COMPARATIVE STUDY OF CHEMOMETRIC APPROACHES AND MACHINE LEARNING FOR MINIATURIZED NEAR-INFRARED (MICRONIR) SPECTROSCOPY IN PLASTIC WASTE SORTING

C.Marchesi[1], M.Rani[1], S.Federici [1*], M. Lancini[2*], L.E. Depero[1]

[1] Department of Mechanical and Industrial Engineering, University of Brescia & UdR INSTM of Brescia, via Branze 38, 25123 Brescia, Italy c.marchesi003@unibs.it, m,rani@unibs.it, stefania.federici@unibs.it, laura.depero@unibs.it
[2] Department of Mechanical and Industrial Engineering, University of Brescia, via Branze 38, 25123 Brescia, Italy, matteo.lancini@unibs.it

*corresponding authors: stefania.federici@unibs.it, matteo.lancini@unibs.it

Abstract:
The plastic recycling industry necessitates fast and reliable methods to recognize the different polymer types to improve the recycling capacity. In this contribution, the coupling of a miniaturized Near-Infrared (NIR) spectroscopy technique with a robust data analysis is presented. Comparison of multiple machine learning techniques, such as Support-Vector Machines (SVM), Fine Tree, Bagged Tree, and Ensemble Learning, and chemometric approaches, such as Principal Component Analysis (PCA) and Partial Least Squares – Discriminant Analysis (PLS -DA), were combined to provide a broad overview and a rational means for selecting the approach in analysing NIR data for plastic waste sorting.

Keywords: Plastic waste sorting; Near-Infrared Spectroscopy (NIRS); Circular Economy; Machine Learning; Chemometrics

1. INTRODUCTION

In the perspective of whole-system economic sustainability, the enormous volume of urban plastic waste and the constant increase in human plastic consumption require a high level of waste valorization. Global plastic production reached 367 million tons in 2021, with Europe accounting for 16% of the total [1]. 9% of plastic is recycled, 12% were incinerated, and 79% ended up in landfills or the natural environment [2]. In this scenario, a key role can be played by the recycling process. Recycling is a technique for plastic product end-of-life waste management [3]. Basically, two types of recycling processes can be distinguished: mechanical and chemical process [4][5]. In both, sorting is the most critical stage in the recycling process, and this is true regardless of how effective the recycling program is [5][6]. The use of automated sorting equipment makes the process more efficient [7]. Usually, these devices rely on vibrational spectroscopic techniques [8][9][10][11] and camera systems for the polymer identification of clear and coloured products [12][13]. Other techniques are based on UV spectroscopy [14][15], X-ray [16], and hyperspectral imaging [17]. Over the years, this strategy has increased the purity of the output plastic, achieving a high percentage of recyclates in the production of secondary materials. However, these systems reach their limits with mixed plastics that require additional sorting elsewhere and can affect the quality of the recyclate if not appropriately allocated. A positive cost-benefit analysis is only possible if the separated polymer fractions have a high purity grade and satisfy the market demand for high-quality recyclates. Therefore, post-consumer recycling consists of many essential steps: collection, sorting, cleaning, size reduction and separation, and/or compatibilization to reduce polymer contamination [3]. In this scenario, the prospect of combining a well-established polymer identification technology with a small, portable, low-cost, real-time spectrometer for local and intermittent semi-automatic sorting is highly desirable, accompanied by robust data analysis [20]. In recent years, chemometric analysis of non-destructive spectroscopic data has been widely investigated as an automated method for improving plastic sorting systems. This improvement has been driven by the need to reduce the environmental impact [21]. Recently, machine learning has attracted considerable attention in plastic waste recognition using spectroscopic techniques [22][23][24]. In this study, we compared machine learning and chemometric techniques for classifying plastic waste data from a portable Near-Infrared (NIR) spectrometer. Comparisons were made between chemometric approaches, Principal Component Analysis (PCA) and Partial Least Squares – Discriminant Analysis (PLS-DA), and machine learning techniques, Support-Vector Machines

(SVM), Fine Tree, Bagged Tree, and Ensemble Learning. A comparison was also made in terms of preprocessing: traditional techniques, such as Standard Normal Variate (SNV) and Savitzky-Golay derivatives were examined in contrast to feature reduction techniques, such as multiple Gaussian Curve Fit based on Radial Basis Functions (RBF). The predictive performances of the tested models were compared in terms of classification parameters, such as Non-Error Rate (NER) and Sensitivity (Sn) with the analysis of confusion matrices, providing a comprehensive overview and a rational means of selecting the approach for the analysis of NIR data for plastic waste sorting.

2. MATERIALS AND METHODS

2.1. Samples Collection

The first batch of plastic samples were collected in the Selection Division of the Montello SpA recovery and recycling plant (Bergamo, Italy), which accepts post-consumer plastic in the form of municipal waste for recycling. Subsequently, the dataset was expanded to include new samples from municipal waste collected before ending up in landfills. A total of 325 samples from a variety of polymer classes were used in this study. Specifically, the products studied were: 75 samples of poly(ethylene terephthalate) (PET), 100 samples of polyethylene (PE), 75 samples of polypropylene (PP), and 75 samples of poly(styrene) (PS). The assortment included bottles, containers, and packaging of various sizes, shapes, and colours.

2.2. NIR Analysis

Plastic samples were analyzed using the MicroNIR On-site (Viavi Solutions Inc., CA, United States) in reflectance mode without pretreatment of the samples. The instrument is a palm-sized, portable spectrometer weighing approximately 250 g and measuring less than 200 mm in length and 50 mm in diameter. Control settings for spectral data acquisition were set to 10 milliseconds integration time and 50 scans, resulting in a short measurement time of 0.25 seconds. A point-and-shoot technique was used to perform 5 replicates for each sample to reduce the effects caused by sample non-uniformity. A total of 1625 spectra were acquired, and acquisition was performed using MicroNIRTM Pro v3.0 software (Viavi Solutions Inc., CA, United States).

2.3. Spectral Pretreatment and chemometrics

Preprocessing of NIR spectral data has become an essential aspect of chemometric modelling. The goal of preprocessing is to eliminate physical events from the spectra to improve subsequent multivariate regression, classification model, or exploratory analysis [25]. In this study, the spectra were retrieved in a single matrix of 1625 x 125 (samples x wavenumbers) and several preprocessing methods were applied. The best results were obtained using the Savitzky-Golay second derivative method with seven data points and a second order polynomial followed by a standard normal variate (SNV). In addition, normalization was performed by mean centering. Different chemometric methods were used for the correct evaluation of the data of all analyzed samples. The first phase was an exploratory analysis by PCA to investigate the data structure. PCA was performed on 1625 NIR spectra from all polymer classes. Then, PLS-DA was applied as a supervised pattern recognition tool to separate the different commodities. Prior to using PLS-DA, data were split into a training set and a test set using a MATLAB proprietary function. The process was repeated 500 times, generating a different training and test set each time (75% of the samples belonged to the training set and 25% to the test set). All chemometric analyses were performed with MATLAB 2021b (The MathWorks, Inc, Natick, MA, USA) using the PLS-Toolbox (Eigenvector Research, Inc. Manson, Washington, USA).

2.4. Machine Learning

Various machine learning algorithms were applied for classification purposes. Support Vector Machine (SVM), Fine Tree, Ensemble Learning, and Bagged Tree In addition, a likelihood-based aggregation procedure (here called Combo) was used to integrate the data into a single predictor, and the same procedure was applied with a Monte Carlo Method (MCM) to make a perturbation on raw data, to improve the generalization performance. The chosen hyperparameters are the following: for Fine Tree Gini's diversity index (gdi) was used as split criterion with 100 maximum number of splits; SVM was performed with a linear kernel function with kernel scale equal to 3. Lastly, Ensemble Learning was performed with the Bagged Tree method with 30 cycles of learning. To test the reliability of the system, 200 random extractions were performed for splitting the training and testing set. Again, 75% of the samples were used for training and the rest for testing. Machine learning methods were performed on the raw data after applying the variable reduction technique based on multiple Gaussian Curve Fit with Radial Basis Functions (RBF) and combining raw and pre-processed data. All calculations were performed using MATLAB and Statistics Toolbox release 2021b (The MathWorks, Inc, Natick, MA, USA). Automation of the procedure was implemented using MATLAB functions created in-house.

3. RESULTS AND DISCUSSION

3.1. Principal Component Analysis

The PCA calculation was performed after the preprocessing described above for the entire spectral range. For data structure analysis, PCA is a useful chemometric method. The goal of PCA is to extract the information stored in many variables into a smaller number of variables, called Principal Components [26]. Figure 1 shows the score plot of the first two components (73.88 of the total explained variability), in which a clear separation between the polymer classes can be seen. Along PC1 PET is distinguished from the other commodities. PET samples show very negative score values, while the other samples show positive score values. On the other hand, along PC2, PS is clearly separated from the other plastics.

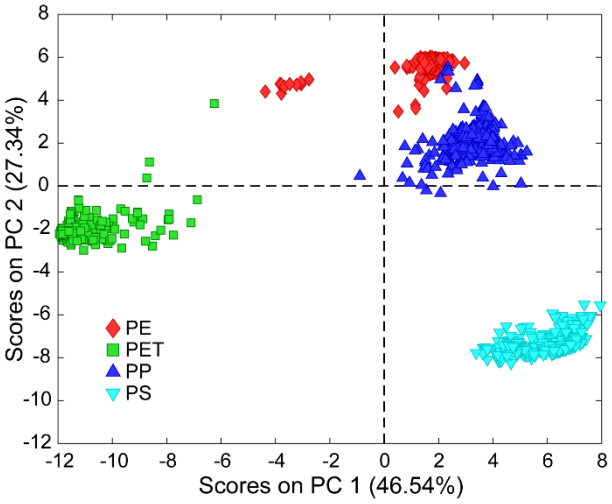

Figure 1: Results of PCA performed with spectral data of different commodities. The score plot of PC1 vs PC2 is presented.

A clear separation between PP and PE can be noticed in the score plot of PC1 vs PC3 in Figure 2. PC3 accounts for 15.83% of the total information and explains the difference of PP from the other class of polymers.

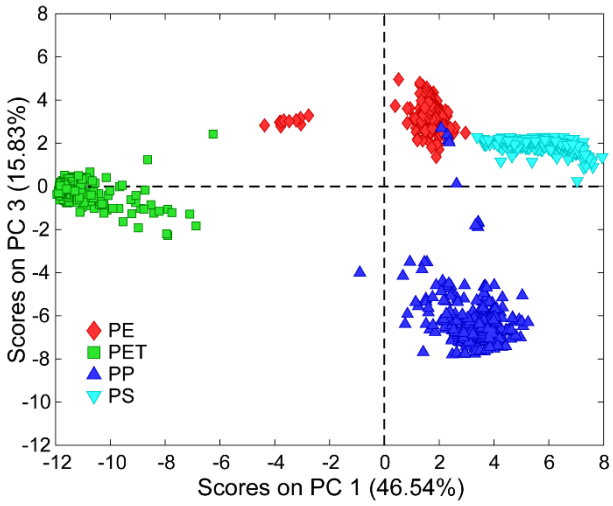

Figure 2: Results of PCA performed with spectral data of different commodities. The score plot of PC1 vs PC3 is presented.

3.2. Partial Least Squares Discriminant Analysis

Following the exploratory PCA analysis, a supervised classification technique was used to distinguish the different plastic groups. In PLS-DA, a classification objective is added to the well-known PLS regression technique. The response variable is categorical and reflects the class to which the statistical units belong. PLS-DA returns the prediction as a vector with values between 0 and 1 and a length equal to the number of classes in the predictor variables [27][28]. Each time PLS-DA was performed, the parameters such as NER and sensitivity were calculated in fitting, in cross-validation (CV), and for the test set. The cross-validation procedure was based on venetian blind approach with 5 groups. CV was also used to determine the optimal number of Latent Variables (LVs) for each PLS-DA model. Figure 3 shows all sensitivities for each class, calculated for training set, CV and for test set. The values are close to 1, indicating a very high classification performance. Moreover, the results are very balanced between training, CV, and test set; therefore, overfitting is completely avoided, and the model can be considered reliable and stable.

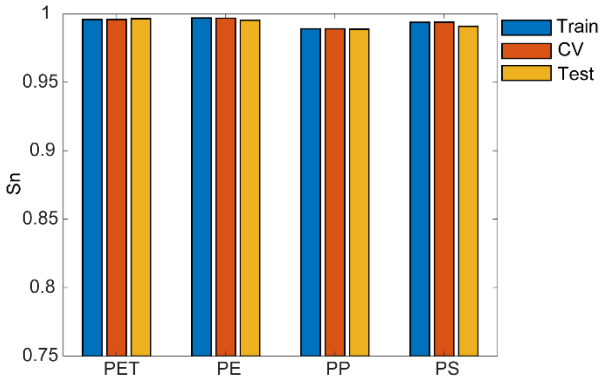

Figure 3: PLS-DA Model. Class sensitivities (Sn) calculated for training set, in cross-validation and for test set.

Table 1 shows NER defined as mean of class sensitivity [29], calculated for the training set, cross-validation set and test set. Overall, 99% of the samples were correctly classified for each of the 500 iterations.

Table 1: PLS-DA Model. Non-Error Rate calculated for training set, in CV and for test set.

	NER
Training	0.99
CV	0.99
Test	0.99

3.3. Machine Learning

Due to the complexity and the large number of results, for the machine learning analysis the classification parameters are presented only for the test set. Figure 4 shows the NER of the classes for each computed model and for each treatment of the data. It is noticeable that the models run on raw data have the worst performances. The NER ranges from 0.74 (Fine Tree) to 0.9 (SVM), indicating a high variability in the results. For raw data only SVM can be considered as a satisfactory model for pattern recognition. Lower variability in the results is observed for pretreated data and for a mixture of pretreated and raw data, where the NER ranges from 0.96 to 0.99 and from 0.96 to 0.98, respectively. Thus, there is no difference in the results between preprocessed data and the combination of raw and pretreated data. These results confirm that feature reduction based on the Gaussian curve with RBF gives high performances for pattern recognition in machine learning analysis.

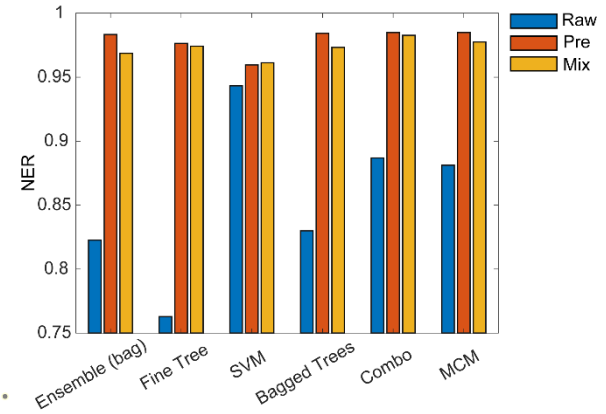

Figure 4: Machine Learning. Comparison of the Non-Error Rate (NER) calculated from the confusion matrices for each model. Results are presented for raw data, pretreated data, and the combination of raw and pretreated data.

In conclusion, model performance is comparable between machine learning and multivariate analysis methods. After random extraction of training and test data repeated 500 and 200 times for chemometrics and machine learning, respectively, the NER calculated for the test set is above 0.95 for both methods. However, the use of chemometrics reduces the computational time, compared to the computationally intensive machine learning algorithms.

4. SUMMARY

This contribution included a side-by-side comparison between conventional chemometric methods and machine learning algorithms for the classification of a dataset obtained from the study of plastic waste with a portable Near-Infrared (NIR) spectrometer. Multivariate methods such as Principal Component Analysis (PCA) and Partial Least Squares - Discriminant Analysis (PLS - DA) were investigated, as well as machine learning methods such as Support Vector Machines (SVM), Fine Tree, Bagged Tree and Ensemble Learning. Results were also compared in terms of data processing: signal preprocessing tools, SNV and Savitky-Golay derivatives were compared with feature reduction approaches such as Multiple Gaussian Curve Fit based on Radial Basis Functions (RBF). In addition, the machine learning algorithms were run on raw data, preprocessed data, and the combination of the two approaches. The results from PLS-DA showed very high performance for pattern recognition; in fact, the NER for the training set, in CV, and for the test set are all equal to 0.99. In contrast, for machine learning, the NER for raw data ranges from 0.74 for Fine Tree to 0.90 for SVM, indicating high variability in the results. The results for the preprocessed data show lower variability with NER value ranging from 0.96 to

0.99, which is also true for the combination of raw data and preprocessed data. This confirms that RBF-based variable reduction is the most crucial point to improve classification performances. We can conclude that the multivariate and machine learning approaches produce comparable results in terms of model performance. The NER estimated for the test set is above 0.95 for both chemometrics and machine learning after randomly extracting the training and test data and repeating them 500 and 200 times, respectively. On the other hand, chemometrics is characterised by a lower computation time compared to machine learning algorithms and it can therefore be considered more advantageous.

5. REFERENCES

[1] Plastics Europe, Plastics the fact 2021, Plast. Eur. Mark. Res. Gr. Conversio Mark. Strateg. GmbH. (2021). https://plasticseurope.org/.

[2] R. Geyer, J.R. Jambeck, K.L. Law, Production, use, and fate of all plastics ever made, Sci. Adv. 3 (2017). https://doi.org/10.1126/sciadv.1700782.

[3] J. Hopewell, R. Dvorak, E. Kosior, Plastics recycling: Challenges and opportunities, Philos. Trans. R. Soc. B Biol. Sci. 364 (2009) 2115–2126. https://doi.org/10.1098/rstb.2008.0311.

[4] K. Ragaert, L. Delva, K. Van Geem, Mechanical and chemical recycling of solid plastic waste, Waste Manag. 69 (2017) 24–58. https://doi.org/10.1016/j.wasman.2017.07.044.

[5] S.M. Al-Salem, P. Lettieri, J. Baeyens, Recycling and recovery routes of plastic solid waste (PSW): A review, Waste Manag. 29 (2009) 2625–2643. https://doi.org/10.1016/j.wasman.2009.06.004.

[6] R. Siddique, J. Khatib, I. Kaur, Use of recycled plastic in concrete: A review, Waste Manag. 28 (2008) 1835–1852. https://doi.org/10.1016/j.wasman.2007.09.011.

[7] S.P. Gundupalli, S. Hait, A. Thakur, A review on automated sorting of source-separated municipal solid waste for recycling, Waste Manag. 60 (2017) 56–74. https://doi.org/10.1016/j.wasman.2016.09.015.

[8] V. Allen, J.H. Kalivas, R.G. Rodriguez, Post-consumer plastic identification using Raman spectroscopy, Appl. Spectrosc. 53 (1999) 672–681. https://doi.org/10.1366/0003702991947324.

[9] V. Ludwig, Z.M. Da Costa Ludwig, M.M. Rodrigues, V. Anjos, C.B. Costa, D.R. Sant'Anna das Dores, V.R. da Silva, F. Soares, Analysis by Raman and infrared spectroscopy combined with theoretical studies on the identification of plasticizer in PVC films, Vib. Spectrosc. 98 (2018) 134–138. https://doi.org/10.1016/j.vibspec.2018.08.004.

[10] O. Rozenstein, E. Puckrin, J. Adamowski, Development of a new approach based on midwave infrared spectroscopy for post-consumer black plastic waste sorting in the recycling industry, Waste Manag. 68 (2017) 38–44. https://doi.org/10.1016/j.wasman.2017.07.023.

[11] A. Vázquez-Guardado, M. Money, N. McKinney, D. Chanda, Multi-spectral infrared spectroscopy for robust plastic identification, Appl. Opt. 54 (2015) 7396. https://doi.org/10.1364/ao.54.007396

[12] J. Hopewell, R. Dvorak, E. Kosior, Plastics recycling: Challenges and opportunities, Philos. Trans. R. Soc. B Biol. Sci. 364 (2009) 2115–2126. https://doi.org/10.1098/rstb.2008.0311

[13] Y. Tachwali, Y. Al-Assaf, A.R. Al-Ali, Automatic multistage classification system for plastic bottles recycling, Resour. Conserv. Recycl. 52 (2007) 266–285. https://doi.org/10.1016/j.resconrec.2007.03.008.

[14] . E. Maris, A. Aoussat, E. Naffrechoux, D. Froelich, Polymer tracer detection systems with UV fluorescence spectrometry to improve product recyclability, Miner. Eng. 29 (2012) 77–88. https://doi.org/10.1016/j.mineng.2011.09.016.

[15] S.M. Safavi, H. Masoumi, S.S. Mirian, M. Tabrizchi, Sorting of polypropylene resins by color in MSW using visible reflectance spectroscopy, Waste Manag. 30 (2010) 2216–2222. https://doi.org/10.1016/j.wasman.2010.06.023.

[16] S. Brunner, P. Fomin, C. Kargel, Automated sorting of polymer flakes: Fluorescence labeling and development of a measurement system prototype, Waste Manag. 38 (2015) 49–60. https://doi.org/10.1016/j.wasman.2014.12.006.

[17] Y. Zheng, J. Bai, J. Xu, X. Li, Y. Zhang, A discrimination model in waste plastics sorting using NIR hyperspectral imaging system, Waste Manag. 72 (2018) 87–98. https://doi.org/10.1016/j.wasman.2017.10.015.

[18] M. Vidal, A. Gowen, J.M. Amigo, NIR Hyperspectral Imaging for Plastics Classification, NIR News. 23 (2012) 13–15. https://doi.org/10.1255/nirn.1285.

[19] M. Moroni, A. Mei, A. Leonardi, E. Lupo, F. La Marca, PET and PVC separation with hyperspectral imagery, Sensors (Switzerland). 15 (2015) 2205–2227. https://doi.org/10.3390/s150102205.

[20] I. Vollmer, M.J.F. Jenks, M.C.P. Roelands, R.J. White, T. van Harmelen, P. de Wild, G.P. van der Laan, F. Meirer, J.T.F. Keurentjes, B.M. Weckhuysen, Beyond Mechanical Recycling: Giving New Life to Plastic Waste, Angew. Chemie - Int. Ed. 59 (2020) 15402–15423. https://doi.org/10.1002/anie.201915651.

[21] Araujo-Andrade, C., Bugnicourt, E., Philippet, L., Rodriguez-Turienzo, L., Nettleton, D., Hoffmann, L., Schlummer, M., 2021. Review on the photonic techniques suitable for automatic monitoring of the composition of multi-materials wastes in view of their posterior recycling. Waste Manag. Res. 39, 631–651. https://doi.org/10.1177/0734242X21997908

[22] Da Silva, V.H., Murphy, F., Amigo, J.M., Stedmon, C., Strand, J., 2020. Classification and Quantification of Microplastics (<100 μm) Using a Focal Plane Array-Fourier Transform Infrared Imaging System and Machine Learning. Anal.

Chem. 92, 13724–13733. https://doi.org/10.1021/acs.analchem.0c01324
[23] Zhu, S., Chen, H., Wang, M., Guo, X., Lei, Y., Jin, G., 2019. Plastic solid waste identification system based on near infrared spectroscopy in combination with support vector machine. Adv. Ind. Eng. Polym. Res. 2, 77–81. https://doi.org/10.1016/j.aiepr.2019.04.001
[24] Michel, A.P.M., Morrison, A.E., Preston, V.L., Marx, C.T., Colson, B.C., White, H.K., 2020. Rapid Identification of Marine Plastic Debris via Spectroscopic Techniques and Machine Learning Classifiers. Environ. Sci. Technol. 54, 10630–10637. https://doi.org/10.1021/acs.est.0c02099
[25] Rinnan, Å., Berg, F. van den, Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC - Trends Anal. Chem. 28, 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007
[26] Ballabio, D., 2015. A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. Chemom. Intell. Lab. Syst. 149, 1–9. https://doi.org/10.1016/j.chemolab.2015.10.003
[27] Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: Linear models. PLS-DA. Anal. Methods 5, 3790–3798. https://doi.org/10.1039/c3ay40582f
[28] Brereton, R.G., Lloyd, G.R., 2014. Partial least squares discriminant analysis: Taking the magic away. J. Chemom. 28, 213–225. https://doi.org/10.1002/cem.2609
[29] Ballabio, D., Grisoni, F., Todeschini, R., 2018. Multivariate comparison of classification performance measures. Chemom. Intell. Lab. Syst. 174, 33–44. https://doi.org/10.1016/j.chemolab.2017.12.004