METHODOLOGY

Development of a machine learning-based method for the analysis of microplastics in environmental samples using μ-Raman spectroscopy

Felix Weber[1*], Andreas Zinnen[1] and Jutta Kerpen[1]

Abstract

This research project investigates the potential of machine learning for the analysis of microplastic Raman spectra in environmental samples. Based on a data set of > 64,000 Raman spectra (10.7% polymer spectra) from 47 environmental or waste water samples, two methods of deep learning (one single model and one model per class) with the Rectified Linear Unit function (ReLU) (hidden layer) as the activation function and the sigmoid function as the output layer were evaluated and compared to human-only annotation. Based on the one-model-per-class algorithm, an approach for human–machine teaming was developed. This method makes it possible to analyze microplastic (polyethylene, polypropylene, polystyrene, polyvinyl chloride, and polyethylene terephthalate) spectra with high recall ($\geq$ 99.4%) and precision ($\geq$ 97.1%). Compared to human-only spectra annotation, the human–machine teaming reduces the researchers' time required per sample from several hours to less than one hour.

Keywords Environmental samples, Machine learning, Microplastics, μ-Raman spectroscopy, Waste water

Introduction

Microplastics as contaminants have been intensively investigated. However, although microplastics were first discovered in the 1970s [1] and the number of studies dealing with microplastics has increased greatly since 2010 [2], much is still unknown. One of the major concerns is the complexity of microplastics analysis. Apart from sampling, sample purification, and the problem of contamination mitigation, the detection of MP is still a challenge. In general, there are two methods for the chemical analysis of microplastics [1]: thermoanalytical and spectroscopic. Thermoanalytical methods such as pyrolysis–gas chromatography-mass spectroscopy (GC–MS) and thermal extraction-desorption (TED)-GC–MS provide information about microplastic masses. They are comparatively fast and are less prone to contamination than spectroscopic approaches. Although they yield information about the concentration and chemical composition of microplastics (mainly the type of polymer), they are not suitable for determining their size [3, 4]. This information is crucial to the risk assessment of microplastics, as (eco-)toxicological studies suggest that the toxicity of microplastics increases with decreasing diameter [5]. Therefore, spectroscopic methods are needed. Using spectroscopic methods (predominantly μ-Raman and (μ)-Fourier-transfrom infrared (FTIR) spectroscopy), microplastic particles and fibers are analyzed on an analysis filter. Usually, elaborate sample preparation is necessary before detection (oxidative or enzymatic digestion, density separation). The results are given in item concentrations [3]. Often only a few selected particles are analyzed [6, 7]. However, this

*Correspondence:
Felix Weber
felix.weber@hs-rm.de
[1] Institute for Environmental and Process Engineering, RheinMain University of Applied Sciences, 65022 Wiesbaden, Germany

is not sufficient, particularly because – even with accurate sample preparation and mitigation of contamination – there are usually numerous non-plastic particles on a filter. It is necessary to analyze several thousand particles per sample to achieve a representative result [8, 9]. Although automatic particle recognition is now available from open-source programs [10] or is incorporated into the instrument software [11], detection is still time-consuming, so the measurement settings are chosen as a compromise between spectra quality and the time required for measurement. This leads to weak spectra quality (including a lower signal-to-noise ratio) [12]. Apart from this, the spectra quality is influenced by the fact that microplastics are aged by environmental influences [13] or contain different concentrations of additives such as plasticizers or pigments. Fluorescence caused by biological-organic matrix debris is a problem, especially for μ-Raman spectroscopy [14]. In addition, particle size influences the spectra quality. As a rule, smaller particles generate lower signal intensity than larger ones. But because the spectra quality also depends on the distance between the particle and the objective (z-hub), the spectra quality also decreases for larger particles, because the z-hub is usually optimized for smaller particles [15]. Hence the analysis and evaluation of the spectra are time-consuming and error-prone, so the automatization of spectra analysis using machine-learning algorithms is one of the major tasks of method development.

A handful of studies have proven the potential of machine learning for microplastics FTIR spectra with a database taken from environmental samples; papers such as Kedzierski et al. [16] have presented a machine-learning algorithm for analyses of FTIR-attenuated total reflection (ATR) spectra. Using the machine-learning method of k-nearest neighbor classification, they trained their algorithm on a database of 969 spectra of marine microplastics. The resulting algorithm is suitable for common polymers such as polyethylene (PE), polypropylene (PP), polyvinyl chloride (PVC), polystyrene (PS), and others. An evaluation was performed with 4000 spectra. In 90.5% the classification was correct. When a human reevaluated the results, this value increased to 97%. Hufnagl et al. [17] also applied machine learning to focal plane array-μFTIR spectra. With model-based machine learning based on random decision forests and Monte Carlo cross-validation for sensitivity, specificity, and precision, they presented an approach for the classification of 20 polymers. Spectra from environmental samples served as the data set. The sensitivity of the algorithm ranged from 0.925 to 1, the specificity from 0.9984 to 1, and the precision from 0.9563 to 0.9965.

Although there are some examples of machine learning involving FTIR spectra, to the authors' knowledge only a few studies have presented machine-learning approaches using Raman spectra from microplastics. Unfortunately, so far spectra from "artificial" microplastics have served as the data basis:

> Lei et al. [18] compiled their data basis for machine-learning algorithms using spectra from purchased microplastic powders or microplastics created from purchased macroplastic samples. They transferred the microplastics onto microscope slides and conducted Raman analysis by mapping. This yielded > 95% classification accuracy using open-source random forest, k-nearest neighbors, and multi-layer perceptron algorithms. These results show the potential of machine-learning applications in microplastics detection. However, the Raman settings applied by Lei et al., especially the mapping of particles, are not comparable to the settings necessary to analyze a large number of particles from environmental samples. Apart from this, the spectra from purchased particles or manufactured microplastics are usually of high quality compared to spectra from environmentally aged microplastics.
> Like Lei et al., Luo et al. [19] used purchased microplastics to generate database spectra. They suspended different concentrations of microplastics in different aquatic environmental media (such as rainwater and surface water) and added a surfactant. Afterwards, the suspensions were filtered on an analysis filter for Raman acquisition. The authors provided no information about the duration of the mixing of microplastics with the media, but it can be assumed that the conditions were not realistic enough to simulate microplastic aging processes or the growth of biofilms, etc. Furthermore, microplastic particles in reality have a variety of chemical composition (e.g. plasticisers, dyes, flame retardants, copolymers). With this method, Luo et al. compiled a database of 3675 Raman spectra for machine learning. Using a coupled sparse autoencoder and a softmax classifier framework for PET, PVC, PP, PS, polycarbonate, and PE, Luo et al. achieved a test accuracy of 99.1%.

As mentioned above, the quality of Raman spectra for microplastics in environmental samples depends on several factors; it is sometimes poor due to environmental and analytical influences. Therefore, the available studies can only be used to a limited extent to draw conclusions about the usability of machine learning for Raman spectra alignment.

This study aims to contribute a further approach to the application of machine-learning algorithms to Raman-spectra analysis for purposes of microplastics detection.

The motivation of this study is to develop a reliable method with the highest possible degree of accuracy that automates the process of spectrum identification as much as possible, thereby reducing the effort required by researchers. To this end, machine-learning algorithms were trained using $> 60,000$ spectra, mainly from microplastic analysis in industrial wastewater. Different algorithms and human–machine teaming were tested for recall and precision, to create an applicable tool for microplastics analysis.

Methods

Different approaches of spectra identification by machine learning were compared with each other and with human annotation. The scheme of the methods is shown in the Fig. 1.

Data set

Source of the data set

The data set used in this study was generated in the research projects EmiStop [20], Eintrag MiPa (2022—2025), and an investigation of microplastics in two German rivers [21]. The set includes data from 47 analysis filters. The samples were taken from six industrial wastewater treatment plants and the Main and Nidda rivers in Germany. Sampling was conducted from January 2021 to April 2022. Sampling, sample preparation, and analysis by µ-Raman spectroscopy were conducted according to the methods used in Weber and Kerpen B [21], Weber and Kerpen A [11], and Barkmann-Metaj et al. [22]: After volume-reduced sampling, the samples were prepared for analysis. Biological-organic matrix components were digested by hydrogen peroxide (323.15 K, 24 h) and sodium hypochlorite (room temperature, 6 d). Inorganic matrix components were separated by density separation in sodium polytungstate ($\rho = 1700$ kg/m^3). If necessary, subsamples were taken from a homogeneously stirred 2-propanol suspension according to Wolff et al. [23]. Where possible, each sampling site was sampled multiple times ($n = 3$). µ-Raman spectroscopy was conducted using a spectroscope (DXR2xi, Thermo Fisher Scientific Inc., Waltham, MA, USA) with a front-illuminated EMCCD detector. For analyses, the electron multiplier (EM) was turned off. All particles and fibers > 20 µm on the analysis filter (silicon) were detected using the automatic particle recognition feature of the instrumental software OMNICxi (v.2.3, Thermo Fisher Scientific Inc., Waltham, MA, USA). Each detected particle was analyzed with a laser wavelength of 785 nm, a laser power of 8 mW, and a total exposure time of 6.75 s (three repetitions of 2.25 s each). The objective used had a $20 \times$ magnification and a numerical aperture of 0.45. Spectra were

recorded in the range of 50—3300 cm^{-1} and with a resolution of 5 cm^{-1}.

Some of the samples were analyzed three times (aliquots), so the 47 samples include several multiple determinations. However, due to limitations in the laboratory and instrument utilization capacities, it was not possible to perform the time-consuming analysis of three subsamples for each sample in the data set presented. In addition, errors in sampling, sample loss in the laboratory, and data loss due to corrupt files meant that not all samples in this data set represent multiple determinations of a sampling point.

In total, 64,301 spectra were generated and used as the data set. On average, each sample contained 1,368 spectra.

The data set is available as electronic supplementary material.

Human annotation (first annotation)

The samples were analyzed and evaluated for the first time in the course of the routine analysis of the research projects. The spectra were presorted using the software OMNICxi (v2.3, Thermo Fisher Scientific Inc., Waltham, MA, USA) being compared to spectra from the reference library P/N L60001 (S.T. Japan Europe GmbH, Cologne, Germany) using correlation. The software was set to evaluate the Raman shift regions 600—950 cm^{-1} and 1000—1800 cm^{-1} (fingerprint region) for comparison with the library. OMNICxi is not a software specifically developed for microplastics analysis. As the results were not sufficient (large numbers of false positives and false negatives, with an estimated 30—50% of false negatives), each spectrum of a sample was evaluated by a domain expert after the presorting. There were no fixed criteria for the classification of a spectrum. The process of evaluation took several hours per sample, depending on the number of spectra per sample and the percentage of microplastics. The domain experts ($n = 3$) reported that the process required a high degree of concentration. Thus, the process was likely prone to error due to a lack of concentration, the domain experts' variations in expertise, and subjective decisions. These data served as the basis for the present study.

Improvements in human annotation and the data set (second annotation)

To improve the data set and generate a reliable ground truth for deep learning, objective criteria for the classification of microplastics (polymers PE, PP, PS, PVC, and PET) were defined. All peaks used for identification must be reliably distinguishable from baseline and noise (sufficient signal-to-noise ratio). Since the criteria were developed for human annotation, no fixed threshold for the
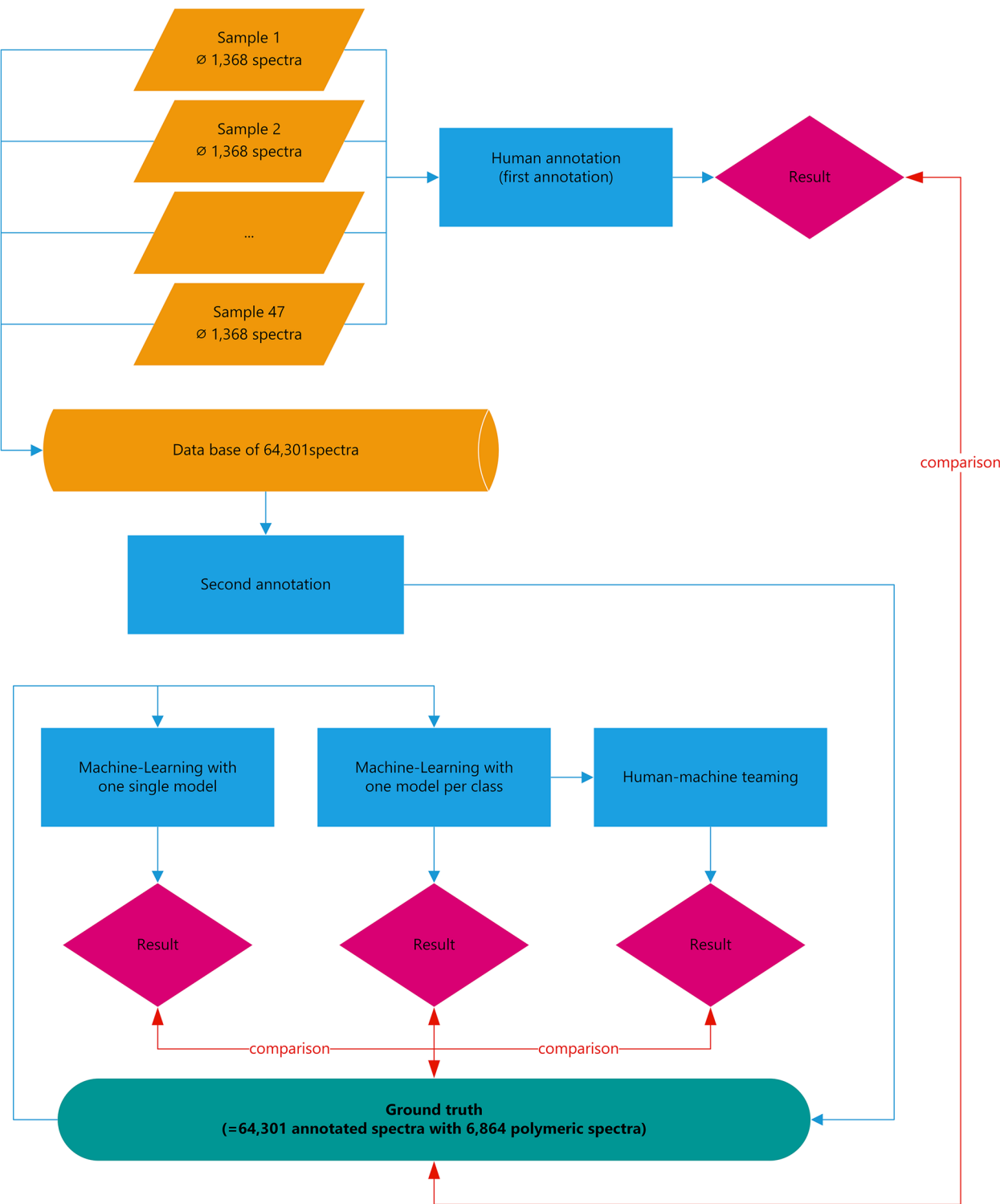
Fig. 1 Schematic overview of the approach and the methods used

signal-to-noise ratio can be given (see Table 1 and Fig. 2). These polymers were found most frequently in all samples of the data set (see the section entitled "Microplastic data set characterization (ground truth)"). In addition, other studies have shown that they are predominant in environmental samples [24–26]. This correlates with the

Table 1 Identification criteria and necessary peaks for polymer classification by domain experts

Polymer	Necessary peaks for identification / cm^{-1}	Further identification criteria
PE	1052, 1120, 1295, 1440	If three peaks are unique, one can be shifted. If four peaks match, they do not require the highest intensity. Two out of four peaks can be wider if the position is correct. The peak at 1440 cm^{-1} can be "smeared" and shifted.
PET	1100, 1278, 1620, 1735	The peaks at 1735 cm^{-1} (which are "hilly" in parts) and 1620 cm^{-1} (prominent) are necessary for identification. The peaks at 1100 cm^{-1} and 1278 cm^{-1} may not be present where there is a low signal-to-noise ratio.
PP	785, 820, 1145, 1328, 1458	The peaks at 785 and 820 are necessary for identification. The peak at 785 is usually higher than the peak at 820 cm^{-1}. A slight shift of the peaks is possible. The peaks at 1145 cm^{-1} and 1328 cm^{-1} may not be present where there is a low signal-to-noise ratio.
PS	990, 1020, 1328, 1607	The peaks at 990 cm^{-1} and 1020 cm^{-1} are necessary for identification. The peak at 1020 cm^{-1} is only marginally visible where there is a low signal-to-noise ratio. In this case, the other peaks are necessary for identification.
PVC	612, 663, 1435	The peaks at 612 cm^{-1} and 663 cm^{-1} are necessary for identification. The peak at 612 cm^{-1} is usually higher. Both peaks are broad and they converge.
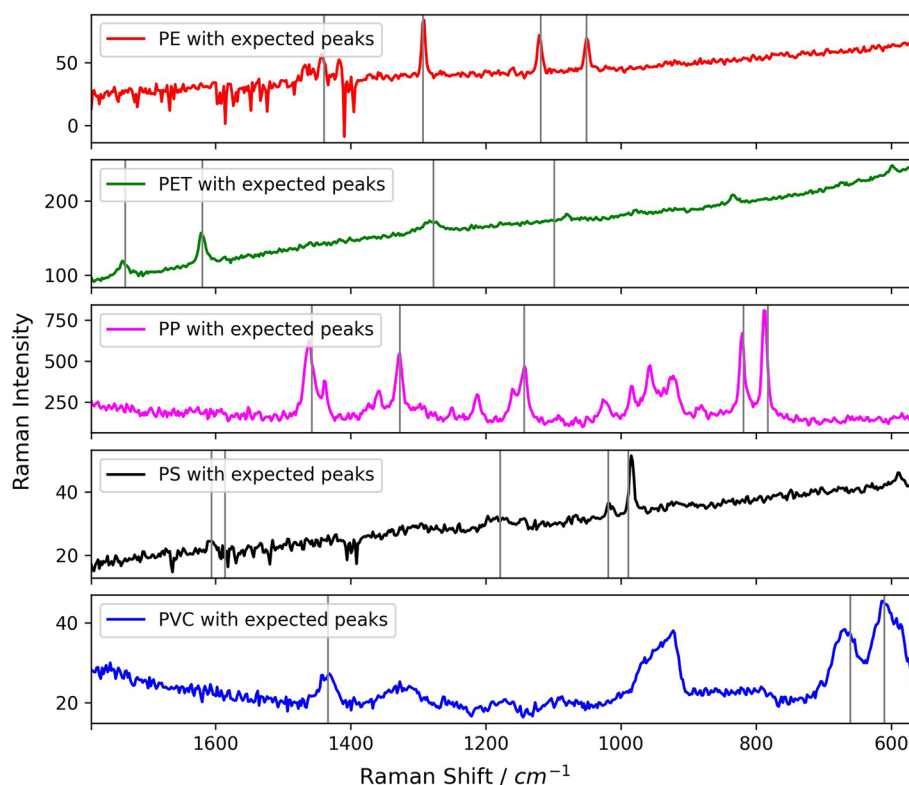
Fig. 2 Examples of high quality PE, PET, PP, PS, and PVC spectra with marked peak positions

production volume of polymers [27]. Polyamides were not included here, because they are not resistant to the sample purification method [28].

In the next step, all spectra that were classified as microplastics in the first evaluation were re-evaluated by one domain expert according to the criteria (second annotation). For this purpose, all Raman spectra were isolated from the analysis data (file format: mapx) and submitted to a domain expert for re-annotation in a custom-programmed user interface (GUI). In the GUI, the domain expert can select classes to annotate. To aid the decision, marking lines were placed where peaks were expected for each class (see Fig. 2). The resulting data were used for the machine-learning models (see the section entitled "Machine learning"). The results of this second annotation were used as the ground truth for the deep-learning experiments.

Weber *et al. Microplastics and Nanoplastics*   (2023) 3:9

Page 6 of 14

Microplastics data set characterization (ground truth)
Based on the first annotation (human annotation), the average size distribution of the microplastics was determined. Because the second annotation did not aim to determine the particle size, this was the only data available for this statistic. Over 60% of the microplastics had a diameter between 20 μm and 50 μm. In the second annotation (using the criteria listed in Table 1), the domain expert classified 6,864 as polymeric (10.7%, classes 1—5) and 57,437 as class 0. PE, PS, and PVC were most frequently detected, followed by PP and PET (see Fig. 3). The results of the second annotation served as the ground truth for this study.

Machine learning
All code was written in Python using Jupyter Notebook (https://jupyter.org). Jupyter Notebook is a web-based interactive computing environment for creating notebook documents. The project can be used to develop open-source software, open standards, and services for interactive computing across multiple programming languages. The code is available as electronic supplementary material.

Definitions and aim of the machine-learning algorithms
The machine-learning algorithm should result in a high recall (as many microplastics as possible being identified) with a very high rate of precision (no false positive identifications) at the same time:

> Precision: False positives lead to an overestimation of microplastics concentrations, and therefore to issues such as the overestimation of waste water treatment plants as emission pathways of MP or the impossibility of source allocation in industrial production plants.

Recall: If microplastics are missed, a false-low result is generated. As a rule, this only leads to significant erroneous detection in plastics classes that account for a small percentage of the total microplastics concentration in a sample. However, frequent plastics are also underreported. This is because a minimum number of particles per analysis and class must be exceeded to be significantly above the blank value of the analysis procedure. While frequently occurring classes (such as PVC in a PVC manufacturing plant) usually well exceed this threshold, a large number of false negatives among the rarer classes can quickly lead to a shortfall. To reduce incidental underreporting in rare classes, good scientific practice recommends the analysis of three independent samples per sampling point and three subsamples per sample.

Deep learning (one single model)
The input values for the single model implemented using Tensorflow were gradients of the Raman shift. Figure 4 shows a sample frequency spectrum with the expected peaks for PE (upper subplot A). The lower subplots (B1 and B2) illustrate the gradients. The domain experts indicated a reasonable range between 562 and 1784. This corresponds to the range highlighted in green, which was selected as input for the deep-learning network.

Since the Raman spectrometer used here stores only every other value between 562 cm^{-1} and 1784 cm^{-1}, this yielded 611 ($= \frac{1784-562}{2}$) as the number of inputs to the neural network. Figure 5 summarizes the single model. The model has 611 input nodes, 128 fully connected
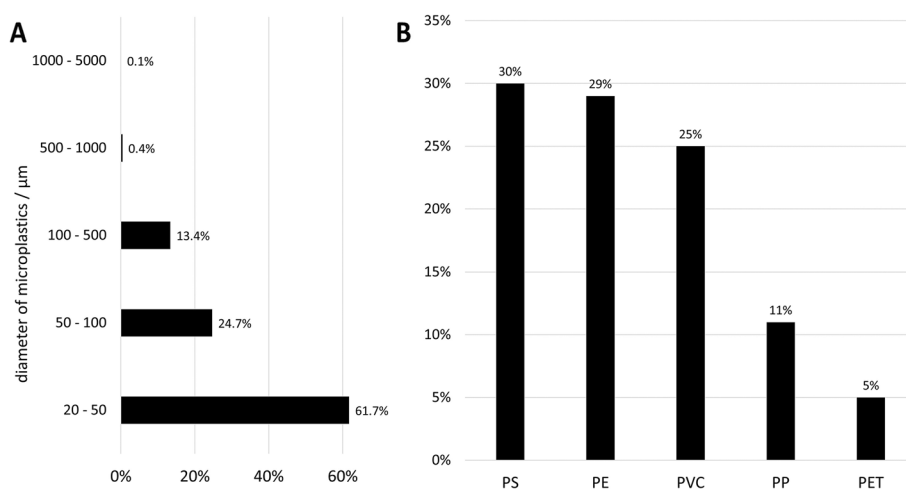

Fig. 3 Size (**A**) and polymer distribution (**B**) of microplastics in the data set

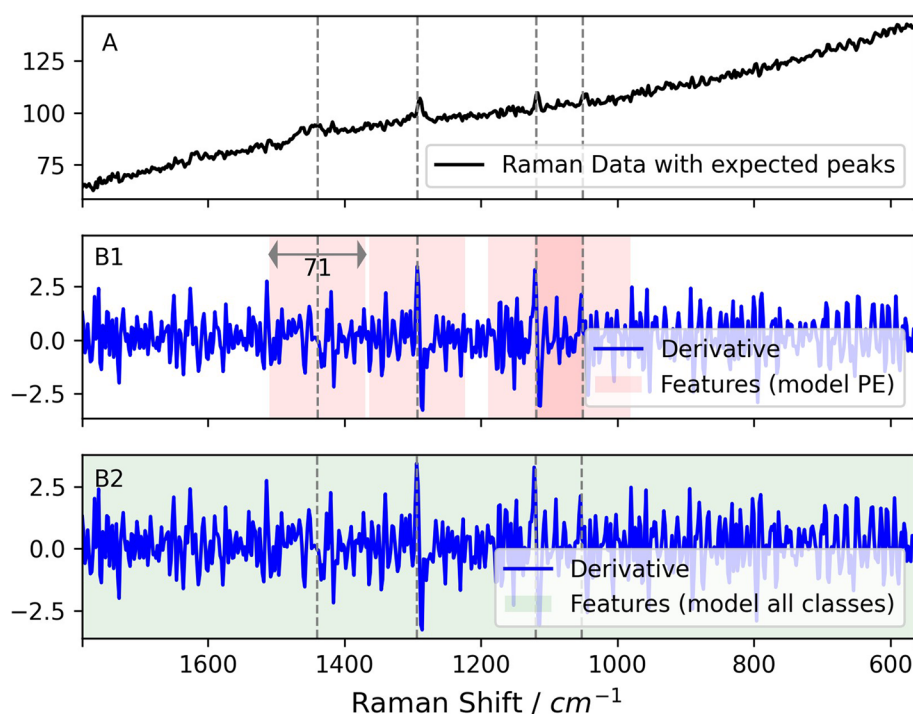**Fig. 4** For one single model, the gradient of the Raman shift (562—1784 cm^{-1}) was included (B2). For one model per class (PE in this case) sections around the peaks were extracted (B2)
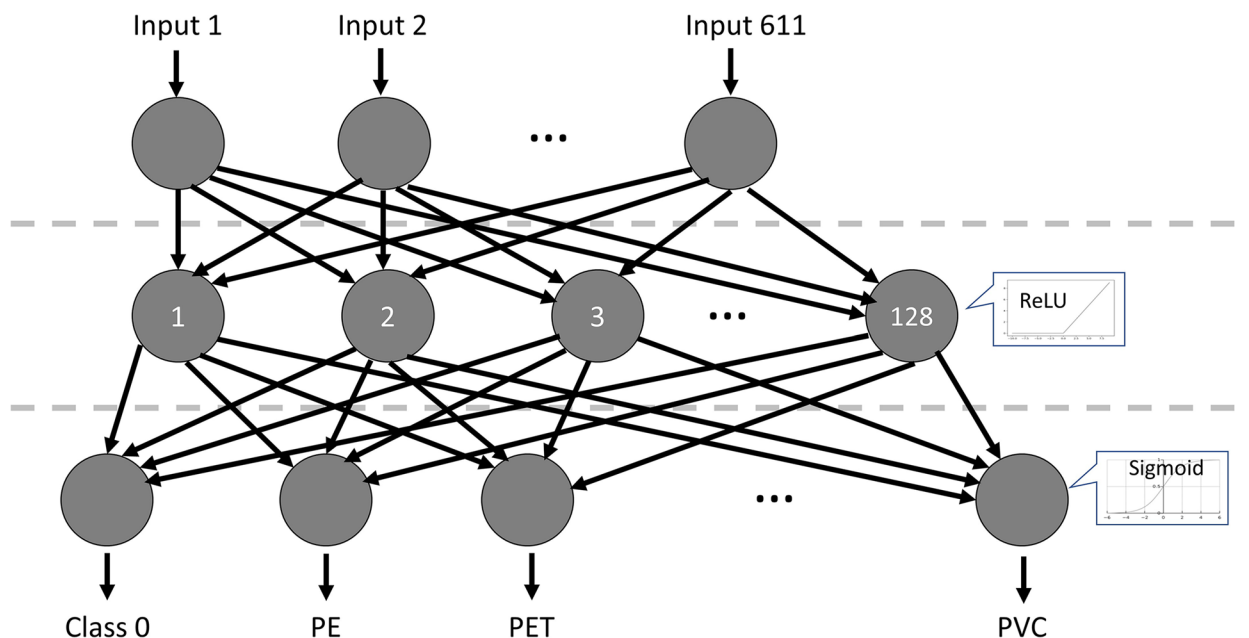

Fig. 5 Deep-learning model for all classes, with 128 fully connected hidden nodes and six outputs

hidden nodes, and 6 output nodes for the 0 class and each class of interest. The ReLU function (hidden layer) and the sigmoid function (output layer) were used as activation functions. Each output node yielded a value between 0 and 1 as a result of the sigmoid activation function. A value of 1 stands for a safe decision for the specific class, and 0 is against it. 0.5 was chosen as a decision limit for a class. L2 regularization ($\lambda = 0.001$) and a dropout layer

Weber *et al. Microplastics and Nanoplastics* (2023) 3:9

Page 8 of 14

(0.5) prevented overfitting. We trained the model with a single batch, 1500 epochs, weighted with respect to the number of instances per class. 47 models were trained to evaluate the deep-learning network that models all classes simultaneously. When training a model, the sample to be tested is always omitted. As there were several multiple determinations of one sample in the data set, these samples would have been omitted as well and were therefore excluded from the training. The training was done on the remaining samples, to ensure that the training was done only with samples that were independent of each other. This procedure enabled us to draw conclusions about the performance beyond foreign samples and was taken into account for the one-model-per-class method as well (see the section entitled "Deep learning (one model per class)").

Deep learning (one model per class)

Not all gradients between $562~\mathrm{cm}^{-1}$ and $1784~\mathrm{cm}^{-1}$ were used as input to the models. Only sections around the peaks named by domain experts (see Table 1) were used. The middle subplot of Fig. 4 illustrates how areas of width 71 were extracted around each expected peak. This led to $284~(=4\times71)$ inputs for the PE model because of the four expected peaks. The number of inputs was different for each class: The PE and PET models had 284 inputs, while PS, and PP had $355~(=5\times71)$ inputs each. Lastly, PVC had only 3 important peaks according to domain experts, resulting in $213~(=3\times71)$ inputs. Figure 6 is an exemplary summary of the model for PE. The

model implemented using Tensorflow had 284 input nodes, 32 fully connected hidden nodes, and one output node for the class of interest. We used the ReLU function (hidden layer) and the sigmoid function (output layer) as activation functions. The output node of a model yielded a value between 0 and 1 as a result of the sigmoid activation function. A value of 1 stands for a safe decision for the class and 0 is against it. For evaluation of the standalone deep-learning method, the decision threshold was set at 0.5. For a combination of deep leaning and human annotation (see the section entitled "Human–machine teaming"), the threshold was reduced to 0.1. L2 regularization ($\lambda=0.001$) and a dropout layer (0.5) prevented overfitting. We trained the model with a single batch, 1500 epochs, weighted with respect to the number of instances per class.

The system was set as a multi-label-classification, where multiple labels were to predicted for each instance. When spectra were classified in more than one class with a value >0.5, both results were accepted. This practice is legitimate because a microplastic particle may have more than one polymer spectrum. This may be due to the composite material from which the microplastic particle originates, or because agglomeration of microplastic particles cannot be completely prevented despite careful sample preparation. As these cases are very rare, they do not have a significant impact on the results.
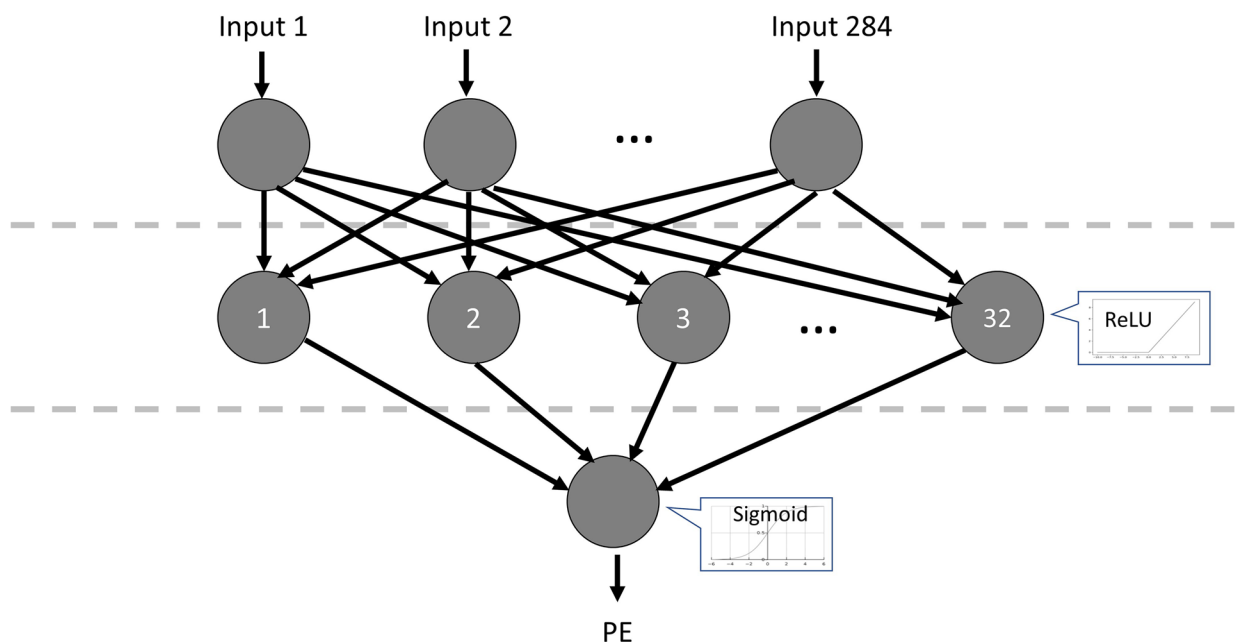

Fig. 6 Deep-learning model for the PE class with 284 inputs, 32 fully connected hidden nodes, and one output

Weber *et al. Microplastics and Nanoplastics*   (2023) 3:9

Page 9 of 14

Human–machine teaming

In this section, we present an approach that combines the methods of human annotation and deep learning (one model per class) to achieve better results. The results of human annotation show that it is trustworthy (see results section). However, in practice, the human error rate increases when the number of spectra examined is very large.

Therefore, the one model per class algorithm was applied as a machine preprocessing in a first step before a domain expert validated the results in a second step.. The decision threshold of the algorithmwas reduced from 0.5 to 0.1. As a result, the recall was artificially increased. This was necessary, because as shown in Table 4, for e.g., PVC, the recall is only about 84.4%. at a threshold of 0.5. 15.6% of the positive instances are filtered out and lead to worse results. By decreasing the threshold to 0.1, all true positives should be recognized as polymers by the network. Table 4 shows that the recall for PVC increases to 99.8%. At the same time, however, more false positives get be misclassified as polymers. However, the precision decreases at the same point. Therefore, in the second step, a domain expert evaluated the results of the network by rejecting false positives and thereby increasing the precision to the level of human annotation. This was done by sorting the results of the machine learning algorithm by polymer class and probability (value between 0 and 1). In a development of the GUI used for the second human annotation, these results were presented to the domain expert. The expert decided whether the classification of the network was correct or incorrect.

Results and discussion

Human annotation (first annotation)

Table 2 summarizes the results of the human annotation method (the method applied before the developments of this study). The annotation was compared to the ground truth (result of the second human annotation). The false negatives indicate how many instances of each class were missed by the domain experts. Compared to the real number of positives, the values were high for all classes except

Table 2 Errors, precision, and recall of the first human annotation in comparison to the second human annotation (ground truth)

Class	PE	PET	PP	PS	PVC
False negatives	392	283	248	324	695
False positives	57	2	12	23	13
Recall in %	83.6	55.7	75.4	84.1	97.0
Precision in %	97.1	99.4	98.4	98.7	99.2

PVC. It was only for PVC that the results were acceptable. The values for PET and PP were particularly high. The results of the false negatives were directly reflected in the recall, which was high for PVC, in the mid-range for PE and PS, and low for PET and PP. If the spectra of these particles are evaluated manually by identifying significant peaks for a substance in the spectra of a measuring point on the particle being analyzed, the evaluation is subject to various sources of error: A. In borderline cases, the domain expert subjectively decides on the class allocation. B. It takes several hours to analyze a sample's 1000—3000 spectra. Loss of concentration increases the risk of false positives and false negatives. C. All polymer classes are analyzed simultaneously so that a sample only requires processing once. This can also lead to false assignments due to concentration problems. D. For PP and PET, there is a further explanation: Several instances of PET and PP were measured on fibers, with several measuring spots. The domain experts identified the fibers as a class only once, meaning many points on the fiber were ignored in the classification.

There were only a few false positives, resulting in a high rate of precision, sufficient for practical applications. However, the low recall means that many instances were undetected and thus were not taken into account, which could influence countermeasures to reduce microplastics emissions.

Deep learning (one single model)

Comparing the results from a single deep-learning model to the results of human annotation (see results in the section entitled "Human Annotation (first annotation)") in Table 3, we see a higher recall, except for the PVC class. The precision is lower for each class. The results for the PET, PP, and PVC classes show that the recall of this automatic decision system is still not sufficient. In practice, this would mean that emissions from wastewater treatment plants were missed, for example, or that concentrations in environmental samples were underdetermined. Furthermore, a precision rate around or below 90% means that emissions would be incorrectly reported. Increasing the number of parameters and epochs would not lead to better results. Increasing the complexity by adding layers even worsened the results. The model chosen is a good choice, although neither its precision nor recall was sufficient. Therefore, the method of one model per class was applied to optimize the results.

Table 3 Precision and recall of deep learning (one single model)

Class	PE	PET	PP	PS	PVC
Recall in %	96.4	91.5	91.3	97.2	87.8
Precision in %	92.1	87.3	90.5	87.6	94.2

Table 4 Precision and recall of deep learning (one model per class)

Class	Threshold	PE	PET	PP	PS	PVC
Recall in %	0.5	99.2	98.8	98.9	99.0	84.4
Precision in %		88.8	74.2	79.5	86.2	86.5
Recall in %	0.1	99.9	99.8	99.4	99.8	99.8
Precision in %		71.7	33.0	52.0	70.5	72.8

Deep learning (one model per class)

Table 4 summarizes recall and precision for the one model per class method, separated by decision threshold. The recall was much higher with separate models than with a single model (see Table 3). This could have been due to additional prior knowledge by domain experts, which restricted the feature space from dimension 611 to 284 (using PE as an example). This meant the algorithm's search range was predefined. For a threshold of 0.5, the recall for all classes was above 98.4%. Reducing the threshold to 0.1 even increased the recall to over 99.4% for all classes. This high recall ensures a high probability that all emissions from a wastewater treatment plant will be identified, among other advantages. However, this high recall came at the expense of precision. The precision dropped below 90% for all classes at a threshold of 0.5. At below 80%, it was lowest for the PET and PP classes. At a reduced threshold of 0.1 the precision rate for all classes was far below 80%, with a precision rate of only 33.0% for PET. Increasing the number of parameters and epochs did not change the results. Increasing the complexity by adding more layers worsened the results (overfitting). Therefore, the degrees of freedom seem to be a good choice.

In summary, the recall achieved with this method was very good for practical applications of deep learning in the field of microplastics detection. However, in the method's present form, the precision rate would be unsuitable in practice. The number of misallocated microplastics was too high to gain reliable analysis data.

Human–machine teaming

The results of human annotation shown in the section entitled "Human annotation (first annotation)" showed a very high rate of precision but a low recall, while the results of the deep learning (one model per class) showed a high recall with low precision (see the results in the section entitled "Deep learning (one model per class)"). The combination of both methods (human–machine teaming, see the method section entitled "Human–machine teaming") resulted in a recall of $\geq 98.4\%$ at a deep-learning threshold of 0.5, and in a recall of $\geq 99.4\%$ and a precision rate of $\geq 97.1\%$ at a threshold of 0.1. A further advantage to this method is the reduction in the time required by human annotation: While human-only annotation requires that all instances be verified, human–machine teaming was used only 18% required verification by a domain expert. However, since 60% of the instances are true positives, the extra effort was actually reduced to approximately 7%. Due to this and the GUI developed with class presorting and peak indication, the evaluation time for one environmental sample was reduced to < 1 h; in most cases it is only a few minutes.

Discussion: data set and methods

As explained in Introduction, the microplastics spectra from environmental samples are often of poor quality, resulting in several problems for the development and application of deep-learning methods.

Annotation

The results of the methods involving deep learning presented here showed that high recall rates were accompanied by low precision, meaning that the number of false positives increased sharply. These instances were closely examined by the domain experts during the second human annotation (see the section entitled "Improvements in human annotation and the data set (second annotation)"). Very often, the expert was unsure whether certain examples belonged to the class or not. Often all the expected peaks were present and were even in the correct positions. However, the expert lacked the unique peak that outperformed the environment. So, to prevent over-finding, the domain expert decided to assign these instances to the 0 class, meaning that many supposed false positives may be correctly classified by deep learning after all. Unfortunately, there is no method requiring reasonable effort that can be used to verify the actual allocation of these particles to classes. Therefore, the annotation in our use case remains uncertain regarding the precision provided by deep learning.

Multi-label and inter-class variance

In practice, one spectrum could consist of peaks from more than one substance. Peaks from the various substances can overlap and are therefore often shifted or smeared. This is of special relevance, because dyed microplastics sometimes show pigment and polymer

spectra as well. Figure 7 illustrates the spectrum of a microplastic particle with peaks resulting from PE, PP, and the pigment copper phthalocyanine (CuPC).

Furthermore, the spectra graphs for the same polymer were often very different. Figure 8 shows three examples of PE spectra, with S1 and S3 differing greatly in intensity: While in S1 the Raman intensity was between 0 and 50, the intensity went up to 1500 in S3. S3 was also characterized by increasing interference (the Raman intensity rose) in the signal as the frequencies progressed from high to low. While some instances re characterized by clear peaks with little fluctuation (such as S1), high fluctuations were evident in S2, which made recognition difficult. Finally, the graphs for different instances varied greatly; for example, marginal

peaks sometimes occurred due to the overlapping of other substances and noise.

Spectra with strong interferences

The methods developed in this study do not allow identification of spectra with strong interfering signals. In cases where polymer peaks are masked by such interfering signals, e.g., fluorescence caused by debris from biological-organic matrix components, no identification can be made by machine learning. However, interfering signals can be avoided by selecting appropriate sample preparation and measurement parameters. However, if interfering signals occur or have occurred in the past, there are already machine learning approaches to deal with them: Brandt et al. [12] developed a method to reconstruct
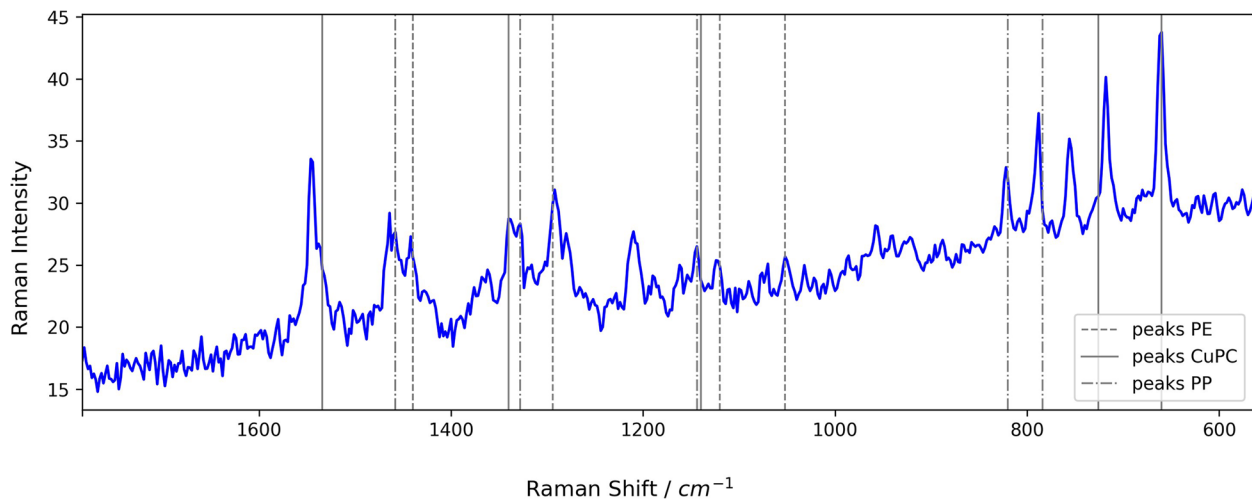

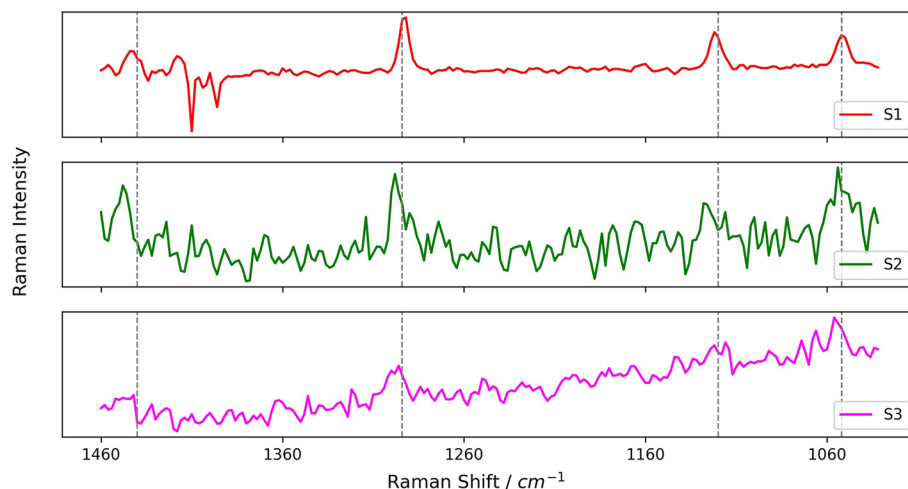**Fig. 7** Spectrum containing peaks from PE, PP, and the pigment copper phthalocyanine

Fig. 8 PE spectra from different microplastic particles: The quality (signal to noise-ratio, peak position, interference) of the spectra varies greatly

low-quality FTIR and Raman spectra with distortions such as fluorescence, interference, or cosmic rays. They validated their neuronal network using spectra from artificially aged (cryo-milling) microplastics (polyethylene terephthalate (PET), PE, poly (methyl methacrylate), PS, PP, PVC) measuring 20—500 µm. Their results show that machine learning is suited to improve poor-quality spectra. As the authors state, the application of their methods may be the reconstruction of existing spectral data. It is not suitable for polymer spectra recognition.

Size of microplastics

The quality of the spectra is strongly dependent on the particle size. In this study, only spectra of microplastic particles ≥ 20 µm were included in the data set. However, this is because this value was defined as the lower limit of quantification in the projects from which the data were obtained. Although the detection of microplastic particles ≥ 1 µm using µ-Raman spectroscopy is not a problem, there are still no methods for reliable sample preparation of environmental or wastewater samples with sufficiently high recovery rates for microplastics < 10 µm [11]. As soon as such methods are available, the spectra identification will probably have to be adapted to smaller microplastics.

Practical relevance of the results

Considering the relatively small dataset, the results of the machine learning models are good. Especially the single model approach seems to be acceptable, as recall and precision are around 90%. However, we decided to increase precision and recall by using the human–machine

teaming approach for several reasons: In spectroscopic microplastic analysis, recovery rates are low [11, 29]. Therefore, any possibility to increase accuracy should be taken. In addition, an error of 10% does not seem that high, but it can increase due to extrapolation: Usually, subsamples are analyzed because the number of particles in a sample is too high for a full µ-Raman spectroscopy analysis. The analysis of these subsamples must be as accurate as possible to avoid extrapolation errors. These arguments are strengthened by the fact that human–machine teaming does not require a lot of human effort. Most samples can be analyzed in < 1 h. This is not much time in microplastics analysis, where sample preparation and analysis of a single sample can take several days.

Therefore, the results of the human–machine teaming method are promising. However, it is important to mention that these results summarize average rates of precision and recall. In practice, it is crucial that microplastics concentrations be detected correctly in all individual samples. Since the domain experts verify all positives in the human–machine teaming method, there is no problem with precision, and the number of false positives should be correspondingly low. The situation with recall rates is different, however. Figure 9 illustrates the recall per sample and class as box plots, separated by threshold. Both the median and the quartiles were close to 1 for all classes. Samples with a recall below 1 lead to falsely low analysis results. In practice, three cases can be distinguished: A. The sample remains in the evaluation despite false negatives. The effect is not large, because this leads only to a slight underestimation of the concentration. Microplastics are at least qualitatively analyzed in this
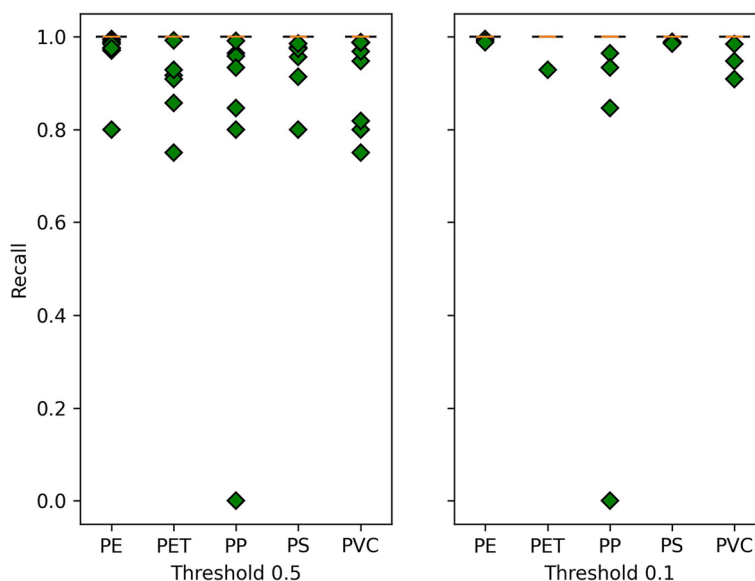

Fig. 9 Recall per sample, class, and threshold

Weber *et al. Microplastics and Nanoplastics*   (2023) 3:9

Page 13 of 14

case. B. The sample is not included because of the false negatives, so the polymer does not appear in the evaluation. However, inclusion of this polymer is important for purposes such as ecotoxicology or wastewater research, since an allocation to the source would require this information. C. The sample is not included anyway, because the number of positive particles is below the blank value. Here, the reduced recall has no effect.

For human–machine teaming, out of 282 ($=47 \times 6$) analyses of the data set with separate deep-learning networks per class, 20 were with a recall < 1. In these, case A occurred 18 times and cases B and C occurred once each. This means that there was only one sample in which one class was not detected. When human annotation alone was used, on the other hand, case 2 applied 27 times. As a result, the number of microplastic particles was underestimated, and the sample was not included in the analysis results. With use of this method, emissions were misreported and would have led to ineffective countermeasures; particularly striking was a sample in which the recall for the PP class was 0. An instance of PP was annotated but not recognized by the algorithm. However, the blank value given by the domain experts was not exceeded for the sample, so the error would have no effect in practice.

Conclusion

Machine-learning algorithms and human–machine teaming are promising methods to improve the quality of spectroscopic analyses of microplastics. With high analytical precision, they can significantly reduce the time required for sample evaluation. However, there are still problems with using deep-learning algorithms as a standalone method to analyze environmental data. Among other factors, this is because the data set from the analyses of environmental samples still presents challenges. The availability of more research data and its use as an extended database could be a solution to this problem. Both the promising approaches and the existing methodological challenges indicate the need for further research in this field.

Abbreviations

ATR	Attenuated total reflection
CuPC	Copper phthalocyanine
FTIR	Fourier transform infrared
GC	Gas chromatography
MS	Mass spectroscopy
PE	Polyethylene
PET	Polyethylene terephthalate
PP	Polypropylene
PS	Polystyrene
PVC	Polyvinyl chloride
ReLU	Rectified Linear Unit
TED	Thermal extraction desorption

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43591-023-00057-3.

> **Additional file 1.**
>
> **Additional file 2.**

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

References
1. Heß M, Völker C, Brennholt N, Herrling PM, Hollert H, Ivleva NP et al. Microplastics in the Aquatic Environment. In: Kramm J, Völker C, Johanna Kramm, Carolin Völker, editors. Living in the plastic age: Perspectives from humanities, social sciences and environmental sciences. Frankfurt, New York: Campus Verlag; 2023. p. 51–86.
2. Florides P, Völker C. Explaining Agenda-Setting of the European Plastics Strategy.: A Multiple Streams Analysis. In: Kramm J, Völker C, Johanna Kramm, Carolin Völker, editors. Living in the plastic age: Perspectives from humanities, social sciences and environmental sciences. Frankfurt, New York: Campus Verlag; 2023, pp. 25–50.
3. Ivleva NP. Chemical analysis of microplastics and nanoplastics: challenges, advanced methods, and perspectives. Chem Rev. 2021;121(19):11886–936.
4. Braun U, Altmann K, Bannick CG, Becker R, Bitter H, Bochow M et al. Analysis of Microplastics - Sampling, preparation and detection methods: tatus Report within the framework program Plastics in the Environment. 2021. https://bmbf-plastik.de/en/publication/analysis-microplastics-sampling-preparation-and-detection-methods. Accessed 23 Feb 2023.

5.   Thornton Hampton LM, Brander SM, Coffin S, Cole M, Hermabes-siere L, Koelmans AA, et al. Characterizing microplastic hazards: which concentration metrics and particle characteristics are most informa-tive for understanding toxicity in aquatic organisms? Micropl Nanopl. 2022;2(1):1–6.
6.   Liu S, Shang E, Liu J, Wang Y, Bolan N, Kirkham MB, et al. What have we known so far for fluorescence staining and quantification of microplas-tics: a tutorial review. Front Environ Sci Eng. 2022;16(1):1–4.
7.   Bayo J, Olmos S, López-Castellanos J. Microplastics in an urban wastewa-ter treatment plant: the influence of physicochemical parameters and environmental factors. Chemosphere. 2020;238: 124593.
8.   Anger PM, von der Esch E, Baumann T, Elsner M, Niessner R, Ivleva NP. Raman microspectroscopy as a tool for microplastic particle analysis. TrAC, Trends Anal Chem. 2018;109:214–26.
9.   Brandt J, Fischer F, Kanaki E, Enders K, Labrenz M, Fischer D. Assessment of subsampling strategies in microspectroscopy of environmental micro-plastic Samples. Front Environ Sci. 2021;8:579676.
10.   Anger PM, Prechtl L, Elsner M, Niessner R, Ivleva NP. Implementation of an open source algorithm for particle recognition and morphological characterisation for microplastic analysis by means of Raman microspec-troscopy. Anal Methods. 2019;11(27):3483–9.
11.   Weber F, Kerpen J. Underestimating microplastics? Quantification of the recovery rate of microplastic particles including sampling, sample preparation, subsampling, and detection using μ-Ramanspectroscopy. Anal Bioanal Chem. 2022. https://doi.org/10.1007/s00216-022-04447-z.
12.   Brandt J, Mattsson K, Hassellöv M. Deep learning for reconstructing low-quality FTIR and Raman Spectra—a case study in microplastic analyses. Anal Chem. 2021;93(49):16360–8.
13.   von der Esch E, Lanzinger M, Kohles AJ, Schwaferts C, Weisser J, Hofmann T, et al. Simple generation of suspensible secondary microplastic refer-ence particles via ultrasound treatment. Front Chem. 2020;8:169.
14.   Ribeiro-Claro P, Nolasco MM, Araújo C. Characterization of Microplastics by Raman Spectroscopy. In: Rocha-Santos TAP, Duarte AC, editors. Char-acterization and analysis of microplastics. Amsterdam: Elsevier; 2017. p. 119–51.
15.   Anger P. Strategien zur Analyse von Mikroplastik mittels RAMAN-Mikrospek-troskopie. Munich Institute of Technology. 2020. https://mediatum.ub.tum.de/doc/1531347/1531347.pdf. Accessed 23 Feb 2023.
16.   Kedzierski M, Falcou-Préfol M, Kerros ME, Henry M, Pedrotti ML, Bruzaud S. A machine learning algorithm for high throughput identification of FTIR spectra: application on microplastics collected in the Mediterranean Sea. Chemosphere. 2019;234:242–51.
17.   Hufnagl B, Stibi M, Martirosyan H, Wilczek U, Möller JN, Löder MGJ, et al. Computer-assisted analysis of microplastics in environmental samples based on μFTIR imaging in combination with machine learning. Environ Sci Technol Lett. 2021;9(1):90–5.
18.   Lei B, Bissonnette JR, Hogan ÚE, Bec AE, Feng X, Smith RDL. Customiz-able machine-learning models for rapid microplastic identification using raman microscopy. Anal Chem. 2022;94(49):17011–9.
19.   Luo Y, Su W, Xu X, Xu D, Wang Z, Wu H, et al. Raman spectroscopy and machine learning for microplastics identification and classification in water environments. IEEE J Select Topics Quantum Electron. 2023;29(4: Biophotonics):1–8.
20.   Barkmann L, Bitter E, Bitter H, Czapla J, Engelhart M, Eslahian KA et al. EmiStop: Identification of industrial plastic emissions using innovative detection methods and technology development to prevent environ-mental input via the wastewater pathway; Final Report. German version. BS Partikel GmbH; EnviroChemie GmbH; Hochschule RheinMain; Inter 3 GmbH; TU Darmstadt. 2021. https://bmbf-plastik.de/en/node/440. Accessed 23 Feb 2023.
21.   Weber F, Kerpen J. Investigation of microplastic sampling in surface waters by means of flow-through centrifuge.: Technical Report; German Version. 2022. https://hlbrm.pur.hebis.de/xmlui/handle/123456789/57. Accessed 20 Feb 2023.
22.   Barkmann-Metaj L, Weber F, Bitter H, Wolff S, Lackner S, Kerpen J et al. Quantification of Microplastics in Wastewater Systems of German Indus-trial Parks. Sci Total Environ. 2023. https://doi.org/10.1016/j.scitotenv.2023.163349.
23.   Wolff S, Weber F, Kerpen J, Winklhofer M, Engelhart M, Barkmann L. Elimination of microplastics by downstream sand filters in wastewater treatment. Water. 2021. https://doi.org/10.3390/w13010033.
24.   Roscher L, Halbach M, Nguyen MT, Hebeler M, Luschtinetz F, Scholz-Böttcher BM et al. Microplastics in two German wastewater treatment plants: year-long effluent analysis with FTIR and Py-GC/MS. Sci Total Environ. 2021. https://doi.org/10.1016/j.scitotenv.2021.152619.
25.   Tamminga M, Hengstmann E, Deuke A-K, Fischer EK. Microplastic concentrations, characteristics, and fluxes in water bodies of the Tollense catchment, Germany, with regard to different sampling systems. Environ Sci Pollut Res. 2021. https://doi.org/10.1007/s11356-021-16106-4.
26.   Mintenig SM, Kooi M, Erich MW, Primpke S, Redondo-Hasselerharm PE, Dekker SC, et al. A systems approach to understand microplastic occurrence and variability in Dutch riverine surface waters. Water Res. 2020;176:115723.
27.   PlasticsEurope. Plastics - the Facts 2020: An analysis of European plastics production, demand and waste data; 2020. https://plasticseurope.org/de/resources/publications/. Accessed 20 Feb 2023.
28.   Wolff S, Kerpen J, Prediger J, Barkmann L, Müller L. Determination of the microplastics emission in the effluent of a municipal waste water treat-ment plant using Raman microspectroscopy. Water Research X. 2019;2: 100014.
29.   Dimante-Deimantovica I, Suhareva N, Barone M, Putna-Nimane I, Aigars J. Hide-and-seek: Threshold values and contribution towards better under-standing of recovery rate in microplastic research. MethodX. 2022;9: 101603.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in pub-lished maps and institutional affiliations.