Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 4
NSCBM 2024

Detection of Nanoplastics Within Complex Environmental and Food Resources Matrices Via Machine Learning

Henry Jin

Greenwich High School, Connecticut, US

henryjin1223@gmail.com

Abstract. The pervasiveness of nanoplastics within the environment underscores the need for robust and accurate methods for their identification and classification. The lightweight and small nanoplastics (NPs) can bypass biological barriers and disperse throughout the environment, posing significant health risks to humans and aquatic life. Typical detection of nanoplastics has relied on cumbersome filtration, and subsequent coloration of the plastics for visualization, once they have been painstakingly separated from their matrix, including fish, sand, and soil using water. Raman spectroscopy, however, offers an alternative, as it effectively detects these particles without the need for separation, with its high resolution (<1μm). Unfortunately, accurate identification and classification are challenging because of the faint Raman scatter of NPs and signal interference from background noise. To address this challenge, this project proposes a method integrating machine learning (ML) with Raman spectroscopy. Multiple ML models were first trained with Raman spectra of 50μg/mL suspensions of PE, PTFE, PS, PMMA, and PVC NPs, and tested against validation data. While ML models achieved an average accuracy of >96%, the Support Vector Machine Classification model reached 99.58% accuracy in NP-identification. These ML models were then validated via analyses of NPs in water. In each case, the NPs were rapidly and successfully identified, while remaining in their glass bottle. The detection of NPs in water, this new Raman-ML model successfully detected as little as 1E5 particles/L, which surpasses new, published detection limits of only a few months ago.

Keywords: Nanoplastics, Raman Spectroscopy, Machine Learning, Support Vector Machine Classification.

1. Introduction

Nanoplastics have permeated throughout the environment due to the large-scale manufacturing and degradation of plastic products into smaller particles. Nanoplastics' small size (<1μm) enables them to permeate through biological barriers and move between tissues causing disruptions in cellular processes. This raises concerns about potential health risks, including blocking essential veins and arteries. Nanoplastics' light weight allows them to spread and circulate throughout the environment by dispersing within airflow and waterflow. Nanoplastics' lightweight nature allows them to spread throughout the environment via both air and water, contaminating ecosystems and food sources. Aquatic organisms can intake nanoplastic particles through ingestion of the particles or other afflicted organisms or through inhalation of the plastic particles. [1], [2], [3], [4], [5]

Human consumption of these afflicted aquatic organisms is harmful due to the toxic elements used in plastic manufacturing. There are also many different types of plastics including: PE (Polyethylene), PS (Polystyrene), PTFE (Polytetrafluoroethylene), PVC (Polyvinyl chloride), and PMMA (Acrylic). These different types of plastics have unique molecular structures and similar chemical compositions, differing with the polarity of atoms giving each plastic polymer different characteristics.[6], [7]

To address the pollution of nanoplastics, it is essential to detect and identify nanoplastics accurately and rapidly. However, nanoplastics' miniscule size makes detection and classification challenging. Traditional methods of detection and classification are burdened with numerous limitations on sensitivity and accuracy.[8] Raman Spectroscopy is used in this research because of its characteristic

of high resolution, high sensitivity, and ability to characterize molecules in otherwise complex background matrices. [9], [10], [11]

2. Engineering Goal

This research will explore the possibility of automatically identifying and classifying nanoplastics through Raman Spectroscopy using different machine-learning algorithms. This research will apply models (e.g., linear regression, random forest, support vector machine, etc.) to efficiently and accurately detect and classify nanoplastics. Machine learning (ML) tools will leverage Raman spectral data of samples to detect nanoplastics within those samples. The completed ML tool will be able to detect nanoplastics in a variety of typical environmental conditions where they persist, including soil, sand, water and aquatic organisms.

3. Current Methods of Nano and Microplastic Detection and Classification

Tarafdar et al. [12] showed a method for detecting microplastics in soil and water. This technique begins by vacuum filtering a sample through a medium with pores small enough to trap micro and nanoplastics, which is a time-consuming process. Next, a specialized dye is applied to the filtered material. This dye selectively binds to the plastics, leaving other materials uncolored. Finally, the sample is examined under a light or fluorescent microscope. Since the dye only colors the plastic, the microplastics are easily identified and counted, providing a qualitative analysis of the microplastic content in the original sample.

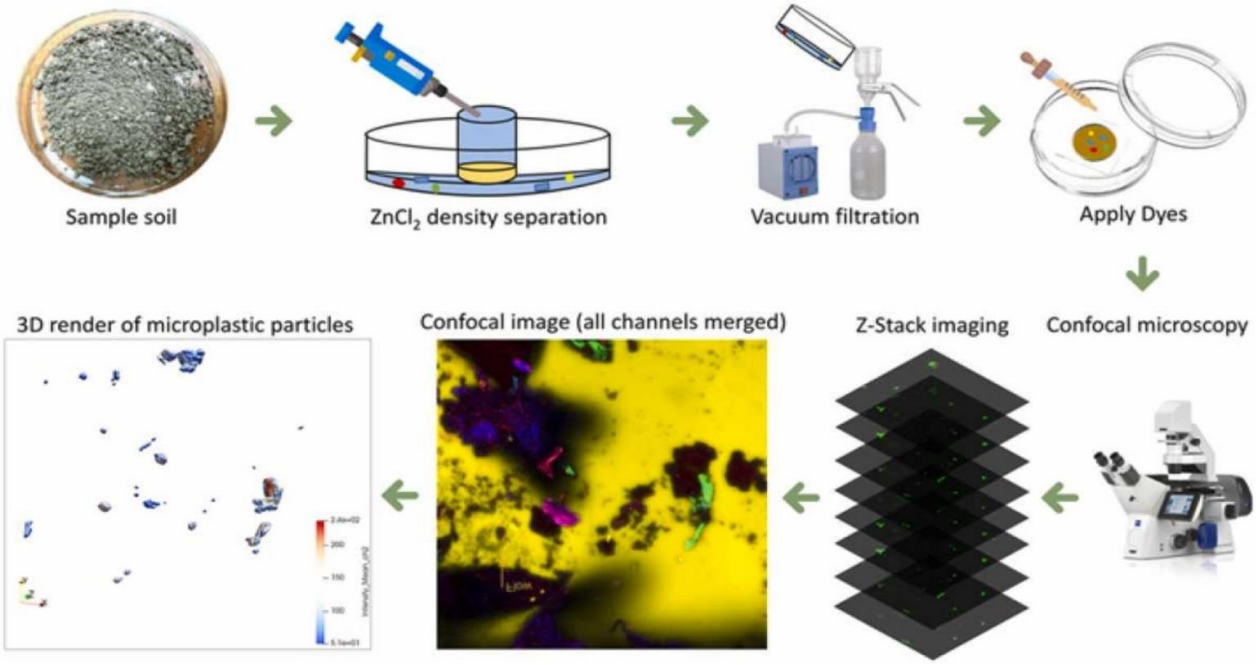

Figure 1. Manual Filtration and Separation of Nano and Microplastics from water or soil, followed by quantitative evaluation of plastics using fluorescent dyes, and microscopic inspection. (Image courtesy of Tarafdar, et al.)

Cai et al. [5] showed another method for detecting microplastics in water, sand, or soil involves stainless steel filtration, Scanning Electron Microscopy (SEM), and Energy Dispersive Analysis (EDS). First, a contaminated sample is mixed with water and vacuum-filtered through a 5-micron stainless steel mesh. This traps microplastics larger than 5 microns. The filtered material is then analyzed using SEM and EDS to confirm if the trapped particles are indeed plastic. However, this method is time-consuming and costly due to specialized equipment required and does not identify the type of plastic present.

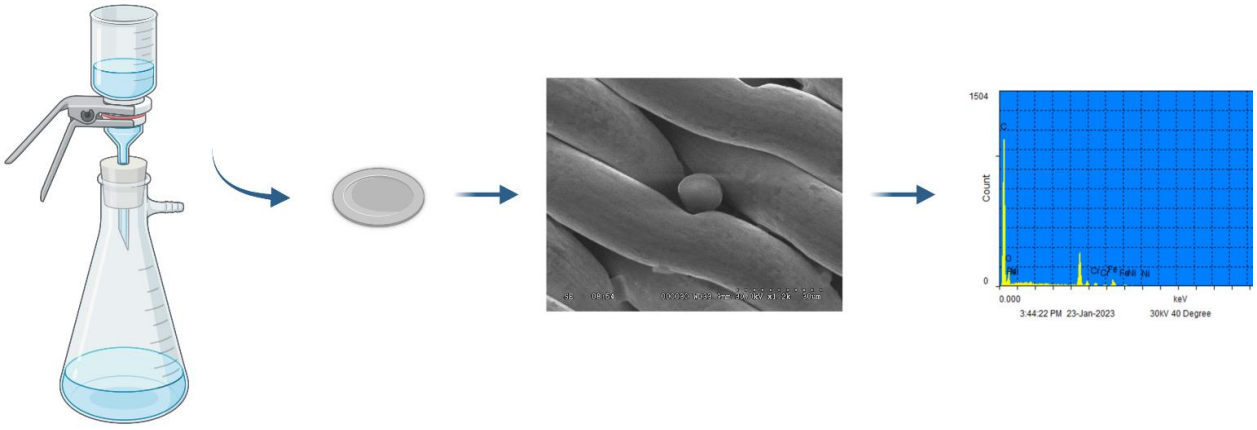

Figure 2. Manual Filtration and Separation of Nano and Microplastics from water, sand, or soil onto a 5um stainless steel filter, followed by inspection via SEM and EDS techniques.

Qian et al. [13] developed a new method for the detection and classification of nanoplastic particles. This involves manual filtration onto a Raman-Specialized membrane, manual inspection on a Stimulated Raman Scattering (SRS Microscopy) Microscope, and "point-by-point" Raman inspection of particles for nanoplastic types of present. This time-consuming method, which still includes separation (via filtration) of the plastic contaminants only in water, is currently the most advanced, published only a few months ago.

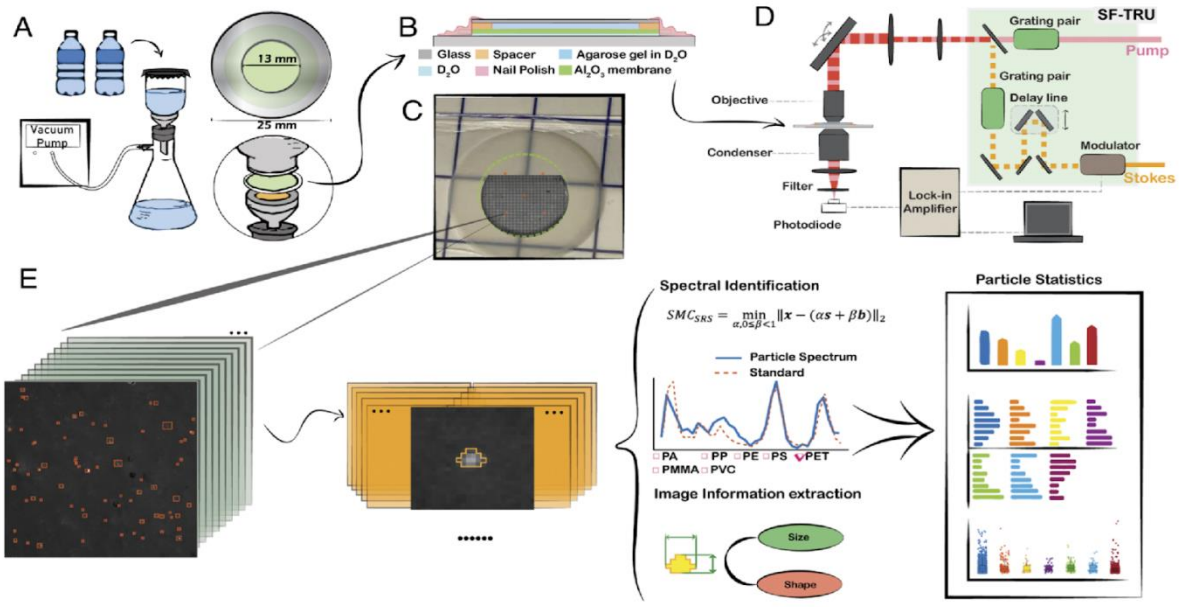

Figure 3. Identification of Nanoplastics and Microplastic in Water (only), via Specialized Manual Filtration, and Point-by-Point inspection of filtration fragments by SRS Microscopy. (Images courtesy of Qian, et al.)

4. Methodology

4.1. Pretreatment of Raman Spectroscopic Data for Each Plastic Contaminant Type

This project's 1600 datasets were obtained from a database at UW-Madison. From this database, the individual Raman spectrum for each type of nanoplastic was obtained, originally measured for nanoparticle-sized (diameter of ≤1 μm) PE, PTFE, PVC, PS, and PMMA pellets, suspended in water at a concentration of 50 μg/mL (Figs. 4-8) respectively, along with Raman spectra of samples without plastics (Fig. 9). Each Raman spectrum was truncated to the range of 500−1800 cm-1 and later

processed by polynomial equations for baseline correction. Weak Raman signal extraction and denoising was required because the inherent noise of the system interferes with the identification of feature peaks. The standardized data was then scaled with intensity to 0-1 (Fig. 10), and it is these spectra that were labeled (by nanoplastic type) and inputted into the ML models.

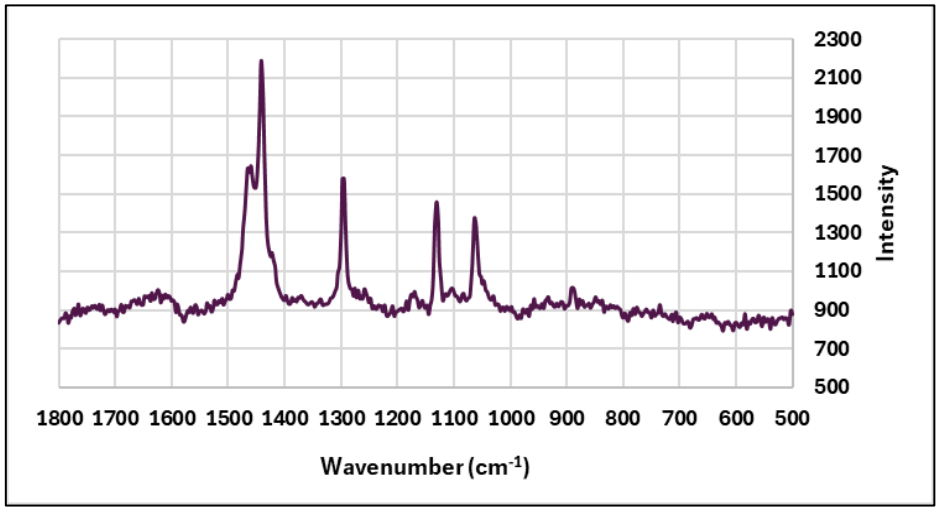

Figure 4. Polyethylene (PE) Raman Spectrum

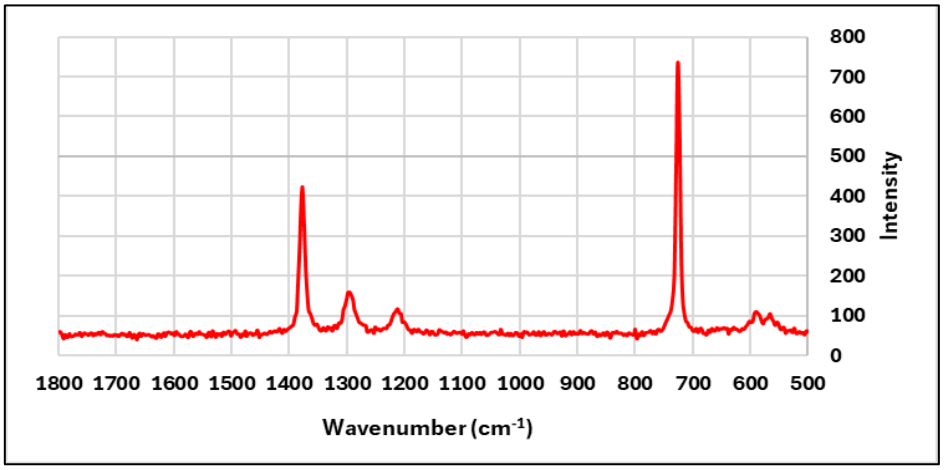

Figure 5. Polytetrafluoroethylene (PTFE) Raman Spectrum

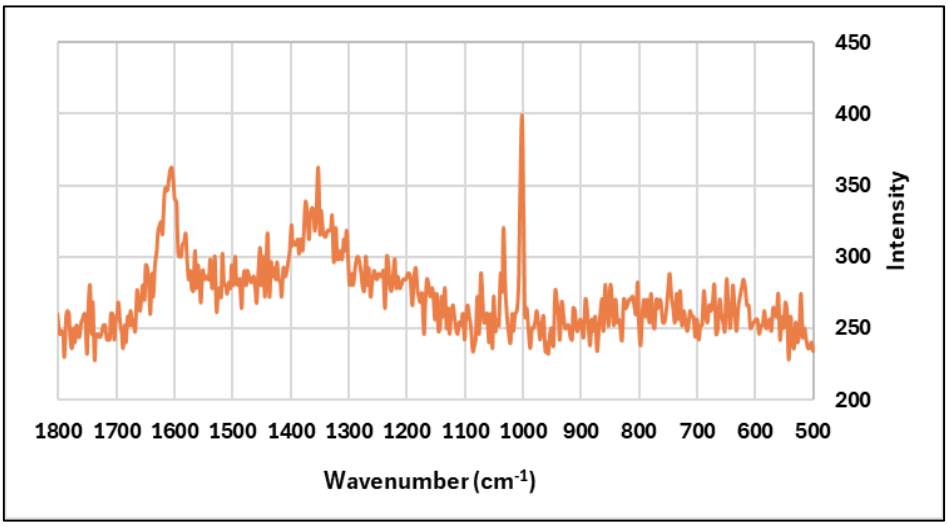

Figure 6. Polystyrene (PS) Raman Spectrum

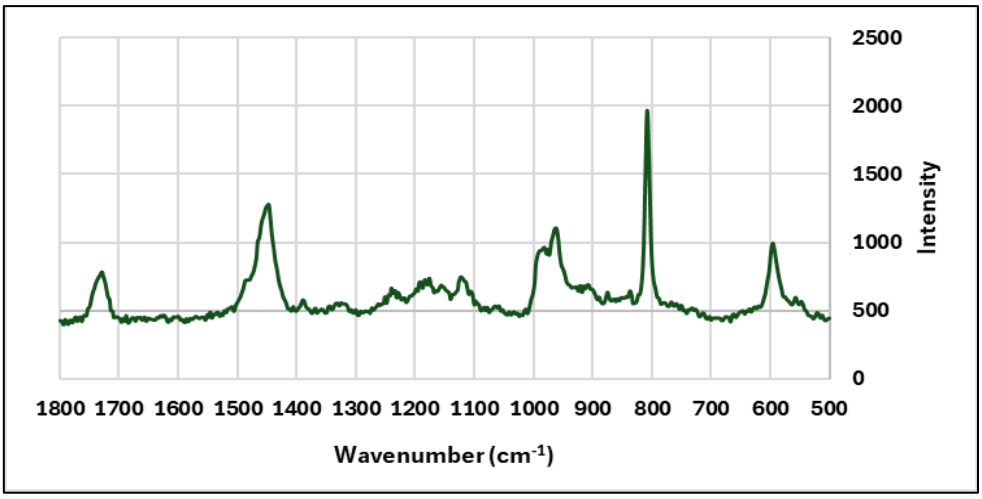

Figure 7. Poly(methyl methacrylate) (PMMA) Raman Spectrum

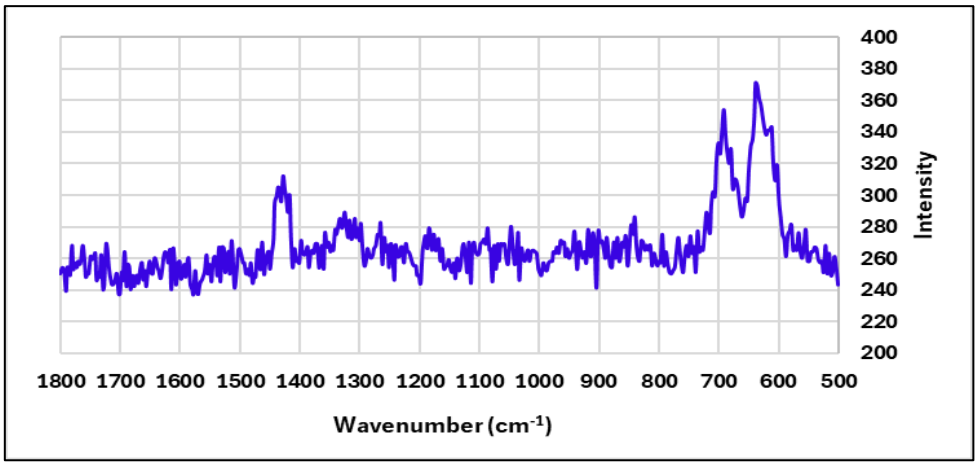

Figure 8. Polyvinyt chloride (PVC)Raman Spectrum

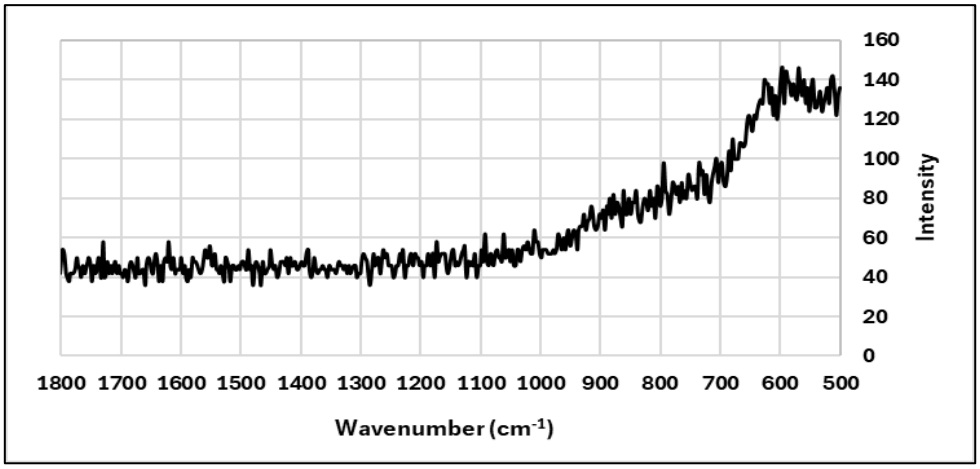

Figure 9. Blank Raman Spectrum

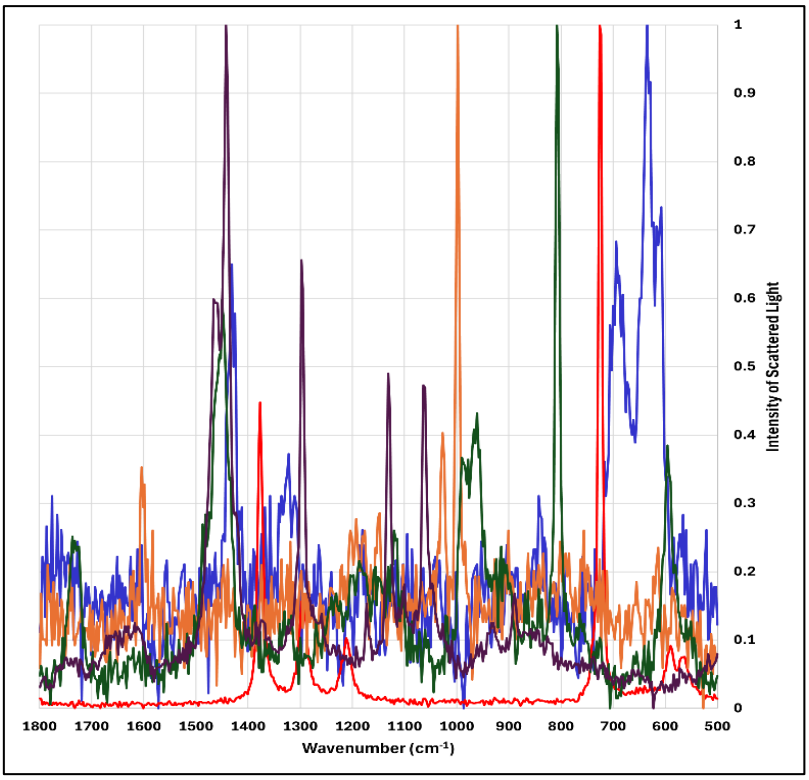

Figure 10. Intensity of Scattered Light-Standarzed

4.2. Machine Learning Models Utilized

Machine learning models comprise an input layer, a hidden layer, and an output layer. The input layer accepts the pretreated Raman spectral data of the samples. The hidden layer calculates the relationship that best fits between the Raman spectral data of the samples and the Raman spectral data of plastics. The output layer produces the final prediction for the classification of the sample. These hidden layers will each be trained using a different machine-learning algorithm written in Python. [14] Each ML algorithm hyperparameter was optimized using a random search with 1000 iterations. The machine learning algorithms that this project utilized include the following (Figs 11 -15).

- AdaBoost improves classification accuracy by iteratively combining weak learners into a strong classifier. It adjusts the focus towards previously misclassified instances by increasing their weights and ensuring subsequent learners pay more attention to these challenging cases. Learners are weighted based on their accuracy, contributing to a collective decision through a weighted vote. [15]

- Decision Tree classifier works by recursively partitioning the feature space into subsets that are increasingly homogeneous with respect to the target variable. At each node, the decision tree selects the feature that best splits the data, aiming to maximize information gain or minimize impurity, until a stopping criterion is met, resulting in a tree structure that can be used to classify new instances by traversing from the root to a leaf node.[8]

- Support Vector Machine (SVM) classifier works by finding the optimal hyperplane that best separates different classes in the input feature space. SVM aims to maximize the margin between classes, with support vectors being the data points closest to the decision boundary, thereby making it effective in handling non-linearly separable data by employing kernel functions to map the input data into higher-dimensional spaces where separation is feasible. [16]

- Random Forest Classifier leverages the power of multiple decision trees for improved accuracy and robustness. It constructs a "forest" of trees, training each tree on a random subset of data and features. This variety prevents overfitting. To make a prediction, the classifier takes input data, runs it through each tree, and the outcome is based on the majority vote across all the trees. [17]

● Gradient Boosting strategically combines weak learners to create a powerful final model. It starts with an initial weak model, identifies prediction errors, and then sequentially adds new trees that focus on correcting those errors. This iterative process of "boosting" the weak learners leads to improved overall accuracy. For classification, the final prediction is determined by a majority vote among the trees in the ensemble. [18]

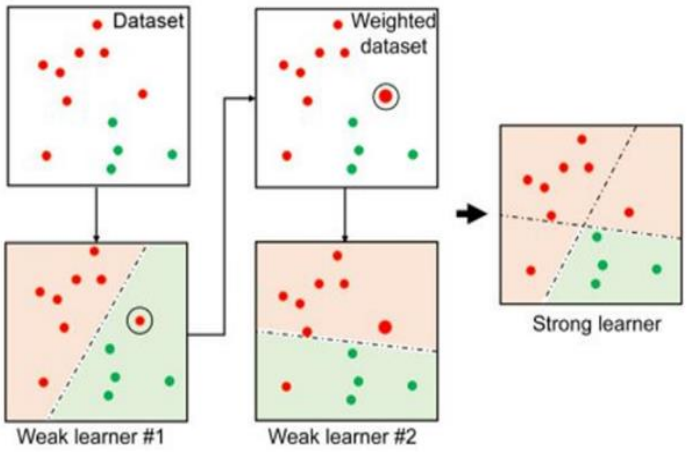

Figure 11: AdaBoost Classifier Credit: Muneer, et al.

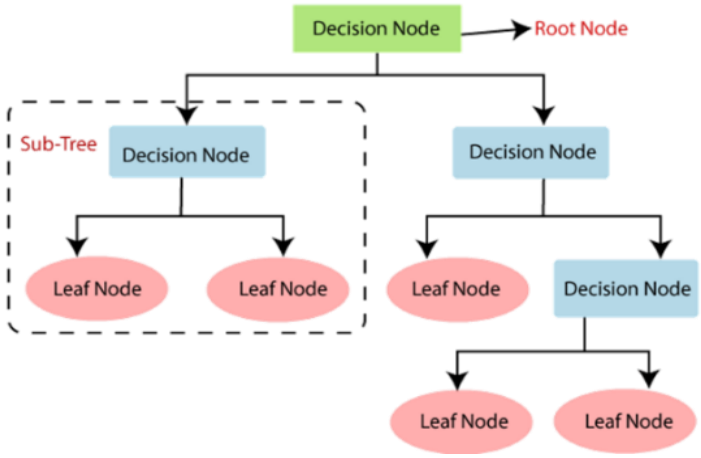

Figure 12: Decision Tree Classifier Credit: Hafeez, et al.

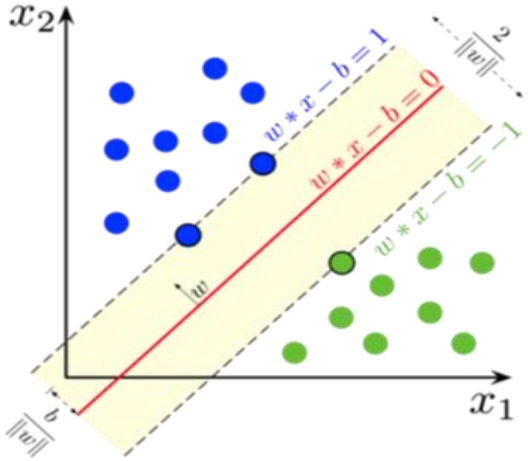

Figure 13: Support Vector Machine Classifier Credit: Larhmam

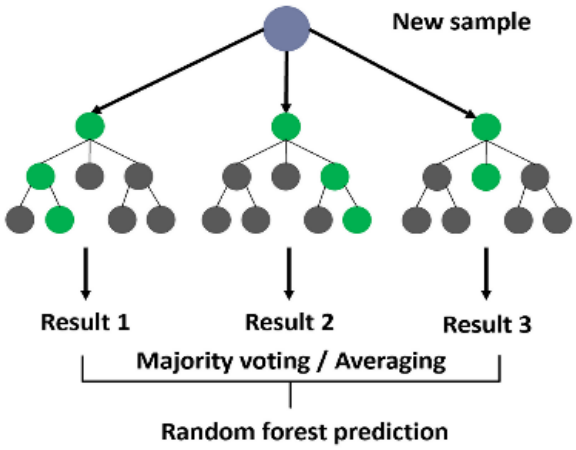

Figure 14: Random Forest Classifier Credit: Roi Yehoshua

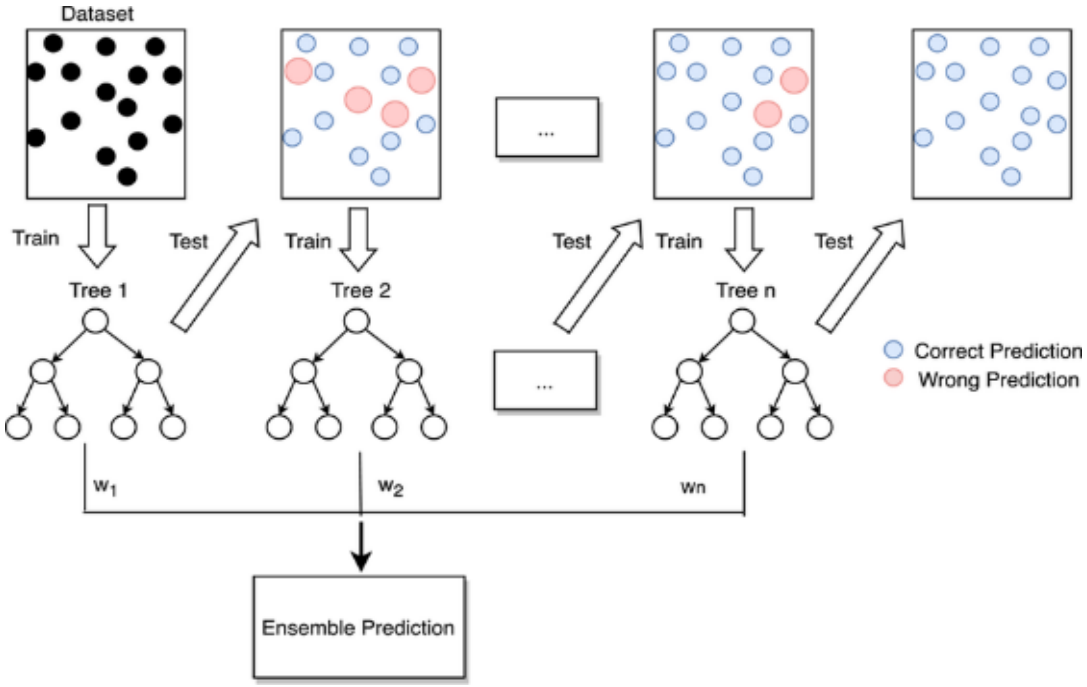

Figure 15: Gradient Boosting Classifier Credit: Zhang, et al.

4.3. Evaluation of ML Models Against Validation Data

The ML models were evaluated against validation data which gives the initial assessment of the accuracy of the model. 80% of the pretreat data from the database at UW-Madison was used to train the ML models. The remaining 20% of the pretreated data was set aside and used as validation data for the ML models. The validation data was given to the models as if it was an end user testing samples as a blind test. The models predict what type of plastic is within the sample or if there is no plastic within the sample. Different types of scores are derived: Precision, Recall, and Accuracy (Fig. 16). To ensure the robustness of the ML models across different data sets, 5-fold cross validation was applied. This technique separates the data into five equal folds. In each iteration, four folds are used for training, and the remaining fold is used for validation. By rotating which fold serves as the test set, every data point contributes to both training and testing at some point, providing a more reliable estimate of the models' accuracy.

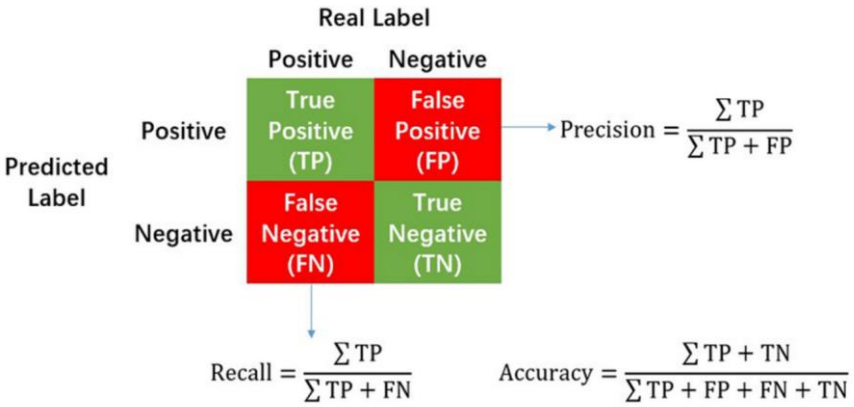

Figure 16: Diagram of how different types of scores are calculated.

True Positive: A result indicating a condition is present when it is present.

True Negative: A result indicating the absence of a condition when it is absent.

False Positive: A result indicating a condition is present when it's absent.

False Negative: A result indicating the absence of a condition when it's present.

Precision is the percentage of predication that are correct. Recall is percentage of predication that are of each specific type. Accuracy is the percentage of all predications that are correct.

Credit Medium (Nicholas Salem)

Each ML model achieved an accuracy of above 96% for detection and classification of nanoplastics in each sample into 6 categories, PE, PTFE, PVC, PS, PMMA, and blank. This segregated testing data was never shown to the ML algorithms during training, demonstrating that the ML algorithms are able to detect and classify nanoplastic particles.

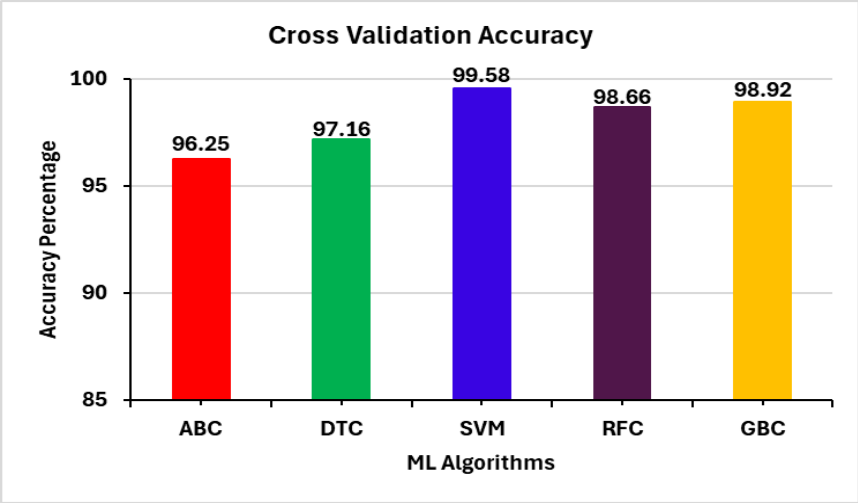

Figure 17: Graph of the cross-validation accuracy of each ML model. Ada Boost Classifier (ABC), Decision Tree Classifier (DTC), Support Vector Machine Classifier (SVMC), Random Forest Classifier (RFC), and Gradient Boosting Classifier (GBC).

4.4. Additional Validation of ML-Models against Actual Samples

To perform additional test of the new ML-model for detection of nanoplastics in any medium, without the need for cumbersome separation, coloration of the plastics, or the use of specialized filters and tedious location of the materials under a microscope, "real-life" samples were made for water, fish, soil, and sand media. The concentrations of the nanoplastics in the respective media is commensurate with literature findings. Specifically, however, the concentration range of nanoplastics in water (1E5

to 20E5 particles/Liter) was created to validate this new, rapid, non-separation Raman data collection against the most recent literature minimum detection limit of 1E5 particles/L. For each sample preparation, scanning electron microscopy (SEM) and energy dispersive spectroscopy (EDS) analyses were carried out on the completed, to-be-evaluated samples, to verify that the nanoplastics were indeed present, and uniform throughout the sample. Figs. 18 below highlights the preparation and analytical results for nanoplastic in water.

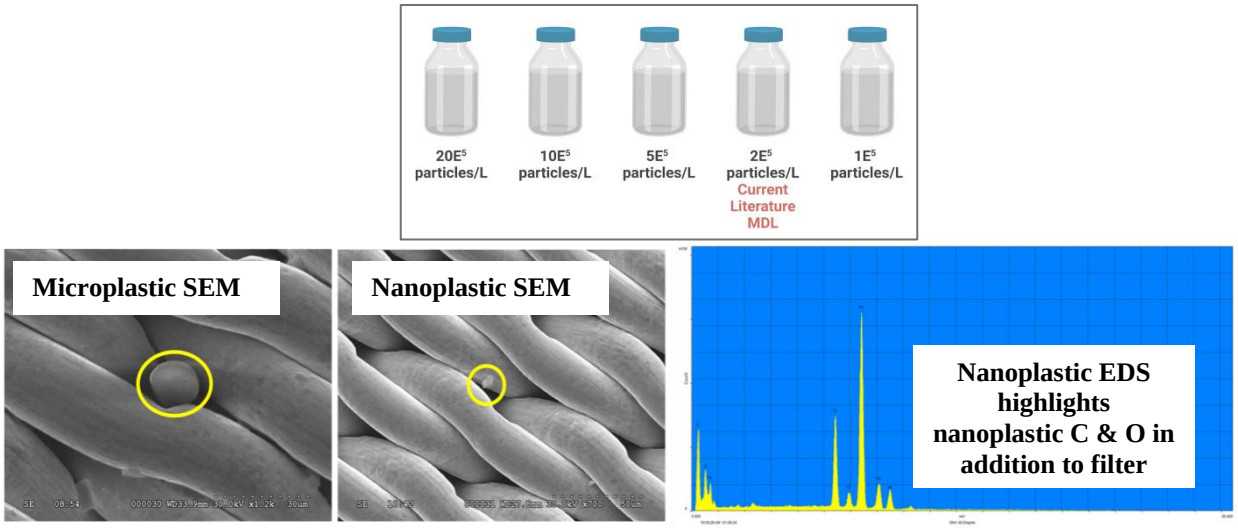

Fig. 18 Nanoplastics in Water Preparation: $1E^5$ to $20E^5$ particles/Liter

To validate the predictive accuracy of the new nanoplastic ML-model, Raman spectra of each of the nanoplastics in water, sand, soil, and aquatic organisms were collected on a PerkinElmer RamanStation 400 (Fig 19 below). The samples were positioned so each Raman spectrum was collected through the open bottle top, alleviating spectral interference from glass.

A preprocessing system was implemented to improve data quality for further analysis. This step aimed to remove unwanted background signals that could potentially mask or distort the true Raman features. Additionally, weak Raman signal extraction and denoising techniques were applied to counteract inherent instrument noise that could interfere with key feature peaks. Finally, the data was scaled to a unified intensity range of 0-1 before being fed into various machine learning algorithms for identification and classification.

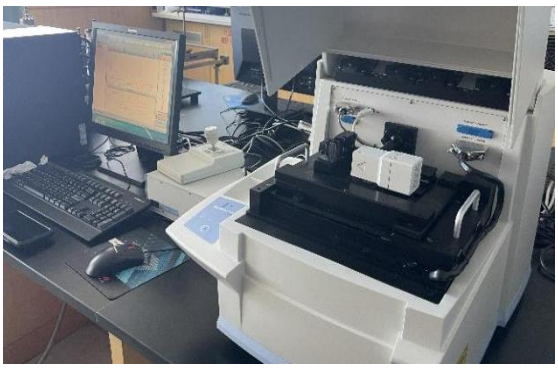

Fig. 19 Collection of the ML-method validation spectra using a PerkinElmer

ML algorithms achieved an average detection accuracy of ~87%. Amongst the evaluated machine learning algorithms, the AdaBoost Classifier emerged as the most effective in handling the detection and classification of nanoplastic particles within water samples. It achieved a commendable detection accuracy of ~70%. Notably, AdaBoost Classifier achieved a classification average of 100%.

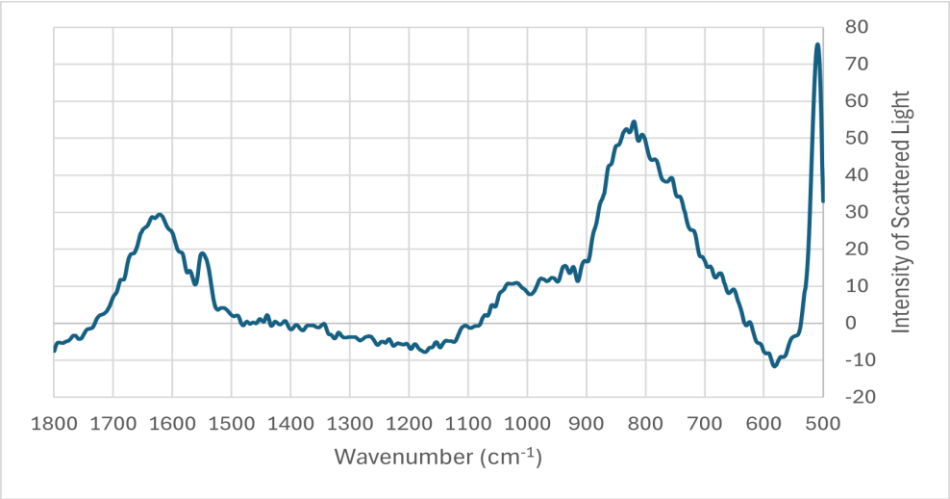

Fig. 20 1E5 Nanoplastic particles/Liter Water

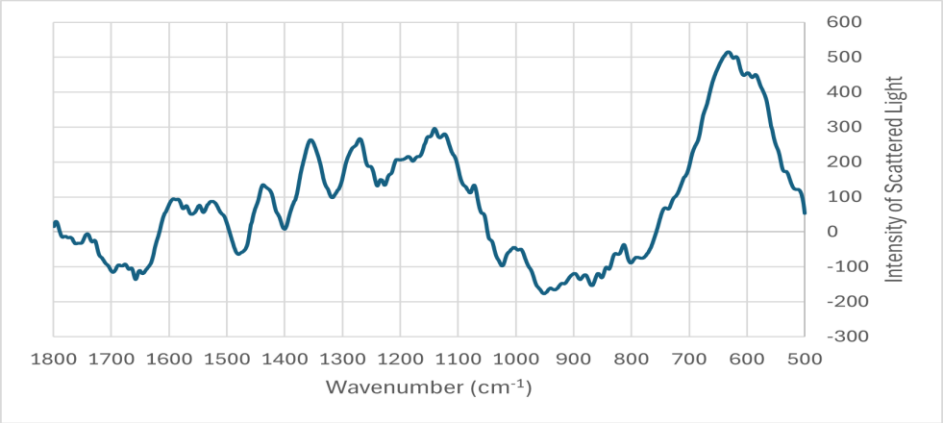

Fig. 21 40mg Nanoplastics/5ml Soil

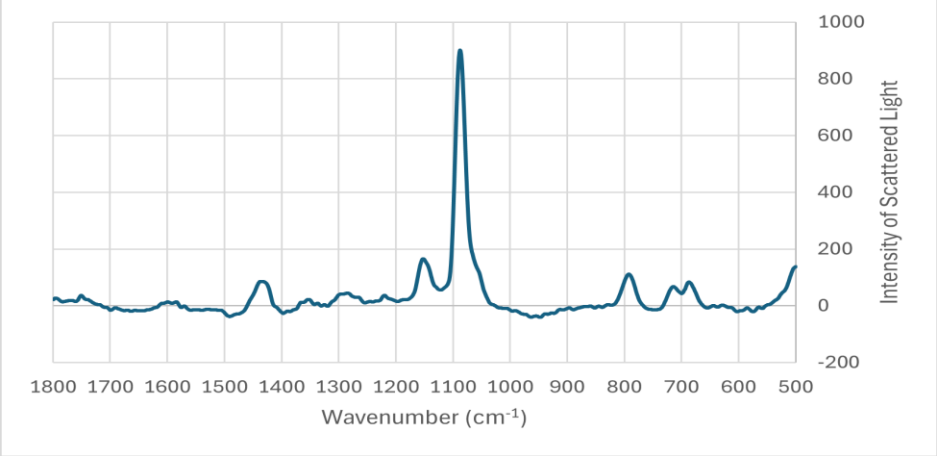

Fig. 22 40mg nanoplastic/5ml Sand

For the samples from other environmental matrices (e.g. soil and sand), the ML algorithms could detect the presence of plastic particles with high accuracy level, but the accuracy for the classification of plastic type was lower due to the similarity of peaks between the different types of plastics as well as inherent instrument noise within these environmental matrices that could interfere with key feature peaks. Clean samples are needed for additional assessment of the accuracy of detection by the ML algorithms.

5. Discussion and Conclusions

This project utilized machine learning algorithms for the detection and classification of nanoplastic particles within environmental and food resource matrices. The key to this approach lies in Raman spectroscopy, a technique that characterizes the samples molecular composition, chemical structure, polymorphism, crystallinity, and intermolecular forces based on their light scattering properties. Leveraging a database of 1600 Raman spectral datasets curated at UW-Madison. This database was pretreated with baseline correction and denoising to characterize the feature peaks of different types of plastic particles. Multiple machine learning algorithms were trained on this data. These machine learning algorithms were then cross validated by training with each part of the data and evaluated against each partition of the data. Then, these machine learning algorithms were evaluated against newly collected Raman spectral data from real life samples from different environmental and food resource matrices, including water, soil, sand, and fish, without cumbersome separation techniques. By evaluating their performance in both detecting and classifying the nanoplastic particles within these samples, it confirms that using machine learning algorithms allows for the accurate detection and classification of samples.

This project leveraged machine learning algorithms and Raman spectroscopy for the accurate detection and classification of nanoplastic particles in environmental and food resources. The successful application of these algorithms on new samples from various matrices such as water, soil, and food demonstrate the potential for scaling up plastic detection and classification processes through automated means. To date, this new and fast Raman-based analysis can simply and easily, without sample preparation, improve the detection of nanoplastics in water, to as little as 1E5 particles/L, which is a 2-fold improvement over the most recent Raman-SRS Microscopy method. The methods of leveraging machine learning algorithms and Raman spectroscopy not only achieved remarkable accuracy levels in the laboratory settings but also demonstrated its efficacy in real-world environmental samples.

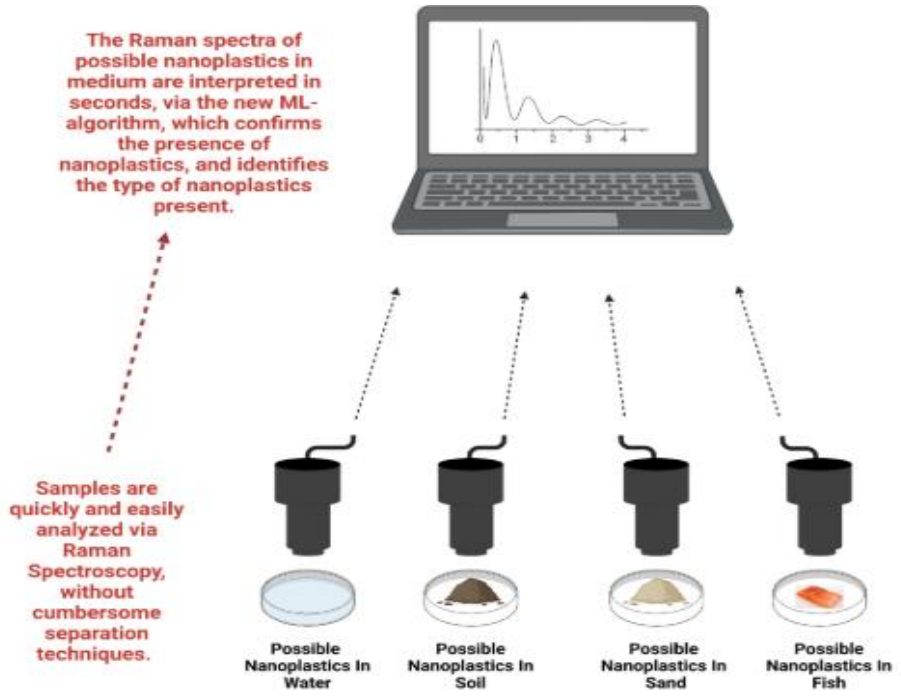

Fig. 23 The New, Raman Spectroscopy-based Rapid and Simple Analysis of Possible Nanoplastics in Typical Media Using the New ML-Algorithm

6. Future Work

Future research should increase training database size, increase types of plastic detectable by machine learning algorithms, and increase clean samples to evaluate machine learning algorithms.

References

[1] N. Ali, J. Katsouli, E. L. Marczylo, T. W. Gant, S. Wright, and J. Bernardino De La Serna, "The potential impacts of micro-and-nano plastics on various organ systems in humans," *eBioMedicine*, vol. 99, p. 104901, Jan. 2024, doi: 10.1016/j.ebiom.2023.104901.

[2] S. Allen *et al.*, "Atmospheric transport and deposition of microplastics in a remote mountain catchment," *Nat. Geosci.*, vol. 12, no. 5, pp. 339–344, May 2019, doi: 10.1038/s41561-019-0335-5.

[3] A. K. Baldwin, S. R. Corsi, and S. A. Mason, "Plastic Debris in 29 Great Lakes Tributaries: Relations to Watershed Attributes and Hydrology," *Environ. Sci. Technol.*, vol. 50, no. 19, pp. 10377–10385, Oct. 2016, doi: 10.1021/acs.est.6b02917.

[4] F. Bessa *et al.*, "Microplastics in gentoo penguins from the Antarctic region," *Sci. Rep.*, vol. 9, no. 1, p. 14191, Oct. 2019, doi: 10.1038/s41598-019-50621-2.

[5] H. Cai, M. Chen, Q. Chen, F. Du, J. Liu, and H. Shi, "Microplastic quantification affected by structure and pore size of filters," *Chemosphere*, vol. 257, p. 127198, Oct. 2020, doi: 10.1016/j.chemosphere.2020.127198.

[6] S. Lin *et al.*, "Metabolomics Reveal Nanoplastic-Induced Mitochondrial Damage in Human Liver and Lung Cells," *Environ. Sci. Technol.*, vol. 56, no. 17, pp. 12483–12493, Sep. 2022, doi: 10.1021/acs.est.2c03980.

[7] D. M. Mitrano, P. Wick, and B. Nowack, "Placing nanoplastics in the context of global plastic pollution," *Nat. Nanotechnol.*, vol. 16, no. 5, pp. 491–500, May 2021, doi: 10.1038/s41565-021-00888-2.

[8] M. A. Hafeez, M. Rashid, H. Tariq, Z. U. Abideen, S. S. Alotaibi, and M. H. Sinky, "Performance Improvement of Decision Tree: A Robust Classifier Using Tabu Search Algorithm," *Appl. Sci.*, vol. 11, no. 15, p. 6728, Jul. 2021, doi: 10.3390/app11156728.

[9] A. Käppler *et al.*, "Analysis of environmental microplastics by vibrational microspectroscopy: FTIR, Raman or both?," *Anal. Bioanal. Chem.*, vol. 408, no. 29, pp. 8377–8391, Nov. 2016, doi: 10.1007/s00216-016-9956-3.

[10] L. Xie *et al.*, "Automatic Identification of Individual Nanoplastics by Raman Spectroscopy Based on Machine Learning," *Environ. Sci. Technol.*, p. acs.est.3c03210, Jul. 2023, doi: 10.1021/acs.est.3c03210.

[11] F. Yu and X. Hu, "Machine learning may accelerate the recognition and control of microplastic pollution: Future prospects," *J. Hazard. Mater.*, vol. 432, p. 128730, Jun. 2022, doi: 10.1016/j.jhazmat.2022.128730.

[12] A. Tarafdar, S.-H. Choi, and J.-H. Kwon, "Differential staining lowers the false positive detection in a novel volumetric measurement technique of microplastics," *J. Hazard. Mater.*, vol. 432, p. 128755, Jun. 2022, doi: 10.1016/j.jhazmat.2022.128755.

[13] N. Qian *et al.*, "Rapid single-particle chemical imaging of nanoplastics by SRS microscopy," *Proc. Natl. Acad. Sci.*, vol. 121, no. 3, p. e2300582121, Jan. 2024, doi: 10.1073/pnas.2300582121.

[14] A. Muneer and S. M. Fati, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter," *Future Internet*, vol. 12, no. 11, p. 187, Oct. 2020, doi: 10.3390/fi12110187.

[15] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[16] V. N. Vapnik, "The Support Vector method," in *Artificial Neural Networks — ICANN'97*, vol. 1327, W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, Eds., in Lecture Notes in Computer Science, vol. 1327. , Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 261–271. doi: 10.1007/BFb0020166.

[17] Tin Kam Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998, doi: 10.1109/34.709601.

[18] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. in Springer series in statistics. New York, NY: Springer, 2009.