Environmental Science & Technology Letters

pubs.acs.org/journal/estlcu

Letter

Identification of Polymers with a Small Data Set of Mid-infrared Spectra: A Comparison between Machine Learning and Deep Learning Models

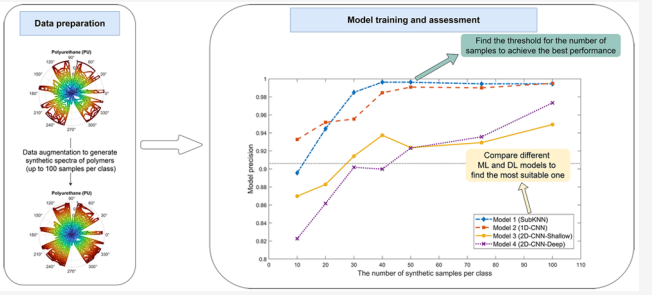
Xin Tian,* Frederic Beén, Yiqun Sun, Peter van Thienen, and Patrick S. Bäuerlein

Cite This: https://doi.org/10.1021/acs.estlett.2c00949

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Identifying environmental polymers and microplastics is crucial for the scientific world, environmental agencies, and water authorities to estimate their environmental impact and increase efforts to decrease emissions. On the basis of different spectroscopy techniques, e.g., laser-directed infrared imaging and Raman spectroscopy, polymers can be observed and represented as spectroscopic signals. The latter can be further analyzed and classified by data science, in particular, machine learning (ML). Past studies applied a variety of ML models to identify polymers from small or large data sets. However, a comprehensive comparison of multiple models across different data set sizes is still needed, which is presented in this study. Furthermore, we also provide a practical data augmentation technique to generate synthetic samples when only a limited number of samples are available. Our results show that the ensemble ML model, compared to neural network models, takes the least training time to achieve the best performance, i.e., a classification accuracy of 99.5%. This study provides a generic framework for selecting ML models and boosting model performance to accurately identify polymers.



KEYWORDS: *microplastics, polymers, LDIR, ensemble-supervised learning, deep learning, data science*

1. INTRODUCTION

Polymers, including microplastics, are ubiquitous in the environment. The new European Drinking Water Directive came into force on January 12, 2021. European Union (EU) member states must transpose the provisions of the revised drinking water directive into their national laws and regulations. Article 13 of the new directive refers to a "watch list" of substances to address growing concerns about new compounds, including explicitly microplastics (MPs), with respect to human health through drinking water. In the United States, the State of California also addresses microplastics in its legislation (Senate Bill 1422).[1] Therefore, reliable methods for identifying polymers and microplastics are needed. Methods describing how to detect nano- and microplastics in the environment include Fourier transform infrared (FTIR) spectroscopy,[2,3] Raman spectroscopy,[4,5] laser-directed infrared (LDIR) spectroscopy,[2,6,7] and thermogravimetric analysis (TGA)/pyrolysis.[8,9] Comprehensive reviews can be found in refs 10 and 11. With spectroscopic methods, polymers are identified mostly through comparison with reference spectra from existing libraries. However, most of these spectra are those of pristine polymers that lack the typical characteristics of polymers that have been exposed to the environment (i.e., degradation through aging and weathering). These effects can alter infrared spectra. Signal intensities change, and additional peaks appear due to the formation of new functional groups (e.g., hydroxyl) in the polymers.[12] This problem can be addressed by adding spectra of weathered polymers to the database.[13,14] However, it is also important to quantify the minimal number of samples required by numeric approaches to distinguished polymers.[15]

Machine learning (ML) models, especially deep neural network (DNN) models, have gained popularity and proven their efficacy in identifying the properties of microplastics,[16−20] minerals,[21−24] and organic samples.[21,22,25] Using LDIR, FTIR, or Raman spectroscopic signals, ML can classify various patterns.[19,20,23,24,26,27] Several studies trained ML models using a large data set, e.g., more than 10 000 samples.[21,22,28] Therefore, the performance of classification models is typically good, with an accuracy of ≤0.98.[25] On the contrary, due to the wide selection of ML algorithms, a few studies also examined and compared conventional ML models (e.g., *k*-nearest neighbor), ensemble ML models, and neural network models.[24,25,29] Noticeably, comparisons were mostly done

Special Issue: Data Science for Advancing Environmental Science, Engineering, and Technology

Received: December 20, 2022
Revised: December 29, 2022
Accepted: January 4, 2023
Published: January 11, 2023

with the data sets of a given size but did not assess model performances using different sizes. This implies that a comprehensive comparison of models with different data set sizes is needed to identify the most appropriate model, given that laboratories usually have different data set sizes for polymer analysis.

Even though ML and DNN models are often trained with a large data set, if there is a large number of categories, some might be underrepresented in terms of the number of spectra available per category.[23] Moreover, data imbalance is also commonly observed in reported studies. In ref 20, an ensemble ML model was developed to address data imbalance. However, when the entire data set is small (e.g., <300 samples in total), no ML model is capable of accurately classifying all categories.[22,29] To tackle this issue, data augmentation is often an effective technique used to generate synthetic samples.[30,31] For instance, the bootstrap method can be adopted to quantify the uncertainty range of prediction accuracy and evaluate identification criteria.[15] Such a method with synthetic samples also has the potential to improve model performance, commonly in deep learning applications.[32,33]

Our study aims (i) to propose a data augmentation method to generate replicate spectra of polymers in cases in which only a small data set is available, (ii) to compare models with different sizes of data sets (from 10 to 100 samples per category), (iii) to propose a framework to determine the most appropriate model and the smallest number of samples needed for classification of polymers, and (iv) to provide a visual transformation, with polarized coordinates, for both algorithms and laboratory analysts to examine the spectra of polymers more easily. To illustrate the approach and performances, a case study identifying 210 samples across 10 polymers is presented.

2. MATERIALS AND METHODS

This study considers a small data set consisting of 210 original samples for 10 polymer classes, eight of which are microplastics. Synthetic samples ($\leq$1000) were generated to improve the model performance. Both data sets were processed as one-dimensional (1D) or two-dimensional (2D) signals for training four models. All models were trained using synthetic samples and evaluated using original samples, in terms of accuracy, precision, recall, and computation time. The detailed design of this research is shown in section S1 of the Supporting Information.

2.1. Measurement, Analysis, and Quality Assurance of Original Samples. Samples were analyzed using the quantum cascade laser (QCL)-based Agilent chemical imaging LDIR system, with wavenumbers from 975 to 1800 cm^{-1} and a resolution of 0.5 cm^{-1}. In total, spectra of 210 particles from various Dutch aqueous matrices (drinking water, surface water, and effluent) and the software's original database (Agilent Clarity version 1.4.10) were used in this study. Details about sampling, sample preparation, and chemicals can be found in previous studies.[6,29]

2.2. Data Processing. *2.2.1. Synthetic Samples Generated by Data Augmentation.* If the trade-off between the model performance and the computational burden is taken into account, N synthetic samples per polymer class were generated on the basis of data augmentation, where N = 10, 20, 30, 40, 50, 75, or 100. In doing so, we attempt to find a Pareto optimal solution, yielding a satisfactory model performance with a reasonably short training time. Specifically, we resampled the replicate spectra by (i) randomly picking two observed spectra of the same specific polymer, V_1 and V_2, (ii) setting two random weights, w_1 and w_2 (*s.t.* $w_1 + w_2 = 1$), (iii) calculating the weighted sum, $w_1V_1 + w_2V_2$, as the new spectra, and (iv) repeating the first three steps N times. Details are provided in section S2 of the Supporting Information.

2.2.2. Training, Validation, and Test Data Sets. The original data set was adopted as the test set to evaluate the model performance, while the synthetic data set was used for training and validation, specifically: training set, a random 70% of the synthetic data set (0.7N samples), used to optimize model parameters to fit the model; validation set, the remaining 30% of the synthetic data set (0.3N samples), used to evaluate or tune model hyperparameters; and test set, the entire original data set (210 samples), used to unbiasedly assess the model performance after completion of training. The model performance on the test set (hereafter termed model performance) is measured by four indicators (introduced in section 2.4).

Note that the (pseudo)randomness used to divide the training and validation sets was based on the same randomness seed. In other words, all proposed models were developed using the same training and validation sets with a given N, allowing for a cross-model comparison.

2.2.3. Conversion of Coordinate Systems. Initially, all spectra were normalized and represented in a Cartesian coordinate system. Such sequence data can be used directly for training machine learning and one-dimensional convolutional neural network models (see section 2.3). Additionally, a new approach is proposed to transform the Cartesian coordinates into polar coordinates, as 2D images. These images were required to train 2D convolutional neural network models (see section 2.3). This study also presents a quick and straightforward way of transforming the 1D vector signal into a two-dimensional image. Specifically, the wavenumbers (975−1800 cm^{-1}) were mapped to 0−360°, and the normalized absorption rate became the magnitude of the polar coordinates, with the color intensified by the value of absorption. Details are provided in section S3 of the Supporting Information.

2.3. Models. This study adopted four widely used models to identify polymers on the basis of their spectra. The first model is one of the simplest machine learning models coupled with the random subspace method.[34] The other models contain 1D and 2D CNN layers.[35] Additional details about these models can be found in ref 36.

2.3.1. Model 1: SubKNN. Model 1 is a k-nearest neighborhood coupled with the random subspace (SubKNN) method. This model was compared with a collection of conventional (ensemble) machine learning models in ref 29 and showed the best performance (accuracy of 90%). Note that SubKNN uses a nonparametric KNN method that can classify samples by computing distance and voting for short-distance neighbors, while the random subspace method is added to KNN to improve performance by randomly selecting inputs for training the KNN model.[37,38]

2.3.2. Model 2: 1D-CNN. The 1D-CNN was simplified from the model architecture proposed in ref 23. Section S4.1 of the Supporting Information provides the details of the model architecture, which consists of a sequence input layer, four 1D-CNN layers, corresponding activation and normalization layers, a fully connected layer, and a classification output layer.

https://doi.org/10.1021/acs.estlett.2c00949
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

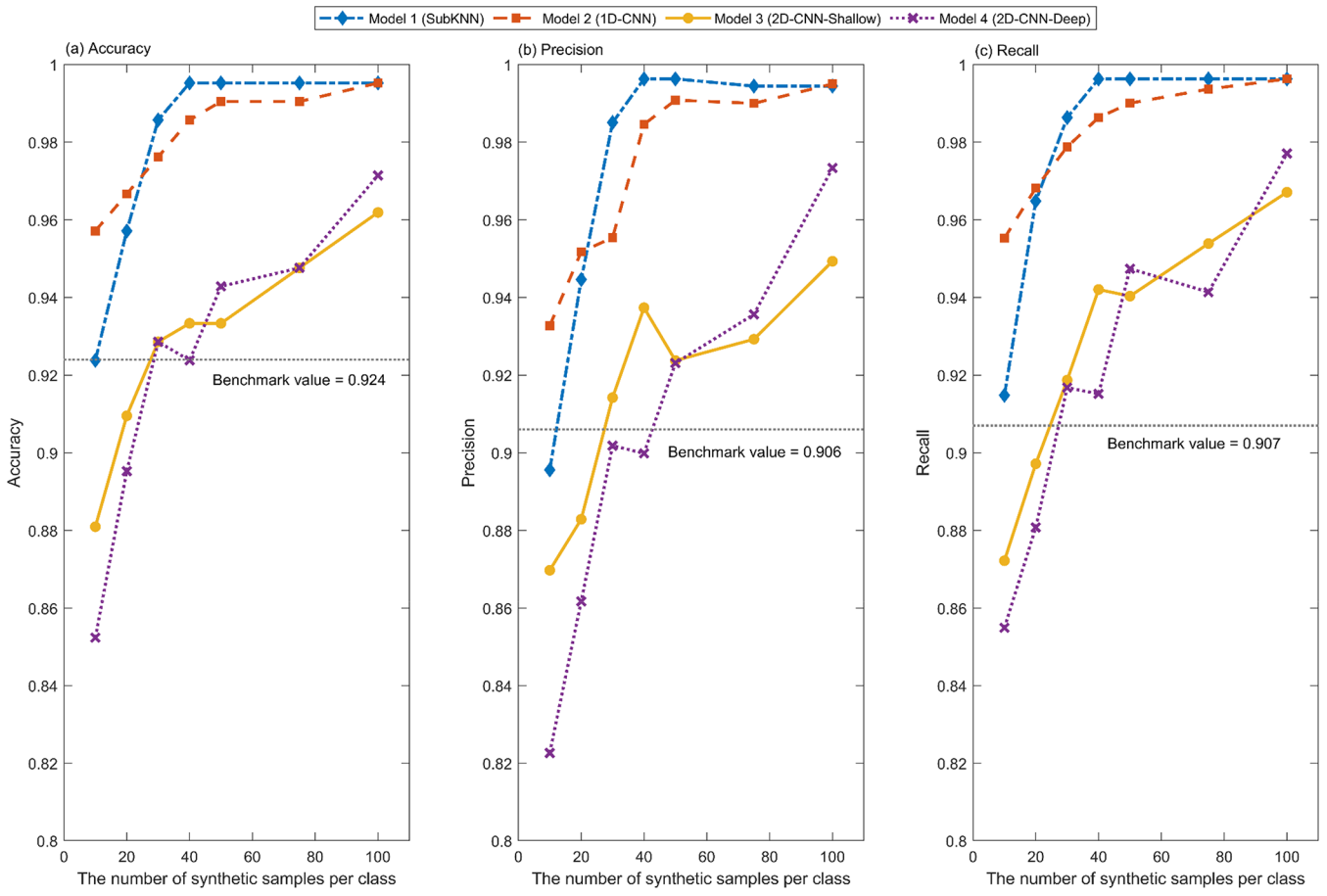


Figure 1. Performances of four adopted models. All of the models were trained on N synthetic samples per class (N = 10, 20, 30, 40, 50, 75, or 100) and tested on 210 original samples. A benchmark value for reference was computed on the basis of the SubKNN model trained and tested on the original samples. The model performance is presented as (a) accuracy, (b) precision, and (c) recall.

2.3.3. Model 3: 2DCNN-Shallow. Model 3 is a shallow 2D-CNN architecture, including an image input layer, three 1D-CNN layers, corresponding activation and normalization layers, a fully connected layer, and a classification output layer. Section S4.2 of the Supporting Information provides the details of the model architecture.

2.3.4. Model 4: 2DCNN-deep. Model 4 was built on the basis of transfer learning. The open-source GoogLeNet is a 22-layer deep neural network model trained and assessed using ImageNet.[39] Model 4 adopted GoogLeNet as the pretrained model by modifying the size of the last CNN layer and the output layers to fit the number of polymers that need to be classified in this study. Section S4.3 of the Supporting Information provides a diagram of the layers involved in the GoogLeNet, and readers can refer to ref 39 for further details.

2.4. Assessment of Model Performance. The model performance is evaluated by three performance indicators, namely, accuracy, precision, and recall. Accuracy stands for the ratio of all correctly identified samples to all samples, precision for the ratio of correctly identified samples to all of the samples predicted as a specific polymer class, and recall for the ratio of the correctly identified samples to all of the actual samples of a given polymer class. Because they are commonly used in machine learning applications,[40] readers can find their detailed definitions in section S5 of the Supporting Information.

2.5. Simulation Settings. By crossing four models and seven values of N, we performed 28 simulations on a compact graphics card (NVIDIA T600) with 640 CUDA cores, based on Intel Core i7-9750H @2.60 GHz. The model parameters used to run simulations are given in section S6 of the Supporting Information. An extra benchmark simulation was also performed to evaluate the use of synthetic samples, which trained model 1 using the original data set. Note that the benchmark was 5-fold cross-validated.

3. RESULTS AND DISCUSSION

This section details the performances of 28 simulation models. It should be noted that all models were trained on N synthetic samples and assessed on the original samples. The latter is presented in this section. On the basis of the cross-comparison of model performances, we also provide recommendations for selecting models, boosting performance with synthetic data, and adapting the proposed method to other applications.

3.1. Model Performance. *3.1.1. Comparison of Different Models.* As shown in Figure 1, the maximum accuracy, precision, and recall values for the four adopted models are all >0.94, when the models are trained with 100 synthetic samples. This implies that each of these models can accurately identify polymers with a 6% false alarm rate at most. The best model with an accuracy score of 0.995 (i.e., model 1 trained on 40 synthetic samples) has only one misclassified sample. In other words, this model is the most suitable one for classifying polymers on the basis of spectral signals. Additionally, although the model architecture of models 1 and 2 is simpler than that

https://doi.org/10.1021/acs.estlett.2c00949
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

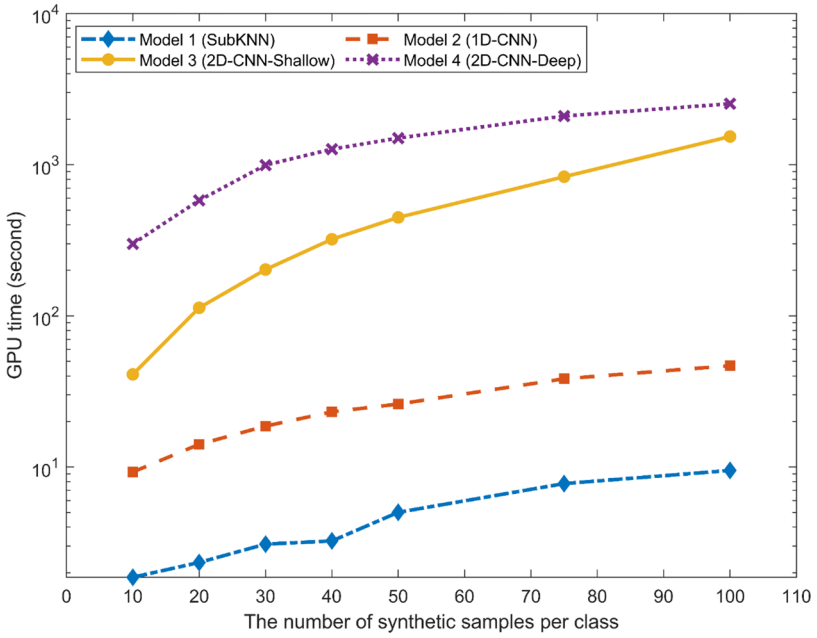


Figure 2. Training times of four adopted models.

of models 3 and 4, the former two outperform the latter two, in terms of all performance indicators.

3.1.2. Model Performance Impacted by the Number of Spectra in Training Sets. In addition, Figure 1 demonstrates that the number of training spectra (the *x*-axis) has a significant impact on the model performance. The performance of all models increases as the number of samples increases, with only a few exceptions between 30 and 50 samples. In other words, optimal model performance can be jeopardized if models are trained with few samples.

Models 1 and 2 can be distinguished from the other two on the basis of their performance. Model 2 outperforms model 1 only when the model is trained on a small data set (<20 samples), whereas model 1 performs best as the number of samples increases beyond 30. Model 1 achieves the best performance with 40 samples, whereas model 2 converges to the same value with 100 samples. Section S7 of the Supporting Information shows the classification results on the basis of model 1 and different sample sizes. Similarly, model 3 marginally outperforms model 4 only when trained on a small data set, but both models seem to require more samples to improve performance. Considering the computation time and the fact that the optimal performance can be achieved by models 1 and 2, model training with more than 100 samples per polymer class was not necessary in this study.

3.1.3. Compared with the Benchmark. The benchmark simulation was conducted using model 1, trained, and cross-validated on the basis of the original data set with 210 samples. Figure 1 shows the benchmark as a reference line. In general, using 50 samples or more per polymer class can result in a performance that is better than the benchmark, for all of the models. In particular, model 1 (the most appropriate model) requires only 20 samples or more to outperform the benchmark. This is due to the uneven distribution of 10 types of polymers in the original data set (see section S2 of the Supporting Information). It implies that it is more effective to improve the model performance by adding (synthetic or newly observed) samples to classes with extremely few samples.

3.1.4. Training Time. Figure 2 shows the training time as a function of data set size. Models 1 and 2 take up to 9.5 and 46.8 s, respectively, for training models, whereas models 3 and 4 take up to 25 and 42 min, respectively. Though the longest period (42 min) is still relatively short, it is important to consider future applications in which the model is used to classify thousands of polymers in an online learning mode, where the model needs to be trained repeatedly with added samples of new polymer types. With regard to the training time, model 1 is still an appropriate approach.

3.2. Discussion of Model Selection. Even though deep learning (DL) has attracted considerable attention in academia and industry over the past few years and has driven artificial intelligence to a new phase, it does not imply that DL can replace conventional machine learning models in every application. In particular, as in this study, DL models require a long training time and more samples to reach comparable performances. This is because DL is often implemented with thousands of samples, e.g., images with context information (not in our case). On the contrary, for application of machine learning models to other applications similar to this study, first testing a simple ML model instead of directly using a DNN model is recommended.

3.3. Discussion of Enhancing Model Performance. In principle, the use of synthetic samples is intended to enlarge data sets and improve model underfitting. This study presents an effective way of synthetically augmenting data, based on the fact that the spectroscopic signals of each polymer showed remarkably distinctive characteristics.

3.4. Implications for Polymer Classification. Our approach can also be applied to other types of spectral data, e.g., ultraviolet, Raman, FTIR, and mass spectra. One or several machine learning and/or deep learning models can be evaluated to determine the most appropriate in terms of performance versus computing time and complexity. If the performances of the model are still below what is deemed to be satisfactory, the data augmentation strategy presented here can be used to incrementally generate synthetic samples and find a threshold value for optimal performance.

https://doi.org/10.1021/acs.estlett.2c00949
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

The model performance also depends on data quality. In this study, the only misclassified spectrum was that of a polyurethane (PU) sample whose spectrum was significantly different from other PU spectra. Upon more detailed inspection, these differences are likely due to errors during the acquisition. Such quality assurance of samples is also recommended when applying the model to other cases because data quality impacts the model performance as much as data quantity.[41]

In conclusion, this study sought an effective method of identifying environmentally exposed polymers on the basis of their mid-infrared spectroscopic signals measured by LDIR. The following findings can be drawn from this investigation.

(1) The simplest model can still be effective. In our case study, a simple sub-KNN model achieved a better model performance than the other three CNN models, with the shortest training time.

(2) The number of (high-quality) samples is a crucial factor for training models. Given our small data set, the proposed data augmentation technique resulted in a remarkable performance improvement.

(3) Although the 2D-CNN models are not optimal for this case study, they seem more useful in dealing with multidimensional spectroscopic signals. Future studies will further investigate their applicability.

(4) Our method also applies to the classification of various polymers whose spectroscopic signals are measurable and of high-quality preparation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.estlett.2c00949.

Polymers, data augmentation, polymer spectra, and model architectures (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Xin Tian − *KWR Water Research Institute, 3433 PE Nieuwegein, The Netherlands;* orcid.org/0000-0002-8696-8527; Email: xin.tian@kwrwater.nl

Authors

Frederic Beén − *KWR Water Research Institute, 3433 PE Nieuwegein, The Netherlands; Chemistry for Environment & Health, Amsterdam Institute for Life and Environment (A-LIFE), Vrije Universiteit Amsterdam, 1081 HZ Amsterdam, The Netherlands;* orcid.org/0000-0001-5910-3248

Yiqun Sun − *School of Earth Sciences and Engineering, Hohai University, Nanjing 210098, China; College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China*

Peter van Thienen − *KWR Water Research Institute, 3433 PE Nieuwegein, The Netherlands*

Patrick S. Bäuerlein − *KWR Water Research Institute, 3433 PE Nieuwegein, The Netherlands;* orcid.org/0000-0002-1110-5997

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.estlett.2c00949

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was funded by the joint research program of the Dutch and Flemish water utilities (BTO 402045/228). The authors thank three anonymous reviewers, who provided helpful comments to improve the quality of this paper.

■ REFERENCES

(1) Coffin, S.; Wyer, H.; Leapman, J. C. Addressing the Environmental and Health Impacts of Microplastics Requires Open Collaboration between Diverse Sectors. *PLOS Biol.* **2021**, *19* (3), e3000932.

(2) Primpke, S.; Godejohann, M.; Gerdts, G. Rapid Identification and Quantification of Microplastics in the Environment by Quantum Cascade Laser-Based Hyperspectral Infrared Chemical Imaging. *Environ. Sci. Technol.* **2020**, *54* (24), 15893−15903.

(3) Primpke, S.; Lorenz, C.; Rascher-Friesenhausen, R.; Gerdts, G. An Automated Approach for Microplastics Analysis Using Focal Plane Array (FPA) FTIR Microscopy and Image Analysis. *Anal. Methods* **2017**, *9* (9), 1499−1511.

(4) Wolff, S.; Kerpen, J.; Prediger, J.; Barkmann, L.; Müller, L. Determination of the Microplastics Emission in the Effluent of a Municipal Waste Water Treatment Plant Using Raman Microspectroscopy. *Water Res. X* **2019**, *2*, 100014.

(5) Schymanski, D.; Oßmann, B. E.; Benismail, N.; Boukerma, K.; Dallmann, G.; von der Esch, E.; Fischer, D.; Fischer, F.; Gilliland, D.; Glas, K.; Hofmann, T.; Käppler, A.; Lacorte, S.; Marco, J.; Rakwe, M. EL; Weisser, J.; Witzig, C.; Zumbülte, N.; Ivleva, N. P. Analysis of Microplastics in Drinking Water and Other Clean Water Samples with Micro-Raman and Micro-Infrared Spectroscopy: Minimum Requirements and Best Practice Guidelines. *Anal. Bioanal. Chem.* **2021**, *413* (24), 5969−5994.

(6) Bäuerlein, P. S.; Hofman-Caris, R. C. H. M.; Pieke, E. N.; ter Laak, T. L. Fate of Microplastics in the Drinking Water Production. *Water Res.* **2022**, *221*, 118790.

(7) Bäuerlein, P. S.; Pieke, E. N.; Oesterholt, F. I. H. M.; ter Laak, T.; Kools, S. A. E. Microplastic discharge from a wastewater treatment plant: long term monitoring to compare two analytical techniques, LDIR and optical microscopy while also assessing the removal efficiency of a bubble curtain. *Water Sci. Technol.* **2022**, wst2022419.

(8) Majewsky, M.; Bitter, H.; Eiche, E.; Horn, H. Determination of Microplastic Polyethylene (PE) and Polypropylene (PP) in Environmental Samples Using Thermal Analysis (TGA-DSC). *Sci. Total Environ.* **2016**, *568*, 507−511.

(9) Gomiero, A.; Øysæd, K. B.; Palmas, L.; Skogerbø, G. Application of GCMS-Pyrolysis to Estimate the Levels of Microplastics in a Drinking Water Supply System. *J. Hazard. Mater.* **2021**, *416*, 125708.

(10) Ivleva, N. P. Chemical Analysis of Microplastics and Nanoplastics: Challenges, Advanced Methods, and Perspectives. *Chem. Rev.* **2021**, *121* (19), 11886−11936.

(11) Cowger, W.; Gray, A.; Christiansen, S. H.; DeFrond, H.; Deshpande, A. D.; Hemabessiere, L.; Lee, E.; Mill, L.; Munno, K.; Ossmann, B. E.; Pittroff, M.; Rochman, C.; Sarau, G.; Tarby, S.; Primpke, S. Critical Review of Processing and Classification Techniques for Images and Spectra in Microplastic Research. *Appl. Spectrosc.* **2020**, *74* (9), 989−1010.

(12) Duan, J.; Bolan, N.; Li, Y.; Ding, S.; Atugoda, T.; Vithanage, M.; Sarkar, B.; Tsang, D. C. W.; Kirkham, M. B. Weathering of Microplastics and Interaction with Other Coexisting Constituents in Terrestrial and Aquatic Environments. *Water Res.* **2021**, *196*, 117011.

(13) De Frond, H.; Rubinovitz, R.; Rochman, C. M. MATR-FTIR Spectral Libraries of Plastic Particles (FLOPP and FLOPP-e) for the Analysis of Microplastics. *Anal. Chem.* **2021**, *93* (48), 15878−15885.

(14) Munno, K.; De Frond, H.; O'Donnell, B.; Rochman, C. M. Increasing the Accessibility for Characterizing Microplastics: Introducing New Application-Based and Spectral Libraries of Plastic Particles (SLoPP and SLoPP-E). *Anal. Chem.* **2020**, *92* (3), 2443−2451.

(15) Morgado, V.; Palma, C.; Bettencourt da Silva, R. J. N. Microplastics Identification by Infrared Spectroscopy − Evaluation of Identification Criteria and Uncertainty by the Bootstrap Method. *Talanta* **2021**, *224*, 121814.
(16) Ng, W.; Minasny, B.; McBratney, A. Convolutional Neural Network for Soil Microplastic Contamination Screening Using Infrared Spectroscopy. *Sci. Total Environ.* **2020**, *702*, 134723.
(17) Ai, W.; Liu, S.; Liao, H.; Du, J.; Cai, Y.; Liao, C.; Shi, H.; Lin, Y.; Junaid, M.; Yue, X.; Wang, J. Application of Hyperspectral Imaging Technology in the Rapid Identification of Microplastics in Farmland Soil. *Sci. Total Environ.* **2022**, *807*, 151030.
(18) Lin, J. yu; Liu, H.-t.; Zhang, J. Recent Advances in the Application of Machine Learning Methods to Improve Identification of the Microplastics in Environment. *Chemosphere* **2022**, *307* (P4), 136092.
(19) Kedzierski, M.; Falcou-Préfol, M.; Kerros, M. E.; Henry, M.; Pedrotti, M. L.; Bruzaud, S. A Machine Learning Algorithm for High Throughput Identification of FTIR Spectra: Application on Microplastics Collected in the Mediterranean Sea. *Chemosphere* **2019**, *234*, 242−251.
(20) Yan, X.; Cao, Z.; Murphy, A.; Qiao, Y. An Ensemble Machine Learning Method for Microplastics Identification with FTIR Spectrum. *J. Environ. Chem. Eng.* **2022**, *10* (4), 108130.
(21) Shen, Z.; Viscarra Rossel, R. A. Automated Spectroscopic Modelling with Optimised Convolutional Neural Networks. *Sci. Rep.* **2021**, *11* (1), 1−12.
(22) Padarian, J.; Minasny, B.; McBratney, A. B. Using Deep Learning to Predict Soil Properties from Regional Spectral Data. *Geoderma Reg.* **2019**, *16*, e00198.
(23) Sang, X.; Zhou, R.-g.; Li, Y.; Xiong, S. One-Dimensional Deep Convolutional Neural Network for Mineral Classification from Raman Spectroscopy. *Neural Processing Letters* **2022**, *54* (1), 677−690.
(24) Liu, J.; Osadchy, M.; Ashton, L.; Foster, M.; Solomon, C. J.; Gibson, S. J. Deep Convolutional Neural Networks for Raman Spectrum Recognition: A Unified Solution. *Analyst* **2017**, *142* (21), 4067−4074.
(25) Neto, H. A.; Tavares, W. L. F.; Ribeiro, D. C. S. Z.; Alves, R. C. O.; Fonseca, L. M.; Campos, S. V. A. On the Utilization of Deep and Ensemble Learning to Detect Milk Adulteration. *BioData Min.* **2019**, *12* (1), 1−14.
(26) Hufnagl, B.; Steiner, D.; Renner, E.; Löder, M. G. J.; Laforsch, C.; Lohninger, H. A Methodology for the Fast Identification and Monitoring of Microplastics in Environmental Samples Using Random Decision Forest Classifiers. *Anal. Methods* **2019**, *11* (17), 2277−2285.
(27) Hufnagl, B.; Stibi, M.; Martirosyan, H.; Wilczek, U.; Möller, J. N.; Löder, M. G. J.; Laforsch, C.; Lohninger, H. Computer-Assisted Analysis of Microplastics in Environmental Samples Based on MFTIR Imaging in Combination with Machine Learning. *Environ. Sci. Technol. Lett.* **2022**, *9* (1), 90−95.
(28) Tsakiridis, N. L.; Keramaris, K. D.; Theocharis, J. B.; Zalidis, G. C. Simultaneous Prediction of Soil Properties from VNIR-SWIR Spectra Using a Localized Multi-Channel 1-D Convolutional Neural Network. *Geoderma* **2020**, *367*, 114208.
(29) Tian, X.; Beén, F.; Bäuerlein, P. S. Quantum Cascade Laser Imaging (LDIR) and Machine Learning for the Identification of Environmentally Exposed Microplastics and Polymers. *Environ. Res.* **2022**, *212* (May), 113569.
(30) Hesterberg, T. Bootstrap. *Wiley Interdiscip. Rev. Comput. Stat.* **2011**, *3* (6), 497−526.
(31) Wong, S. C.; Gatt, A.; Stamatescu, V.; McDonnell, M. D.Understanding Data Augmentation for Classification: When to Warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*; IEEE, 2016; pp 1−6.
(32) Shorten, C.; Khoshgoftaar, T. M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6* (1), 60.
(33) Taylor, L.; Nitschke, G.Improving Deep Learning with Generic Data Augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*; IEEE, 2018; pp 1542−1547. DOI: 10.1109/SSCI.2018.8628742
(34) Ho, T. K. Nearest Neighbors in Random Subspaces. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **1998**, *1451*, 640−648.
(35) Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86* (11), 2278−2324.
(36) Géron, A.*Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; 2017.
(37) Caradot, N.; Granger, D.; Chapgier, J.; Cherqui, F.; Chocat, B. Urban Flood Risk Assessment Using Sewer Flooding Databases. *Water Sci. Technol.* **2011**, *64* (4), 832−840.
(38) Peterson, L. K-Nearest Neighbor. *Scholarpedia* **2009**, *4* (2), 1883.
(39) Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.*Going Deeper with Convolutions*; 2014.
(40) Géron, A.*Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed.; O'Reilly Media, Inc., 2019.
(41) Ng, A.*MLOps: From Model-Centric to Data-Centric AI*; 2021.

NOTE ADDED AFTER ASAP PUBLICATION

This article originally published with a missing affiliation for author Frederic Been. The affiliation to Vrije Universiteit Amsterdam was added and the article reposted January 13, 2023.

https://doi.org/10.1021/acs.estlett.2c00949
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX