut tissues. *Mycological Research, 109*(1), 96–102.

Washington, W., Allen, A., & Dooley, L. (1997). Preliminary studies on Phomopsis castanea and other organisms associated with healthy and rotted chestnut fruit in storage. *Australasian Plant Pathology, 26*(1), 37–43.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1–3), 37–52.