Machine Learning-Based Terahertz Spectroscopy for Starch Concentration Prediction in Biofilms

Juan-Jesús Garrido-Arismendis[1], Jimy Oblitas[2],
César Niño[3], Himer Avila-George[4]*, Wilson Castro[5]
Facultad De Ingeniería De Industrias, Alimentarias Y Biotecnología, Universidad Nacional De Frontera, Sullana, Perú[1,5]
Facultad De Ingeniería, Universidad Privada Del Norte, Cajamarca, Perú[2]
Facultad De Ingeniería Industrial, Universidad Nacional De Piura, Piura, Perú[3]
Departamento De Ciencias Computacionales E Ingenierías, Universidad De Guadalajara, Ameca, México[4]

***Abstract*—Food preservation and safety require advanced detection methods to ensure transparency in supply chains. Terahertz (THz) spectroscopy has emerged as a powerful, non-invasive tool for material characterization. This study explores the integration of THz spectroscopy and machine learning for accurately quantifying maize starch adulteration in bioplastics derived from potato starch. Bioplastic samples with varying concentrations of maize starch were prepared, molded into three different thicknesses, and subjected to a two-stage drying process, resulting in 81 samples (27 treatments with three replicates each). The spectral profiles at THz (0.5 to 2 THz) were recorded and analyzed using three regression models: support vector regression, partial least squares regression, and multiple linear regression. The models were evaluated using the coefficient of determination (R^2), Root Mean Square Error (RMSE), and the Residual Predictive Deviation (RPD). The results showed R^2 values ranging from 0.7283 to 0.9495, RMSE between 0.0594 and 0.1393, and RPD values from 1.8753 to 4.4479, demonstrating strong predictive performance. These findings highlight the potential of THz spectroscopy and machine learning in the noninvasive detection of starch adulterants in bioplastics, paving the way for future research to enhance model robustness and applicability.**

***Keywords*—Terahertz spectroscopy; machine learning; chemometrics; starch detection; biofilms**

I. Introduction

Every year, approximately 1.3 billion tons of by-products from the global agri-food industry pile up, creating substantial economic and environmental pressures [1]. Many of these by-products hold untapped potential, containing valuable bioactive compounds such as starch—a carbohydrate recognized by [2] and [3] as essential for human and animal nutrition. The versatility of starch, mainly composed of amylose and amylopectin, significantly influences its industrial applications due to distinct functional properties highlighted in studies by [4], [5], and [6]. Yet, despite its promise, starch faces inherent limitations, including low thermal stability and pronounced hydrophilicity, restricting its broader industrial adoption [7].

Responding to escalating environmental concerns, starch-based bioplastics have surfaced as compelling alternatives to traditional petroleum-derived plastics. These innovative materials, praised by researchers like [8], [9], and [10] for their biodegradability and compostability, offer practical, eco-friendly solutions particularly suited for food packaging. Nonetheless, maintaining high-performance standards in bioplastics is complex, as accurate assessments of their composition [11] and structural integrity [12] are critical.

Traditional methods for starch characterization are often invasive and labor-intensive, risking alteration or damage to sample integrity. Terahertz (THz) spectroscopy, as presented in works by [13], [14], and [15], emerges as a promising alternative, operating in the unique 0.1–10 THz frequency range and providing insightful, non-destructive material characterization. Specifically, Time-Domain Terahertz Spectroscopy (THz-TDS) has garnered attention within food science, enabling detailed biopolymer analysis without sample degradation, as shown by [16] and [17]. Chemometric techniques, integrating statistical and machine learning methods, significantly improve the interpretation of complex spectral data, thereby dramatically enhancing starch identification and quantification in bioplastics [18].

Complementing traditional chemometric approaches, recent breakthroughs in deep learning are revolutionizing analysis across various sectors. Intelligent methods have significantly improved waste management by optimizing material classification [19]. Likewise, advancements in agricultural practices have been achieved through sophisticated algorithms and IoT integration [20]. Metaheuristic approaches have accelerated neural network hyperparameter tuning [21], and innovative machine learning techniques have enhanced cybersecurity through efficient data filtering [22]. Additionally, machine learning advancements continue refining the precision of GPS positioning [23]. While our current study employs traditional machine learning frameworks, future integration of advanced AI methods could further refine THz spectral analyses, optimizing feature selection and enhancing predictive accuracy.

Considering this context, THz-TDS spectroscopy integrated with chemometric methods has proven effective in the non-invasive characterization of polymers, though its application to starch-based biopolymers remains limited. To our knowledge, this research represents the first effort to combine THz spectroscopy with machine learning to predict potato and maize starch concentrations in bioplastics. Here, we propose an approach integrating spectral analysis of THz signals with three machine learning models: Support Vector Regression (SVR), Partial Least Squares Regression (PLSR), and Multiple Linear Regression (MLR). Furthermore, a feature selection method was employed to optimize these models, aiming to enhance

*Corresponding authors.

predictive accuracy. This new approach expects to improve quality assessment in sustainable packaging, contributing to advancements in environmentally friendly industrial materials.

The remainder of this paper is organized as follows: Section II details the methodology for the preparation of bioplastic samples and the application of THz spectroscopy. Section III presents the experimental results, including the performance of the regression models used for starch concentration prediction. Finally, Section IV provides concluding remarks on the implications of this study for sustainable bioplastic development.

II. Material and Methods

This section describes the methodology employed for preparing and characterizing bioplastics, primarily using starch and polyvinyl alcohol as foundational materials. The bioplastics were synthesized through the solution casting method, adapting protocols previously detailed by [24], [25], and [26]. An overview of the methodological steps is illustrated in Fig. 1, with each stage further detailed in subsequent subsections.

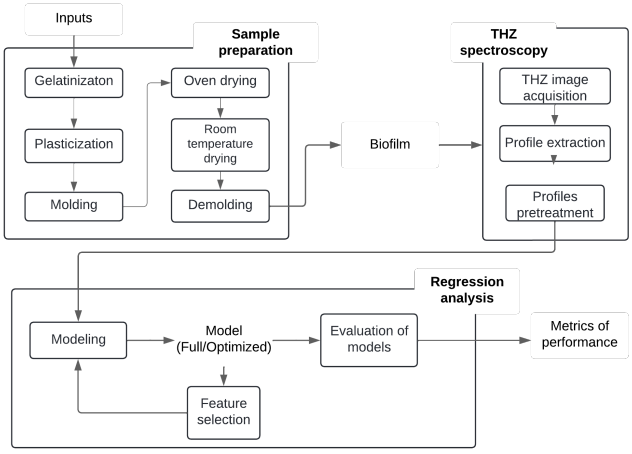


Fig. 1. Flow Diagram of the experimental methodology used for the preparation and analysis of starch-based bioplastics.

A. Sample Preparation

The inputs used to prepare the samples were high-purity potato starch and maize starch, purchased online from Peruvian suppliers through the Mercado Libre platform. Additionally, distilled water, technical grade glycerin (97% purity), and laboratory-grade polyvinyl alcohol (98% purity) were used, all purchased from a laboratory supply store located in the district of Sullana, province of Sullana, department of Piura. All activities of the experimental scheme were carried out in the food safety research laboratory of the National University of the Frontier.

For the preparation of the bioplastics, 12 grams of starch were used in each formulation. In this study, a base bioplastic with potato starch was formulated (control sample), and eight additional bioplastics were made, in which potato starch was partially substituted with maize starch in proportions ranging from 10% to 80% in increments of 10%. These starch mixtures were manually and meticulously prepared to ensure a uniform distribution of the components, thus guaranteeing consistent and reproducible results in subsequent experiments [27].

The initial chemical process for sample preparation involved gelatinization. In this step, 12 grams of the starch mixture were dissolved in 400 ml of distilled water and heated to 100°C for 45 minutes. According to the methodology described by [28], these conditions are optimal for breaking down starch granules without causing their denaturation, thereby enabling them to swell and rupture to form a gelatinous paste. Subsequently, for plasticization, the temperature was lowered to 80°C for an additional 15 minutes. At this stage, 7 ml of glycerin and 8 grams of polyvinyl alcohol, previously dissolved in 100 ml of water, were added. This combination, as highlighted by [29] and [30], effectively reduces material fragility, enhances flexibility, and improves tensile strength. The mixture was stirred to distribute the plasticizers evenly.

Subsequently, the molding process followed, where the plasticized mixture was poured into Petri dishes with a diameter of 9 cm in amounts of 12 ml, 15 ml, and 18 ml. The precision in the molding is crucial to obtain comparable samples and avoid unwanted variations in experimental results [31]. The samples were dried in an oven at 45°C for 22 hours to reduce the water content. This step is important to prevent cracking or rapid deformation [32]. Subsequently, they were subjected to a second drying at room temperature (24°C) for 48 hours in a silica gel desiccator to remove residual moisture, ensuring dimensional stability and suitable mechanical properties for analysis [33].

After the second drying process, the samples were carefully demolded using a scalpel, tweezers, and surgical gloves to avoid damage or deformation, resulting in smooth and defect-free samples ready for evaluation. A total of 81 bioplastic sheets were produced (27 treatments with three replicas each). Each sheet was cut into rectangles of 15 mm x 45 mm, and their thickness was measured using a Dasqua digital micrometer with a range of 0-25 mm and a resolution of 0.001 mm. Five measurements were taken at different points on each sheet, and the values were averaged.

B. THz Spectroscopy

The bioplastic sheets were placed on a polylactic acid (PLA) sample holder for analysis. A TeraSmart Compact Industry-Proven THz spectrometer of German origin was used in transmission mode. This system has a scanning range of 850 ps and includes a compact spectrometer with a spectral range of 6 THz and a resolution of 1.2 GHz; it is equipped with an ultrafast laser that emits femtosecond pulses, and the signal is directed through a system of nonlinear cyclic optical mirrors (Fig. 2), connected to the spectrometer via a fiber optic cable. A tower with vertical and horizontal displacement capabilities was used to move the sample. Image acquisition was controlled using TeraImage and Scam Control software, which allows defining and adjusting the appropriate scanning range. Menlo Systems provides both the equipment and the software.

The experimental phase was conducted under normal atmospheric conditions, which generated many peaks due to the strong absorption characteristics of water vapor in the THz range, which can interfere with measurements [34]. It

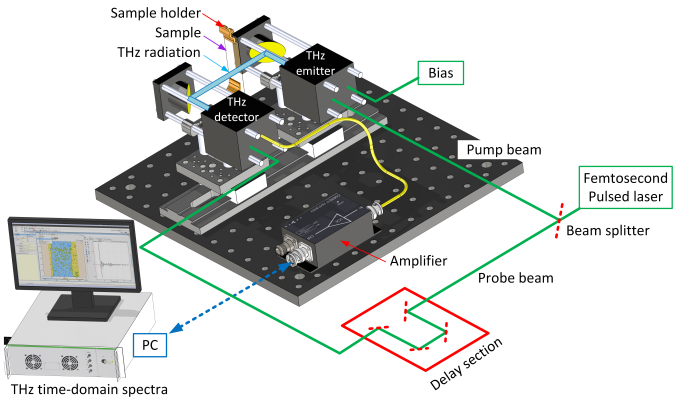


Fig. 2. Schematic representation of the transmission-mode THz-TDS system used in this study.

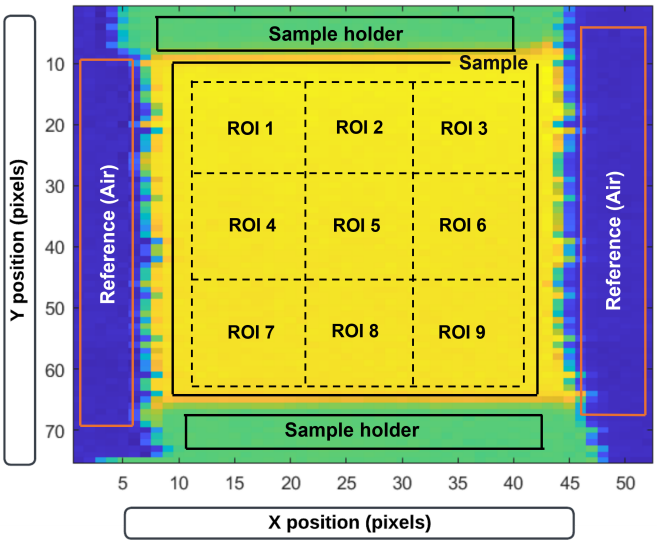


Fig. 3. Representative transmittance image showing the contrast between the bioplastic sample area and the reference.

is important to note that the signals obtained were measured with a relative humidity close to 50%.

Initially, the THz spectrometer generated files in the `IGTIFF` format, which were converted to the `MAT` format using Epina ImageLab software. The resulting files were loaded into Matlab (version R2024a, The MathWorks, Inc., USA), where high-contrast images were generated (Fig. 3) to distinguish the sample, the sample holder, and the air. This facilitated the acquisition of profiles for the sample (bioplastic film) and reference (air). To obtain the profiles of interest, the THz image was divided into nine equal-sized subareas (ROIs), from which the average profile was extracted for further processing. A total of 729 profiles were generated in the time domain, which was then transformed into the frequency domain using a Fourier transform, employing Eq. 1. These profiles in the frequency domain were used for the regression analysis.

$$E(t) \rightarrow \text{FFT} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E(t) e^{-i\omega t}\, dt = E(\omega), \quad (1)$$

where $E(t)$ represents the signal function in the time domain, $e^{-i\omega t}$ is the kernel of the transform, and $E(\omega)$ denotes the signal function in the frequency domain.

C. Regression Analysis

The THz profiles in the frequency domain (predictor variable: X) and the maize starch concentration values (response variable: Y) were used to train three regression models: MLR, SVR, and PLSR. The models used are detailed below.

1) MLR: It is a statistical technique that estimates the relationship between a dependent variable and several independent variables using a linear equation [35]. This multivariate statistical method restructures the original dataset into linear combinations of the variables, creating independent new variables known as principal components that capture most of the variability [36]. This model is based on Eq. 2.

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon, \quad (2)$$

where Y represents the starch percentage, β_0 is the constant term, $\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients, $X_1, X_2, \ldots, X_n$ correspond to the THz profiles in the frequency domain, and ϵ denotes the error term.

2) SVR: The Support Vector Machine for Regression examines the relationship between variables using a subset of data, balancing the complexity of the model with the precision of prediction in complex scenarios [37]. Unlike conventional machine learning approaches, the SVR model effectively handles issues related to small sample sizes, high dimensionality, and local minima and is noted for its remarkable ability to generalize [38].

3) PLSR: This technique, common in multivariate analysis, simplifies the relationship between multiple variables by projecting them onto orthogonal vectors, thus facilitating understanding [39]. It is used primarily in chemometrics to investigate how spectral data correlate with reference indicators [40]. PLSR transforms predictor variables (X) into response variables (Y). It decomposes X and Y and projects them into new directions to capture joint variability [41]. Then, a regression is performed with these decomposed variables, as shown in the model of Eq. 3.

$$Y = \beta X + e, \quad (3)$$

where Y represents the starch concentration in the bioplastics, X is the intensity data matrix (n observations × m frequencies), β is the coefficient matrix, and e denotes the error term.

It is essential to eliminate irrelevant spectral information, as this complicates the development of simple and effective models [42]. For this reason, the method of feature selection using beta coefficients (β) was chosen, which are associated

with frequency values and absolute loadings in regression models. These coefficients were selected for their ability to adequately represent the dependent variable, contributing to improved model accuracy [41].

The performance of the MLR, SVR, and PLSR models was evaluated using the metrics R^2, RMSE, and RPD (see Eq. 4, 5, and 6).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}, \tag{4}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2}, \tag{5}$$

$$\text{RPD} = \frac{\mathcal{S}}{\text{RMSE}}, \tag{6}$$

where Y_i represents the reference concentration of the i-th instance, $\overline{Y}$ is the mean value of the reference concentrations, $\hat{Y}_i$ denotes the predicted concentration of the i-th instance, N is the number of instances, and $\mathcal{S}$ corresponds to the standard deviation of the reference values.

Finally, cross-validation was implemented using a five-fold strategy with 30 iterations. In each iteration, the dataset was partitioned into five subsets: one used for testing and the remaining four for training. Performance metrics were calculated for each iteration, following the procedure described in [42], [43]. This validation approach is essential for assessing model performance, as it ensures robustness and generalization by training and evaluating the models across different data splits [44]. Using multiple partitions reduces the risk of overfitting and prevents dependency on a specific training–validation division [45]. Additionally, this strategy increases the consistency and reliability of the predictive results in diverse scenarios [46].

III. Results

A. Bioplastic Obtention

Fig. 4 shows the control bioplastic and its variants with different levels of maize starch adulteration (10% — 80%) and thicknesses (E1 = 0.12 mm, E2 = 0.15 mm, E3 = 0.18 mm). Visually, the samples appear similar, although this uniform appearance does not necessarily reflect their differences in biodegradability. Previous studies indicate that bioplastics made solely with maize starch tend to degrade more slowly than those made with other types of starch [47]. Furthermore, the choice of starch and plasticizers can significantly affect the physicochemical properties of bioplastics [48]. This is consistent with similar research that also used potato starch and found variations in physical properties based on the formulation [49]. Therefore, while the appearance may be uniform, the properties and degradation can vary depending on the composition and plasticizers used.

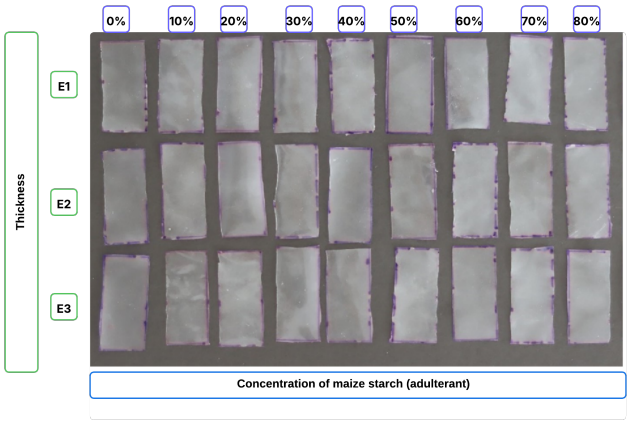
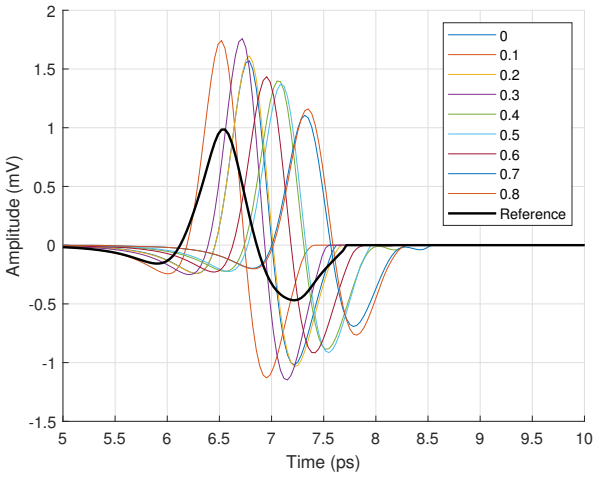


Fig. 4. Bioplastic sheets formulated with varying maize starch concentrations and molded at three different thicknesses.

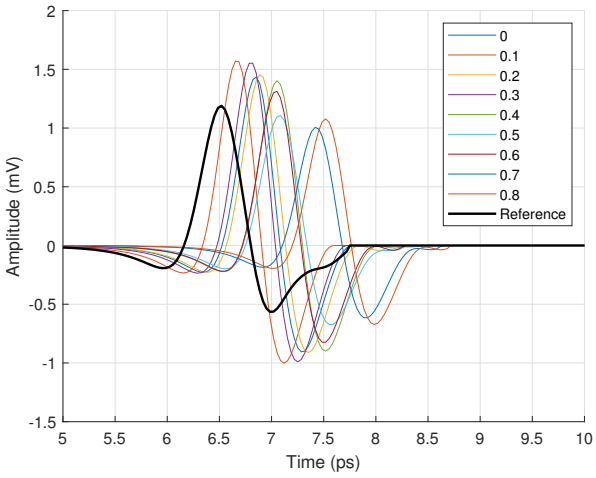
B. THz Spectral Analysis

1) Profiles in the time domain: Fig. 5 presents the average profiles in the time domain of bioplastics with nine concentrations and three thicknesses in the range of 5 to 10 picoseconds. These graphs show how starch concentrations affect the amplitude and arrival time of THz pulses. The echoes generated by multiple and internal reflections within the sample were removed to analyze the main signal free of interference. These reflections are related to the Fabry-Pérot effect [50]. After removing echoes from multiple reflections, it was observed that as the thickness increases, the absorption of the THz signal rises along with the attenuation, indicating a more effective interaction between the THz signal and starch. The length of the THz signals obtained in the experiments ranged between five and nine picoseconds, with each signal averaged from three measurements to improve the signal-to-noise ratio.

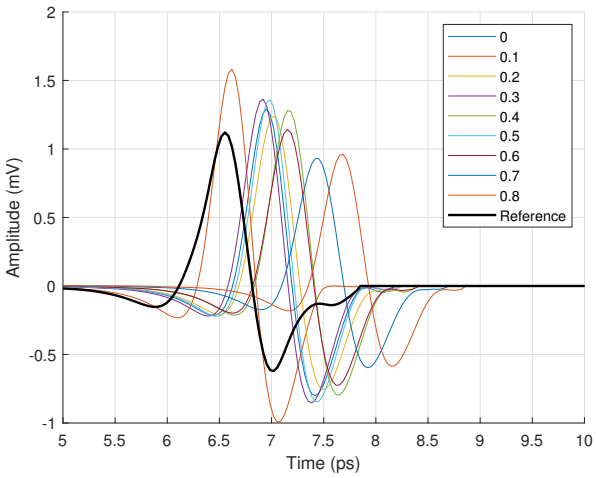
2) Profiles in the frequency domain: Fig. 6 presents the average frequency domain profiles of bioplastics with nine different concentrations and three thicknesses, covering the range from 0.5 to 2 Terahertz. These profiles were obtained by removing the Fabry-Pérot term from the time-domain data, as noted by [51], where reflection signals can merge in thin samples, and the main reflection echoes may be lost. The greater differentiation observed in the 0.5 to 2 THz range aligns with previous reports on sensitivity in thin samples, ranging from 0.5 to 1.5 THz [52], as well as with studies on organic samples reporting sensitivity to THz between 0.1 and 1.4 THz [53]. This frequency range was also chosen in similar studies such as [54], which analyzed bacterial cellulose films from 0.3 to 2.8 THz, and [55], which evaluated food-grade oils from 0.5 to 3 THz. However, other studies have used different ranges, such as [56], which measured the elasticity of poly-l-proline helices from 0.6 to 4.5 THz, and [57], which classified inorganic pigments in 0.1 to 1.2 THz. This indicates that while the 0.5 to 2 THz range is typical, the choice of frequency range is tailored to the specific characteristics of the material and the objectives of the study.



(a) E1

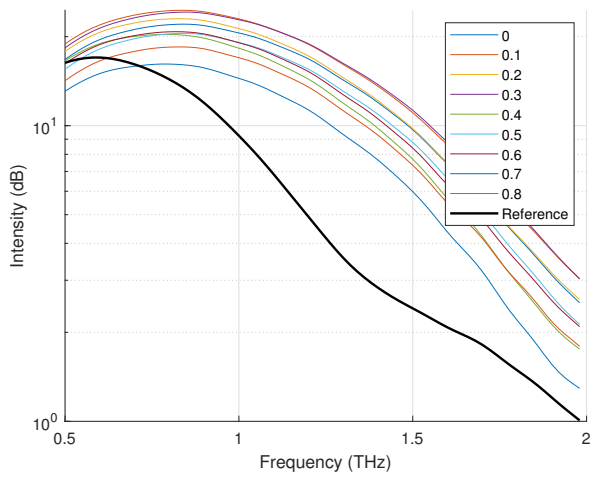


(b) E2

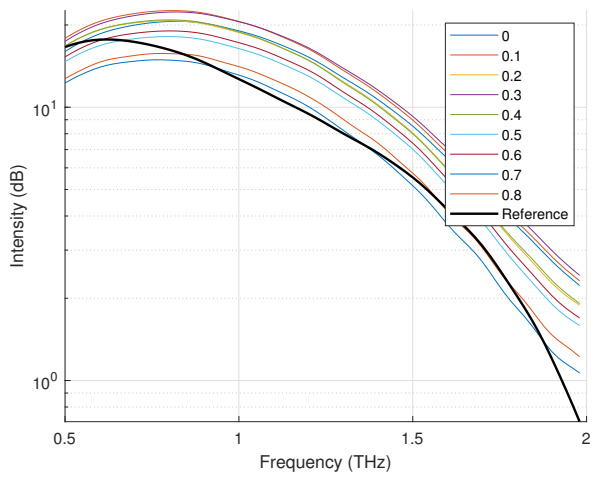


(c) E3

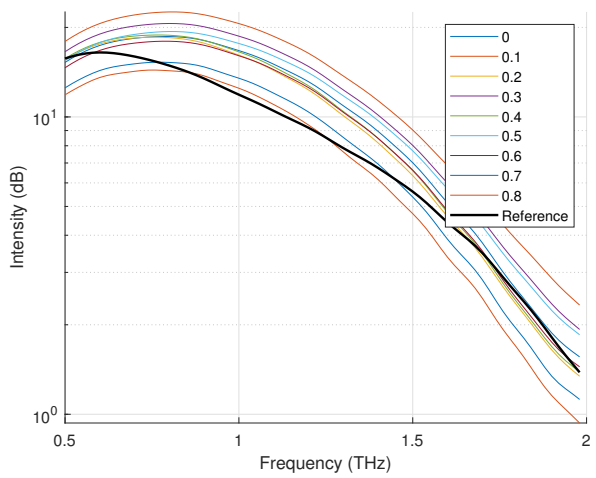
Fig. 5. Average time-domain THz profiles of bioplastic samples with varying maize starch concentrations and three film thicknesses.



(a) E1



(b) E2



(c) E3

Fig. 6. Average frequency-domain THz profiles of bioplastic samples with varying maize starch concentrations and three film thicknesses.

C. THz Profile Modeling

Table I presents the plots of the actual vs. predicted values for the full and optimized models for each thickness. For SVR, the R^2 values are 0.9158 ± 0.0041 (full model) and 0.8149 ± 0.0024 (optimized model) for E1, 0.9028 ± 0.0032 and 0.8041 ± 0.0022 for E2, and 0.8733 ± 0.0042 and 0.7283 ± 0.0016 for E3. Similarly, for MLR, the R^2 values are 0.9495 ± 0.0041 (full model) and 0.8245 ± 0.0025 (optimized model) for E1, 0.9191 ± 0.0066 and 0.8528 ± 0.0030 for E2, and 0.8841 ± 0.0160 and 0.7807 ± 0.0033 for E3. Likewise, for PLSR, the R^2 values are 0.8503 ± 0.0005 (full model) and 0.8213 ± 0.0004 (optimized model) for E1, 0.8379 ± 0.0007 and 0.8517 ± 0.0003 for E2, and 0.8342 ± 0.0007 and 0.7768 ± 0.0004 for E3.

All three models demonstrated good performance in predicting the maize starch concentration in bioplastic samples, particularly in the lower thicknesses. These models, combined with THz spectroscopy, have been widely applied in various studies of functional and organic material characterization. For instance, in the work of [58], SVR ($R^2 = 0.9793$) was used to predict bovine serum albumin concentration in thin films. Similarly, in [59], PLSR and SVR ($R^2 = 0.994$ for both models) were applied to analyze the amount of α-lactose in a lotus root starch mixture. Furthermore, [60] evaluated the microstructural characteristics of thermal coatings (MLR, $R^2 = 0.97$). On the other hand, [61] used SVR to analyze porosity in fiberglass-reinforced polymers ($R^2 = 0.976$), and [62] employed MLR to measure the coating thickness in nifedipine tablets ($R^2 = 0.99$). Furthermore, [63] used MLR to predict the density ($R^2 = 0.97$) and moisture content ($R^2 = 0.78$) in wood, using refractive indices and absorption coefficients. Finally, [64] applied PLSR to predict glycerol concentration in liquid solutions (RPD = 6.095). In most cases, the models demonstrated high performance, confirming the feasibility of using chemometric models combined with THz spectroscopy for material characterization.

The results demonstrate competitive performance compared to previous studies that have used machine learning models combined with THz spectroscopy for analyzing organic samples, reinforcing both the applicability and robustness of the proposed approach. While some prior studies reported slightly superior outcomes, these differences can be primarily attributed to lower variability in the composition of their samples. Nevertheless, the high precision achieved in our study for predicting starch concentrations in bioplastics clearly illustrates the effectiveness of the proposed methodology. Recent research by [65] indicates that integrating deep learning methods can substantially enhance predictive accuracy and improve interpretability by identifying informative spectral bands. Furthermore, [66] highlights the capability of deep learning to effectively model complex data structures, suggesting promising potential for further improvements in predictive precision in future THz spectroscopy applications.

D. Performance Metrics

Table II shows the average performance metrics (R^2, RMSE and RPD) with their standard deviations for the SVR, MLR, and PLSR models, both in their complete and optimized versions, applied to three different thicknesses of bioplastic films. The complete SVR, MLR, and PLSR models accurately predicted maize starch concentration. For thickness E1, R^2 values ranged from 0.8503 ± 0.0005 to 0.9495 ± 0.0041, RMSE values from 0.0594 ± 0.0025 to 0.0999 ± 0.0002, and RPD values from 2.5847 ± 0.0044 to 4.4479 ± 0.1761. For E2, R^2 values ranged from 0.8379 ± 0.0007 to 0.9191 ± 0.0066, RMSE values from 0.0754 ± 0.0032 to 0.1040 ± 0.0002, and RPD values from 2.4835 ± 0.0053 to 3.5020 ± 0.1546. For E3, R^2 values ranged from 0.8342 ± 0.0007 to 0.8841 ± 0.0160, RMSE values from 0.0891 ± 0.0057 to 0.1051 ± 0.0002, and RPD values from 2.4556 ± 0.0050 to 3.1874 ± 0.1488. Among these models, the complete MLR model performed the best in all thicknesses.

For the optimized models, both PLSR and MLR showed strong performance, with MLR performing slightly better in most cases. For E1 (R^2 = 0.8245 ± 0.0025, RMSE = 0.1091 ± 0.0007, RPD = 2.4029 ± 0.0174); for E2 (R^2 = 0.8528 ± 0.0030, RMSE = 0.0996 ± 0.0011, RPD = 2.6311 ± 0.0284); and for E3 (R^2 = 0.7807 ± 0.0033, RMSE = 0.1218 ± 0.0011, RPD = 2.1491 ± 0.0157). Interestingly, optimizing the models using beta coefficients sometimes led to slightly decreased performance metrics. Although these coefficients are useful for selecting important variables, as mentioned in [42], they can slightly lower the performance of the model.

In general, the study highlights the impact of the selection of features on the effectiveness of starch prediction models in bioplastics. Although the MLR and PLSR models showed promising results, the drop in performance metrics after optimization suggests that exploring other feature selection methods could be beneficial. Trying different approaches may improve the models and help them find broader use in industrial applications, ultimately advancing bioplastic analysis and production.

IV. Conclusion

This study demonstrates that integrating THz spectroscopy with machine learning offers a promising, non-invasive approach for predicting bioplastic starch concentration. By applying regression models such as PLSR, SVR, and MLR, we achieved high predictive accuracy—particularly with the optimized MLR model, which performed well even with a relatively small dataset. Nevertheless, due to the variability in starch formulations and the precision required for industrial applications, more extensive and diverse datasets will be essential to enhance the generalizability of the models.

The use of beta coefficients in the spectral analysis proved effective for identifying key frequency features in the THz spectrum. This approach supports the potential development of compact, cost-effective systems for real-time starch monitoring during bioplastic production. Such feature selection methods are especially useful in the packaging industry, where rapid and accessible quality control tools are highly valuable. Future work could involve implementing more advanced feature selection strategies to improve model performance further.

Moreover, the proposed methodology can be extended to other quality control applications involving bioplastics and biodegradable materials, contributing to developing sustainable and high-performance industrial solutions.

TABLE I. COMPARING REAL VERSUS PREDICTED MAIZE STARCH CONCENTRATIONS IN BIOPLASTIC SAMPLES, USING SVR, MLR, AND PLSR MODELS UNDER THREE THICKNESS CONDITIONS: E1 (0.12 MM), E2 (0.15 MM), AND E3 (0.18 MM)

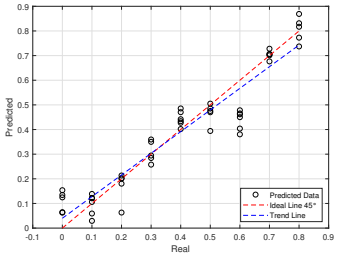
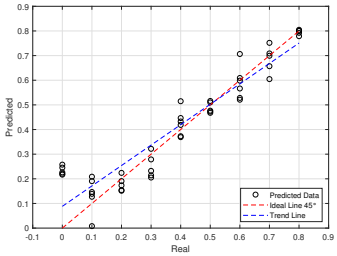
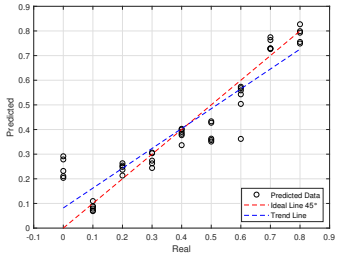
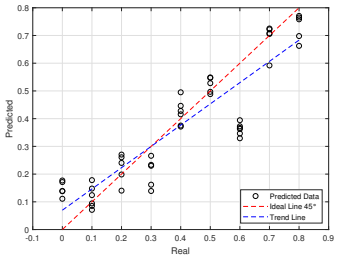
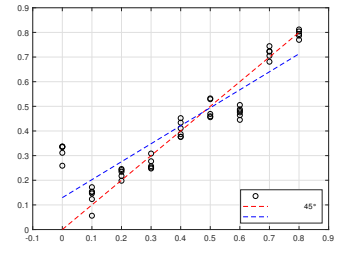
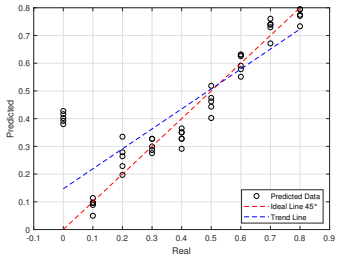
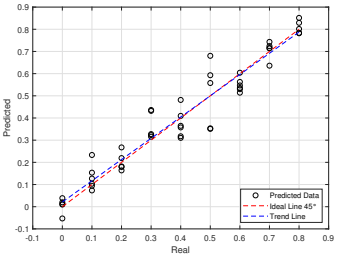
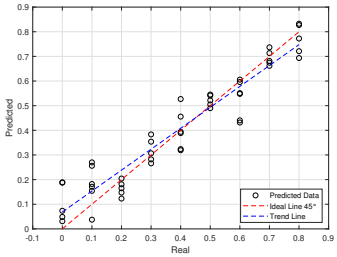
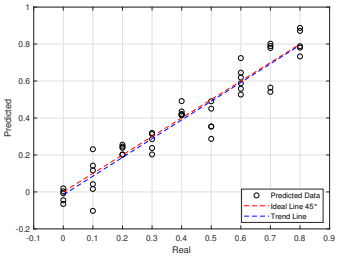
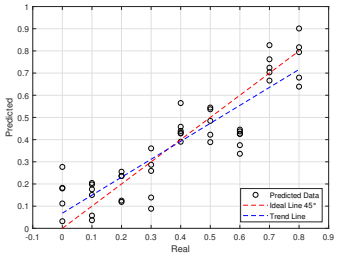
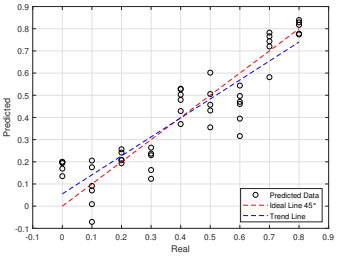
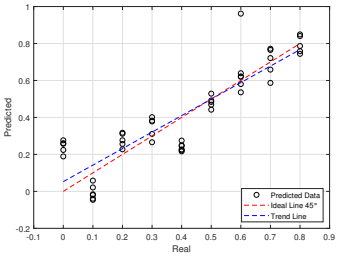
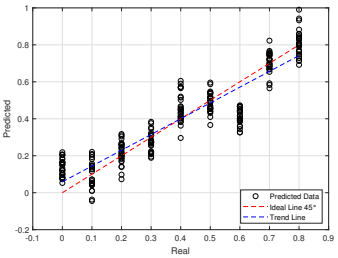
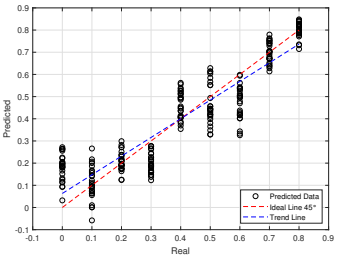
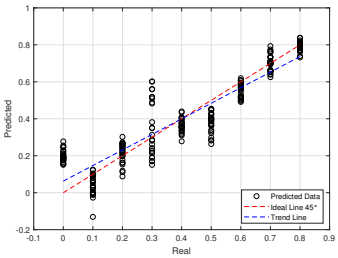
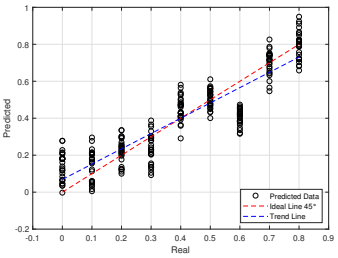
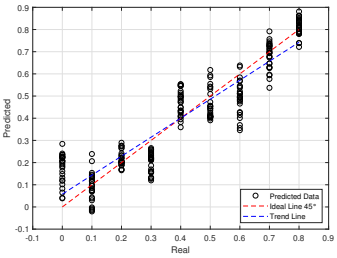
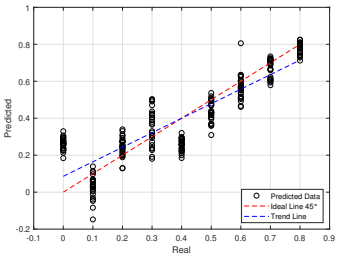
Models	E1	E2	E3
SVR Full			
SVR Optimized			
MLR Full			
MLR Optimized			
PLSR Full			
PLSR Optimized			

TABLE II. SUMMARY OF PERFORMANCE METRICS—R^2, RMSE, AND RPD—FOR THE SVR, MLR, AND PLSR MODELS APPLIED TO PREDICT MAIZE STARCH CONCENTRATION IN BIOPLASTIC SAMPLES OF THREE DIFFERENT THICKNESSES

Thickness	Model	Type	R^2	RMSE	RPD
E1	PLSR	Full	0.8503 ± 0.0005	0.0999 ± 0.0002	2.5847 ± 0.0044
		Optimized	0.8213 ± 0.0004	0.1092 ± 0.0001	2.3654 ± 0.0028
	SVR	Full	0.9158 ± 0.0041	0.0761 ± 0.0019	3.4498 ± 0.0881
		Optimized	0.8149 ± 0.0024	0.1126 ± 0.0006	2.3228 ± 0.0137
	MLR	Full	0.9495 ± 0.0041	0.0594 ± 0.0025	4.4479 ± 0.1761
		Optimized	0.8245 ± 0.0025	0.1091 ± 0.0007	2.4029 ± 0.0174
E2	PLSR	Full	0.8379 ± 0.0007	0.1040 ± 0.0002	2.4835 ± 0.0053
		Optimized	0.8517 ± 0.0003	0.0994 ± 0.0001	2.5969 ± 0.0029
	SVR	Full	0.9028 ± 0.0032	0.0828 ± 0.0013	3.1705 ± 0.0509
		Optimized	0.8041 ± 0.0022	0.1185 ± 0.0006	2.2075 ± 0.0117
	MLR	Full	0.9191 ± 0.0066	0.0754 ± 0.0032	3.5020 ± 0.1546
		Optimized	0.8528 ± 0.0030	0.0996 ± 0.0011	2.6311 ± 0.0284
E3	PLSR	Full	0.8342 ± 0.0007	0.1051 ± 0.0002	2.4556 ± 0.0050
		Optimized	0.7768 ± 0.0004	0.1220 ± 0.0001	2.1161 ± 0.0020
	SVR	Full	0.8733 ± 0.0042	0.0942 ± 0.0015	2.7838 ± 0.0446
		Optimized	0.7283 ± 0.0016	0.1393 ± 0.0004	1.8753 ± 0.0056
	MLR	Full	0.8841 ± 0.0160	0.0891 ± 0.0057	3.1874 ± 0.1488
		Optimized	0.7807 ± 0.0033	0.1218 ± 0.0011	2.1491 ± 0.0157

Finally, while this study prioritized traditional regression models for their interpretability and robustness, future research will explore using more sophisticated techniques, such as artificial neural networks and ensemble methods like XGBoost, to better capture non-linear patterns in THz spectral data and potentially boost predictive power.

ACKNOWLEDGMENT

This project was funded by the Programa Nacional de Investigación Científica y Estudios Avanzados (PROCIENCIA) through the Tesis de Pregrado y Posgrado en Ciencia, Tecnología e Innovación Tecnológica 2023 competition, under the project titled Evaluación Del Espesor Y Ratio De Contenido De Dos Almidones En El Perfil THz De Biopelículas, contract number PE501085439-2023-PROCIENCIA.

REFERENCES

[1] K. B. Arun, A. Madhavan, R. Sindhu, P. Binod, A. Pandey, R. Reshmy, and R. Sirohi, "Remodeling agro-industrial and food wastes into value-added bioactives and biopolymers," *Industrial crops and products*, vol. 154, p. 112621, 2020, https://doi.org/10.1016/j.indcrop.2020.112621.

[2] A. Apriyanto, J. Compart, and J. Fettke, "A review of starch, a unique biopolymer–Structure, metabolism and in planta modifications," *Plant Science*, vol. 318, p. 111223, 2022, https://doi.org/10.1016/j.plantsci.2022.111223.

[3] X. Tan, A. Sun, F. Cui, Q. Li, D. Wang, X. Li, and J. Li, "The physicochemical properties of Cassava Starch/Carboxymethyl cellulose sodium edible film incorporated of Bacillus and its application in salmon fillet packaging," *Food Chemistry: X*, p. 101537, 2024, https://doi.org/10.1016/j.fochx.2024.101537.

[4] G. Cheng, M. Zhou, Y.-J. Wei, F. Cheng, and P.-X. Zhu, "Comparison of mechanical reinforcement effects of cellulose nanocrystal, cellulose nanofiber, and microfibrillated cellulose in starch composites," *Polymer Composites*, vol. 40, no. S1, pp. E365–E372, 2019, https://doi.org/10.1002/pc.24685.

[5] P. R. Fitch-Vargas, I. L. Camacho-Hernández, F. J. Rodríguez-González, F. Martínez-Bustos, A. Calderón-Castro, J. de Jesús Zazueta-Morales, and E. Aguilar-Palazuelos, "Effect of compounding and plastic processing methods on the development of bioplastics based on acetylated starch reinforced with sugarcane bagasse cellulose fibers," *Industrial Crops and Products*, vol. 192, p. 116084, 2023, https://doi.org/10.1016/j.indcrop.2022.116084.

[6] N. Piñeros-Guerrero, J. P. Fernández-Trujillo, R. Pamies, and Y. Piñeros-Castro, "Evaluation and optimization of esterified starch and Canna edulis ker fiber films for food packaging applications," *Future Foods*, p. 100432, 2024, https://doi.org/10.1016/j.fufo.2024.100432.

[7] M. Lenti, D. Parisi, and P. Raffa, "Starch benzoylation in supercritical CO_2. A novel sustainable route towards biodegradable hydrophobic polymeric materials," *Carbohydrate Polymer Technologies and Applications*, vol. 7, p. 100483, 2024, https://doi.org/10.1016/j.carpta.2024.100483.

[8] I. Cacciotti, S. Mori, V. Cherubini, and F. Nanni, "Eco-sustainable systems based on poly (lactic acid), diatomite and coffee grounds extract for food packaging," *International journal of biological macromolecules*, vol. 112, pp. 567–575, 2018, https://doi.org/10.1016/j.ijbiomac.2018.02.018.

[9] H. Liu, L. Wei, L. Ba, Q. Yuan, and Y. Liu, "Biopolymer production in microbiology by application of metabolic engineering," *Polymer Bulletin*, vol. 79, no. 8, pp. 5773–5794, 2022, https://doi.org/10.1007/s00289-021-03820-9.

[10] P. Rodsamran and R. Sothornvit, "Rice stubble as a new biopolymer source to produce carboxymethyl cellulose-blended films," *Carbohydrate polymers*, vol. 171, pp. 94–101, 2017, https://doi.org/10.1016/j.carbpol.2017.05.003.

[11] X. Wu, M. V. Galkin, T. Stern, Z. Sun, and K. Barta, "Fully lignocellulose-based PET analogues for the circular economy," *Nature Communications*, vol. 13, no. 1, p. 3376, 2022, https://doi.org/10.1038/s41467-022-30735-4.

[12] Y. Wang, X. Zhang, L. Kan, F. Shen, H. Ling, and X. Wang, "All-biomass-based eco-friendly waterproof coating for paper-based green packaging," *Green Chemistry*, vol. 24, no. 18, pp. 7039–7048, 2022, https://doi.org/10.1039/D2GC02265F.

[13] K. Wang, D.-W. Sun, and H. Pu, "Emerging non-destructive terahertz spectroscopic imaging technique: Principle and applications in the agri-food industry," *Trends in Food Science & Technology*, vol. 67, pp. 93–105, 2017, https://doi.org/10.1016/j.tifs.2017.06.001.

[14] J. El Haddad, B. Bousquet, L. Canioni, and P. Mounaix, "Review in terahertz spectral analysis," *TrAC Trends in Analytical Chemistry*, vol. 44, pp. 98–105, 2013, https://doi.org/10.1016/j.trac.2012.11.009.

[15] J. Neu and C. A. Schmuttenmaer, "Tutorial: An introduction to terahertz time domain spectroscopy (THz-TDS)," *Journal of Applied Physics*, vol. 124, no. 23, 2018, https://doi.org/10.1063/1.5047659.

[16] W. Liu, C. Liu, J. Yu, Y. Zhang, J. Li, Y. Chen, and L. Zheng, "Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics," *Food chemistry*, vol. 251, pp. 86–92, 2018, https://doi.org/10.1016/j.foodchem.2018.01.081.

[17] M. Naftaly and R. E. Miles, "Terahertz time-domain spectroscopy for material characterization," *Proceedings of the IEEE*, vol. 95, no. 8, pp. 1658–1665, 2007, https://doi.org/10.1109/JPROC.2007.898835.

[18] F. Qu, L. Lin, C. Cai, B. Chu, Y. Wang, Y. He, and P. Nie, "Terahertz fingerprint characterization of 2, 4-dichlorophenoxyacetic acid and its enhanced detection in food matrices combined with spectral baseline correction," *Food Chemistry*, vol. 334, p. 127474, 2021, https://doi.org/10.1016/j.foodchem.2020.127474.

[19] S. Neelakandan, M. Prakash, B. Geetha, A. K. Nanda, A. M. Metwally, M. Santhamoorthy, and M. S. Gupta, "Metaheuristics with Deep Transfer Learning Enabled Detection and Classification Model for Industrial Waste Management," *Chemosphere*, vol. 308, p. 136046, 2022, https://doi.org/10.1016/j.chemosphere.2022.136046.

[20] F. Kiani, G. Randazzo, I. Yelmen, A. Seyyedabbasi, S. Nematzadeh, F. A. Anka, F. Erenel, M. Zontul, S. Lanza, and A. Muzirafuti, "A Smart and Mechanized Agricultural Application: From Cultivation to Harvest," *Applied Sciences*, vol. 12, no. 12, 2022, https://doi.org/10.3390/app12126021.

[21] M. Kaveh and M. S. Mesgari, "Application of Meta-Heuristic Algorithms for Training Neural Networks and Deep Learning Architectures: A Comprehensive Review," *Neural Processing Letters*, vol. 55, no. 4, pp. 4519–4622, 2023, https://doi.org/10.1007/s11063-022-11055-6.

[22] B. Arasteh, B. Aghaei, B. Farzad, K. Arasteh, F. Kiani, and M. Torkamanian-Afshar, "Detecting SQL injection attacks by binary gray wolf optimizer and machine learning algorithms," *Neural Computing and Applications*, vol. 36, no. 12, pp. 6771–6792, 2024, https://doi.org/10.1007/s00521-024-09429-z.

[23] M. Zontul, Z. G. Ersan, I. Yelmen, T. Cevik, F. Anka, and K. Gesoğlu, "Enhancing GPS Accuracy with Machine Learning: A Comparative Analysis of Algorithms," *Traitement du Signal*, vol. 41, no. 3, pp. 1441–1450, 2024.

[24] A. A. Arrieta, P. F. Gañán, S. E. Márquez, and R. Zuluaga, "Electrically conductive bioplastics from cassava starch," *Journal of the Brazilian Chemical Society*, vol. 22, pp. 1170–1176, 2011, https://doi.org/10.1590/S0103-50532011000600024.

[25] A. Sultan, H. Sultan, W. Shahzad, A. Kareem, A. Liaqat, Z. Ashraf, A. Shahid, A. Rauf, S. Saeed, T. Mehmood *et al.*, "Comparative analysis of physical and mechanical properties of starch based bioplastic derived from the pulp and peel of potatoes," *Journal of the Indian Chemical Society*, p. 101301, 2024, https://doi.org/10.1016/j.jics.2024.101301.

[26] R. Jimenez, G. Sandoval-Flores, S. Alvarado-Reyna, S. E. Aleman-Castillo, R. Santiago-Adame, and G. Velazquez, "Extraction of starch from hass avocado seeds for the preparation of biofilms," *Food Science and Technology*, vol. 42, p. e56820, 2021, https://doi.org/10.1590/fst.56820.

[27] M. Paluch, J. Ostrowska, P. Tyński, W. Sadurski, and M. Konkol, "Structural and thermal properties of starch plasticized with glycerol/urea mixture," *Journal of Polymers and the Environment*, pp. 1–13, 2022, https://doi.org/10.1007/s10924-021-02235-x.

[28] M. de Oliveira Barros, A. L. A. Mattos, J. S. de Almeida, M. de Freitas Rosa, and E. S. de Brito, "Effect of ball-milling on starch crystalline structure, gelatinization temperature, and rheological properties: Towards enhanced utilization in thermosensitive systems," *Foods*, vol. 12, no. 15, p. 2924, 2023, https://doi.org/10.3390/foods12152924.

[29] F. Kahvand and M. Fasihi, "Plasticizing and anti-plasticizing effects of polyvinyl alcohol in blend with thermoplastic starch," *International journal of biological macromolecules*, vol. 140, pp. 775–781, 2019, https://doi.org/10.1016/j.ijbiomac.2019.08.185.

[30] R. Lim, P. L. Kiew, M. K. Lam, W. M. Yeoh, and M. Y. Ho, "Corn starch/PVA bioplastics—the properties and biodegradability study using Chlorella vulgaris cultivation," *Asia-Pacific Journal of Chemical Engineering*, vol. 16, no. 3, p. e2622, 2021, https://doi.org/10.1002/apj.2622.

[31] J. Venugopal, B. Dhanasakkaravarthi, R. Surakasi, M. L. Rinawa, L. Manjunatha, R. A. Alshgari, S. M. Wabaidur, M. A. Islam, and I. Jenish, "Effect on compression molding parameters in mechanical properties of MWCNT/glass fiber/epoxy composites," *Advances in Polymer Technology*, vol. 2022, no. 1, p. 9295407, 2022, https://doi.org/10.1155/2022/9295407.

[32] Y. N. Jo, B.-D. Park, and I. C. Um, "Effect of storage and drying temperature on the gelation behavior and structural characteristics of sericin," *International journal of biological macromolecules*, vol. 81, pp. 936–941, 2015, https://doi.org/10.1016/j.ijbiomac.2015.09.016.

[33] S. Sängerlaub, E. Kucukpinar, and K. Müller, "Desiccant films made of low-density polyethylene with dispersed silica gel—water vapor absorption, permeability (h2o, n2, o2, co2), and mechanical properties," *Materials*, vol. 12, no. 14, p. 2304, 2019, https://doi.org/10.3390/ma12142304.

[34] H. Arteaga, N. León-Roque, and J. Oblitas, "The frequency range in THz spectroscopy and its relationship to the water content in food: A first approach," *Scientia Agropecuaria*, vol. 12, no. 4, pp. 625–634, 2021, https://doi.org/http://dx.doi.org/10.17268/sci.agropecu.2021.066.

[35] P. F. Smith, S. Ganesh, and P. Liu, "A comparison of random forest regression and multiple linear regression for prediction in neuroscience," *Journal of neuroscience methods*, vol. 220, no. 1, pp. 85–91, 2013, https://doi.org/10.1016/j.jneumeth.2013.08.024.

[36] S. Sousa, F. G. Martins, M. C. Alvim-Ferraz, and M. C. Pereira, "Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations," *Environmental Modelling & Software*, vol. 22, no. 1, pp. 97–103, 2007, https://doi.org/10.1016/j.envsoft.2005.12.002.

[37] F. Zhang and L. J. O'Donnell, "Support vector regression," in *Machine learning*. Elsevier, 2020, pp. 123–140.

[38] Y. Cheng, Q. Zhu, Y. Peng, X.-F. Huang, and L.-Y. He, "Multiple strategies for a novel hybrid forecasting algorithm of ozone based on data-driven models," *Journal of Cleaner Production*, vol. 326, p. 129451, 2021, https://doi.org/10.1016/j.jclepro.2021.129451.

[39] Z. Ye, X. Tan, M. Dai, X. Chen, Y. Zhong, Y. Zhang, Y. Ruan, and D. Kong, "A hyperspectral deep learning attention model for predicting lettuce chlorophyll content," *Plant methods*, vol. 20, no. 1, p. 22, 2024, https://doi.org/10.1186/s13007-024-01148-9.

[40] A. M. Mulowayi, Z. H. Shen, W. J. Nyimbo, Z. F. Di, N. Fallah, and S. H. Zheng, "Quantitative measurement of internal quality of carrots using hyperspectral imaging and multivariate analysis," *Scientific Reports*, vol. 14, no. 1, p. 8514, 2024, https://doi.org/10.1038/s41598-024-59151-y.

[41] N. Vásquez, C. Magán, J. Oblitas, T. Chuquizuta, H. Avila-George, and W. Castro, "Comparison between artificial neural network and partial least squares regression models for hardness modeling during the ripening process of swiss-type cheese using spectral profiles," *Journal of Food Engineering*, vol. 219, pp. 8–15, 2018, https://doi.org/10.1016/j.jfoodeng.2017.09.008.

[42] V. Tirado-Kulieva, C. Quijano-Jara, H. Avila-George, and W. Castro, "Predicting the evolution of pH and total soluble solids during coffee fermentation using near-infrared spectroscopy coupled with chemometrics," *Current Research in Food Science*, p. 100788, 2024, https://doi.org/10.1016/j.crfs.2024.100788.

[43] W. Castro, J. Oblitas, L. Nuñez, I. Yoplac, H. Avila-George, and M. De-la Torre, "Adulterant estimation in paprika powder using deep learning and chemometrics through near-infrared spectroscopy," *Neural Computing and Applications*, pp. 1–11, 2024, https://doi.org/10.1007/s00521-024-09830-8.

[44] C. B. Pande, N. Radwan, S. Heddam, K. O. Ahmed, F. Alshehri, S. C. Pal, and M. Pramanik, "Forecasting of monthly air quality index and understanding the air pollution in the urban city, India based on machine learning models and cross-validation," *Journal of Atmospheric Chemistry*, vol. 82, no. 1, 2025, https://doi.org/10.1007/s10874-024-09466-x.

[45] D. Agyapong, J. R. Propster, J. Marks, and T. D. Hocking, "Cross-validation for training and testing co-occurrence network inference algorithms," *BMC Bioinformatics*, vol. 26, no. 1, 2025, https://doi.org/10.1186/s12859-025-06083-7.

[46] S. Ariccio, O. Mosca, F. Dessi, F. Fornara, and M. Bonaiuto, "Cross-validation of the biofuels beliefs scale (BBS) on a european sample: A tool to measure the perception of the technological and contextual features of biofuels," *Technology in Society*, vol. 81, 2025, https://doi.org/10.1016/j.techsoc.2024.102780.

[47] Y. Zoungranan, E. Lynda, K. K. Dobi-Brice, E. Tchirioua, C. Bakary, and D. D. Yannick, "Influence of natural factors on the biodegradation of simple and composite bioplastics based on cassava starch and corn

starch," *Journal of Environmental Chemical Engineering*, vol. 8, no. 5, p. 104396, 2020, https://doi.org/10.1016/j.jece.2020.104396.

[48] A. Shafqat, A. Tahir, W. U. Khan, A. Mahmood, and G. H. Abbasi, "Production and characterization of rice starch and corn starch based biodegradable bioplastic using various plasticizers and natural reinforcing fillers," *Cellulose Chemistry and Technology*, vol. 55, pp. 867–881, 2021, https://doi.org/10.35812/CELLULOSECHEMTECHNOL.2021.55.73.

[49] H. Yuan, Q. Liu, A. Hrymak, M. Thompson, and J. Ren, "Thermoplastic potato starch blends and bioplastic films," in *Annual Technical Conference-ANTEC, Conference Proceedings*, vol. 2, 2010, pp. 1463–1467.

[50] K. Sulovská and M. Lehocký, "Terahertz spectroscopy characterization of antibacterial surfaces prepared via multistep physicochemical procedure," *Optical Engineering*, vol. 54, no. 3, pp. 034 107–034 107, 2015, https://doi.org/10.1117/1.OE.54.3.034107.

[51] D. Liu, T. Lu, and F. Qi, "A reliable method for removing Fabry–Perot effect in material characterization with terahertz time-domain spectroscopy," *IEEE Transactions on Terahertz Science and Technology*, vol. 10, no. 5, pp. 443–452, 2020, https://doi.org/10.1109/TTHZ.2020.3001508.

[52] D. I. Ramos-Soto, A. K. Singh, E. Saucedo-Casas, E. Castro-Camus, and M. Alfaro-Gomez, "Visualization of moisturizer effects in stratum corneum in vitro using THz spectroscopic imaging," *Applied optics*, vol. 58, no. 24, pp. 6581–6585, 2019, https://doi.org/10.1364/AO.58.006581.

[53] J. F. O. Cruz, "Classification of chocolate according to its cocoa percentage by using Terahertz time-domain spectroscopy," *Food Science and Technology*, vol. 43, p. e89222, 2022, https://doi.org/10.1590/fst.89222.

[54] A. V. Andrianov, A. N. Aleshin, A. K. Khripunov, and V. N. Trukhin, "Terahertz properties of bacterial cellulose films and its composite with conducting polymer PEDOT/PSS," *Synthetic Metals*, vol. 205, pp. 201–205, 2015, https://doi.org/10.1016/j.synthmet.2015.04.016.

[55] M. Karaliūnas, K. E. Nasser, A. Urbanowicz, I. Kašalynas, D. Bražinskienė, S. Asadauskas, and G. Valušis, "Non-destructive inspection of food and technical oils by terahertz spectroscopy," *Scientific reports*, vol. 8, no. 1, p. 18025, 2018, https://doi.org/10.1038/s41598-018-36151-3.

[56] M. T. Ruggiero, J. Sibik, R. Orlando, J. A. Zeitler, and T. M. Korter, "Measuring the elasticity of poly-l-proline helices with terahertz spectroscopy," *Angewandte Chemie International Edition*, vol. 55, no. 24, pp. 6877–6881, 2016, https://doi.org/10.1002/anie.201602268.

[57] A. Sarjaš, B. Pongrac, and D. Gleich, "Automated inorganic pigment classification in plastic material using terahertz spectroscopy," *Sensors*, vol. 21, no. 14, p. 4709, 2021, https://doi.org/10.3390/s21144709.

[58] Y. Sun, P. Du, X. Lu, P. Xie, Z. Qian, S. Fan, and Z. Zhu, "Quantitative characterization of bovine serum albumin thin-films using terahertz spectroscopy and machine learning methods," *Biomedical optics express*, vol. 9, no. 7, pp. 2917–2929, 2018, https://doi.org/10.1364/BOE.9.002917.

[59] Y. Gao, Y. Zhou, and K. Xu, "Quantitative analysis of materials based on terahertz spectroscopy," in *2019 18th International Conference on Optical Communications and Networks (ICOCN)*. IEEE, 2019, pp. 1–3, https://doi.org/10.1109/ICOCN.2019.8934897.

[60] D. Ye, W. Wang, H. Zhou, H. Fang, J. Huang, Y. Li, H. Gong, and Z. Li, "Characterization of thermal barrier coatings microstructural features using terahertz spectroscopy," *Surface and Coatings Technology*, vol. 394, p. 125836, 2020, https://doi.org/10.1016/j.surfcoat.2020.125836.

[61] X. Lu, Y. Shen, T. Xu, H. Sun, L. Zhu, J. Zhang, T. Chang, and H.-L. Cui, "Accurate detection of porosity in glass fiber reinforced polymers by terahertz spectroscopy," *Composites Part B: Engineering*, vol. 242, p. 110058, 2022, https://doi.org/10.1016/j.compositesb.2022.110058.

[62] N. Odani, S. Mohan, E. Kato, H. Feng, Y. Li, M. N. Hossain, J. K. Drennen III, and C. A. Anderson, "Determining the effect of photodegradation on film coated nifedipine tablets with terahertz based coating thickness measurements," *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 145, pp. 35–41, 2019, https://doi.org/10.1016/j.ejpb.2019.09.024.

[63] M. Kashima, S. Tsuchikawa, and T. Inagaki, "Simultaneous detection of density, moisture content and fiber direction of wood by THz time-domain spectroscopy," *Journal of wood science*, vol. 66, pp. 1–8, 2020, https://doi.org/10.1186/s10086-020-01874-3.

[64] W. Liang, J. Zuo, Q. Zhou, and C. Zhang, "Quantitative determination of glycerol concentration in aqueous glycerol solutions by metamaterial-based terahertz spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 270, p. 120812, 2022, https://doi.org/10.1016/j.saa.2021.120812.

[65] A. Arefi, B. Sturm, and T. Hoffmann, "Explainability of deep convolutional neural networks when it comes to NIR spectral data: A case study of starch content estimation in potato tubers," *Food Control*, vol. 169, p. 110979, 2025, https://doi.org/10.1016/j.foodcont.2024.110979.

[66] A. Sonthalia, J. Femilda Josephin, E. G. Varuvel, A. Chinnathambi, T. Subramanian, and F. Kiani, "A deep learning multi-feature based fusion model for predicting the state of health of lithium-ion batteries," *Energy*, vol. 317, p. 134569, 2025, https://doi.org/10.1016/j.energy.2025.134569.