TRAINING AND EVALUATING MACHINE LEARNING ALGORITHM FOR MICROPLASTICS CLASSIFICATION THROUGH RAMAN SPECTROSCOPY

Authors: P Subanaveen, Pooja Bhoomigha S
SRM University AP, Amaravathi, Andhra Pradesh, India.

Research Supervisor: Dr Rajapndiyan Paneerselvam,
Raman Research Laboratory (RaRe Lab), Department of Chemistry, SRM University – AP, Andhra Pradesh, India .

.

ABSTRACT: For the detection of chemical elements and compounds with their own spectral fingerprints, **Raman spectroscopy** has been an excellent tool. However, it can be difficult to classify the Raman spectrum in a precise way, especially when dealing with overestimated data sets and complex parameter spaces. **ADASYN**, a synthetic oversampling strategy, balances the dataset to address the issue of class imbalance. **Random Forest Classifier**—a well-liked option for multiclass classification tasks—we utilise **GridSearchCV**. Tested on a test dataset, the modified Random Forest Classifier shows increased accuracy along with detailed insights from a confusion matrix and a classification report. The finished model can correctly identify elements from Raman spectra. This research contributes to the field of Raman spectroscopy by offering a robust solution for element identification,.It showcases the effectiveness of machine learning techniques and data preprocessing in enhancing the accuracy and reliability of element classification in Raman spectra.

KEYWORDS: Microplastic, Raman Spectroscopy, Machine Learning, Random Forest, Spectral analysis, Indentification.

INTRODUCTION

Microplastics are basically tiny particles of plastic with size range of nanometers to a few millimetres in diameter have arisen as a major global environmental concern these microscopic persistent contaminants are a serious hazard to human beings health and wildlife because they have permeated all kinds of habitats from terrestrial landscapes to marine settings as a result developing efficient techniques for microplastic detection identification analysis has emerged as a top priority in fields such as environmental preservation and scientific research.

The ability of raman spectroscopy to reveal precise chemical and structural details about a wide range of materials has made it a potent and non-destructive analytical tool that has been at the forefront of microplastics investigation raman spectroscopy provides the ability to characterise microplastics and differentiate them among other compounds in complicated environmental samples by producing spectra that function as chemical fingerprints but using raman spectroscopy for analysing microplastics has its own set of difficulties mostly differences in tools data collection methods the properties of microplastics spectra themselves study the microplastic mixtures where many forms of microplastics coexist and conventional analytical methods are insufficient is a major barrier in this sector the spectral data obtained from mixes can be quite complicated necessitating lengthy preprocessing procedures like baseline correction smoothing and filtering.

In order to make the spectra readable although beneficial these preprocessing procedures may need more man power and it can be time-consuming which may restrict scalability and usefulness of microplastics study this research report presents a novel method that combines machine learning approaches with high-resolution full-window raman spectra to overcome the difficulties associated with microplastics characterization the main novelty lies in using machine learning to overcome the requirement for complex spectrum preprocessing.

In this paper we show how to build machine learning classification models capable of recognising analysing the chemical components of microplastics in mixes without any spectral preprocessing this method has a substantial benefit in terms of simplifying microplastic application and examination procedure to generate this type of model we use an ensemble of open-source machine learning methods including random forest these models routinely achieve classification accuracies surpassing 95 percent when trained on high-resolution spectra even when spectral data is reduced to 1 ,2, 4, or 8 cm $^{-1}$ spacings in raman shift importantly.

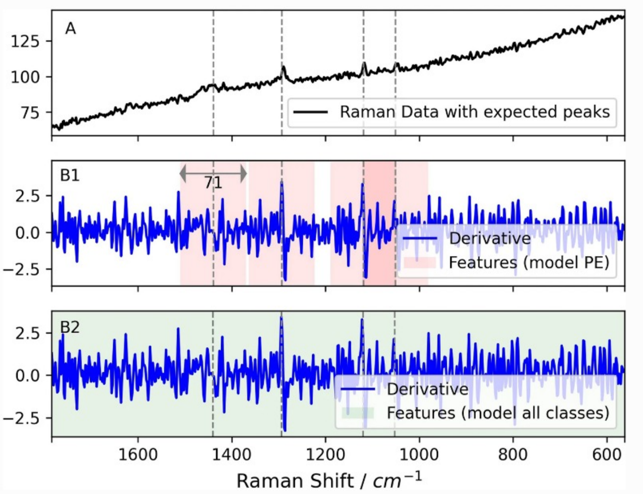
These models show resilience even under non-ideal situations such as low spectroscopic sampling rates or particles located outside the lasers focus plane this study marks a significant advancement in microplastics analysis and environmental science fields not only it presents a transformational answer to the spectral preprocessing difficulty but it also demonstrates the adaptability and durability of machine learning algorithms

in reference to the study of the environment this method offers up new avenues for classification models development that can handle various spectrometer setups and fulfil the changing demands of environmental researcher
.

1 DATA, SAMPLING AND METHODOLOGY

1.1 DATA

A set of unique datasets has been acquired by utilizing Raman spectroscopy. The initial dataset, named Slopp Spectral Libraries of Plastic Particles, encompasses roughly 148 spectral references that specifically relate to plastics. Conversely, the second dataset called Slopp-E Spectral Library of Plastic Particles Aged in the Environment contains 113 spectral data points. By employing a combination approach involving our internal Raman laboratory and leveraging the BW Tek portable Raman spectrometer alongside the aforementioned Slopp dataset shown in fig(1), we have achieved successful identification results for specific components. Furthermore, descriptive tags have been incorporated to facilitate future applicability in our modeling and coding endeavors.



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)

Figure 1: example how the peaks are extracted(PE)

1.2 SAMPLING:Adaptive Synthetic Sampling (ADASYN)

When it comes to microplastics classification datasets, class imbalance is a common problem. It frequently happens when the number of samples in one class—the presence of microplastics, for example—is significantly greater than the number of samples in another class—the lack of microplastics. An important barrier to the creation of precise machine learning models is this class imbalance. Inadvertently developing a bias in favour of the majority class might cause the model to perform less well than ideal when it comes to identifying the existence of microplastics.

We used the Adaptive Synthetic Sampling (ADASYN) algorithm, a sophisticated synthetic oversampling method, in our study to try and solve this class imbalance problem. In order to perform its job, ADASYN creates artificial examples of the minority class—in this case, microplastic-containing samples.

Our research's strategic use of ADASYN aims to improve our machine learning model's robustness and efficacy while reducing the negative effects of class imbalance. This all-encompassing strategy is essential to guarantee that the unequal class distribution does not impair our model's predictive power. Through a methodical approach to this problem, we hope to lower the possibility of false positives and negatives, and to improve the dependability of our model and potential use in the precise categorization of microplastics.
.

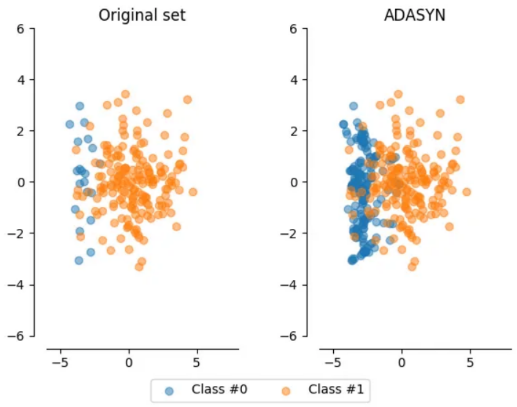


Figure 2: Imbalanced data after using ADASYN

.
.

2 MODEL

2.1 Random Forest : to reduce overfitting of data

this study focuses on the development of a machine learning model specifically utilizing the random forest classifier to address the growing concern regarding microplastic pollution microplastics are tiny plastic particles ranging from nanometers to a few millimeters in size and their pervasive presence poses significant challenges for environmental preservation and scientific investigation these minuscule contaminants have infiltrated various ecosystems including terrestrial landscapes and marine environments posing serious threats to human health and wildlife welfare consequently it is crucial to develop effective techniques for detecting identifying and analyzing microplastics the random forest classifier was selected as the cornerstone of this model due to its exceptional capability in managing complex data sets with imbalances within them

.

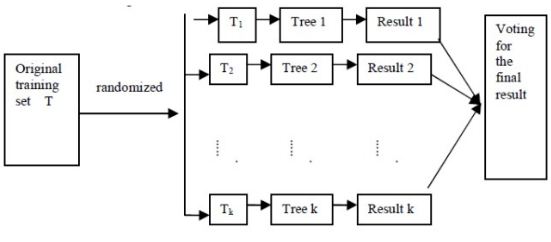


Fig. 1. Random forest schematic

Figure 3: Enter Caption

- decision trees the fundamental building blocks of random forest are decision trees which aim to divide the data into subsets that maximize purity or minimize impurity within each subset the objective of decision trees is to identify optimal splits nodes using different criteria
- bootstrapping random forest creates multiple decision trees by utilizing bootstrapped subsets of the data introducing randomness into the process by training each tree on a unique subset diversity is added to the model
- in the random forest model for creating each decision, tree random feature selection is used. it is a very important step which helps in counteracting overfitting and it aslo decreases the correlation between

trees this feature selection technique which plays a very important part in improving the efficiency of this random forest model

- voting or averaging in classification tasks random forest employs majority voting to combine the predictions from multiple decision trees for regression tasks it takes an average of each trees prediction
- variance-bias tradeoff the collection of decision trees in random forest aims to strike a balance between bias and variance in the model while averaging across multiple trees helps reduce variance individual decision trees can be susceptible to overfitting due to their high variance nature

3 Model Training and Performance Evaluation: Methodology and Metrics

3.1 Data preprocessing and Data Cleaning

In this study, the dataset contains Raman shift and intensity features which are important for the microplastic classifications. Prior to analysis, the dataset is expected to be thoroughly cleaned to exclude any outliers or missing values or irrelevant information, as clean data is essential for the accuracy of the model. The StandardScaler is used to normalize the features, which is essential for the performance of the machine learning model. Class imbalance is a common issue in microplastic classification datasets, and ADASYN is used to overcome this problem by oversampling the minority class. This technique helps to reduce the impact of the class imbalance on the model training.

3.2 Machine learning Model

The Random Forest Classifier is our main machine learning algorithm. The Random Forest is an ensemble learning method that combines several decision trees to generate predictions. This method is well suited for the use of complex, highly dimensional data for the use of microplastic classifying.

3.3 Hyperparameter tuning

In order to enhance the performance of the Random Forest model, we utilize hyperparameter tuning. The GridSearchCV technique is employed to explore the optimal combination of hyperparameters, which include the number of estimators, maximum depth of the trees, and minimum samples required to split a node. After selecting the hyperparameters through grid search, the Random Forest model is trained on the training data to learn the underlying patterns in the data.

3.4 ModelPerformance and Evaluation

To evaluate the effectiveness of the model in microplastic classification, we use a range of performance metrics such as accuracy, precision, recall, and F1-score. The **accuracy** measures the proportion of correctly classified samples in the testing set, while **precision** quantifies the ratio of true positive predictions to the total positive predictions. **Recall** assesses the proportion of true positive predictions to the actual positive samples, and the **F1-score** provides a balance between precision and recall. To assess the model's performance, we conduct an extensive evaluation on the testing set and calculate the accuracy, precision, recall, and F1-score. Additionally, a classification report is generated to provide a comprehensive overview of these metrics for each class in the dataset. Finally, a confusion matrix is constructed to visualize the true positives, true negatives, false positives, and false negatives.

$$Accuracy = TP + TN/TP + TN + FP + FN \tag{1}$$

$$Precision = TP/TP + FP \tag{2}$$

$$Recall = TP/TP + FN \tag{3}$$

$$F1Score = 2 * Precision * Recall/Precision + Recall \tag{4}$$

By following this rigorous methodology and using a diverse set of performance metrics, we ensure a comprehensive evaluation of the Random Forest classifier for microplastic classification using Raman spectroscopy data. This enables us to draw meaningful conclusions about the model's effectiveness and its potential implications for environmental research and preservation.

4 Results and Significance: Enhancing Model Performance with ADASYN

The study reveals a significant enhancement in the performance of a classsifaction model using the ADASYN algorithm to rectify class imbalance. Key performance metrics improved, including accuracy, precision, recall, and F1 score. The model's accuracy increased, resulting in higher correct classification rates. Precision reduced false positives, reducing anxiety and follow-up procedures. Recall improved, reducing the risk of wrong prediction. The F1 score balanced precision and recall, ensuring a balanced model. The study's results were validated through 5-fold cross-validation, confirming the model's reliability across various data subsets. These results have implications for clinical practice, providing a reliable tool for microplastic identification

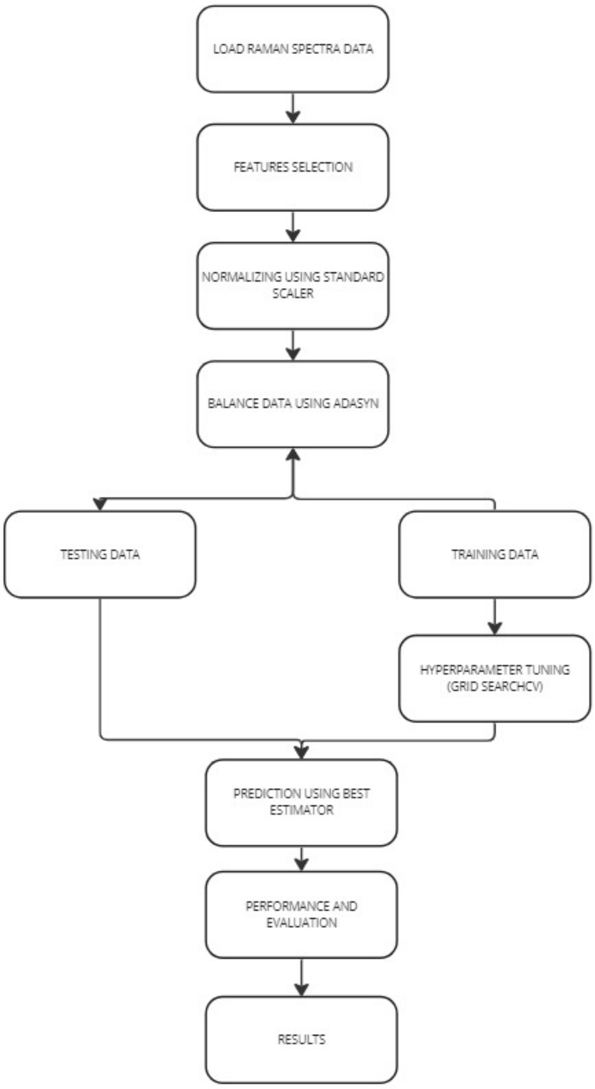

LOAD RAMAN SPECTRA DATA
FEATURES SELECTION
NORMALIZING USING STANDARD SCALER
BALANCE DATA USING ADASYN
TESTING DATA
TRAINING DATA
HYPERPARAMETER TUNING (GRID SEARCHCV)
PREDICTION USING BEST ESTIMATOR
PERFORMANCE AND EVALUATION
RESULTS

Figure 4: example how the peaks are extracted(PE)

5 Conclusion

Raman Spectroscopy is effective when combined with machine learning classification models such as Random Forest Where we used data sets from SLoPP and data from Raman Lab using BW Tek portable Raman spectrometer to get an accuracy of around 80 percent . This helps the model analyze real-life problems or daily situations to identify which microplastic it is among PS, PC, PVC, PE

5.1 Limitations and Recommendations for further work

For this research paper, it is mandatory to consider people from other fields who are not so comfortable with working in a coding environment so we can hide the implementation and keep it more user-friendly by creating software or app or a website.

References

Lei, B., Bissonnette, J. R., Hogan, Ú. E., Bec, A. E., Feng, X., Smith, R. D. (2022). Customizable machine-learning models for rapid microplastic identification using Raman microscopy. Analytical Chemistry, 94(49), 17011-17019.

Liu, Y., Wang, Y., Zhang, J. (2012). New machine learning algorithm: Random forest. In Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3 (pp. 246-252). Springer Berlin Heidelberg.

Weber, F., Zinnen, A., Kerpen, J. (2023). Development of a machine learning-based method for the analysis of microplastics in environmental samples using µ-Raman spectroscopy. Microplastics and Nanoplastics, 3(1), 9.

Chaczko, Z., Wajs-Chaczko, P., Tien, D., Haidar, Y. (2019, July). Detection of microplastics using machine learning. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC) (pp. 1-8). IEEE.

Yan, X., Cao, Z., Murphy, A., Qiao, Y. (2022). An ensemble machine learning method for microplastics identification with FTIR spectrum. Journal of Environmental Chemical Engineering, 10(4), 108130.

Hudspeth, E. D., Cleveland, D., Batchler, K. L., Nguyen, P. A., Feaser, T. L., Quattrochi, L. E., ... Lombardi, D. (2006). Teaching Raman spectroscopy in both the undergraduate classroom and the laboratory with a portable Raman instrument. Spectroscopy letters, 39(1), 99-115.

Zada, L., Leslie, H. A., Vethaak, A. D., Tinnevelt, G. H., Jansen, J. J., de Boer, J. F., Ariese, F. (2018). Fast microplastics identification with stimulated Raman scattering microscopy. Journal of Raman spectroscopy, 49(7), 1136-1144.

Yang, S. J., Feng, W. W., Wang, Q., Cai, Z. Q., Liu, Q. Y., Hou, Y. B., Zhang, Q. Q. (2020, October). Rapid identification of microplastic using portable Raman system and extra trees algorithm. In Real-time Photonic Measurements, Data Management, and Processing V (Vol. 11555, pp. 70-77). SPIE.

Lin, J. Y., Liu, H. T., Zhang, J. (2022). Recent advances in the application of machine learning methods to improve identification of the microplastics in environment. Chemosphere, 136092.