Training and evaluating machine learning algorithms for ocean microplastics classification through vibrational spectroscopy

Abstract

Microplastics are contaminants of emerging concern - not only environmental, but also to human health. Characterizing them is of fundamental importance to evaluate their potential impacts and target specific actions aiming to reduce potential harming effects. This study extends the exploration of machine learning classification algorithms applied to FTIR spectra of microplastics collected at sea. A comparison of successful classification models was made in order to evaluate prediction performance for 13 classes of polymers. A rigorous methodology was applied using a pipeline scheme to avoid bias in the training and selection phases. The application of an oversampling technique also contributed by compensating unbalanceness in the dataset. The log-loss was used as the minimization function target and to assess performance. In our analysis, Support Vector Machine Classifier provides a good relationship between simplicity and performance, for a fast and useful automatic characterization of microplastics.

Keywords

Microplastics; Marine Pollution; Chemical Identification; Vibrational Spectroscopy; FTIR; Machine Learning

1. Introduction

Plastic debris is today found in virtually every habitat on earth (Free et al., 2014; Lebreton et al., 2018; Saito et al., 2018; Wang et al., 2019; Zhang et al., 2020). Yet, major scientific concern has been given to plastics at sea, where they are almost omnipresent and when found in large quantities and/or concentrations their negative impacts may be serious (Barnes, 2002; Chae and An, 2017). Originated mainly on land as a result of large consumption of disposable items and poor waste management, plastics compose most marine debris (Galgani et al., 2015).

According to the International Organization for Standardization (ISO), microplastics are any solid plastic particle insoluble in water with any dimension from 1 μm to 1 000 μm (1 mm) and large microplastics between 1 mm and 5 mm. (ISO/TR 21960:2020). They can originate from the degradation and fragmentation of larger plastic debris when exposed to environmental conditions or can be directly emitted in their form as, for example, microbeads, pellets, or textile fiber (Fendall and Sewell, 2009; GESAMP, 2019; Hidalgo-Ruz et al., 2012; Thompson et al., 2009). Their wide occurrence and physical characteristics, such as density and chemical composition, have contributed to consider these as emerging environmental contaminants (Sauvé and Desrosiers, 2014).

An adequate characterization of these contaminants can provide substantial information on the inputs and transport in the oceans, rates of degradation and fragmentation, interaction with biota, consequences of their presence in natural habitats, and, finally, risk assessment and management. Multiple parameters are relevant for these analyses, such as size, mass, sampling site, and DNA from the plastisphere (Zettler et al., 2013). Determination of the chemical composition of sampled plastics has been used to suppose where they originated from, what human use they might have had, and even estimate their age (Song et al., 2015; Turner and Holmes, 2011).

While diversity is wanted in the production and use phases of polymer products, it adds complexity to the identification process. For macroplastics, characterization may still be done by recycling codes in the product or by physical characterization. Yet, for microplastics, these procedures have been shown to result in high rates of false positives when not followed by any chemical analysis (GESAMP, 2019; Hidalgo-Ruz et al., 2012)Thus, a chemical characterization of microplastics is fundamental to more adequately assess the sources, fate, and impacts of microplastics.

Vibrational Spectroscopy, namely Raman and Fourier Transform Infrared Spectroscopy (FTIR), are established techniques to assess chemical composition (Käppler et al., 2016). Both analyses provide information on specific chemical bonds and functional groups, albeit by different methods (Kuptsov and Zhizhin, 1998). They allow differentiation between synthetic and natural polymers, identification of polymer type, and degree of weathering, hence their wide utilization in microplastic identification methods (Andrady, 2017).

FTIR analysis can be done visually by an expert, but it is often more convenient to use a peak matching dedicated software with the available databases (Li et al., 2006). They perform what is called a *library search*, going through all spectra in the database and showing those with better ranking. This strategy allows a faster polymer identification, but since sampled microplastics are somehow dirty and weathered, peaks absent in the same non-weathered polymers can be seen in microplastics. (Jung et al., 2018; Xu et al., 2019). Therefore, their chemical signature (spectrum) can be significantly different from virgin polymers that commonly constitute databases (GESAMP, 2019; Primpke et al., 2018). Procedural factors can also influence the quality of measurements and impact identification accuracy and time. The size of identifiable microplastics, for example, is limited by the FTIR setup and should not be less than 500 μm (unless the spectrometer is coupled with a microscope) (Käppler et al., 2016; Shim et al., 2017). Weathered microplastics could be fragile and shatter while being manipulated, which also hinders spectrum acquisition (Shim et al., 2017).

A few alternatives have been proposed and tested to surpass these limitations. Pre-processing of the raw data, the usage of specifically designed databases, and dual database searches are some of the approaches that have been demonstrated to improve match results and allow library searches to be used with good confidence (Primpke et al., 2018; Primpke et al.2020; Renner et al., 2019).

Most notably, µ-FTIR and µ-Raman techniques are being used to address common limitations on microplastics analysis. Coupling a spectrometer and a microscope, allows the characterization of smaller particles and with an additional Focal Plane Array (FPA) detector, several particles can be characterized at once (Cabernard et al. 2018; Brandt et al. 2020). In this sense, these techniques can provide a more accurate picture of the evaluated microplastic environment. Yet, this kind of equipment may be less available mainly due to cost issues.

Improving methodologies for conventional vibrational spectroscopy could provide better alternatives to marine pollution researchers in places where financial support is insufficient, mainly the global south. Having said that, the proposed methodology could be used to evaluate the performance of classification algorithms using data from different spectroscopic techniques, especially FTIR imaging where data volumes are significantly larger (~ 106 spectra per image) and spectra quality tends to be lower (Primpke et al. 2017).

Sampling microplastics at sea produces hundreds or even thousands of individual particles, which by manual methodologies would take far too long to be characterized. Increased automation of the characterization process of microplastic samples with higher confidence in the results could quicken information acquisition on this emerging contaminant, filling multiple knowledge gaps and significantly advancing understanding on the field. Indeed, that is the main objective of the study presented in this paper. A machine learning pipeline was proposed for the selection of the best among a few machine learning algorithms to classify microplastics spectra, then discuss the main findings.

ML algorithms have recently been proposed as faster and more accurate methods to analyze spectra from marine microplastics and have been successfully used in chemometrics (and many other fields) for more than a decade (Conroy et al. 2005). Most efforts have been applied to imaging techniques, probably due to high computational demands associated with higher data volumes - thus, the necessity to improve efficiency (Hufnagl et al., 2019). Nevertheless, conventional spectroscopic techniques could benefit from these improvements.

In machine learning, classification algorithms are a category of prediction models used to attribute unidentified samples to a given class based on a set of variables. These algorithms have been applied by Hufnagl (Random Forest Classifier) and Kedzierski (K-nearest neighbors) to spectral data of microplastics and resulted in expressive classification accuracies (Hufnagl et al., 2019; Kedzierski et al., 2019). However, a comparison of different algorithms hasn't been made. Given the already demonstrated potential of these techniques, further investigation could lead to even better results. Since the learning process highly depends on the database available for training, comparable results should ideally derive from the same dataset. In the present study, data previously published by Kiedzierski were used for this purpose (Kedzierski et al., 2019).

Methodologies focusing on the selection of relevant attributes on microplastics spectra have been proposed by Renner 2017, Hufnagl 2019, and da Silva 2020. Indeed, it is the approach experts take when visually interpreting a spectrum. In this paper, a *dimensionality reduction (DR)* technique called Principal Component Analysis (PCA) was applied to extract the most relevant features in the whole dataset. Some machine learning models lose performance when the data is represented in a high dimension space, falling into the so-called curse of dimensionality (Trunk, 1979). Feature selection or extraction is used to reduce the number of variables that describe a certain set of instances (samples) while retaining most of the information. This can, sometimes, reduce the time of implementation without significant loss of information and,when the curse of dimensionality is observed, even improve prediction performance.

The main disadvantage of the machine learning approach to this classification task is that it is limited to analyzing classes of polymers that are well represented in the training database. Since there is an enormous variety of plastics, setting up rather complete databases would be laborious. However, from all the known range of polymers, there are only a few commonly found as marine microplastics, making the setup of representative databases a lot more feasible. The method described in this paper goes further in the process of setting up this type of database by using an artificial oversampling technique to mitigate the effects of imbalanced datasets in the learning process, which can produce models that are biased towards the majority class.

2. Materials and Methods

This study extends previous efforts by combining several machine learning techniques and comparing algorithms, to find, among them, the best method to automatically classify microplastic spectra. It differs from other approaches by proposing a pipeline methodology to train, evaluate and select classification models. The methodology described next was rigorously defined to justify choosing one model over all others. Once the model is chosen, it can be trained and called to make a prediction. A researcher interested in reproducing the experiment must follow the entire methodology, whereas a potential user of the proposed predictive tool can skip to section 2.9 for a short explanation on how it can be used.

All programming was made in a Core i5-7200U with 16 GB ram, using only 1 core of the processor to avoid issues related to parallel programming. We used Python programming language and some of its data analysis and machine learning libraries, namely: scikit-learn (Pedregosa et al., 2011), scipy (Virtanen et al., 2020), and imblearn (Lemaître et al., 2017). All programming is available in a supplementary file (Appendix D) so that the methodology described in this paper can be audited, reproduced and even improved by peers. It is also available online at GitHub (https://github.com/EdsonCilos/mp_classification).

2.1. The dataset

The data used to train the algorithms were previously published by Kedzierski et al., 2019, and were generated from the Attenuated Total Reflection FTIR spectroscopy of samples collected during expeditions in the Mediterranean Sea. Spectra were recorded in absorbance mode in the range of 4000 to 600 cm^{-1} with 4 cm^{-1} resolution and 16 scans. The labeled dataset was constructed using the raw data consisting of 958[1] spectra previously identified and assigned to 17 different classes.

2.2. Dataset pre-processing pipeline

Typically in the spectra evaluation by an expert, the analysis of some frequencies is not considered due to the presence of noise or just because the knowledge in the field prescribes that samples can be distinguished by some specific peaks. Our perspective, however, stands in the Machine Learning background, assigning to the algorithm the task to find the most successful pattern matching. As pointed by Vapnik: "In a wide philosophical sense, predictive models do not necessarily connect prediction of an event with an understanding of the law that governs the event; they are just looking for a function that explains the data best." (Vapnik, 2006)

In other words, our goal is to look for a good explanation of the available data, and therefore, no feature selection based on expert knowledge was done, nor seemed to be necessary. As a consequence, all recorded wavelengths were used for the analysis. Additionally, a major drawback of a priori feature selection is that it may be suboptimal and should be revised when new classes are introduced, while PCA is embedded in the pipeline.

Since the original dataset was highly imbalanced, the first step in pre-processing consisted of renaming the classes that were underrepresented in the dataset, assigning them to a generic class called "unknown". Samples from the 4 least represented classes, namely, Polyurethane, Animal fiber-like, Poly(vinyl chloride), and acrylic (PMMA) were moved to the unknown class, effectively removing 4 classes. The latter having only 3 samples, while the others only 1 sample, making it impossible to train a model using the methodology adopted in this paper. The dataset then consisted of 958 samples from 14 different classes (13 microplastic classes + unknown).

It is important to note that apart from relabeling underrepresented classes and correcting the baseline, all preprocessing steps must be done after the train/test split, to avoid the so-called "snooping bias/data leakage" (see "Results and Discussion" for further details). It is a common mistake when coding machine learning algorithms. It happens when some information that typically would be available only in subsequent steps (like in a "production environment"), is introduced in the learning process.

The remaining pre-processing steps were assembled in a pipeline, which we will call, for short, a Pre-pipeline (Later on in Section 2.6 we shall discuss the full Pipeline of the methodology, which includes the Pre-pipeline). All steps were tested in the "turn on/ turn off" configuration, to evaluate which of those was able to improve the final model's performance. It consisted of the application of the following techniques in the raw data (in this order):

1. Baseline correction using Asymmetric Least Squares (Eilers and Boelens, 2005). Parameters were: lamba (2^{nd} derivative constraint) = $1e^5$; p (weighting deviations) = 0.05; itermax (number of iterations to perform) = 10
2. Standard scaler (Z-transform), since the features are arranged at different scales, which could affect the model performance (Géron, 2017);
3. Principal component analysis (PCA), as a method to verify if higher dimensions could impact the algorithm's performance (Jolliffe, 2002). PCA is an unsupervised learning technique that creates a new linear space with orthogonal variables, called principal components (PCs), which are the directions of most variance in the dataset. If the original features are uncorrelated, the new space contains the same number of dimensions as the original space, but, generally, a much smaller number of variables

is sufficient to describe the data without much loss of information. In our analysis, the first *n* PCs containing 99% of the explained variance were used.

4. Oversampling to mitigate the effects of imbalance in the dataset. Despite having assembled underrepresented classes, the dataset was still highly imbalanced, as the 2 most populated classes contained 45% of all samples in the dataset. This can cause distortions in the learning process, as the algorithm would be more likely to assign a new sample to these classes (Prati et al., 2009). This was done using the Imbalanced Learning API inside a Pipeline scheme (Figures 1 and 3). It generates new instances by randomly sampling the under-represented classes, effectively copying existing spectra (Lemaître et al., 2017).

The four pre-processing techniques described (Baseline correction, Standard Scaler, PCA, and Oversampling) were subjected to analysis to evaluate if including them would improve the performances of the classifiers. Several combinations were tried for each classifier (e.g. kNN without PCA, with Standard Scaler, and with oversampling). The best combination was forwarded to another trial to examine if a frequency smoothing strategy could improve the performance likely by suppressing features that arise from non-ideal instrument and sample conditions (Renner et al., 2019; Zimmermann and Kohler, 2013). Ergo, a Savitzky-Golay filter was included before the standard scaler step, in which an evaluation of 115 combinations of windows and degrees was performed. This step was done separately in order to speed the evaluation, considering it already took 4 days.

2.3. The holdout strategy

Initially, the dataset was split into training and test sets (holdout strategy), stratifying proportionally to the number of classes, to make sure every class is present both in the train and test sets. The test set ratio choice was 25%. To allow reproducibility, we fixed the seed, as well as all random states, equal to 0. It is worth mentioning that the authors used the holdout strategy to be able to evaluate, under mathematical assumptions, an “unbiased”[2] estimate of the generalization error. The holdout strategy allows the trained algorithm to be tested on an “external” dataset, completely unknown to the algorithm. This can be done using mathematical tools like Hoeffding's inequality (Hoeffding, 1963), Vapnik-Chervonenkis theory (Vapnik, 2000) or usual statistical inference techniques. Thereby, a unique and final evaluation in the test set is capable of providing a confidence interval for the proposed algorithm, as we shall see in the Results/Discussion section.

2.4. The classification algorithms

According to the "No Free Lunch Theorem" (Wolpert, 1996): there is no a priori reason to prefer one machine learning model over another without making any assumption about the data. Therefore, in practice, we must select some models and test them on the problem that we aim to solve. For this reason, we chose some of the most popular and relevant models in machine learning: Decision Tree (DT) (Moore, 1987; Murphy, 2012, p. 544), Gaussian Naive Bayes (GNB) (Murphy, 2012, p. 82), k-Nearest Neighbor (kNN) (Nordhausen, 2009, p. 14), Random Forest (RF) (Nordhausen, 2009, p. 587), Logistic Regression (LR) (Murphy, 2012, p. 14; Nordhausen, 2009, p. 119), Support Vector Machine Classifier (SVC) (Murphy, 2012, p. 496; Nordhausen, 2009, p. 417), and Neural Networks (Goodfellow et. al, 2016).

Each model has general characteristics that may be more suited to one application than the other. A comparison of 5 of the models proposed here is presented in Table 1, adapted from Kotsiantis (Kotsiantis, 2007). For instance, GNB has a high speed of learning and classification but is highly affected by interdependent attributes. SVC, on the other hand, has a high speed of classification and is very tolerant to irrelevant attributes, but is slow to learn and falls short in explainability (Kotsiantis, 2007).

	Decision Trees	Neural Networks	Naive Bayes	kNN	SVC
Accuracy in general	**	***	*	**	****

Speed of learning	***	*	****	****	*
Speed of classification	****	****	****	*	****
Tolerance to missing values	***	*	****	*	**
Tolerance to irrelevant attributes	***	*	**	**	****
Tolerance to redundant attributes	**	**	*	**	***
Tolerance to highly interdependent attributes	**	***	*	*	***
Dealing with discrete/binary/continuous attributes	****	***	***	***	**
Tolerance to noise	**	**	***	*	**
Dealing with danger of overfitting	**	*	***	***	**
Attempts for incremental learning	**	***	****	****	**
Explanation ability	****	*	****	**	*
Model parameter handling	***	*	****	***	*

Table 1 - A comparison of models' characteristics (Adapted from Kotsiantis, 2007)

The algorithms have been configured to return probabilities (provided by Sklearn; *sklearn.calibration*) rather than just a deterministic output. This seems to be suitable for the current task, allowing the researcher to be more confident when the probability score is higher. In this context, the cross-entropy (or log-loss) of a multinoulli distribution was chosen as an objective function to be minimized, to make the probability distribution of the model as close as possible to the empirical distribution (see Appendix B for further details) (Goodfellow et al., 2016).

2.5. Gridsearch

Every machine learning model has a set of parameters that must be predefined by the user and are not learned during training. These are called "hyperparameters". Grid Search is an efficient tool to find the best combination of hyperparameters, given a predefined grid.

All combinations of hyperparameters can be found in the project's file (Appendix D) (refer to param_grid.py). Just to mention a few, Logistic Regression, for example, was tested using diverse penalty schemes (l1, l2 and Elastic Net with different l1 ratios ranging from 0.1, 0.2,..., 0.9), different regularization parameter C (C = 10^j for j = -2, -1, .., 3) and several solvers depending on the penalty (newton-cg, sag, saga, lbfgs and/or liblinear).

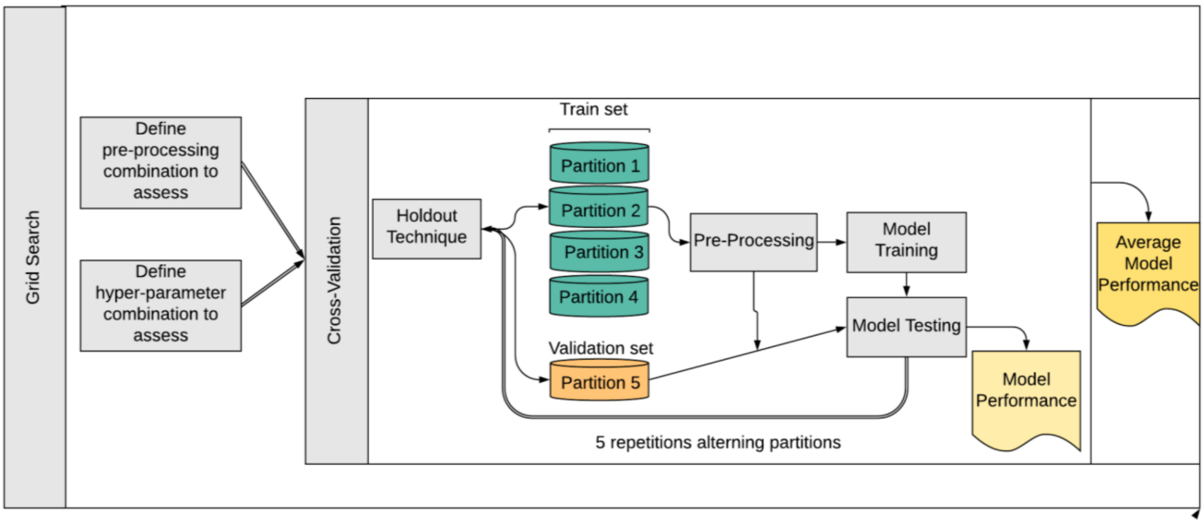

Figure 1 - Graphical representation of the pipeline used with Grid Search for assessing multiple classifiers.

In a cross-validation we split the data in k folds (or partitions), for each fold we train the model in the remaining data (all data except the fold) and evaluate the model in the fold; in the end the k results are averaged producing a score. Such a strategy was used to evaluate the grid search, in a 5-fold cross-validation scheme (Figure 1). We fixed the seed that generated the partition in the cross-validation, therefore a reproduction of this experiment is likely to achieve the same result.

Throughout this section whenever we refer to a "training set" we are referring to a new training set created by cross-validation or another validation mechanism. Otherwise, when referring to the initial training set, we will refer to it as the "original training set".

Cross-validation was performed within Grid Search to avoid overfitting on the validation set - because these hyperparameters can be tuned optimally for a specific validation set. That is, for a certain combination of hyperparameters, 5 different training and test sets are used to train and test the model's performance. The final result for that specific combination is the average of these tests (standard deviations were also computed).

2.6. Monte Carlo Cross-Validation

After the GridSearch, we looked for the best scores. Decision Tree e Gaussian Naive Bayes classifiers performed very poorly. Among the remaining models, we selected the most promising configuration in each case: the top 2 Neural Networks (with less or more neurons), top 2 Support Vector Machine (linear kernel and 'rbf' kernel), the best Random Forest, the best kNN and the best Logistic Regression. These models were subjected to subsequent analysis considering a Monte Carlo Cross-Validation (MCCV) (briefly described next). The remaining classifiers showed a considerably higher log-loss and/or its parameters were unlikely to improve the performance; therefore, they were not forwarded to the next step. The selected model will be referred to as "the final hypothesis".

An MCCV consists in randomly splitting the data in train and validation sets, several times. It's a kind of holdout strategy with many trials, also known as "repeated learning-test". This technique allows us to draw a "monte carlo picture" and check a model's performance in a histogram, allowing us to compute some statistics like mean and standard variation. In our case, the MCCV consisted of 1700 trials. Since the grid search was done with a fixed seed, it can produce models that perform well in rare cases. By introducing randomness through MCCV, we can make a second filter in the models, in an attempt to avoid "rare best performing models".

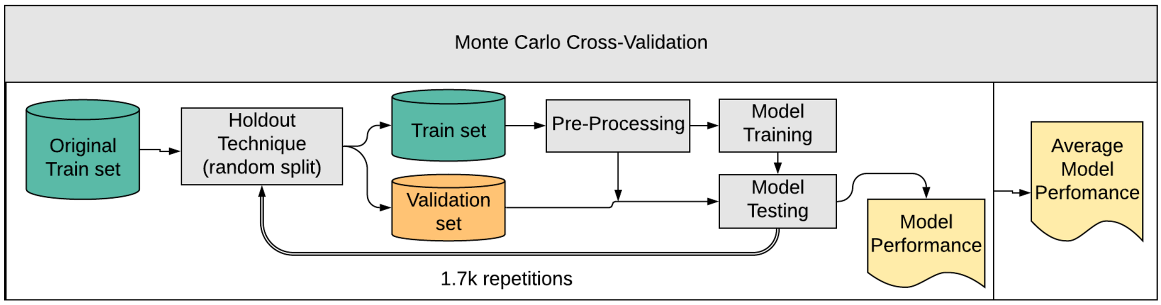

Figure 2 - Graphical representation of the pipeline used in the Monte Carlo Cross-Validation. After 1,700 repetitions the results are averaged. The process was repeated on the seven most promising models.

2.7. The Pipeline

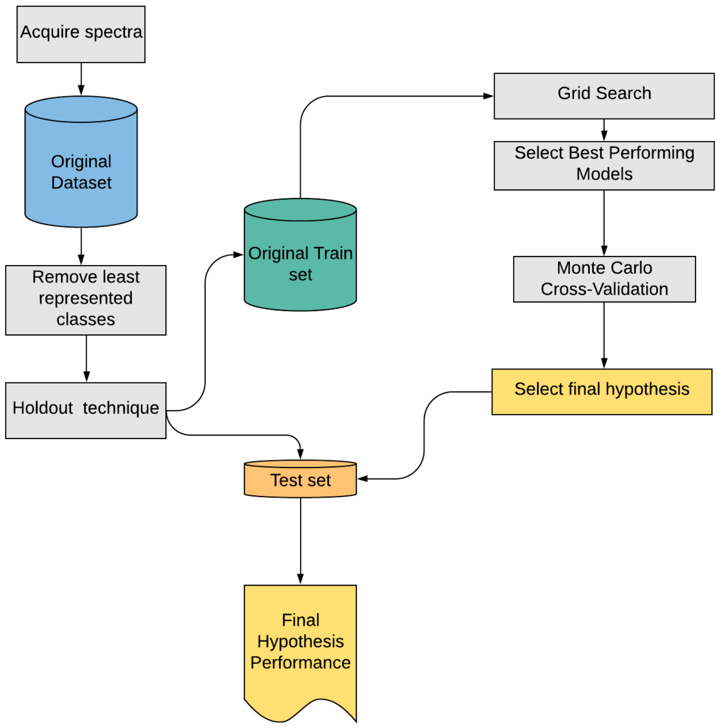

Figure 3 - Graphical representation of the full pipeline methodology.

The application of the tools described until now was done by means of a pipeline methodology. In total, 878 models under 5 combinations of pre-processing techniques were subjected to the same workflow. Thus, using a pipeline was useful to automate the sequence of operations that we proposed (Figure 3).

One important step to mention is the pre-processing of the data, which happens after the train/test split, something that is commonly overlooked, but that avoids the snooping bias. Notice, as well, that all steps concerning the selection of the best model are done using the "original training set", not the "original dataset". Keeping part of the original dataset aside, we can use it in the end to test the selected model on a dataset completely new to the algorithm.

We suggest this pipeline to be used to evaluate and compare the performances of different models in the task of microplastics classification with ATR-FTIR spectroscopy. It can be equally used with different datasets, containing, for example, other polymer classes.

2.8. Comparability and External Validation

In order to be able to compare our methodology with previous works, we applied, to the final hypothesis, the training methodology described by Kedzierski et al., 2019 (which consisted essentially of a MCCV with 1000 slightly different splits) with three differences: Firstly, we used a stratified split. This is important because our dataset is highly unbalanced with some classes underrepresented, therefore it is possible that in certain splits some classes remain only in the training set (more probable) or only in the test set (less probable), which can, in any case, introduce bias in our model. Secondly, Kedzierski et al. assign all unclear predictions (less than 3 votes in kNN) to the "unknown" class, which the present methodology does not. Finally, and most importantly, it seems that the authors included a standardization pre-process in the entire dataset, which introduces a "snooping" bias in the model. This means, for example, that the model "already knows" the mean and standard deviation of the features of the validation set before even being trained. On the other hand, our pre-processing methodology is built inside a pipeline that uses, for example, the standardization learned in the training set and applies it in the validation set, which avoids such bias.

A further verification was included to prove that the approach proposed here is usable outside the Kedzierski dataset. To do so, a different environmental dataset was used to train and evaluate the final hypothesis, which was done by MCCV, as described in section 2.6. There was no tuning of hyperparameters at this step. The dataset contained 800 FTIR spectra of polymers ingested by turtles and was published in a previous study by Jung et al. (2018). There were initially 9 classes, including a differentiation between high and low density poly(ethylene). As previously, the less representative classes were suppressed and spectra moved to the "unknown" class, which resulted in 5 classes.

2.9. Application of the classifier

If the reader is particularly interested in using the proposed final hypothesis (classifier), they can do so by running the code "final_*classifier.py*" in a python console, which is located in the root folder of the project repository.

It should be stressed that the data provided for classification should not have been pre-processed at all, as the appropriate pre-processing will be performed by the algorithm considering the results presented in the next section of this paper. Additionally, the spectral data should have been collected at the same frequency range described in section 2.1.

Given the limitation of the classifier to recognize only the classes it was trained on, the classification result will be given as a probability of that sample belonging to the assigned class. If the user is then unsure of the assignment, it can be visually confirmed. An example of this application and output is given in Appendix C.

3. Results/Discussion

Using the methodology described in the previous section, 28096 models were evaluated in the 5-fold cross-validation. Results (log-loss and standard deviation) for the best combination of hyperparameters and pre-processing techniques for each different model are presented in Figure 4.

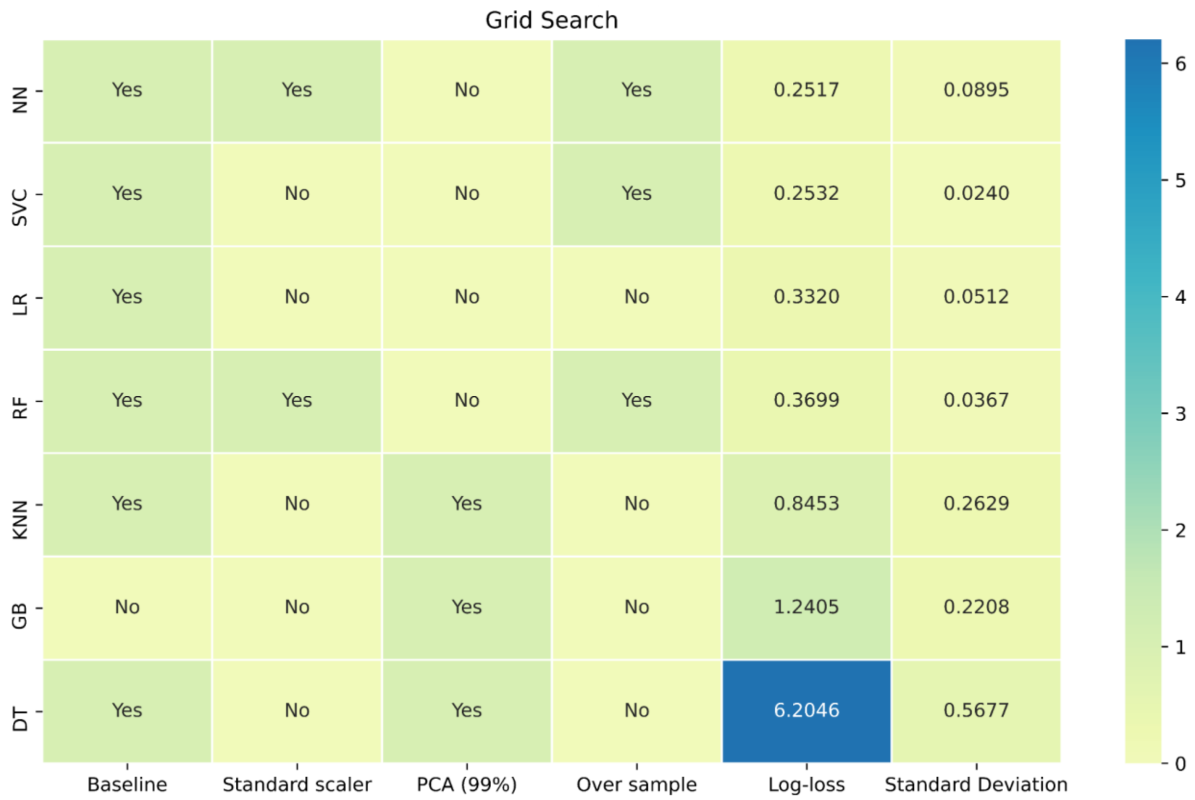

Figure 4 - Best performance found within the gridsearch for each evaluated algorithm, namely Random Forests (RF), Support Vector Machine Classifier (SVC), Logistic Regression (LR), K-Nearest Neighbours (kNN), Decision Trees (DT), Gaussian Naive Bayes (GB), considering the best combination of hyperparameters and pre-processing techniques - the baseline correction, Standard scaler, Principal Component analysis (PCA) and (Naive) oversampling. The log-loss shown here is the mean of the five results (cross-validation). The correspondent standard deviation is also presented.

For each simulation, a more detailed table can be found in Appendix D (paths can be found in Appendix A). Analyzing the GridSearch, the seven most promising models, based on the proposed metrics, were: LR with baseline correction; Sigmoid NN with baseline correction, standard scaler, and oversampling; Sigmoid NN with baseline correction and standard scaler; SVC with linear kernel, baseline correction, and oversampling; SVC with rbf kernel, baseline correction, and oversampling; RF with baseline correction, standard scaler, and oversampling; kNN with baseline correction and PCA. The reader can refer to the "*gridsearch"* subfolder in the project repository for further details.

The pre-processing combination that improved the performance was different for each model. For instance, SVC did not benefit from PCA, while kNN, DT, and GB did.

When comparing several models with and without PCA, most of them present little scoring difference when PCA is applied. Meanwhile, we verified significant differences in the learning time. For example, baseline + oversample took around 5 hour to run, whereas baseline + pca + oversample took about 17 min. It should be noticed, however, that this is not a detailed picture of which algorithm could benefited from PCA, but an experimental evidence to affirm that future analyses, mainly when they involve the comparison of several algorithms, can be done with PCA (99%) without significant loss of information but with a saving time bonus. All grid search and MCCV (next) simulations have their time computed and stored in the file "*time.csv*" in the "*results*" project's folder.

Previous results suggested the use of peak smoothing techniques over the spectra (da Silva et al. 2020; Zimmermann et al. 2013). In our case, however, a Savitzky-Golay filter did not improve the performance of any model. For this reason, the procedure was not included in the project. Such a phenomenon probably happens because smoothing peaks favours experts' visual inspection, while in a pattern recognition this is likely to be irrelevant.

The next step was to subject the most promising models to an MCCV in order to validate the consistency of their performances when randomness is allowed. The resulting histogram of each model is presented in Figure 5.

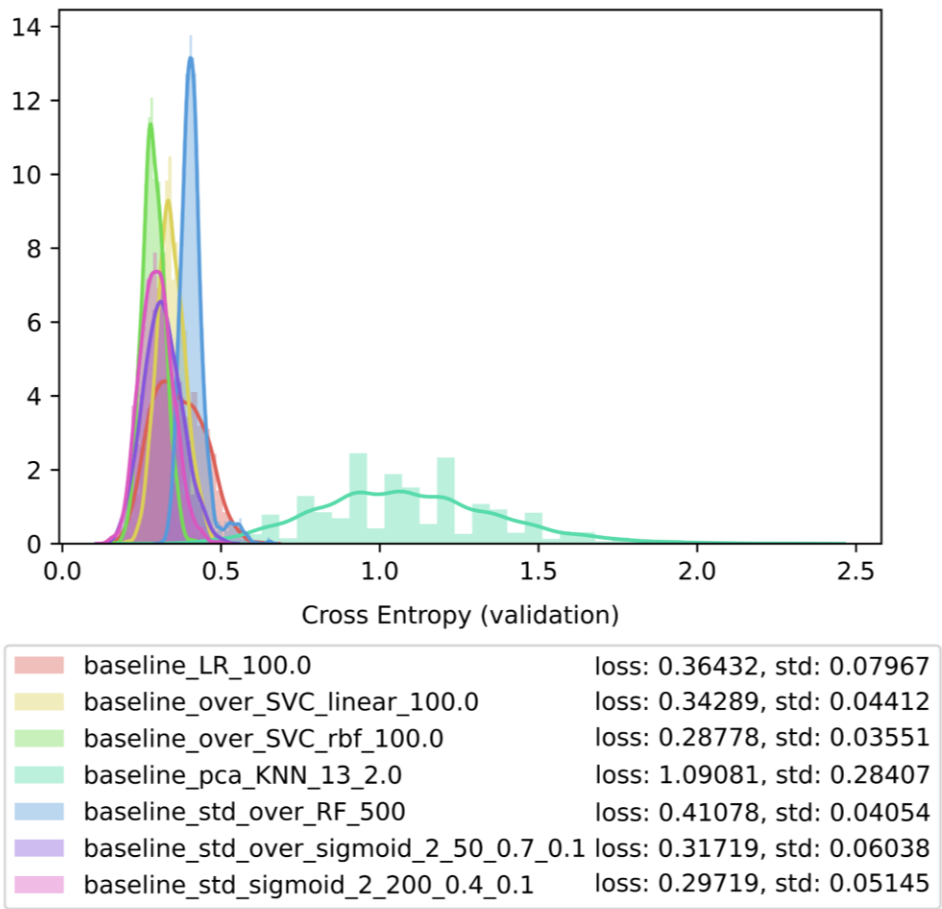

Figure 5 - Cross-entropy log-loss for the seven most promising models: Logistic Regression with baseline correction; SVC with linear kernel, baseline correction, and oversampling; SVC with rbf kernel, baseline correction, and oversampling; kNN with baseline correction and PCA; RF with baseline correction, standard scaler, and oversampling; Sigmoid NN with baseline correction, standard scaler, and oversampling; Sigmoid NN with baseline correction and standard scaler

Overall, the seven models had a decrease in performance (increased log-loss) but it could be said that they are, in fact, robust against randomness. Even though the results indicate that kNN and RF are the worst models, only by looking at these results it is not possible to choose the best model. To extend the analysis, the authors examined their accuracies and the results were very similar between models, so no further conclusion could be made. Considering the methodology proposed by (Renner et al., 2019), which was also used by (Kedzierski et al. 2019), we evaluated, for each model, how many classes were "working well" (sensitivity greater than 75%) and equal conclusions were possible. Therefore, since none of these metrics provided clear boundaries for a final decision, we shall take into account Occam's Razor, which states that "the simplest model that fits the data is also the most plausible" (Abu-Mostafa et al., 2012, p. 167). Appropriately, SVC with a linear kernel was chosen as the final hypothesis. The reader can refer to the "*results/graphics"* subfolder in the project repository for further graphs in this matter.

More complex, non-linear models, such as neural networks, can be tuned to fit the data well. Nonetheless, if the data were collected under different circumstances, or other pre-processing techniques or parameters were changed, this would cause minor changes in the final performance of every model. The neural networks, for instance, would probably still perform well, but there would be many more hyper-parameters to tune again in order to obtain optimal results. Linear SVC has only one such parameter (C) that could be readily tuned in case new reference samples or a whole new training set was provided. Having considered this, SVC is expected to perform efficiently with low tuning and prediction times.

Additionally, regarding the similarity in the performance of distinct models under various criteria, it is the authors' opinion that the "ideal theoretical structure" of the data is indeed linearly separable. By "ideal theoretical structure" of the data, we mean: no error in data labeling, materials without wear, and all micro-plastic polymers well represented, thus we conjecture that under FTIR the microplastics exhibits an almost linear pattern, with linearity suffering small violations due to the presence of these "non-ideal elements". Another important remark is that our model improved performance when combined with oversample technique, agreeing with the fact that in an imbalanced problem, a SVC classifier can produce suboptimal models that are biased toward the majority class (He, H., Ma, Y., 2013, p. 83). Lastly, Linear SVC can be implemented using the LIBLINEAR library, which is capable of handling large data sets (Fan et al., 2008). Therefore, we expect that the model could be implemented in a massive dataset, in which we expect similar results[3].

Since the methodology describes supervised learning procedures, it can be expected to observe where the algorithm makes mistakes when tested on a validation set. This is possible by looking at the confusion matrix shown in Figure 6, generated by averaging results obtained for every split in the MCCV for the selected model.

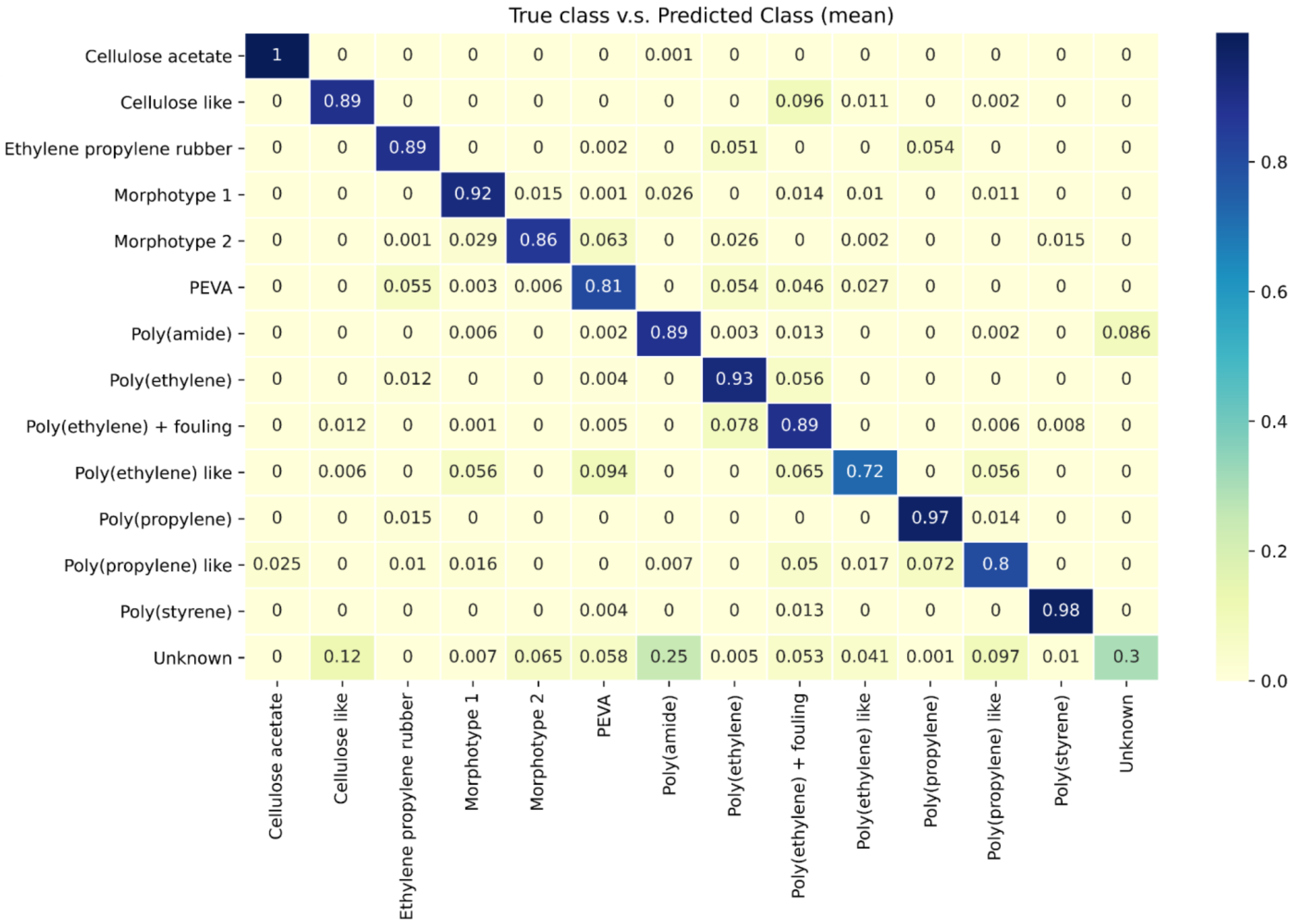

Figure 6 - Confusion Matrix for the results of MCCV for the final hypothesis. Each row represents the true class, meanwhile each column is the class predicted proportion, on average, by baseline_over_SVC_linear_100.0. Ideally, the matrix should be an identity matrix (perfect match).

A confusion matrix compares the true class in the validation set with the class predicted by the algorithm for the same sample (Ballabio et al., 2018). Considering this, it can be observed, for instance, that samples from the “Cellulose Acetate” class were predicted to belong to their true class every time, whereas for the “Cellulose Like” class, 9.6% of all predictions were incorrectly assigned to the “Poly(ethylene) + fouling” class and 1.1% were incorrectly assigned to the “Poly(ethylene) like” class. The same kind of error was observed by Kedzierski et al. (2019) using a kNN classifier. Since the algorithm is trained to distinguish between classes by identifying patterns in the provided dataset, which only contains information on spectral features and polymer classes, this could indicate similarities between spectra of both classes. As pertinently pointed out by the referred authors, spectral bands associated with aging and biofouling of microplastics could be the cause, with a preponderance of the latter. A visual examination of spectra from the “Poly(ethylene) + fouling” and “Cellulose Like” classes was done to check for specific bands related to these changes in their molecular fingerprint. However, it was not possible to visually establish any specific relations that could be causing the algorithm to mistake both classes, as spectra had too many overlapping peaks. We emphasize that the results are promising, nonetheless.

It may be clearer to depict faults made by the algorithm by observing the “Ethylene Propylene Rubber” (EPR) class, where the algorithm wrongly predicted the “Poly(ethylene)” (PE) and “Poly(propylene)” (PP) classes. This is comprehensible since EPR is made of the same monomers as PE and PP and share their chemical characteristics, hence, their spectral features. For classes such as “Poly(ethylene) like”, “Poly(ethylene) + fouling” and “Poly(propylene) like”, the main error made by the algorithm was to assign samples truly belonging to these classes to their regular counterparts (i.e. “Poly(ethylene)” and “Poly(propylene)”.

In the case of “Poly(propylene) like”, which had one of the worst results, most incorrect predictions (7,2%) were made to the “Poly(propylene) class. Since they were originally assigned to these classes by an expert due to their spectral similarities, this could be expected - and, in fact, was also observed by Kedzierski et al (2019). However, despite having similar spectra, the algorithm was able to differentiate them with considerable accuracy, showing better class results than the k-nearest neighbour (k-NN) method originally proposed by the referred authors. Further research on the quality of the data can respond more assertively if errors of the algorithm are indeed mistakes or indicative of a human error in labelling the data. The class “Morphotype 2” was mistaken mostly with “Poly(amide)”, however, it returned excellent results meaning this grouping is probably very concise and should contain samples from remarkably similar polymers, or even, mostly samples from a single polymer type. It is also unlikely that this polymer is one of the others already in this database.

Concerning the oversample methodology, adding new samples in the distribution of the training set can “impose non-uniform error costs, causing the learner to be biased in favour of predicting the rare class” (He and Ma, 2013, p. 37). As we can see in Figure 6, the proposed model does not suffer from this problem, since the less representative classes are not “stealing” sensitivity of the more representative ones, like (PE) or (PP). This indicates, within the results shown, that the model properly handled the unbalanced dataset problem.

After these considerations, the sensitivity calculated from results shown in Figure 6 is presented in Figure 7, along with results from a previous paper and the final test results.

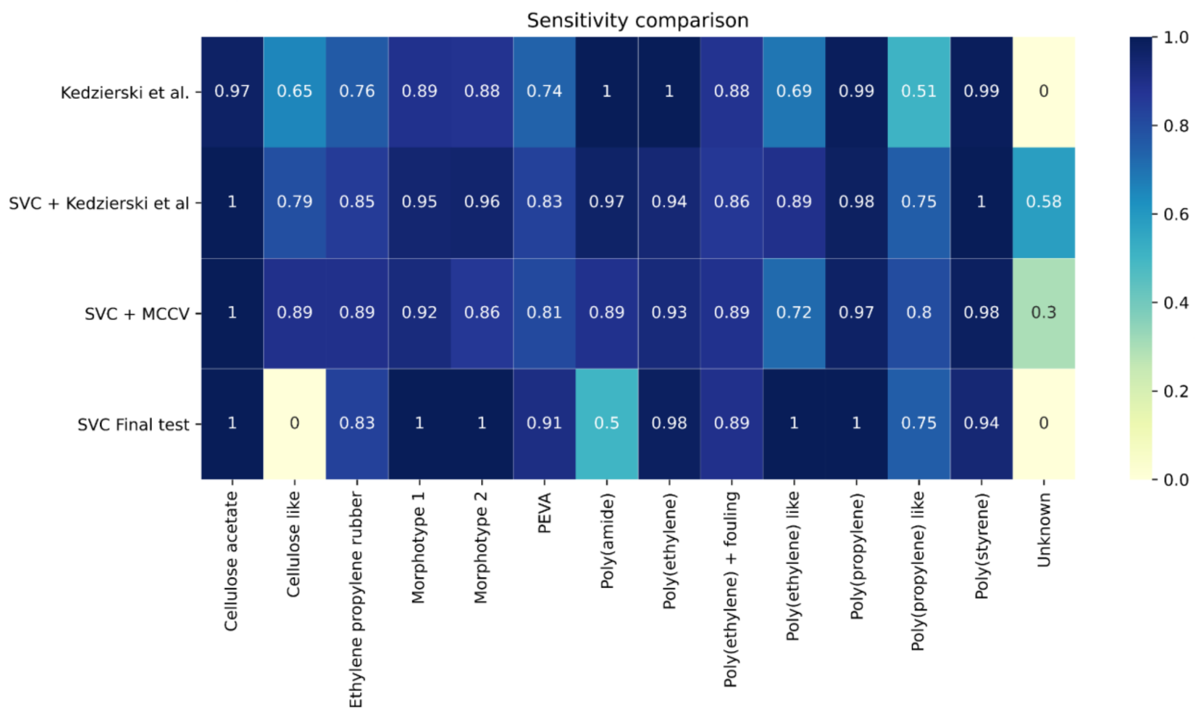

Figure 7 - Sensitivity for every class in the dataset according to results published by Kedzierski et al. (2019) using a k-Nearest Neighbour algorithm, Support Vector Machine Classifier (SVC) in the methodology proposed by Kedzierski et al. (2019), SVC in the Monte Carlo Cross-Validation and SVC in the final test.

Assuming that the dataset is statistically representative, the Final test results are what could be approximately expected of the performance of the algorithm in “real” conditions, that is, with the application of the deployed model to non-labeled or unknown spectra. Contrarily to the results given by the MCCV and the methodology adapted from Kedzierski et al. (2019) (refer to section 2.7. for a comparison between methodologies), which used the entire dataset for training, having test samples used for training even if in different splits/simulations, the test set in this case was completely new to the algorithm, giving an “unbiased” prediction.

All methodologies had comparable sensitivities, with those from Kedzierski et al. (2019) being lower for some classes (Figure 7). Nonetheless, the referred author had shown that conventional machine learning algorithms could be powerful tools for classifying microplastics spectra. Interestingly, classes that did not perform as well with the kNN algorithm, like “PEVA”, “Poly(ethylene) like” and Poly(propylene) like” also had worse sensitivities with SVC in the MCCV, but presented a slight improvement. In general, this comparison indicates that Support Vector Machine Classifiers offer more appropriate classification boundaries for this specific task with even better scalability, since kNN algorithms are highly computationally demanding with large datasets (Gutierrez et al., 2016).

The “Cellulose like” class presented no sensitivity, however. Among 240 samples, 4 of them belong to this class, therefore this undesirable event is plausible to happen and should not be interpreted as the actual performance for that class. In this regard, we expect that more samples in this class could improve the model performance.

Assuming the data is large enough, independent and identically distributed (i.i.d), the overall accuracy (non class-specific) of the model (which can be given with a 95% probability) is approximately 91.25% +/- 3.6% (Mitchell, 1997, p. 132, formula 5.1). The condition “large enough data” ensures that the sample mean will be approximately normal, due to the Central Limit Theorem (Mitchell, 1997, p. 142). A more conservative confidence interval without such assumption can be computed by using Hoeffding's Inequality (Hoeffding, 1963), which can establish the following bound for the generalization accuracy (Abu-Mostafa, 2012, p. 40):

$$|A_{out}(g) - A_{in}(g)| = |E_{out}(g) - E_{in}(g)| \leq \sqrt{\frac{1}{2N} ln \frac{2M}{\delta}},$$

with probability greater than $1 - \delta$ where M is the number of final models to be evaluated in the test set, $A_{out}(g)$is the generalization accuracy of the model g, $A_{in}(g)$ is the in-sample accuracy, similarly $E_{out}(g)$ the generalization error (the ratio of misclassified labels) and $E_{in}(g)$ the in-sample error. With $N = 240, M = 1$ and δ= 0.05 and assuming the data is i.i.d, the confidence interval is 91.25% +/- 8.8% with 95% of probability.

In regard to the model tested on the Jung dataset, a detailed score is presented in Figure 8. The lower sensitivities for “LDPE” and “Mixture” are attributed to errors similar to those found in the confusion matrix in Figure 6.

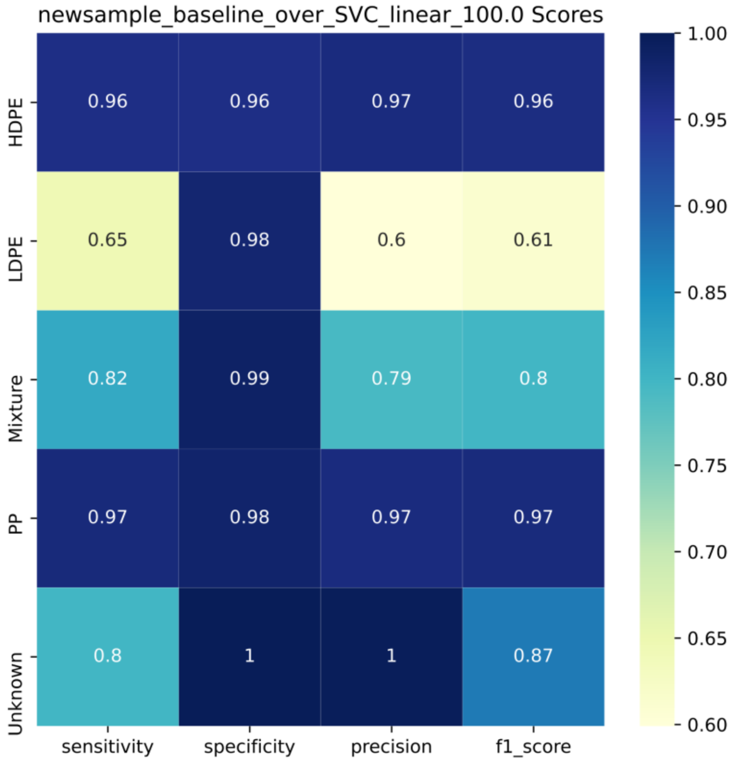

Figure 8 - Detailed score for the MCCV applied to the Jung dataset considering five classes: High-density Poly(ethylene) (HDPE), Low-density Poly(ethylene) (LDPE), a mixture of PE and PP, Poly(propylene) (PP) and Unknown

The model's cross-entropy log loss was 0.24554 (with a standard deviation of 0.0662) and the accuracy was 94% (with a standard deviation of 0.01209). This result is in accordance with those from the Kedzierski dataset in the MCCV (shown in Figure 5), although their comparison is not straightforward since the number of classes is different. The excellent performance of the Linear SVC classifier on a different dataset is another evidence of the linear separability of MP spectra.

4. Conclusion

The present study was able to demonstrate the performance of different machine learning classification algorithms to the classification of ocean microplastics using previously identified samples' ATR-FTIR spectra as input data. For this purpose, we proposed log-loss to measure models' multi-class probabilities, an approach that has not been used previously in machine learning methods for microplastics characterization.

We presented a machine learning combination not yet proposed for classification of microplastics, choosing linear SVC as the final classifier after thoroughly evaluating multiple conventional classification algorithms. The rigorous pipeline methodology described in this paper is essential to avoid introducing bias in the model training and evaluation, which supports the selection of the algorithm and substantiates the performance obtained.

The best pre-processing of the raw spectra was also evaluated for each classifier. We identified that for the selected model, only a baseline correction and a naive oversampling was more effective.

Linear SVC is well suited for scaling up. After being trained, the classifier can be directly deployed and applied to classify unknown spectra. The user doesn't have to upload the database or train the SVC. Given a spectra, the proposed model would return a probability score for each class, rather than simply a deterministic output. This procedure seems to be more realistic with practice, since the researcher may have greater or lesser confidence in the resulting model's evaluation, depending on the probabilities returned.

This study attests for the use of this methodology applied to ATR-FTIR data. In many cases, other vibrational spectroscopic techniques may have been used, namely Raman and microspectroscopy. Despite the new challenges these techniques impose, such as particle morphology, moisture, and blank spectra, in the case of an FPA detector, due to the similarities between them, similar results could be expected. This hypothesis could be verified by applying the same evaluation methodology described in this paper.

Machine learning models work within the learned database, which means that the database must be representative of the required task. Thus, it should be applied with caution to spectra collected under different circumstances. Beyond that, the major limitation of the learning approach is that the final model cannot predict samples for a class it does not know, or doesn't know well, which could restrict the full deployment of this procedure, seeing that polymers can be of many kinds.

The disproportionate amount of samples in some classes in the dataset was addressed in this study, but not solved. Regardless of the method adopted, "artificially balancing the training distribution may help with the effects of class imbalance, but does not remove the underlying problem" (He and Ma, 2013). In other words, the most appropriate solution to the problem needs to be solved through some specialized algorithm, which does not seem to be the case, since the statistical support will be missing anyway, or through the acquisition of new data, which seems to be the critical point to solve the issue of the unbalanced dataset and even improve the results in this and future works.

Data sharing is necessary for the improvement of machine learning algorithms as it increases the size and diversity of training databases. Nonetheless, the authors came across some obstacles, such as: different file formats, huge files with lots of redundant metadata, missing class labels on identified spectra and data that were already pre-processed. We recommend the publication of raw spectra, as the prior pre-processing may introduce bias. The JCAMP-DX file format is a standard defined by the The International Union of Pure and Applied Chemistry (IUPAC) and is considered the optimal format for sharing spectroscopy data. Also, providing the polymer class labels of every spectra should reduce redundant work and speed the collective learning process.

Since identification is a critical step in the study of these contaminants, improving confidence and speeding up the process are crucial to the advancement of the area. Despite not yet being able to stand alone as a method to automatically classify every conceivable sample, given its limitations, this study presents robust statistics to support the utilization of machine learning methods to the problem of automatic classification of microplastics.

[1] Comparing with Kedzierski et al., 2019 dataset, we deleted 12 spectra, since they were not available as raw data.. Spectras related to Cellulose Acetate are also unavailable in the raw data, but we requested to Kedzierski, which gently provided us the data.

[2] Quotes were used here because, despite every effort to minimize bias, machine learning models cannot really be unbiased.

[3] As long as this massive dataset adequately represents the distribution of MP classes in a way that the imbalance is less deleterious. Even in a scenario with little data, Linear SVC without oversample is still highly competitive.

Acknowledgements

This work was funded by the brazilian government through the National Council for Scientific and Technological Development (CNPq). Thanks are also due to Mikael Kedzierski, author of a previous paper, who was considerate and transparent in regard to his own work.

References

Abu-Mostafa, Y.S., Magdon-Ismail, M., Lin, H.-T., 2012. Learning From Data.

Andrady, A.L., 2017. The plastic in microplastics: A review. Marine Pollution Bulletin 119, 12–22. https://doi.org/10.1016/J.MARPOLBUL.2017.01.082

Barnes, D.K.A., 2002. Biodiversity: Invasions by marine life on plastic debris. Nature 416. https://doi.org/10.1038/416808a

Ballabio, D., Grisoni, F., & Todeschini, R. 2018. Multivariate comparison of classification performance measures. Chemometrics and Intelligent Laboratory Systems, 174, 33-44.

Brandt J, Bittrich L, Fischer F, et al. High-Throughput Analyses of Microplastic Samples Using Fourier Transform Infrared and Raman Spectrometry. Applied Spectroscopy. 2020; 74(9):1185-1197. https://doi.org/10.1177/0003702820932926

Burman, P., 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika 76. https://doi.org/10.1093/biomet/76.3.503

Cabernard, L., Roscher, L., Lorenz, C., Gerdts, G., Primpke, S., 2017. Comparison of Raman and Fourier Transform Infrared Spectroscopy for the Quantification of Microplastics in the

Aquatic Environment. Environmental Science & Technology 2018 52 (22), 13279-13288. https://doi.org/10.1021/acs.est.8b03438

Chae, Y., An, Y.J., 2017. Effects of micro- and nanoplastics on aquatic ecosystems: Current research trends and perspectives. Marine Pollution Bulletin 124. https://doi.org/10.1016/j.marpolbul.2017.01.070

Classifier comparison [WWW Document], n.d. . https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html.

Conroy, J., Ryder, A. G., Leger, M. N., Hennessey, K., Madden, M. G., 2005. Qualitative and quantitative analysis of chlorinated solvents using Raman spectroscopy and machine learning", Proc. SPIE 5826, Opto-Ireland 2005: Optical Sensing and Spectroscopy. https://doi.org/10.1117/12.605056

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(20), 13724-13733.

Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C. 2008. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research

Fendall, L.S., Sewell, M.A., 2009. Contributing to marine pollution by washing your face: Microplastics in facial cleansers. Marine Pollution Bulletin 58. https://doi.org/10.1016/j.marpolbul.2009.04.025

Forests of randomized trees [WWW Document], n.d. . https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees.

Free, C.M., Jensen, O.P., Mason, S.A., Eriksen, M., Williamson, N.J., 2014. High-levels of microplastic pollution in a large , remote , mountain lake. Marine Pollution Bulletin 85, 156–163. https://doi.org/10.1016/j.marpolbul.2014.06.001

Galgani, F., Hanke, G., Maes, T., 2015. Global distribution, composition and abundance of marine litter, in: Marine Anthropogenic Litter. https://doi.org/10.1007/978-3-319-16510-3_2

Géron, A. 2017. Hands-on Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media

GESAMP, 2019. Guidelines for the monitoring and assessment of plastic litter in the ocean.

Ghosal, S., Chen, M., Wagner, J., Wang, Z., Wall, S., 2018. Molecular identification of polymers and anthropogenic particles extracted from oceanic water and fish stomach: A Raman micro-spectroscopy study *. Environmental Pollution 233, 1113–1124. https://doi.org/10.1016/j.envpol.2017.10.014

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Gutierrez, P., Lastra, M., Bacardit, J., Benitez, J., Herrera, F., 2016. GPU-SME-kNN: scalable and memory efficient kNN and lazy learning using GPUs. Inf. Sci., 373 (2016), pp. 165-182

He, H., Ma, Y., 2013. Imbalanced learning: Foundations, algorithms, and applications, Imbalanced Learning: Foundations, Algorithms, and Applications. https://doi.org/10.1002/9781118646106

Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*. **58** (301): 13–30. doi:10.1080/01621459.1963.10500830.

Hidalgo-Ruz, V., Gutow, L., Thompson, R.C., Thiel, M., 2012. Microplastics in the Marine Environment: A Review of the Methods Used for Identification and Quantification. Environmental Science Technology 46, 3060–3075. https://doi.org/10.1021/es2031505

Hufnagl, B., Steiner, D., Renner, E., Löder, M.G.J., Laforsch, C., Lohninger, H., 2019. A methodology for the fast identification and monitoring of microplastics in environmental samples using random decision forest classifiers. Analytical Methods 11, 2277–2285. https://doi.org/10.1039/C9AY00252A

Jolliffe, I. T., 2002. Principal Component Analysis. Springer

Melissa R. Jung, George H. Balazs, Thierry M. Work, T. Todd Jones, Sara V. Orski, Viviana Rodriguez C., Kathryn L. Beers, Kayla C. Brignac, K. David Hyrenbach, Brenda A. Jensen, and Jennifer M. Lynch., 2018. Polymer Identification of Plastic Debris Ingested by Pelagic-Phase Sea Turtles in the Central Pacific. Environmental Science & Technology 52 (20), 11535-11544 DOI: 10.1021/acs.est.8b03118

Jung, M.R., Horgen, F.D., Orski, S. v., Rodriguez C., V., Beers, K.L., Balazs, G.H., Jones, T.T., Work, T.M., Brignac, K.C., Royer, S.J., Hyrenbach, K.D., Jensen, B.A., Lynch, J.M., 2018. Validation of ATR FT-IR to identify polymers of plastic marine debris, including those ingested by marine organisms. Marine Pollution Bulletin 127, 704–716. https://doi.org/10.1016/j.marpolbul.2017.12.061

Käppler, A., Fischer, D., Oberbeckmann, S., Schernewski, G., Labrenz, M., Eichhorn, K.-J., Voit, B., 2016. Analysis of environmental microplastics by vibrational microspectroscopy: FTIR, Raman or both? Analytical and Bioanalytical Chemistry 408, 8377–8391. https://doi.org/10.1007/s00216-016-9956-3

Kedzierski, M., Falcou-Préfol, M., Kerros, M.E., Henry, M., Pedrotti, M.L., Bruzaud, S., 2019. A machine learning algorithm for high throughput identification of FTIR spectra. Chemosphere.

Kotsiantis, S.B., 2007, Supervised machine learning: A review of classification techniques. Informatica (Ljubljana), 31 (3), 249-268.

Kuptsov, a. h., Zhizhin, g. n., 1998. Handbook of Fourier Transform Raman and Infrared Spectra of Polymers, Bioseparation.

Kyoung, Y., Hee, S., Jang, M., Myung, G., Rani, M., Lee, J., Joon, W., 2015. A comparison of microscopic and spectroscopic identification methods for analysis of microplastics in environmental samples. Marine Pollution Bulletin 93, 202–209. https://doi.org/10.1016/j.marpolbul.2015.01.015

Lebreton, L., Slat, B., Ferrari, F., Aitken, J., Marthouse, R., Hajbane, S., 2018. Evidence that the Great Pacific Garbage Patch is rapidly accumulating plastic. Scientific Reports 1–15. https://doi.org/10.1038/s41598-018-22939-w

Lemaître, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research 18.

Li, J., Hibbert, D.B., Fuller, S., Vaughn, G., 2006. A comparative study of point-to-point algorithms for matching spectra. Chemometrics and Intelligent Laboratory Systems 82. https://doi.org/10.1016/j.chemolab.2005.05.015

Mitchell, T. M., 1997. Machine Learning. McGraw-Hill Science

Moore, D.H., 1987. Classification and regression trees, by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Brooks/Cole Publishing, Monterey, 1984,358 pages, Cytometry 8. https://doi.org/10.1002/cyto.990080516

Murphy, K.P., 2012. Machine learning: a probabilistic perspective (adaptive computation and machine learning series), Mit Press. ISBN.

Nordhausen, K., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. International Statistical Review 77. https://doi.org/10.1111/j.1751-5823.2009.00095_18.x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12.

Prati, R.C., Batista, G.E.A.P.A., Monard, M.C., 2009. Data mining with unbalanced class distributions: Concepts and methods, in: Proceedings of the 4th Indian International Conference on Artificial Intelligence, IICAI 2009.

Primpke, S., Lorenz, C., Gerdts, G., 2017. An automated approach for microplastics analysis using focal plane array (FPA) FTIR microscopy and image analysis. Analytical Methods 1499–1511. https://doi.org/10.1039/c6ay02476a

Primpke, S., Wirth, M., Lorenz, C., Gerdts, G., 2018. Reference database design for the automated analysis of microplastic samples based on Fourier transform infrared (FTIR) spectroscopy. Analytical and Bioanalytical Chemistry 5131–5141.

Primpke, S., Cross, R.K., Mintenig, S.M., Simon, M., Vianello, A., Gerdts, G. and Vollertsen, J. 2020. Toward the Systematic Identification of Microplastics in the Environment: Evaluation of a New Independent Software Tool (siMPle) for Spectroscopic Analysis. Appl. Spectrosc. 74(9), 1127-1138.

RandomForestClassifier [WWW Document], n.d. . https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Renner, G., Nellessen, A., Schwiers, A., Wenzel, M., Schmidt, T.C., Schram, J., 2019. Data preprocessing & evaluation used in the microplastics identification process : A critical review & practical guide. Trends in Analytical Chemistry 111, 229–238. https://doi.org/10.1016/j.trac.2018.12.004

Rocha-Santos, T., Duarte, A.C., 2015. A critical overview of the analytical approaches to the occurrence, the fate and the behavior of microplastics in the environment. TrAC - Trends in Analytical Chemistry. https://doi.org/10.1016/j.trac.2014.10.011

Saito, H., Fletcher, R., Yogi, T., Kayo, M., Miyagi, S., Ogido, M., Fujikura, K., 2018. Human footprint in the abyss : 30 year records of deep-sea plastic debris Sanae Chiba. Marine Policy 96, 204–212. https://doi.org/10.1016/j.marpol.2018.03.022

Sauvé, S., Desrosiers, M., 2014. A review of what is an emerging contaminant. Chemistry Central Journal. https://doi.org/10.1186/1752-153X-8-15

Shim, W.J., Hong, S.H., Eo, S.E., 2017. Identification methods in microplastic analysis: A review. Analytical Methods 9, 1384–1391. https://doi.org/10.1039/c6ay02558g

sklearn.calibration.CalibratedClassifierCV. viewed 08 June 2021, <*https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html#sklearn.calibration.CalibratedClassifierCV*>

Song, Y.K., Hong, S.H., Jang, M., Han, G.M., Shim, W.J., 2015. Occurrence and Distribution of Microplastics in the Sea Surface Microlayer in Jinhae Bay, South Korea. Archives of Environmental Contamination and Toxicology 69. https://doi.org/10.1007/s00244-015-0209-9

Support Vector Machines [WWW Document], n.d. . https://scikit-learn.org/stable/modules/svm.html.

Thompson, R.C., Swan, S.H., Moore, C.J., vom Saal, F.S., 2009. Our plastic age. Philosophical Transactions of the Royal Society B: Biological Sciences. https://doi.org/10.1098/rstb.2009.0054

Trunk, G. v, 1979. A Problem of Dimensionality: A Simple Example given. IEEE Transactions on Pattern Analysis and Machine Intelligence 306–307.

Turner, A., Holmes, L., 2011. Occurrence, distribution and characteristics of beached plastic production pellets on the island of Malta (central Mediterranean). Marine Pollution Bulletin 62. https://doi.org/10.1016/j.marpolbul.2010.09.027

Vapnik, V., Estimation of Dependencies Based on Empirical Data. Empirical Inference Science: Afterword of 2006, New York: Springer, 2006.

Vapnik, V. The Nature of Statistical Learning. Information Science and Statistics. New York: Springer-Verlag, 2000.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. pietro, Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D. v., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T., Jones, T.R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., Vázquez-Baeza, Y., 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods 17. https://doi.org/10.1038/s41592-019-0686-2

Wang, J., Liu, X., Li, Y., Powell, T., Wang, X., Wang, G., Zhang, P., 2019. Microplastics as contaminants in the soil environment : A mini-review. Science of the Total Environment 691, 848–857. https://doi.org/10.1016/j.scitotenv.2019.07.209

Wolpert, D. H., 1996. The Lack of A Priori Distinctions between Learning Algorithms, Neural Computation 8(7), 1341–1390.

Xu, Q.S., Liang, Y.Z., 2001. Monte Carlo cross-validation. Chemometrics and Intelligent Laboratory Systems 56. https://doi.org/10.1016/S0169-7439(00)00122-2

Zettler, E.R., Mincer, T.J., Amaral-Zettler, L.A., 2013. Life in the “plastisphere”: Microbial communities on plastic marine debris. Environmental Science and Technology 47. https://doi.org/10.1021/es401288x

Zhang, Y., Kang, S., Allen, S., Allen, D., Gao, T., Sillanpää, M., 2020. Atmospheric microplastics: A review on current status and perspectives. Earth-Science Reviews. https://doi.org/10.1016/j.earscirev.2020.103118

Zimmermann, B., Kohler, A., 2013. Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy. Applied Spectroscopy 67, 892–902. https://doi.org/10.1366/12-06723