proceedings

MDPI

Proceedings

Multivariate Analysis for the Classification of Chocolate According to Its Percentage of Cocoa by Using Terahertz Time-domain Spectroscopy (THz-TDS) †

Jimy Oblitas [1,*], Jorge Ruiz [2]

[1] Universidad Privada del Norte; jimy.oblitas@upn.edu.pe
[2] Universidad Privada del Norte; jorge,ruiz@upn.pe
* Correspondence: jimy.oblitas@upn.edu.pe; Tel.: +51 976647476
† Presented at the 1st International Electronic Conference on Food Science and Functional Foods, 10–25 November 2020; Available online: https://foods_2020.sciforum.net/

Submitted: date; Accepted: date; Published: date

Abstract: Terahertz Time-domain Spectroscopy is a useful technique for determining some physical characteristics of materials, which is based on selective frequency absorption of a broad-spectrum electromagnetic pulse. In order to investigate the potential of this technology to classify cocoa percentages in chocolates, the terahertz spectra (0.5-10 THz) of 5 chocolate samples (50%, 60%, 70%, 80% and 90% of cocoa) were examined. The acquired data matrices were analyzed with the MATLAB 2019b application where the dielectric function was obtained along with the absorbance curves and was classified by using 24 mathematical classification models, achieving differentiations of around 93% obtained by the Gaussian SVM algorithm model with a kernel scale of 0.35 and a one-against-one multiclass method. It was concluded that the combined processing and classification of images obtained from the Terahertz Time-domain Spectroscopy and the use of machine learning algorithms can be used to successfully classify chocolates with different percentages of cocoa.

Keywords: Terahertz Spectroscopy; Multivariate analysis; Cocoa; chocolate

1. Introduction

The different spectroscopy techniques used in organic products have always explored ranges within the spectra: visible, ultraviolet and infrared, assessing how light-sensitive photoreceptors control many crucial biological processes [1], This boom in studies at this frequency is due to access to instruments that are available, but some spectra of the intermediate band or terahertz region (THz) are not totally studied and defined yet [2] showing great potential for uses in products of biological origin.

The so-called non-contact and non-destructive methods, such as NIR spectroscopy [3] and multispectral / hyperspectral images [4] , images in the visible range [5], RAMAN spectroscopy [6], have been widely used in the food sector as they are sensitive to intra-molecular vibration [7] and have increasingly been applied as a powerful analytical tool for determining food quality, as well as for identifying the geographical origin.

Although many of the spectroscopic techniques mentioned above have been used in the application of food detection, little attention has been paid to the use of Terahertz spectroscopy (THz), which is in a relatively unexplored range of the electromagnetic spectrum ranging from 0,1 to 10 THz, which lies between the mid-infrared and microwave ranges [8].

The composition of cocoa beans is directly influenced by genetic variability, geographical origin and processing. Therefore, chemical and biochemical characteristics and their relationship to external parameters are key characteristics for quality control and technological aspects [9], Currently, there are studies using near-infrared spectroscopy (NIRS) in the cocoa and chocolate industry [10], showing that it can detect differences but that there are still points for improvement, such as exploring other spectra. Here THz spectroscopy could provide information on time and frequency domains while being insensitive to background thermal radiation [11].

Looking for the applicability of this technology in the chocolate industry, the objective is to determine the level of differentiation of chocolate bars based on their percentage of cocoa in their composition by using THZ spectroscopy and multivariate analysis.

2. Materials and Methods

2.1. Raw Material

The cocoa genotype (Theobroma cacao L.) that was used is called "Marañon Native" and comes from the area of Cajamarca -Peru, which was used to make chocolate bars with 50%, 60%, 70%, 80% and 90% of cocoa in their composition. For this process, 10 samples were used for each percentage used. The bars used had dimensions of 10 cm × 10 cm with 0.5 mm of depth. This gave us an image for each sample. In total, 50 images were taken, with 2048 wavelengths, which gave us an average of 600 Mb per square centimeter analyzed.

2.2. Imaging Equipment in the THZ Range

Terahertz time-domain measurements were obtained by using a Terapulse 4000 spectrometer (Teraview Ltd., Cambridge, UK) in transmission mode. The transmission chamber, the operating scheme are shown in Figure 1. For its operation it was purged with dry nitrogen gas throughout the measurement and the noise was reduced with an average of 10 measurements. Each wave form in the time domain covered a range of 150 ps using a resolution of 0.1 ps. Images were built with equipment scanner.

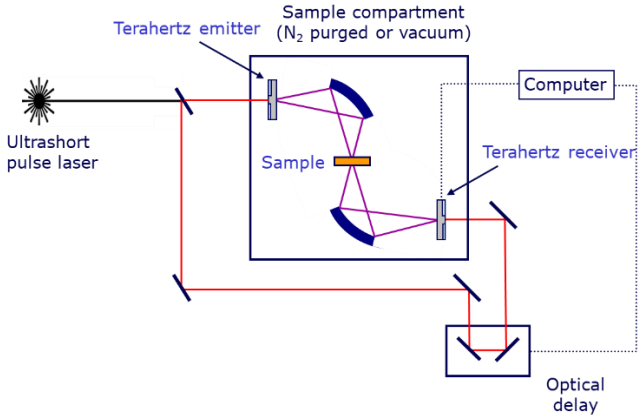


Figure 1. Terahertz pulsed spectroscopy operation schematic.

The data acquisition was performed in the TPRJ format and the images were analyzed by using codes internally developed in Matlab v.2019b (Mathworks, Massachusetts, USA).

2.3. Multivariate Analysis

The flow chart presented in Figure 2 shows the methodology used to determine the best classifier of chocolates based on their cocoa content. To identify the best classifier, 7 classification algorithms were used: *Decision Trees*, *Linear Discriminant Analysis* (LDA), *Quadratic Discriminant Analysis* (QDA), *Support Vector Machines* (SVM), *Nearest Neighbor Classifiers*, *Ensemble Classifiers* and *Naive Bayes*

Classifiers. Each of these algorithms was combined with its level of interpretability and flexibility, obtaining 24 models that are used in this research. The selection of discriminating variables was carried out using a feature selection technique based on the staggered decorrelation of variables.

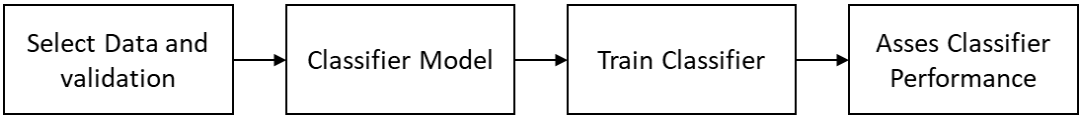


Figure 2. Supervised Classification.

Cross-validation was used to randomly divide the original data set of THZ spectra into a training set and test set, measuring the mean cross-validation error as a performance indicator. For the other parameters, a heuristic procedure was used to select the scale value based on the Kernel function to calculate the best classifier. The best model will be determined based on its Accuracy. Finally, once the best model is determined, the characteristics will be transformed with a PCA to reduce its dimensionality.

3. Results

3.1. Terahertz Imaging Analysis

In the experiment, chocolate samples were ordered in 10 cm × 10 cm trays and were subjected to reflection image measurements. Figure 3(a) show the transmitted time domain impulse and the Absorbance values for each type of chocolate. All images were pretreated with a linear filter to reduce image noise, as recommended by Shen [13]. The THz spectral image data set of the sample is based on specific parameters, such as the time interval, amplitude or the phase of the THz wave, and, then, it builds the refractive index, the spatial density distribution, the thickness distribution, and the sample contour.

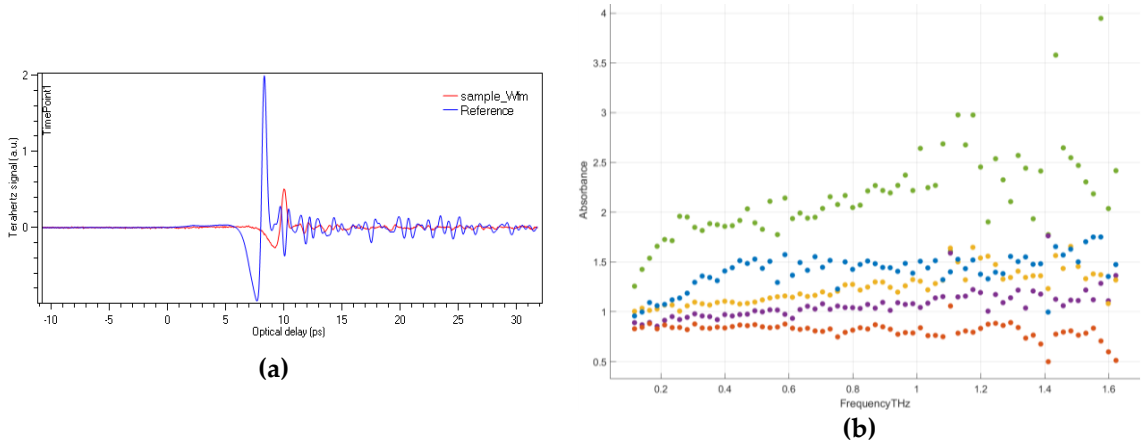


Figure 3. (**a**) Time-domain THZ Pulse; (**b**) Absorbance of the samples.

From each sample its absorption coefficient was obtained versus the THz frequency, which is shown in Figure 3(a), the absorption spectrum of chocolate samples was measured from 0.1 to 8 THz. In the respective analysis it was observed that the region of absorption spectra from 0.1 to 2.0 THz show a greater difference among the samples, and this was used for the multivariate analysis. It is also possible to observe in Figure 3(b) that the samples are divided into 5 groups, except for a slight overlap of the frequency of 1.6 THz. This confirms that terahertz spectra have enough information to classify different products based on their cocoa percentage.

3.2. Multivariate analysis

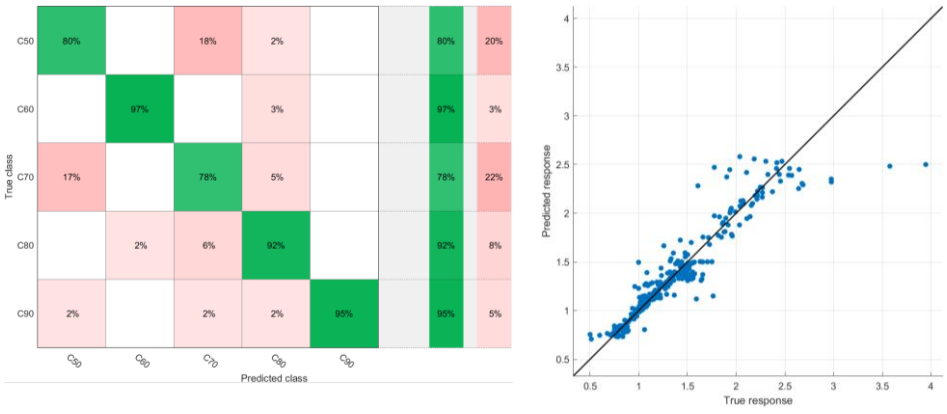
To achieve a proper classification, linear (LDA) and nonlinear (SVM) classification models were used. These models were made by using the Matlab Machine Learning application, which allowed us to explore the data set interactively, the selection of characteristics and specification of validation schemes. The training accuracy of the used models was evaluated by using the accuracy indicator (%). All models used a cross validation (15 folds). This PCA multivariate analysis generated the test of 24 models. The models with the best Accuracy are shown in Table 1.

Table 1. Models with the best accuracy.

Model	Accuracy (%)
Fine Gaussian SVM	91
Medium Gaussian SVM	90
Quadratic Discriminant	89
Optimizable SVM	93

The best model was the optimized model of Fine Gaussian SVM which obtained an Accuracy of 93%, with a Kernel Scale of 1 and cubic function type and a Multiclass Method One vs One, optimized with a Bayesian function of 30 interactions. This type of model has been reported many times in research works on the use of Machine Learning for image recognition [14].

Figure 3(a) shows the confusion matrix of the model. The coefficients for this PCA application are PC1 (63.8%) and PC2 (36.2%). In addition, a prediction model was adjusted, obtaining a RMSE of 0.171751 with a function of SVM type, which is shown in Figure 4(b).



(**a**) (**b**)

Figure 4. (**a**) Confusion Matrix; (**b**) Prediction

The use in the field of chocolate production has been given as a great potential in THZ spectroscopy. A study by Catapano [15] shows work on quality control of chocolate bars contaminated with foreign objects, where THz technology showed a great ability to detect and discriminate different types of materials based on their composition. Techniques for carrying out these quality processes, especially based on their composition, have always used techniques such as mass spectroscopy [16], so the importance of evaluating novel techniques such as Time-domain Spectroscopy become necessary, especially by assessing the importance of non-destructive inspection for the food industry, meeting the needs of modern and rapid techniques for international trade in food.

THz waves have the ability to penetrate a wide variety of materials, and vibration and rotational energy levels of most biological systems are in the THz band [17]. Currently, THz-TDS remains an expensive technology. However, the recent and rapid development of THz systems for agro-

industrial research opens up real possibilities for these costs to be significantly reduced in the coming years.

4. Conclusion

The overall results show that Terahertz time-domain spectroscopy together with classification modeling can successfully identify the composition of chocolate bars based on their cacao percentage. Along with this, the ability of this technique to characterize the molecular structure of many biological substances, makes them an attractive analytical process tool for better monitoring in food quality control. But while this Terahertz time-domain spectroscopy is demonstrating efficiency in classification methods, as in chocolate, there are still many parameters to take into account in the use of this type of technology.

Author Contributions: Individual contributions for the authors are as follows: conceptualization, J.O. and J.R.; methodology, J.O.; software, J.R. and J.O.; validation, J.R.; formal analysis, J.O.; resources, J.R.; data curation, J.R.; writing–original draft preparation, J.O.; writing–review and editing, J.O and J.R.; visualization, J.O and J.R.; supervision, J.O.; project administration, J.O.; funding acquisition, J.O.

Funding: The autors acknowledges the financial support of the Project Concytec - World Bank "Development of Predictive Models of Food Quality Based on THz Imaging Technology", through its executing unit Fondecyt. [contract number 006-2018-FONDECYT/BM-Mejoramiento de la infraestructura para la investigación (equipamiento)]

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Song, S.-H.; Dick, B.; Penzkofer, A.; Pokorny, R.; Batschauer, A.; Essen, L.-O. Absorption and fluorescence spectroscopic characterization of cryptochrome 3 from Arabidopsis thaliana. *J. Photochem. Photobiol. B: Biol.* **2006**, *85*, 1–16, doi:10.1016/j.jphotobiol.2006.03.007.
2. Singh, A.K.; Pérez-López, A.V.; Simpson, J.; Castro-Camus, E. Three-dimensional water mapping of succulent Agave victoriae-reginae leaves by terahertz imaging. *Sci. Report* **2020**, *10*, 1404, doi:10.1038/s41598-020-58277-z.
3. Uddin, M.N.; Ferdous, T.; Islam, Z.; Jahan, M.S.; Quaiyyum, M.A. Development of chemometric model for characterization of non-wood by FT-NIR data. *J. Bioresour. Bioprod.* **2020**, *5*, 196–203, doi:10.1016/j.jobab.2020.07.005.
4. De-la-Torre, M.; Avila-George, H.; Oblitas, J.; Castro, W. Selection and Fusion of Color Channels for Ripeness Classification of Cape Gooseberry Fruits. *Advance. Intell. Syst. Comput.* **2020**, *1071*, 219–233, doi:10.1007/978-3-030-33547-2_17.
5. Castro, W.; Oblitas, J.; Chuquizuta, T.; Avila-George, H. Application of image analysis to optimization of the bread-making process based on the acceptability of the crust color. *J. Cereal Sci.* **2017**, *74*, 194–199, doi:10.1016/j.jcs.2017.02.002.
6. Xu, Y.; Zhong, P.; Jiang, A.; Shen, X.; Li, X.; Xu, Z.; Shen, Y.; Sun, Y.; Lei, H. Raman spectroscopy coupled with chemometrics for food authentication: A review. *TrAC Trend Anal. Chem.* **2020**, *131*, 116017, doi:10.1016/j.trac.2020.116017.
7. Wang, K.; Sun, D.-W.; Pu, H. Emerging non-destructive terahertz spectroscopic imaging technique: Principle and applications in the agri-food industry. *Trend Food Sci. Technol.* **2017**, *67*, 93–105, doi:10.1016/j.tifs.2017.06.001.
8. Ferguson, B.; Zhang, X.-C. Materials for terahertz science and technology. *Nat. Mat.* **2002**, *1*, 26–33, doi:10.1038/nmat708.
9. Ferreira de Oliveira, A.P.; Milani, R.F.; Efraim, P.; Morgano, M.A.; Tfouni, S.A.V. Cd and Pb in cocoa beans: Occurrence and effects of chocolate processing. *Food Control.* **2021**, *119*, 107455, doi:10.1016/j.foodcont.2020.107455.
10. Barbin, D.F.; Maciel, L.F.; Bazoni, C.H.V.; Ribeiro, M. da S.; Carvalho, R.D.S.; Bispo, E. da S.; Miranda, M. da P.S.; Hirooka, E.Y. Classification and compositional characterization of different varieties of cocoa beans by near infrared spectroscopy and multivariate statistical analyses. *J. Food Sci. Technol.* **2018**, *55*, 2457–2466, doi:10.1007/s13197-018-3163-5.

11. Liu, W.; Liu, C.; Yu, J.; Zhang, Y.; Li, J.; Chen, Y.; Zheng, L. Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics. *Food Chem.* **2018**, *251*, 86–92, doi:10.1016/j.foodchem.2018.01.081.
12. Oblitas, J.; De-la-Torre, M.; Avila-George, H.; Castro, W.; De-la-Torre, M.; Avila-George, H.; Castro, W. The Use of Correlation, Association and Regression Techniques for Analyzing Processes and Food Products Available online: https://www.taylorfrancis.com/ (accessed on Sep 28, 2020).
13. Shen, X.; Dietlein, C.R.; Grossman, E.; Popovic, Z.; Meyer, F.G. Detection and Segmentation of Concealed Objects in Terahertz Images. *IEEE Trans. Image Process.* **2008**, *17*, 2465–2475, doi:10.1109/TIP.2008.2006662.
14. Loussaief, S.; Abdelkrim, A. Machine Learning framework for image classification. *Advance Sci. Technol. Eng. Syst.* **2018**, *3*, 1–10, doi:10.25046/aj030101.
15. Catapano, I.; Soldovieri, F. Chapter 11 - THz imaging and data processing: State of the art and perspective. In *Innovation in Near-Surface Geophysics*; Persico, R., Piro, S., Linford, N., Eds.; Elsevier, 2019; pp. 399–417 ISBN 978-0-12-812429-1.
16. Humston, E.M.; Knowles, J.D.; McShea, A.; Synovec, R.E. Quantitative assessment of moisture damage for cacao bean quality using two-dimensional gas chromatography combined with time-of-flight mass spectrometry and chemometrics. *J. Chromatogr. A* **2010**, *1217*, 1963–1970, doi:10.1016/j.chroma.2010.01.069.
17. Wang, C.; Zhou, R.; Huang, Y.; Xie, L.; Ying, Y. Terahertz spectroscopic imaging with discriminant analysis for detecting foreign materials among sausages. *Food Control.* **2019**, *97*, 100–104, doi:10.1016/j.foodcont.2018.10.024.