Random Forest Microplastic classification using spectral subsamples of FT-IR hyperspectral images

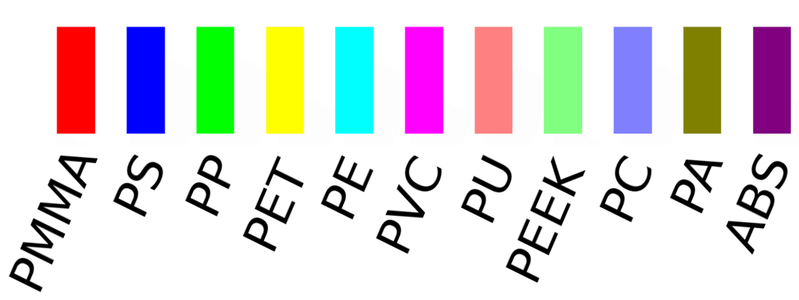
Jordi Valls-Conesa[1,2], Dominik Winterauer[1], Niels Kröger-Lui[1], Sascha Roth[1], Stephan Lüttjohann[1], Roland Harig[1], Jes Vollertsen[2]

[1]Bruker Optics GmbH & Co. KG, Rudolf-Plank-Str. 27, 76275 Ettlingen, Germany

[2]Department of the Built Environment, Aalborg University, Thomas Manns Vej 23, 9220 Aalborg Øst, Denmark

Introduction

Microplastics (MPs) are microscopic polymer particles present almost everywhere. To accurately understand the effects of these particles in ecosystems and organisms the composition, physical structure and abundance of them needs to be accurately analysed. **Micro-Fourier Transform Infrared** (μFT-IR) using Focal Plane Array (FPA) facilitates the MP spectra recollection by simultaneously recollecting full chemical images. The classification model implemented on the recollected spectra is **Random Decision Forest**[1] (RDF), a supervised Machine Learning classification algorithm using an ensemble of Decision Trees (tree-like structures).



Materials and Methods

RDF is trained with spectra from 11 different MP polymer types. Pure-type MP are deposited over an Aluminium oxide anodisc, and the spectra are selected from the μFTIR image using **Fast Background Correction and Identification**[2] (FBCI). The average spectra of each polymer type selected are shown in *Fig.1*.

The average time to train the RDF with an average of 1000 spectra per class in the training data set is 45 *s*.

The RDF validation data set determines the accuracy of the model and provides a visual interpretation. A testing **Ground Truth** (GT) procedural image (*Fig.2 Left*) provides validation information.

For a fast classification in QCL-based microscopes range, the most relevant signal wavenumbers (channels) are chosen using a decision tree and calibrated using **Area Under the Curve** (AUC) validation. The selected channels work in conjunction with a baseline wavenumber. The channels selected are represented in *Fig.1*.

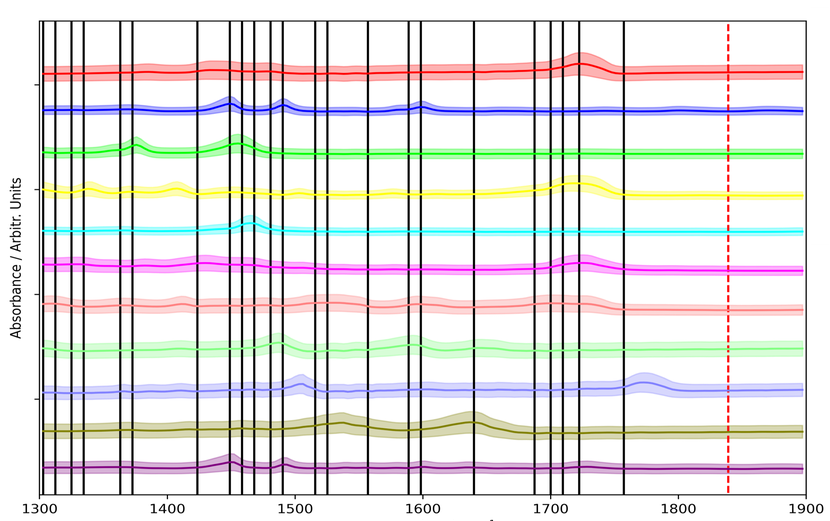


***Figure 1**: Channel-wise average and standard deviation of the spectra for each MP type used in the training data set. The black lines correspond to the most relevant channels selected by a DT classifier. The dashed red line represents the channel chosen for baseline correction.*

Sample analysis

The RDF prediction is depicted as a colour-map and the relevant label is selected with a winner-takes-all representation (*Fig.2 Right*).

An alternative way to display the accuracy of an RDF is by means of a **Confusion Matrix** (CM), a MP class cross-correlation visualization.

Environmental samples (*Fig.3*) are a common application scenario to demonstrate a MP classifier's fitness for purpose. The accuracy of a MP classifier cannot be assessed due to the lack of a GT, validation data provides a trusted comparison, however, it does not accurately depict all details from environmental samples[3].

The required time to for *Fig.3* RDF prediction, assuming 12 polymer classes and 15 channels selected, is approximately 30 *s*.

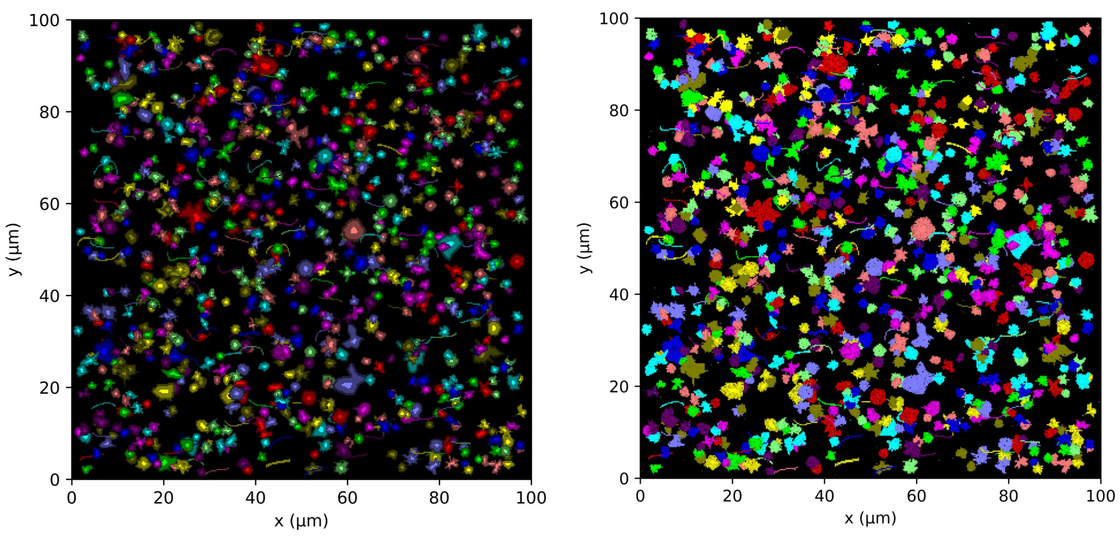


***Figure 2**: **Left**) Simulated test image created using MP spectra and environmental background spectra from the test data set. The color brightness represents the thickness, which decreases towards the particles edges, approximating the geometry of real MP particles. **Right**) Visualization of the RDF output for MP classification applied to the IR hyperspectral image from artificial sample shown on the left.*

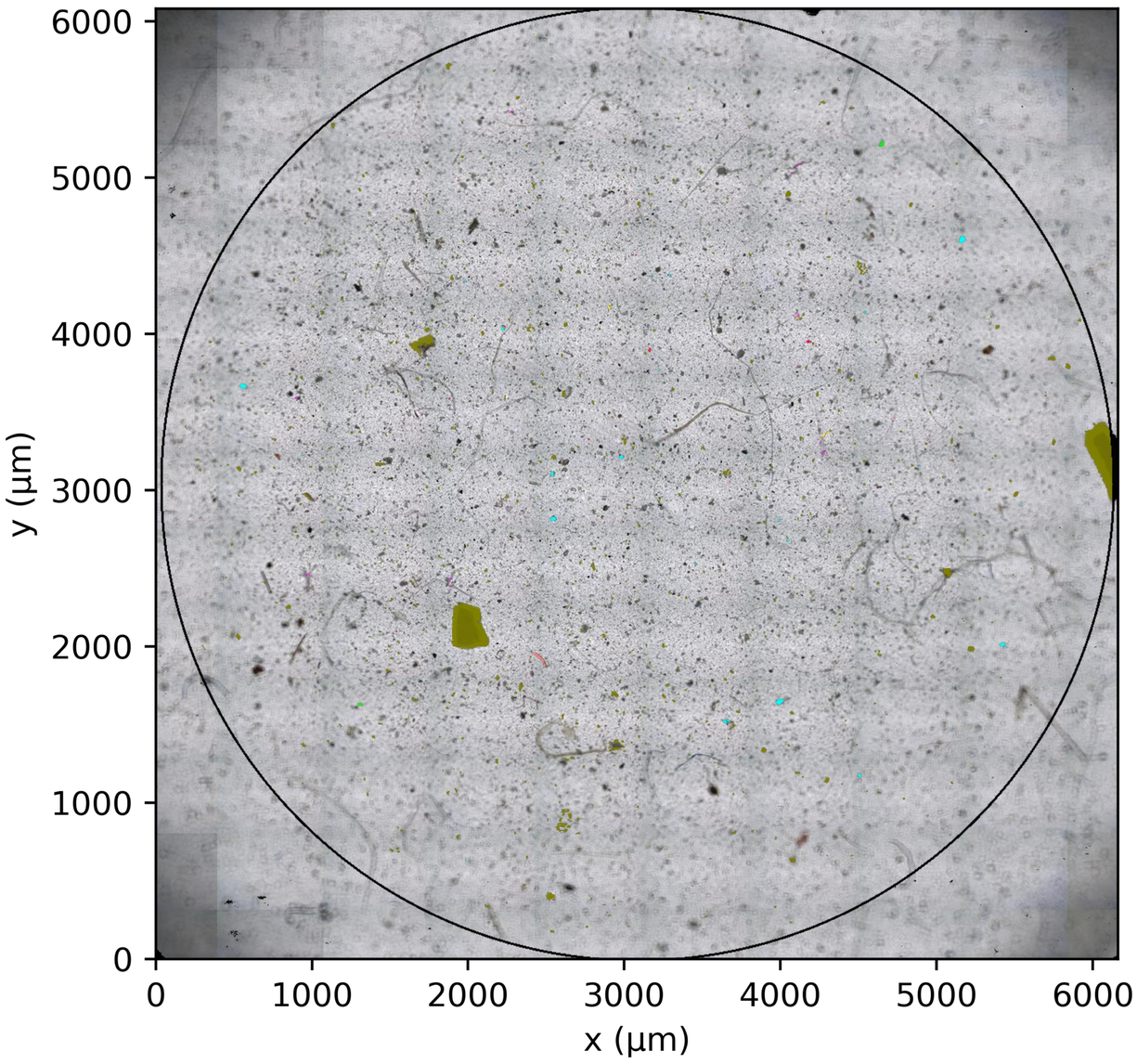


***Figure 3**: RDF classification over a sea-salt sample. The background and non polymer regions are invisible, showing the visible image of the sample. The RDF model is trained with 15 channels and a subtraction wavenumber at 1839 cm^{-1}.*

Conclusion

Training and test data sets have been created based on purpose-made pure-type MP samples using a spectra selection methodology applicable to other spectra classification problems of similar nature. Validation is done with AUC and CM to fit all possible parameters into the optimal version of the RDF. The classifier results over sea-salt and river sediment samples show that the analysis time typical from MP classification techniques can be reduced to below one minute.

The results achieved show that a RDF model operating with highly discriminative single channels is an effective tool to classify microplastic hyperspectral data.

Contact

Jordi.Valls-Conesa@bruker.com

Acknowledgements

The authors thank Wessling for the collaboration by sending grinded MP for the training of the RF dataset. This project has received funding from the European Union's Horizon 2020, under the Marie Skłodowska-Curie Innovative Training Networks (MSCA-ITN-ETN) grant agreement 860775.

Funded by the European Union

References

[1] *Benedikt Hufnagl, et al*. **2019**. *A methodology for the fast identification and monitoring of microplastics in environmental samples using random decision forest classifiers*. **Anal. Methods**.

[2] R. Harig, *et al*. **2001**. *Toxic cloud imaging by infrared spectrometry: A scanning μFTIR system for identification and visualization*. **Field Analytical Chemistry & Technology**.

[3] S. Primpke, *et al*. **2017**. *An automated approach for microplastics analysis using focal plane array (FPA) FTIR microscopy and image analysis*. **Anal. Methods**.