Advances in Artificial Intelligence and Machine Learning; Research 3 (1) 647-668 Received 13-01-2023; Accepted 06-02-2023; Published 15-02-2023

Machine Learning of Polymer Types From the Spectral Signature Of Raman Spectroscopy Microplastics Data

Sheela Ramanna s.ramanna@uwinnipeg.ca
FDepartment of Applied Computer Science
University of Winnipeg, Winnipeg,
Manitoba R3B 2E9, Canada

Danila Morozovskii morozovskii-d@webmail.uwinnipeg.ca
Department of Applied Computer Science
University of Winnipeg,
Winnipeg, Manitoba R3B 2E9, Canada

Sam Swanson sam@compoundconnect.ca
Compound Connect Winnipeg,
Manitoba Canada

Jennifer Bruneau jennifer@compoundconnect.ca
Compound Connect Winnipeg,
Manitoba Canada

Corresponding Author: Sheela Ramanna

Abstract

The tools and technology that are currently used to analyze chemical compound structures that identify polymer types in microplastics are not well-calibrated for environmentally weathered microplastics. Microplastics that have been degraded by environmental weathering factors can offer less analytic certainty than samples of microplastics that have not been exposed to weathering processes. Machine learning tools and techniques allow us to better calibrate the research tools for certainty in microplastics analysis. In this paper, we investigate whether the Raman shift values are distinct enough such that well studied machine learning (ML) algorithms can learn to identify polymer types using a relatively small amount of labeled input data when the samples have not been impacted by environmental degradation. Several ML models were trained on a well-known repository, Spectral Libraries of Plastic Particles (SLOPP), that contain Raman shift and intensity results for a range of plastic particles, then tested on environmentally aged plastic particles (SloPP-E) consisting of 22 polymer types. After extensive preprocessing and augmentation, the trained random forest model was then tested on the SloPP-E dataset resulting in an improvement in classification accuracy of 93.81% from 89%.

Citation: Sheela Ramanna, et al Machine Learning of Polymer Types From the Spectral Signature of Raman Spectroscopy Microplastics Data. Advances in Artificial Intelligence and Machine Learning. 2023;3(1):44.

Keywords: Environmental weathering, Raman Spectroscopy, Microplastics Data, Machine Learning, Polymer types, Feature Engineering, Supervised Learning

1. INTRODUCTION

Plastic pollution is exclusively the result of anthropogenic activities, with the majority of plastic entering the environment through land-based activities but ending up far from their source, having travelled though atmospheric and riverine pathways and degrading through multiple processes [1]. The durability and strength of plastics that make them suitable for a broad range of applications are also what cause them to disperse easily and have led to them becoming a global pollution problem. The primary reason they pose such a threat to the environment is their resistance to degradation, allowing them to persist for hundreds or thousands of years. However, their exposure to a variety of factors will result in them breaking down from macroplastics to microplastics [2].

Microplastics (less than 5mm) are composed of various polymers and include a broad array of chemical additives [3]. It is understood that microplastics can decay at different rates depending on climate conditions, and that different stages of decay pose differing levels of toxicity to plant and animal life [4]. Thus, the chemical diversity of microplastics is an important consideration. The impacts of microplastics range from those on marine, freshwater, and terrestrial ecosystems [1], on human health through ingestion of beverages and contamination in food and food packaging [5], and on microorganisms through uptake by zooplankton in freshwater ecosystems or interference with nutrient production and cycling in aquatic ecosystems. Finally, consumption of microplastics by humans through the food chain raises concerns about possible health risks and effects on the human body [2].

The two most promising techniques for microplastics analysis, are Raman and Fourier transform infrared (FTIR) spectroscopy [6]. The preferred method for identifying microplastics is Raman spectroscopy which is an indispensable tool for the analysis of very small microplastics less than 20 `m [7]. This is a vibrational spectroscopy technique based on the inelastic scattering of light [8, 9]. When laser light is inelastically scattered from a chemical, the inelastic energy, which indicates an energy difference, represents a change of the vibrational energy level in the bond or bonds in a molecule.

A Raman spectrum of the sample is created by plotting the Raman shift against the light frequency[1]. For example, in Fig 1, the y-axis gives the intensity of the scattered light, and the x-axis gives the energy of light. The specific type of material is marked with peaks in Raman spectroscopy. One of the primary advantages of Raman spectroscopy is that even after exposure to ultraviolet (UV) light, the Raman spectra of microplastics are not so altered that it no longer provides a polymer signal. This is significant as microplastics typically experience multiple forms of degradation with the majority of microplastics samples being degraded [10].

Machine learning (ML) algorithms such as Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), K-Neighbour methods (KNN), Artificial Neural Networks (ANN) have been successfully applied to Raman spectra data in diverse areas of science. In [11], ANN and KNN methods were used to predict the concentration of cocaine using Raman spectroscopy. The

[1] https://www.utsc.utoronto.ca/~traceslab/PDFs/raman_understanding.pdf

authors [12], use Raman spectroscopy to detect chemical changes in melanoma tissue of patients and achieve 85% (sensitivity) and 99% (specificity) results with ANNs. In [13], several well-known ML algorithms were applied to determine the mine of origin and extraction depth of samples by finding Raman spectral differences for variscite (phosphate mineral) specimens from the Gavà mining complex where the SVM classifier gave the best result of almost 90% classification accuracy. In [14], the SVM model was able to achieve a diagnostic accuracy of 92% for tuberculosis patients using Raman spectra of blood sera. RF classifier was used in the analysis of spectral information of various cultural heritage materials by [15]. The authors [16], classify seven types of oils using Raman spectroscopy: sunflower, sesame, hemp, walnut, linseed (flaxseed), sea buckthorn and pumpkin seeds where a subspace KNN ensemble classifier gave the best classification accuracy of 88.9%. In [17], the authors explore association between Raman spectroscopy and machine learning to differentiate fruit distillate samples (alcoholic beverage) to determine trademark, geographical and botanical origin. The best geographical classification of the fruit distillates was obtained with the ensemble (subspace KNN) method resulting in an accuracy of 90.9% for 30 samples. In [18], ANN and SVM algorithms were used to diagnose biochemical composition of biological fluids of patients with Alzheimer's disease based on near infrared (NIR) Raman spectroscopy with 84% sensitivity and specificity values. The authors [19], present deep learning methods to extract and analyze chemical information in big and complex datasets derived from Raman and surface-enhanced Raman scattering (SERS) techniques. In [20], 230 Raman spectra samples of high dimensional solvent and solvent mixtures (chemicals) were classified with deep neural networks using a locally connected architecture, resulting in a mean accuracy of 96.0%. In [21], Raman spectra of oral tongue squamous cell carcinoma and para-carcinoma tissues of 24 patients were analyzed. A convolutional neural network model was used to extract features, which were then input to an SVM classifier resulting in a 99.96% accuracy. AlexNet deep learning model was used to classify chronic renal failure using serum Raman spectra of 100 patients with an accuracy of 95.22% [22].

In [23], six types of common household plastics using Raman spectroscopy were evaluated to demonstrate the potential of machine learning methods such as principle component analysis, KNN as well as regression models for classification and prediction. In [24], hyperspectral imaging was used to detect micro plastic contamination in soils. Classification precision of 86% for polymers containining microplastics particles of size between 1-5 mm and about 99% precision for microplastics particles of size between 0.5-1 mm were obtained. In [25] laser-induced breakdown spectroscopy was used to create plastic samples containing different additives such as flame retardants. Principle component analysis (PCA) and Linear Discriminant Analysis (LDA) were used to discriminate 11 different types of additives with LDA achieving almost 100% accuracy. In [26], 4000 images belonging to the five categories of plastic resin codes from a public database were classified using convolutional neural networks with an accuracy of 99.79%. In [27], micro Fourier Transform Infrared (`-FTIR) hyper-spectral imaging with Partial least squares discriminant analysis (PLS-DA) and soft independent modelling of class analogy (SIMCA) which is based on PCA, were used to classify nine of the most common polymers in microplastics found on seabed sediment samples. A review of polymer informatics is presented in [28]. In [29], PCA and clustering with K-means on short wave infrared hyperspectral data prepared using reflection imaging with a hyperspectral camera was used to analyze and classify 13 commercially available plastics.

Our work differs from the more recent work where either hyperspectral imaging, digital images or laser-induced breakdown spectroscopy of plastics were used with machine learning models including deep learning models. The datasets used were either prepared by the authors or included

large image repositories suitable for deep learning. On the other hand, Raman spectroscopy data typically consist of approximately 1,000 to 3,000 data points. It is difficult and expensive to obtain the spectroscopy data, and only a limited amount of data is available online. Additionally, each sample might contain not one type of microplastic, but rather a combination of materials. To this end, in our work, several machine learning models were trained on a well-known repository, Spectral Libraries of Plastic Particles (SLOPP) containing 148 samples and 158 samples from Mendeley [2]. The SLOPP library contains Raman shift and intensity results of Raman spectroscopy laboratory analyses conducted at the Rochman Lab [3] in the Department of Ecology and Evolutionary Biology at the University of Toronto for a range of plastic particles. This library also includes environmentally aged plastic particles (SloPP-E) containing 97 samples.

The SLOPP/SLOPP-E is a single dataset with the only distinction being the categorization as weathered or not-weathered respectively. The dataset was collected in an otherwise uniform process with the specific purpose of having comparable datasets. A combined dataset of SLoPP and Mendeley (second dataset) was used as our training dataset, since the standard Mendeley data shared similar characteristics with SLOPP (non-weathered) data. Although the peaks and valleys of the SLOPP and Mendeley datasets did differ due to the non-identical collection, we were able to make comparisons on a rate of change basis nonetheless. SLoPP-E (weathered) was used as the testing dataset. After extensive preprocessing and augmentation, the trained random forest model was then tested on the SloPP-E dataset resulting in an improvement in classification accuracy of 93.81% from 89%. This work contributes to the understanding of environmental polymers by validating the machine learning methods that improve the predictive capability of Raman spectroscopy data analysis.

Our paper is organized as follows: In section 2, we give a description of the open source spectroscopy datasets considered in this work. In section 3, we present a detailed discussion of the preprocessing and augmentation techniques used in this research for generating training and testing examples. In section 4, we give an in-depth analysis of the classification results of our final model followed by concluding remarks in section 5.

2. MATERIALS- SPECTROSCOPY DATA

A Raman spectrum can provide molecular bond information on a particular substance and may be described as a "fingerprint" of the substance due to its uniqueness [30]. Raman spectra are a plot of scattered intensity as a function of the energy difference between the incident and scattered photons and are obtained by pointing a monochromatic laser beam at a sample [31]. The resultant spectra are characterized by shifts in wave numbers (inverse of wavelength in $2< \ 1$) from the incident frequency. The frequency difference between incident and Raman-scattered light is termed the Raman shift, which is unique for individual molecules. For this research, the following datasets have been used SLoPP, SLoPP-E, Mendeley. A combined dataset of SLoPP and Mendeley was used as our training dataset, while SLoPP-E was used as the testing dataset.

SLOPP: SLoPP is a spectral library of microplastic particles with 148 samples, having different polymer types (shown in TABLE 1), colours and morphologies. Examples of colours are

[2] https://data.mendeley.com/datasets/kpygrf9fg6/1
[3] https://rochmanlab.wordpress.com/spectral-libraries-for-microplastics-research/

turqoise, orange, green, white, grey, black, light brown and clear. Examples of morphologies include: fragments, sphere, film, foam, and fiber. SLoPP was collected in the range of 100-3500 $2<$ 1 and was created to include commonly used plastics.

FIGURE 1, illustrates the Raman spectra for one type of polymer (polypropylene). The y-axis shows the intensity of the scattered light, and the x-axis shows the energy (frequency) of light. Different colours represent different samples in the dataset. It can be observed, that the most distinguishing feature of the Raman spectroscopy is peaks on different energies of the light.

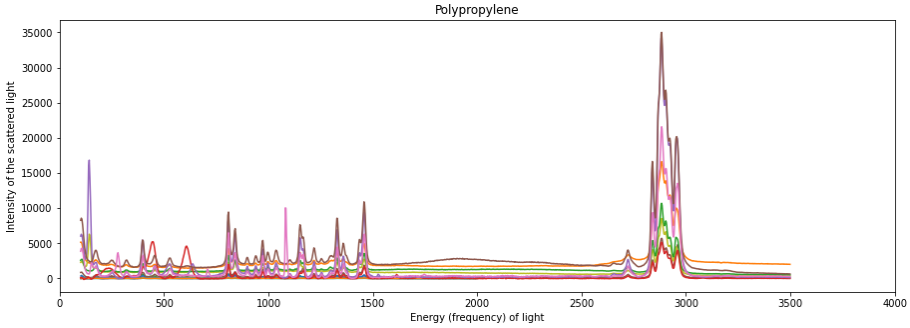

Figure 1: Raman spectra of **Polypropylene** from the SLoPP dataset.

SLOPP-E: SLoPP-E dataset is similar to the SLoPP dataset, however, it includes samples exposed to a variety of environmental conditions (e.g., some samples have undergone some chemical degradation, ageing). The microplastics in this library SLoPP-E include environmental samples obtained across a range of matrices, geographies, and time. FIGURE 2, illustrates the Raman spectra for the same type of polymer (polypropylene) as the one shown in FIGURE 1. Different colours represent different samples in the dataset. It can be observed that these two datasets share the same values on the x-axis (frequency) and similar intensities (peaks on the y-axis) for the same polymer type.

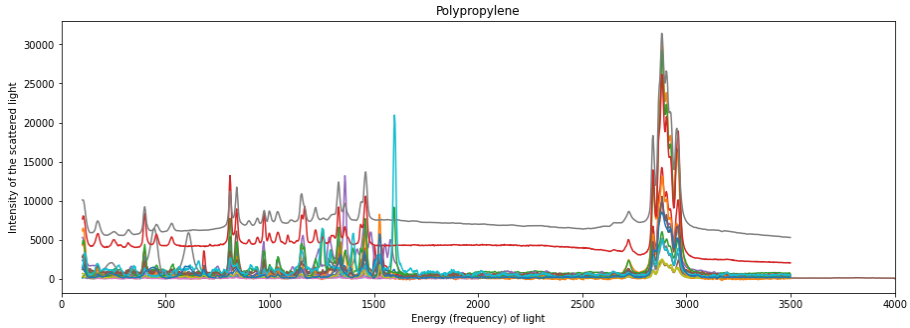

Figure 2: Raman spectra of **Polypropylene** from the SLoPP-E dataset.

TABLE 1, shows the distribution of polymers types for SLoPP and SLoPP-E. It can seen that some types are either missing in training (SLoPP) or testing (SLoPP-E) datasets. However, the training dataset includes many more types that are missing compared to the testing dataset.

Polymer Types	SLoPP samples	SLoPP-E samples
Acrylic	10	3
Acrylonitrile Butadiene Styrene	10	1
Cellulose Acetate	4	3
Cotton	16	-
Polyamide	7	7
Polycarbonate	7	2
Polyester	10	12
Polyethylene	24	26
Polyethylene Terephthalate	9	1
Polyethylene Vinyl Acetate	5	-
Polymethyl Methacrylate	1	3
Polypropylene	17	21
Polystyrene	11	9
Polyurethane	6	6
Polyvinyl Chloride	11	3
Dyed Cellulose	-	5
Polybutylene Terephthalate	-	1
Polyethylene Terephthalate-co-Polycarbonate	-	1
Polyethylene-co-Polypropylene	-	3
Polystyrene-co-Polyvinyl Chloride	-	1
Polysulfone	-	1
Rubber	-	4

Table 1: Data Distribution for SLoPP and SLoPP-E.

Mendeley: This dataset has two variations of microplastics: standard and weathered. The standard data is similar to SLoPP and the weathered data is similar to SLoPP-E (by description), subjected to environmental conditions.

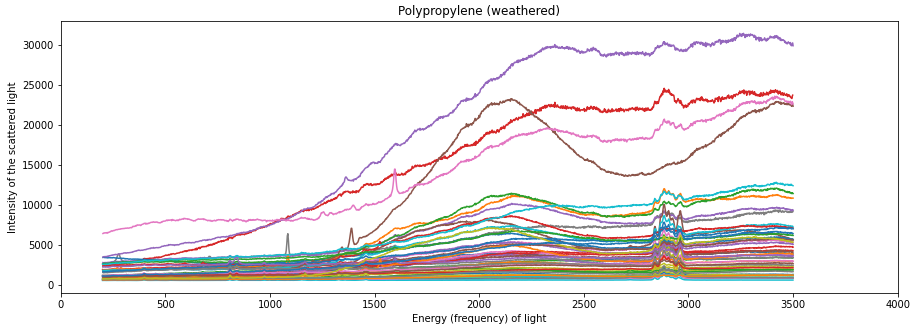

Figure 3: Raman spectra of **Polypropylene** from the Mendeley dataset.

A plot of the Raman spectroscopy for polypropylene is shown in FIGURE 3. From this plot, one can observe the following : i) some samples in the Mendeley dataset have a wave-like structure, and ii) the peaks (intensities of scattered light) are not as sharp and separated as

in the SLoPP or SLoPP-E datasets. TABLE 2, shows the data distribution for Mendeley dataset. The majority of samples in this dataset belong to two polymer types: polypropylene and polyethylene and can be used in the training dataset, as they are also present in the SLoPP dataset.

Polymer Types	Mendeley samples
Not detected	8
Acrylonitrile Butadiene Styrene	1
Nitrocellulose	1
Polyamine (nylon)	6
Polycarbonate	2
Polyethylene	74
Polyester	16
Polypropylene	54
Polystyrene (maybe)	2
Polyvinyl chloride	9

Table 2: Data Distribution for Mendeley.

2.1 Final Dataset

The final dataset used in our experiments is shown in TABLE 3. Note, that only the polymer types that are present in SLoPP are used. The majority of samples come from SLoPP, however, Mendeley contains a lot of samples for the polyester, polyethylene and polypropylene polymer types. The test dataset consists of only SLoPP-E, which was reduced to match the classes (SLOPP polymer types) present in the training set. 16 samples from 7 different types of plastic have been removed resulting in a combined dataset of **306** training samples and **97** testing samples.

3. METHODS: FEATURE ENGINEERING AND PREPROCESSING

As has been discussed in the previous section, different polymer types can be identified by the location of peaks on the x-axis (energy). Before this data can be used for classification learning, feature engineering which includes data transformation as well as preprocessing techniques such as normalization and discretization have been used. These techniques are described below.

3.1 Normalization

Here we discuss scaling methods for normalizing both the intensity (y-axis) and energy (x-axis) feature values since there are multiple problems with the feature values such as: varying ranges,

Polymer Types	SLoPP	SLoPP-E	Mendeley
Acrylic	10	3	-
Acrylonitrile Butadiene Styrene	10	1	1
Cellulose Acetate	4	3	-
Cotton	16	-	-
Polyamide	7	7	-
Polycarbonate	7	2	2
Polyester	10	12	16
Polyethylene	24	26	74
Polyethylene Terephthalate	9	1	-
Polyethylene Vinyl Acetate	5	-	-
Polymethyl Methacrylate	1	3	-
Polypropylene	17	21	54
Polystyrene	11	9	2
Polyurethane	6	6	-
Polyvinyl Chloride	11	3	9

Table 3: Final dataset (SLoPP, SLoPP-E, Mendeley).

varying step values, integer vs. real values as well negative values. Algorithm 1 gives the pseudo-code for scaling the energy values.

- **Energy values**: Each sample in the dataset has a different x-axis range (i.e., one sample might have y-axis values between 100 and 1200 on the x-axis and another one between 300 and 3000). Furthermore, each sample's range between individual points on the x-axis is different as well (i.e., one sample can have a step value of 2 and another sample with a step value of 3). Therefore, scaling of the x-axis should be performed, where all values would be mapped to the corresponding points. Additionally, x-axis values are continuous values (ex: real value of 101.23), therefore, x-axis values should be mapped to integer values.

  Scaling works as follows: firstly, as each sample has a different x-axis range, these values are mapped to the same range, by finding the minimum and maximum value of x for all samples (shown as parameter **min_range, max_range** in Algorithm 1). In the case of the combined dataset, these parameter values are set to 0 and 3500 respectively. Then the values are populated by either the first value if the values occur at the beginning of the dataset, or by the last value if the values occur at the end. For example, if a sample has values on the x-axis ranging between 100 and 3000, then the values between 0-99 are populated with 100, and the values between 3001-3500 are populated with the value 3000.

  Secondly, as samples have a different step between each value, the gaps between these values are populated with the value which is at the beginning of a gap (i.e., if the x-values of two samples are 100 and 103, then all x-values having either 101 and 102 are replaced with value 100). As a result, this function produces 3501 points (3500 - 0 + 1), which are populated using the information from the original sample.

- **Intensity values**: Some samples in the SLoPP-E test set have negative values for the intensity (y-axis). Hence, all values have been scaled by adding a constant factor of one unit, which

Algorithm 1 Scaling(Dataset, min_range, max_range)

```
B20;43_30C0      382C1º
for ?;0BC82_CH?4, 83GB in    0C0B4C•8C4<B1º do
    B20;43_30C0»?;0BC82_CH?4¼      »¼
    for 83G in 83GB do
        2 0=643_30C0        0C0 A0<41f
            0G0 : A0=641<8=_A0=64-<0G_A0=64 ¸ 1º-
            0H0 : »0•0¼  1<0G_A0=64 ¸ 1ºgº
        ;0BC_83G        1
        for 8=34G, ?>8=C in 83G•8C4AA>FB1º do
            83G_>5_4;      8=C1?>8=C»0G0¼º
            if 83G_>5_4; ¡ <0G_A0=64 then
                break
            if ;0BC_83G != 1 then
                for 8 in A0=641;0BC_83G ¸ 1-83G_>5_4;º do
                    2 0=643_30C0•0C»8-0H0¼      2 0=643_30C0•0C»;0BC_83G-0H0¼
            else
                for 8 in A0=641 83G_>5_4;º do
                    2 0=643_30C0•0C»8-0H0¼      ?>8=C»0H0¼
            2 0=643_30C0•0C»83G_>5_4;-0H0¼      ?>8=C»0H0¼
            ;0BC_83G      83G_>5_4;
        for 8 in A0=641;0BC_83G ¸ 1-<0G_A0=64 ¸ 1º do
            2 0=643_30C0•0C»8-0H0¼      2 0=643_30C0•0C»;0BC_83G-0H0¼
        B20;43_30C0»?;0BC82_CH?4¼      B20;43_30C0»?;0BC82_CH?4¼ ¸
                                       2 0=643_30C0
return B20;43_30C0
```

is the minimum negative value on the y-axis plus 1. This also ensures that there are no zero values. It should be noted that it is the relative difference between the peaks in a sample that are important, not the absolute intensity values.

3.2 Data Transformation

Two well-known data transformation techniques were used: Rate of Change (ROC) and Percentage Change (PC) shown in Eqns. 1 and 2. Both these techniques modify the original data by making sharp changes in the original dataset more visible.

- **Rate of Change (ROC)**:

$$' \$ \quad = \frac{5\,{}^{1}1^{o} \quad 5\,{}^{1}0^{o}}{1 \quad 0} \tag{1}$$

where f(a) and f(b) are values on the y-axis and a and b are their corresponding values on the x-axis.

- **Percentage Change (PC)**:

$$\% \quad = \frac{5\,{}^{1}0^{o}}{<40=^{1}5\,{}^{1}0 \quad 1^{o}\text{-}\bullet\bullet\bullet\bullet\text{-}\ 5\,{}^{1}0 \quad =^{oo}} \tag{2}$$

where f(a) is the current value of the intensity (y-axis) and n is the number of values on the y-axis.

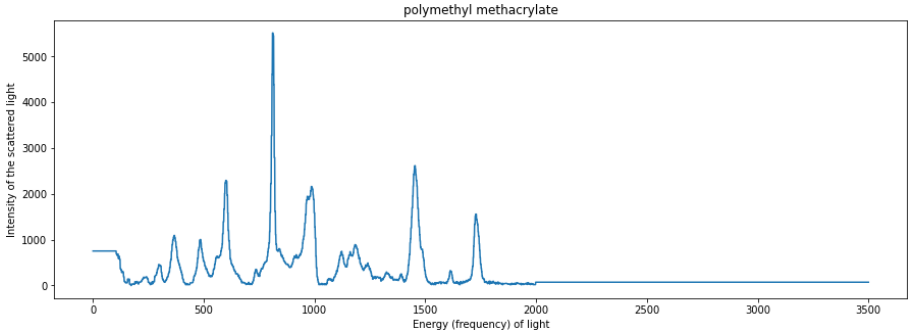

Figure 4: Sample for polymethyl methacrylate.

Since the PC function did not give good classification results, the ROC function was used as the main data transformation technique. However, the PC function was used in the augmentation of the training set, which is described later in Sec. 3.4. As an illustration of this technique, we present two figures. FIGURE 4, shows the plot for a single sample of type polymethyl methacrylate. FIGURE 5, shows the transformed plot. The ROC transformation was applied to the intensity values (y-axis) which results in sharp peaks and preserves the changes in intensity values at the same energy (x-axis) co-ordinate. It should be noted that, the values on the y-axis can be either positive or negative, meaning a positive or negative rate of change.

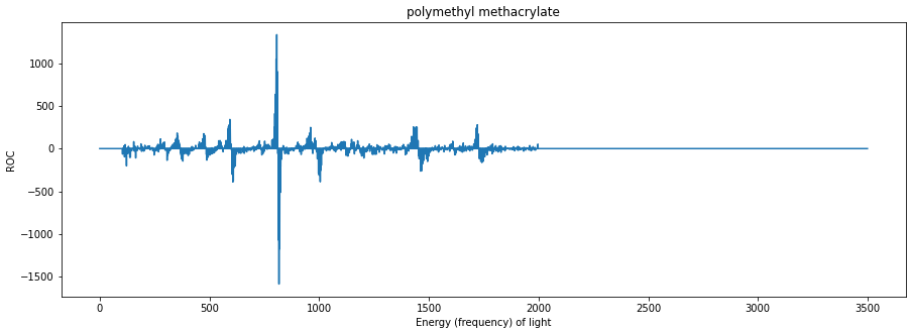

Figure 5: ROC processed sample for polymethyl methacrylate.

3.3 Discretization- Smoothing by Bin-means

Since Raman spectroscopy data has the characteristics of time-series data, the peaks in the distribution are the most important patterns that can be extracted from the samples. An equal-width binning technique was used in this research. Then a smoothing by bin-means technique is applied where the average of the values in a bin is calculated and each bin is now represented using the average value. FIGURE 6, shows the results of this technique applied to a single sample for polymethyl methacrylate type with a bin width of 11. That is, every 11 values are mapped to the same bin, and the average value of the bin is calculated. One can also observe the compressed scale on the x-axis as compared to the scale in FIGURE 5.

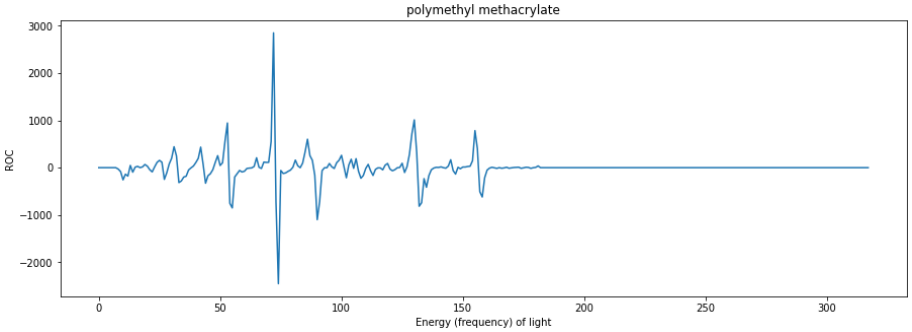

Figure 6: Binning technique of ROC processed sample for polymethyl methacrylate.

3.4 Augmentation

As the dataset is very small, the data augmentation function has been implemented to populate the training dataset with more samples. The pseudo-code for the augmentation process is given in Algorithms 2-5.

The augmentation function works the following way: firstly, the function iterates over a polymer (plastic) type that needs to be augmented, and the **pct_change** function is used to calculate the

change between the current and the previous value of a sample, by dividing the two numbers. This helps to keep information about the changing values.

Secondly, a random uniform distribution is applied, where random values between -0.05 and 0.05 are chosen. This is a user-defined parameter **random_change** and controls how much the augmented dataset differs from the original sample.

The last step is to reverse the percentage change function, by multiplying the original value with the new percentage change value. As the percentage change value has been changed slightly, each generated value is different from the original value. However, such a change leads to a problem of rapidly increasing or decreasing graph fluctuations. These sharp fluctuations are controlled by the **max_pct_change** parameter. This parameter value is set to 99, meaning that the generated value could be up to 99% more than the original value or 99% less than the original value. Additionally, Algorithm 2 includes parameter **shift**. This parameter is meant to shift the values on the y-axis (higher or lower). However, this value was set to 0, as it does not change the test accuracy significantly.

Algorithm 2 Generate_Augmented_Data(train_dataset, plastic_type_list, min_num, random_change=0.05, shift=0, max_pct_change=99)

```
?;0BC82_CH?4_;8BC »4;•;>F4A1º for 4; in ?;0BC82_CH?4_;8BC¼
CA08=_0D6< CA08=_30C0B4C
for ?;0BC82_CH?4, 83GB in CA08=_30C0B4C•8C4<B1º do
    8C4A0C4 0
    if ?;0BC82_CH?4 in ?;0BC82_CH?4_;8BC then
        while ;4=1CA08=_0D6<»?;0BC82_CH?4¼º Ÿ <8=_=D< do
            2DA_83G 8C4A0C4 % ;4=183GBº
            ?2C_2 0=64_;8BC ?2C_2 0=641830»2DA_83G¼»0H0¼º
            0D6<_4G0<?;4 64C_0D6<_4G0<?;41?2C_2 0=64_;8BC-
                                                A0=3><_2 0=64º
            8=8C_E0;D4 83GB»2DA_83G¼»0H0¼»0¼
            if <8=183GB»2DA_83G¼»0H0¼º 0 then
                8=8C_E0;D4 8=8C_E0;D4 ¸ 01B1<8=183GB»2DA_83G¼»0H0¼°° ¸ 1
            0D6<_30C0 64C_5D;;_0D6<_4G0<?;4183GB»2DA_83G¼»0H0¼-
                                0D6<_4G0<?;4-8=8C_E0;D4-
                                B 85C-<0G_?2C_2 0=64º
            CA08=_0D6<»?;0BC82_CH?4¼ CA08=_0D6<»?;0BC82_CH?4¼ ¸
                                0C0 A0<41f0G0:83GB»2DA_83G¼»0G0¼-
                                                0H0:0D6<_30C0gº
            8C4A0C4 8C4A0C4 ¸ 1
CA08=_30C0B4C CA08=_0D6<
```

Example of augmented data is shown in FIGURE 7. The line which is coloured red is the original sample, and a blue line is the augmented sample. It can be seen, that the augmented sample keeps the same trajectory as the original sample, but introduces some changes on the y-axis values. Peaks on generated samples are retained on same x-axis values, however, the intensity of such peaks is different.

Algorithm 3 pct_change(dataframe)

```
?2C_2 0=64_;8BC »¼
<8=_E0;D4 <8=¹30C0 5A0<4º
if <8=_E0;D4 0 then
    30C0 5A0<4 »8 ¸ 01B¹<8=_E0;D4º ¸ 1 for 8 in 30C0 5A0<4¼
for =D<¸ _ in 4=D<4A0C4¹30C0 5A0<4»:;4=¹30C0 5A0<4º 1¼º do
    ?2C_2 0=64_;8BC ?2C_2 0=64_;8BC ¸
                    ¹30C0 5A0<4»=D< ¸ 1¼•30C0 5A0<4»=D<¼º
return ?2C_2 0=64_;8BC
```

Algorithm 4 get_augm_example(pct_change_list, random_change=0.2)

```
0D6<_?2C_2 0=64_;8BC »¼
for 4; in ?2C_2 0=64_;8BC do
    C<?_4; 4; ¸ A0=3><•D=8 5>A<¹ A0=3><_2 0=64-A0=3><_2 0=64º
    if C<?_4; 0 then
        C<?_4; 4;                                  • el is ¡ 0, because of ?2C_2 0=64 function
    0D6<_?2C_2 0=64_;8BC 0D6<_?2C_2 0=64_;8BC ¸ C<?_4;
return 0D6<_?2C_2 0=64_;8BC
```

Algorithm 5 get_full_augm_example(original_dataset, pct_change_list, init, shift=0, max_pct_change=10)

```
?A4E8>DB_E0;D4 8=8C ¸ B 8 5C
0D6<_?2C_2 0=64_;8BC »¼
0D6<_?2C_2 0=64_;8BC 0D6<_?2C_2 0=64_;8BC ¸ ?A4E8>DB_E0;D4
<8=_E0;D4 <8=¹>A868=0;_30C0B4Cº
if <8=_E0;D4 0 then
    >A868=0;_30C0B4C »8 ¸ 01B¹<8=_E0;D4º ¸ 1 for 8 in >A868=0;_30C0B4C¼
for =D<¸ 4; in 4=D<4A0C4¹?2C_2 0=64_;8BCº do
    ?A4E8>DB_E0;D4 ?A4E8>DB_E0;D4 4;
    if ?A4E8>DB_E0;D4 ¡ >A868=0;_30C0B4C»=D<¸ 1¼
                        ¹1 ¸ <0G_?2C_2 0=64•100º then
        ?A4E8>DB_E0;D4 >A868=0;_30C0B4C»=D<¸ 1¼
                        ¹1 ¸ <0G_?2C_2 0=64•100º
    if ?A4E8>DB_E0;D4 Ÿ >A868=0;_30C0B4C»=D<¸ 1¼
                        ¹1 <0G_?2C_2 0=64•100º then
        ?A4E8>DB_E0;D4 >A868=0;_30C0B4C»=D<¸ 1¼
                        ¹1 <0G_?2C_2 0=64•100º
    0D6<_?2C_2 0=64_;8BC 0D6<_?2C_2 0=64_;8BC ¸ ?A4E8>DB_E0;D4
return 0D6<_?2C_2 0=64_;8BC
```

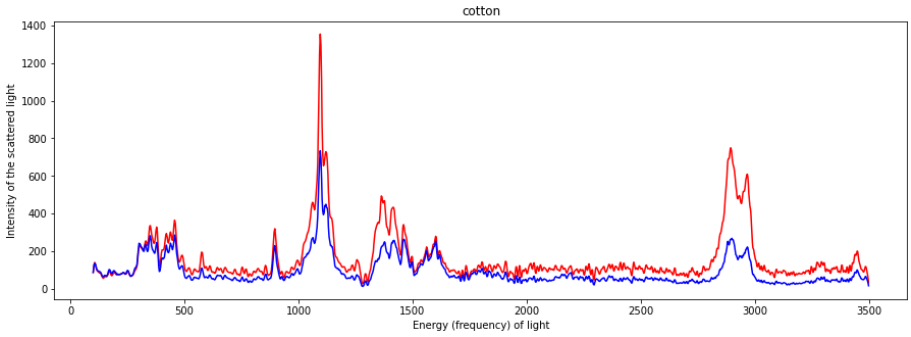

Figure 7: Improved augmented data (red – original; blue – augmented).

In our research, we have augmented polymer types that have either a small number of samples (e.g., less than 5) or have performed poorly on test results.

4. RESULTS AND DISCUSSION

In this section, we analyze the results of the experiments. The following ML algorithms have been used in this research using the scikit-learn workbench[4]: support vector machines (SVM), random forest (RF), decision tree (DT), k-nearest neighbours (KNN), and artificial neural network (ANN). The RF model is the only model that gives high classification accuracy. Hence, in our discussions related to the analysis of the effect of different preprocessing, discretization as well as augmentation techniques, we will use the RF model as our baseline model.

Parameters of all models have been chosen based on the performance of algorithms, however, as the dataset is small, it is hard to identify the best parameters, as results might be fluctuating after each training step. For RF model, we used entropy and gini parameters as a criterion. Changing the number of estimators for RF does not change the accuracy a lot, and 150 n-estimators are used. For DT we used the entropy parameter for the criterion. SVM uses a linear kernel, as it produced the best accuracy, compared to other kernels. For KNN we used 3 neighbours, as it produced the best results and increasing the number would not change results significantly. For ANN, we used 4 layers of 128, 64, 32 and sparse categorical cross-entropy with the adam optimizer, inner layer relu activation function and sigmoid final activation function. We have tried to change the number of layers or the size of the layer, but the chosen parameters produced the highest results.

In an effort to increase the training set size, we also experimented with another microplastic dataset **Open Specy** [5]. This dataset contains a total of 183 examples and 137 polymer types [32]. We observed that most of the polymer types in this dataset do not appear in SLoPP, therefore, cannot be used. Additionally, the intensities of the scattered light (y-axis) for Open Spacy dataset is normalized to values between 0 and 1, and there the original values for the intensities cannot be reconstructed. As a result, this dataset was not used in our final model training experiments.

[4] https://scikit-learn.org/stable/
[5] https://doi.org/10.1021/acs.analchem.1c00123.s001

Experiment	Preprocessing Methods	RF Accuracy (%)
1	scaling (x-axis), ROC	**79.38**
2	scaling (x-axis), no ROC	61.85
3	no scaling (x-axis), ROC	72.16
4	no scaling(x-axis), no ROC	53.61

Table 4: Model accuracy with different variations of preprocessing functions.

TABLE 4, presents experiments using different preprocessing functions and with no scaling of the y-axis values. All experiments were conducted with a combination of the scaling and ROC transformation methods described in sections 3.1 and 3.2. The best result (accuracy of 79.38% highlighted in blue) was achieved with scaling energy values and using the rate of change transformed feature.

FIGURE 8, shows the performance of the RF model using different bin sizes ranging from 2 to 50. The discretization technique which achieves the best result (classification accuracy 86.59% with information gain criteria) is when the bin size is between 10 and 20. This experiment does not use any augmentation method.

Figure 8: Accuracy of the model with different bin sizes.

However, if the training dataset is augmented, the accuracy increases dramatically (see FIGURE 9). Augmentation has been applied to the following polymer types: Cellulose Acetate, Polyamide, Polymethyl Methacrylate and Polyurethane, where each type has been augmented up to 15 examples. Experiments were conducted with two different criteria for the RF model, information gain (in FIGURE 9) and gini (in FIGURE 10). One can observe that the classification accuracy is not as good with the gini criteria as with the information gain criteria.

The best result of 91.75% classification accuracy was obtained with a bin size of 12 and information gain (entropy) as the criteria for tree construction.

FIGURE 11, gives the confusion matrix with classification details for each polymer type in the SLOPP-E dataset. It can be seen, that the model detects most of the samples correctly. However, it misclassifies a few samples, especially the Polyurethane polymer where 4 out of 6 samples were misclassified. The training accuracy of this model is 100%, which signifies that the model overfits. This is due to the fact that that the model was trained on just 306 samples.

Figure 9: Accuracy of the model with different bin size with augmented data with **information gain** criteria.

Figure 10: Accuracy of the model with different bin size with augmented data with **gini** criteria.

TABLE 5, shows the most commonly occurring misclassified samples. The model always predicts the same polymer type for these samples, irrespective of how the data has been processed. Upon examination of the corresponding Raman spectroscopy plots, it is hard to detect whether the sample is mislabeled or the model predicts the result wrongly (FIGURE 12-13). FIGURE 12 shows a sample of a type cellulose acetate, plotted with one example from a train dataset of cotton type (incorrect type). The peak around 1000 value on the x-axis has the same shape as other peaks which do not fully correspond to this type (cotton). FIGURE 13, shows a sample of cellulose acetate type, plotted with one example from a train dataset of the same type (correct type). It can be observed, that these samples have a different shape compared to the one in FIGURE 12. Hence these samples do not match.

Since the SLoPP-E test set was subject to weather and ageing, another experiment was conducted by adding some noise to the training (SLoPP) dataset to introduce some non-linearity. For each value on the x-axis, a small random value was either added or subtracted. However, the addition of noise did not change the accuracy in any significant way (see FIGURE 14).

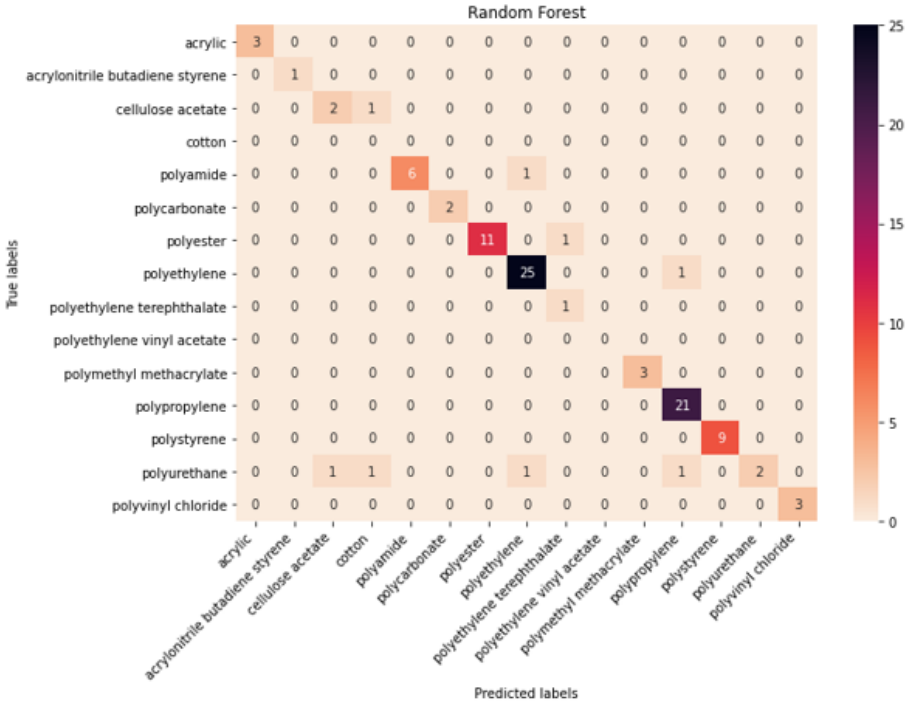

Figure 11: Confusion matrix for the model that achieved 91.75% accuracy.

Sample number	Model predicted	Actual label
6	Cotton	Cellulose Acetate
12	Polyethylene	Polyamide
24	Polyethylene Terephthalate	Polyester
50	Polyurethane	Polyethylene
88	Cellulose Acetate	Polyurethane
89	Polyamide	Polyurethane

Table 5: Misclassified cases.

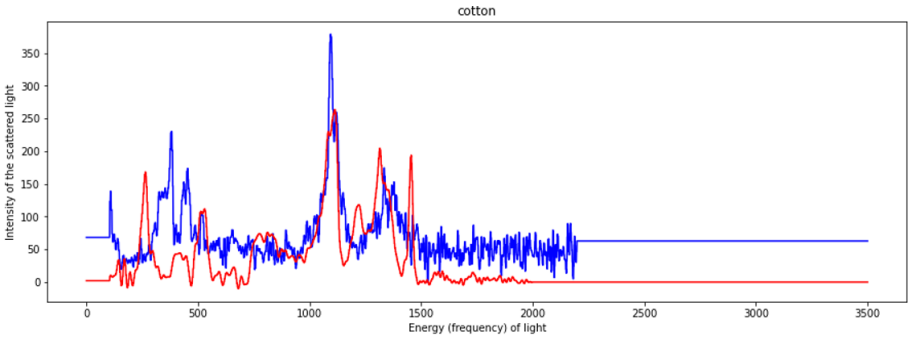

Figure 12: Wrongly detected sample (red) plotted on a wrongly predicted type (blue).

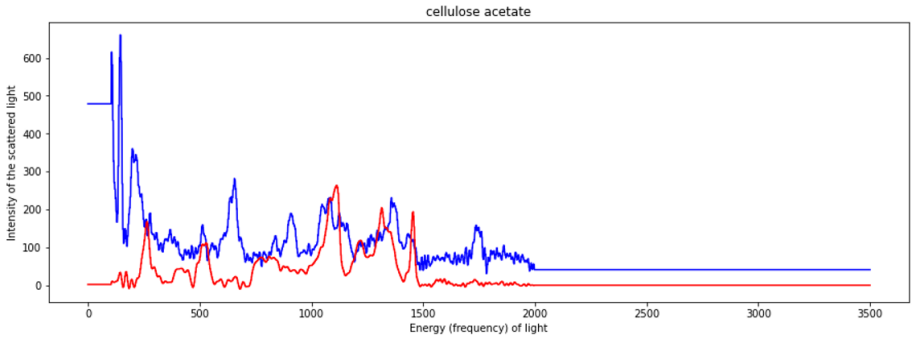

Figure 13: Wrongly detected sample (red) plotted on a correct type (blue).

Figure 14: Accuracy of the model with noise added to the training dataset.

The final model was trained on augmented data, which was preprocessed using the following functions: ROC, scaling x-axis (0-3500), discretization with the window size 12, no y-axis rescaling. The following polymer types were augmented: Cellulose Acetate, Polyamide and Polyurethane (30 samples), Polyester (40 samples), Polymethyl Methacrylate (10 samples) and Polystyrene (20 samples).

Figure 15, gives the confusion matrix where the trained model detects most of the samples correctly, and only a few samples are mislabeled. The model mostly performs poorly on the Polyurethane type, as it misclassifies 3 out of 6 test samples. The Acrylonitrile Butadiene Styrene type also performs poorly, as it misclassifies a single test sample. Since there is only 1 test sample available, this classification could be misleading.

TABLE 6, gives the results of experiments with other models. However, none of the other models achieved the same accuracy on the test dataset as the random forest model. The ANN model with 4 layers of 128, 64, 32 and sparse categorical cross-entropy was used with the adam optimizer. SVM with the linear kernel (it produced the best accuracy, compared to other kernels), DT (with entropy) and KNN with 3 nearest neighbours were used.

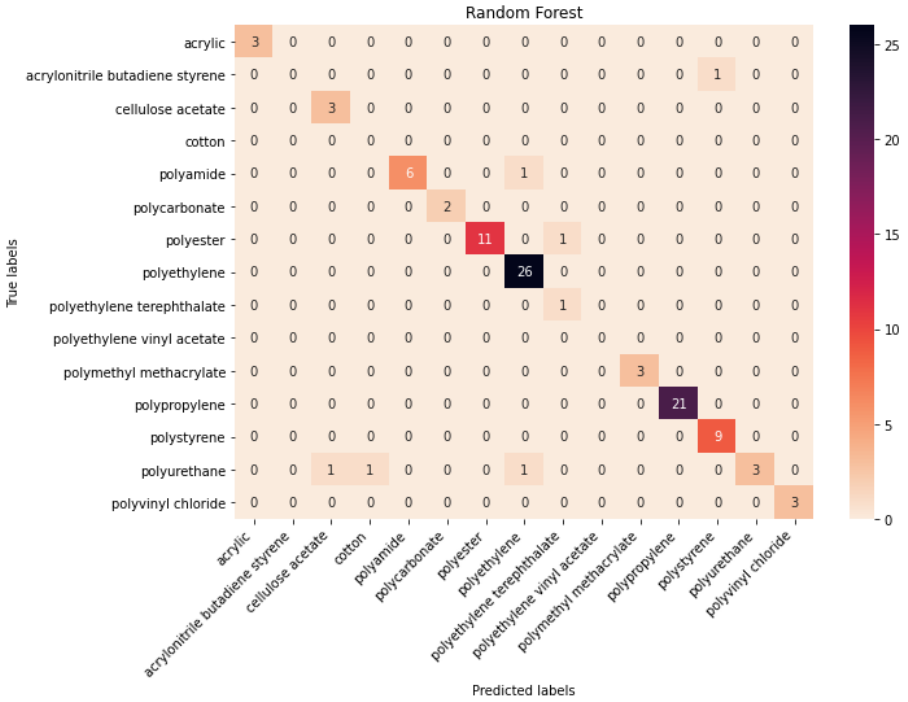

Figure 15: Confusion matrix for the model that achieved 93.81% accuracy.

Models	Classification Accuracy
ANN	71.13%
SVM (linear kernel)	73.19%
DT	69.07%
KNN	73.19%

Table 6: Accuracies of different machine learning models.

In summary, our experiments demonstrate that there is a significant improvement in the classification accuracy (from 89% to 93.81%) when the dataset is augmented. This shows that a larger data set with more training and balanced samples can improve the classification performance beyond 94% and learn from environmentally degraded samples. The other important issue is that there is some concern that the original sample maybe mislabeled. This is because the predicted type (by the model) is not similar to the actual type (visually). Another observation is that even when wave-like samples (from the Mendeley dataset) were excluded from the training set, the classification accuracy was around 90%. This shows that adding the samples (even though some of the shapes were different) may have in fact helped the model to learn, or at least, did not have a negative effect on the model. One reason could be that SLoPP-E (test dataset) does not have similar wave-like samples.

5. CONCLUSION

In this work, we were primarily interested in detecting polymer types from the spectral signature of Raman spectroscopy microplastics data which were environmentally aged from a well-known dataset. Environmental weathering occurs from exposure to temperature extremes, UV radiation, wind, water erosion in freshwater environments, and saltwater erosion in marine environments, in addition to other factors in localized ecosystem contexts. Exposure of microplastics to the environment affects their spectrographic output data, making spectrographic analysis results less reliable than unaffected samples. Different normalization methods as well as data transformation methods for preprocessing and feature engineering were applied. Since the number of training samples in certain polymer types were limited, a data augmentation method was used. Different ML models were trained with the random forest model giving the best result with an improvement in classification accuracy of 93.81% from 89%. A detailed discussion of the results is presented in an effort to contribute to the understanding chemical compounds of plastics that have been weathered by various environmental processes. The significance of this research project is to strive for a measurably improved predictive capacity of Raman spectroscopy data to help classify polymer types through an applied machine learning process. This work can lead to applications in ecotoxicology and environmental research, the circular economy for plastics recycling processes, water quality testing and treatment processes, food and beverage quality control testing, to name a few.

References

[1] Booth A, Sørensen L. Microplastic Fate and Impacts in the Environment. In: Rocha-Santos CMT, Costa M, editors. Handbook of microplastics in the environment. Cham: Springer International Publishing; 2020:1-24.

[2] Conesa J, Iñiguez M. Analysis of Microplastics in Food Samples. In: Rocha-Santos CMT, Costa M, editors. Handbook of microplastics in the environment, springer. Cham: International Publishing; 2020:1-16.

[3] Rochman CM, Brookson C, Bikker J, Djuric N, Earn A, et al. Rethinking Microplastics as a Diverse Contaminant Suite. Environ Toxicol Chem. 2019;38:703-711.

[4] Pflugmacher S, Huttunen JH, von Wolff MV, Penttinen OP, et al. Enchytraeus Crypticus Avoid Soil Spiked With Microplastic. Toxics. 2020;8:10.

[5] Fournier E, Etienne-Mesmin L, Blanquet-Diot S, Mercier-Bonin M. Impact of Microplastics in Human Health. In: Rocha-Santos CMT, Costa M, editors. Handbook of microplastics in the environment. Cham: Springer International Publishing; 2021:1-25.

[6] Cabernard L, Roscher L, Lorenz C, Gerdts G, Primpke S. Comparison of Raman and Fourier Transform Infrared Spectroscopy for the Quantification of Microplastics in the Aquatic Environment. Environ Sci Technol. 2018;52:13279-13288.

[7] Araujo CF, Nolasco MM, Ribeiro AMP, Ribeiro-Claro PJA. Identification of Microplastics Using Raman Spectroscopy: Latest Developments and Future Prospects. Water Res. 2018;142:426-440.

[8] Zhu X, Nguyen B, You JB, Karakolis E, Sinton D, et al. Identification of Microfibers in the Environment Using Multiple Lines of Evidence. Environ Sci Technol. 2019;53(20):11877-11887.

[9] Popov MN, Spitaler J, Veerapandiyan VK, Bousquet E, Hlinka J, et al. Raman Spectra of Fine-Grained Materials From First Principles. npj Comp Mater. 2020;6.

[10] Liu M, Lu S, Chen Y, Cao C, Bigalke M, et al. Analytical Methods for Microplastics in Environments: Current Advances and Challenges. In: v. D. He, Costa M, editors. The handbook of environmental chemistry: microplastics in terrestrial environment. Cham: Springer International Publishing; 2020:95:3-24.

[11] Madden MG, Ryder AG. Machine Learning Methods for Quantitative Analysis of Raman Spectroscopy Data. Proc SPIE Int Soc Opt Eng. 2002;4876.

[12] Gniadecka M, Philipsen PA, Sigurdsson S, Wessel S, Nielsen OF, et al. Melanoma, Diagnosis by Raman Spectroscopy and Neural Networks: Structure Alterations in Proteins and Lipids in Intact Cancer Tissue. J Invest Dermatol. 2004;122:443-449.

[13] DiezPastor J, JorgeVillar S, ArnaizGonzalez A, GarciaOsorio C, DiazAcha Y, et al. Machine Learning Algorithms Applied to Raman Spectra for the Identification of Variscite Originating From the Mining Complex of Gava. Raman Spectrosc. 2018:1-12.

[14] Khan S, Ullah R, Shahzad S, Anbreen N, Bilal M, et al. Analysis of Tuberculosis Disease Through Raman Spectroscopy and Machine Learning. Photodiagn Photodyn Ther. 2018;24:286-291.

[15] Sevetlidis V, Pavlidis G. Effective Raman Spectra Identification With Tree-Based Methods. J Cult Herit. 2019;37:121-128.

[16] Berghian-Grosan C, Magdas DA. Raman Spectroscopy and Machine-Learning for Edible Oils Evaluation. Talanta. 2020;218:121176.

[17] Berghian-Grosan C, Magdas DA. Application of Raman Spectroscopy and Machine Learning Algorithms for Fruit Distillates Discrimination. Sci Rep. 2020;10:21152.

[18] Ryzhikova E, Ralbovsky NM, Sikirzhytski V, Kazakov O, Halamkova L, et al. Raman Spectroscopy and Machine Learning for Biomedical Applications: Alzheimers Disease Diagnosis Based on the Analysis of Cerebrospinal Fluid. Spectrochim Acta A Mol Biomol Spectrosc. 2021;248:119188.

[19] Lussier F, Thibault V, Charron B, Wallace GQ, Masson J-F. Deep Learning and Artificial Intelligence Methods for Raman and Surface-Enhanced Raman Scattering. Trends Anal Chem. 2020;124.

[20] Houston J, Glavin FG, Madden MG. Robust Classification of High-Dimensional Spectroscopy Data Using Deep Learning and Data Synthesis. J Chem Inf Model. 2020;60:1936-1954.

[21] Xia J, Zhu L, Yu M, Zhang T, Zhu Z, et al. Analysis and Classification of Oral Tongue Squamous Cell Carcinoma Based on Raman Spectroscopy and Convolutional Neural Networks. J Mod Opt. 2020;67:481-489.