Journal of Environmental Chemical Engineering 10 (2022) 108130

Contents lists available at ScienceDirect

Journal of Environmental Chemical Engineering

journal homepage: www.elsevier.com/locate/jece

An ensemble machine learning method for microplastics identification with FTIR spectrum

Xinyu Yan [a,b], Zhi Cao [c], Alan Murphy [c], Yuansong Qiao [a,*]

[a] Software Research Institute, Technological University of the Shannon: Midlands Midwest, Ireland
[b] Luoyang Institute of Science and Technology, China
[c] Materials Research Institute, Technological University of the Shannon: Midlands Midwest, Ireland

ARTICLE INFO

Editor: Luigi Rizzo

Keywords:
Microplastics identification
Machine learning
FTIR
Deep learning
Data pre-processing

ABSTRACT

Microplastics (MPs) (size < 5 mm) marine pollution have been investigated and monitored by many researchers and found in many coasts around the world. These toxic chemicals make their way into human diet through food chain when aquatic organisms ingest MPs. Attenuated Total Reflection Fourier transform infrared spectroscopy (ATR-FTIR) is a very effective method to detect MPs. To provide the automatic detecting method for MPs, Numerous studies have proposed Machine Learning (ML) based methods, such as Support Vector Machines, K-Nearest Neighbours, and Random Forests, for identification and classification of MPs through using the ATR-FTIR data. The evaluations of these ML based methods primarily focus on the average scores across all types of MPs. However, the existing FTIR datasets are normally imbalanced. Furthermore, some MPs contain the identical functional group, and some MPs may be fouled or contaminated, which will reduce the quality of FTIR data samples (e.g. lacking of peaks or creating noises). These factors will interfere the ML classification algorithms and cause the algorithms to perform differently while identifying different MPs. Hence, this work proposes an ensemble learning algorithm to exploit the advantage of different ML algorithms based on a systematic evaluation of the existing ML based MP identification approaches. A neural network is employed to fuse the outputs of chosen ML algorithms to improve the overall metrics. The evaluation results show that the proposed algorithm outperforms existing single ML based approaches.

1. Introduction

Plastic products consumption has been increasing exponentially [1]. There are various sources of plastic pollution in the marine environment, including marine and coastal ships and sailboats, commercial fishing operations, and land-based sources including trash and manufacturing waste [2]. Microplastics (MPs) arising from these specific activities and products has been the most prevalent forms of manmade pollution in the marine environment [3]. Due to the increased plastics consumption and poor waste management, this form of contamination is now not only widely spread, but also is eaten by a wide range of animals from various environments and feeding approaches, including pelagic and benthic fish, filter-feeding fish, and benthos [4,5]. MPs are ingested by fish and other aquatic creatures, rendering them a potential vector of hazardous substances into the human food [6,7]. As there are normally substantial concentrations of MPs (both natural and manufactured fibres) around urbanized coastlines, More research is conducted in these locations to measure and estimate the quantity of MPs absorbed by marine organisms and the ecological importance of this phenomena. Identifying MPs has become increasingly crucial [8,9].

Chemical examination of environmental materials is often confined to general properties, such as polymer abundance. To evaluate the particle size distribution of MPs, Attenuated Total Reflection Fourier Transform Infrared Spectroscopy (ATR-FTIR) technology [10], which can also be coupled with microscopes named Micro Fourier Transform Infrared Spectroscopy (μ-FTIR) [5], is becoming noticeably popular

Abbreviations: MP, Microplastic; ATR, Attenuated Total Reflection; FTIR, Fourier transform infrared spectroscopy; ML, Machine Learning; SVM, Support Vector Machines; KNN, K-Nearest Neighbours; RF, Random Forests; PLSDA, Partial Least Squares Discriminant Analysis; SIMCA, Soft Independent Modelling of Class Analogies; MLP, Multilayer Perceptron; ANN, Artificial Neural Network; PCA, Principal Component Analysis; LDA, Linear Discriminant Analysis; EPR, Ethylene propylene rubber.

* Corresponding author.
E-mail addresses: xinyuyan1989@gmail.com (X. Yan), zhi.cao@tus.ie (Z. Cao), alanj.murphy@tus.ie (A. Murphy), ysqiao@research.ait.ie (Y. Qiao).

https://doi.org/10.1016/j.jece.2022.108130
Received 9 March 2022; Received in revised form 13 June 2022; Accepted 17 June 2022
Available online 20 June 2022

because the sample preparation process is simple, and the detection precision is high. A single spectral measurement can conveniently detect minute microscopic particles. Hence, many methods have been proposed for automatic and semi-automatic identification of microplastics with ATR-FTIR spectral data, e.g., Derivative approaches, Euclidean distance, and Peak searching algorithms [11]. These techniques are commonly used in commercial equipment. In these approaches, the spectrums are compared with standard spectrums of samples that are pure or unadulterated. However, the similar characteristic bands resulting from plastic aging can cause aged polymers to be matched to an erroneous reference spectrum [9,12]. For example, Figure A.1 (A.x denotes Additional figures or tables presented in Appendix A) depicts the spectrums of poly(ethylene) + fouling (blue line), PEVA (purple line), and standard polyethene (red line). Suppose poly(ethylene) + fouling (blue line) is the sample to be classified. It appears that it is more similar to PEVA instead of the standard polyethene. Therefore, the traditional library search is error prone in identifying the MPs.

With the development of Machine Learning (ML) technologies, ML based MP identification methods become increasingly popular for extracting the characteristics from the MP ATR-FTIR data to minimize the aged or polluted bias without expert knowledge, e.g. the methods based on Support Vector Machines (SVM) [13], K-Nearest Neighbours (KNN) [12], Random Forests (RF) [14], Partial Least Squares Discriminant Analysis (PLSDA) and Soft Independent Modelling of Class Analogies (SIMCA) [5]. Existing ML based approaches are typically based on a single ML model. Majority of existing literature is focused on proposing new ML based identification methods and compare their proposed methods with a limited number of other algorithms using a limited number of evaluation metrics. [12–15]. The performances of these methods are commonly evaluated using the average scores of all types of MPs [16].

In currently available MP FTIR datasets, the sample sizes of the MPs are imbalanced and usually have significant differences, e.g., the MP sample sizes in the Kedzierski dataset [12] vary from 25 % to 2 %. In addition, some MP samples are tampered or contaminated. These FTIR spectrums will have lack of peaks or contain unexpected noise [17]. Hence, the characteristics of these MP FTIR datasets will affect the performance of different ML methods, i.e. different ML algorithms may perform better in identifying different types of MPs. For example, the test results in the later sections show that, comparing KNN with SVM, tampering affects more on the SVM, whereas dataset imbalance affects are more on KNN (SVM outperforms KNN when the sample size is small).

This has motivated us to identify the limitations of different ML based MP identification methods by performing a comprehensive evaluation of the existing algorithms through a standardized identification procedure, and then propose an ensemble learning algorithm to exploit the advantages of different ML algorithms. The contribution of the paper can be summarized into 2 parts:

- This paper provides a systematic evaluation on a broad range of ML based MP identification methods from both macroscopic (using radar charts) and microscopic (using confusion matrices and class reports) perspectives. The evaluation results reveal that, for different types of MPs, different ML based methods display different performances. A summarisation of the performance is presented.
- Based on the evaluation results, the ML algorithms that performed the best in different MPs identification scenarios are chosen as the components in creating the proposed ensemble learning algorithm. The main concept is to aggregate the outputs of the chosen ML methods through a neural network model. The evaluation results show that the proposed ensemble learning method outperforms the single ML model based approaches for identifying every type of MPs in the dataset.

The rest of the paper is structured as follows. Section 2 introduces the background knowledge on machine learning and discusses the state of art of related work. As the proposed ensemble learning algorithm is based on the results of the systematic evaluation on the existing ML based algorithms, a separate set of "Materials and methods" and "Results and discussion" are provided the systematic evaluation and the ensemble learning algorithm, i.e. Section 3 and Section 4 present "Materials and methods" and "Results and discussion" for the systematic evaluation of ML based methods, and Section 5 and Section 6 introduce the "Materials and methods" and "Results and discussion" for the proposed ensemble learning algorithm. Finally, Section 7 provides conclusion & future work of the paper.

2. Related work

This section provides a brief overview of the theoretical foundations of spectral data processing, background on ML algorithms and ML based MP identification methods used in previous works.

2.1. Spectral data processing & evaluation workflow

For ML based MP identification methods, data processing is very important [18]. The outline of the standard workflow of data processing for identifying MP as follows:

2.1.1. Capturing the microplastic spectrum data

To minimize the measurement variations, the sample is purified physically, chemically, or biologically to remove the matrix. The operation steps of the FTIR instrument should be standardized [18].

2.1.2. Reducing the noise intensity (if necessary) and correcting the baseline

If the process of capturing MP spectra generates excessive noise in the original data, it will hinder the identification of MPs. Hence, the spectral data needs to be denoised by curve smoothing algorithms. The Savitzky-Golay (SG) [19] smoothing method to process spectra data with noise as shown in Figure A.2. As denoising method will result in loss of spectral characteristics, if not necessary, this method should be avoided in data pre-processing.

The baseline effects [20] are derived from non-uniform optical properties and non-planar geometry of the samples, which will cause data curves to slope and drift [21], as depicted in Figure A.3.

2.1.3. Normalizing the data

Due to the geometrical properties of samples and measurement process, the FTIR signal intensity will vary greatly. These variations will adversely influence the ML model training and the correspondent MP identification [22]. Therefore, the intensities should be normalized to an identical scale before training the ML models. The normalized FTIR data has been shown in Figure A.4.

2.1.4. Extracting features from the data

This step is to extract the features from the dataset (before training the ML models). Traditional spectra identification is based on FTIR experts to analyse most prominent peaks of specific wavelength. Some methodologies with manually extracting relevant spectral data have been proposed [15,23,24], which can be used by human operators to compare with a standard spectral curve. For ML based MP identification methods, the original data dimensions can be too high and cause the dimensional curse problem. Principal Component Analysis (PCA) is a common tool for the data dimensional reduction [25,26]. Hence, this work will evaluate whether data dimension will affect the MP identification, by using the ML models with and without PCA.

2.1.5. MP identification

The traditional method is to search a spectrum library to match the extracted features [18]. For ML based approaches, the pre-processed dataset will be used as the training and validation dataset to train and

validate the ML models. This is the focus of this paper.

2.2. Machine learning background

This section will introduce the ML algorithms used in this paper.

Principal Component Analysis (PCA) [25–28] is to project high-dimensional data into low-dimensional space and capture the data's greatest information (variance) on the projected dimension. It can utilize fewer data dimensions while keeping the majority characteristics of the original data points.

Multilayer Perceptron (MLP) is also called Artificial Neural Network (ANN). It's based on biological neural networks, which are used to mimic animal brains. It's composed of artificial neurons, which are a network of linked units or nodes that imitate neurons in a biological brain. [29].

Random Forests (RF) classifier contains a set of tree structured classifiers (decision trees) [30,31]. For categorization, the sample is fed into each branch. A classification result will be generated for each tree. The random forest combines all of the voting results, and the eventual output is the category with the most voting times [32].

The K-Nearest Neighbors (KNN) classification method [33] trains a model for calculating the distance between an unknown sample and all previously measured samples. The top K known samples with the highest similarity to the unknown sample are chosen. Assuming K equals 3, if there are 2 samples in these 3 known samples belonging to the same category, it means that the unknown sample is classified to this category. If the 3 known samples are different, the unknown sample will belong to the nearest one [34].

The concept of Linear Discriminant Analysis (LDA) [13] is quite simple: the labelled data (points) are projected into a lower dimensional space by a transformation function similar to PCA. However, LDA is a supervised learning method [35]. Compared with PCA, this method aims to obtain max variance between each class instead of the max variance between each data. In this low dimensional space, similar samples (samples with the same labels) are as close as possible, and heterogeneous samples are as far away as possible. In another word, after the projection, the variance of an identical class is the smallest, and the variance between different classes is the largest. For PCA, the variance of each data is the largest [27].

Support Vector Machines (SVM) [30] are a set of methods that map all the points to a "high-dimensional space", and then locates a "hyperplane" in the high-dimensional space that can split these points (for the two-dimensional plane, the hyperplane is a straight line, and so on).

Partial Least Squares Discriminant Analysis (PLSDA) [5] is based on Partial Least Squares (PLS) regression to a linear model for classification or prediction [36]. PLS looks for a linear regression model by projecting the prediction variable and the observed variable into a new space, respectively.

Soft Independent Modelling of Class Analogies (SIMCA) [37] focuses on the analogy between specific class samples. SIMCA is a type of PCA based model. Basically, for each specified class, SIMCA will develop one model using PCA to extract features for each class [38].

2.3. ML based MP identification approaches

Kedzierski et al. [12] present an automated method using KNN to classify the spectrum data. In their work, results reveal that KNN performs well in identifying spectra of conventional polymers such as polyethylene. They set up a standard database for training the KNN model. Michel et al. [13] compare the MP identification spectroscopic techniques e.g., ATR-FTIR, near-infrared (NIR) reflectance spectroscopy combined with different ML methods. The results show that FTIR performs better than other spectral technologies. Their work adopts SVM, LDA KNN, PCA_SVM (Using PCA to reduce data dimension and then classify the data with SVM), PCA_LDA, PCA_KNN for classifying the MP FTIR data. The LDA and KNN based approach achieves the highest level of precision in their work based on a proprietary database. Hufnagl et al. [14] presented an automated method based on reduced dimensional spectral data and the RF methods for the classification, according to their dataset. Their work achieves an average precision score of 96.6 %. Da Silva et al. [5] adopt the PCA to reduce FTIR data dimension and use the PLS-DA and SIMCA models for MP characterization. The PLS-DA model performs better than SIMCA with higher recall scores. Back et al. [16] design an automatic method for comparing different ML pipelines (including SVM, MLP, KNN etc) to search the best score and best procedure for each ML pipeline. Their results discover that the SVM pipeline without using a baseline correction algorithm performs the best while compared with the others (including the ones with baseline corrections). However, different instruments may induce different baseline effects. The baseline effects must be eliminated if the model is to be reused for analysing data from different instruments.

The above discussion shows that many ML based MP identification methods have been proposed. However, they are not benchmarked with the same set of standards and consequently it is difficult for users to choose the ML methods for their usage scenarios. The different results imply that diverse datasets will have a significant impact on ML methods. Because their datasets may lack of peaks or contain unexpected noise or imbalanced sample quantities.

3. Materials and methods for systematic evaluation of ML based methods

3.1. ATR-FTIR

FTIR is a popular technology to analyse the structure of compounds. Attenuated Total Refraction (ATR) technology applied on FTIR simplifies the measurement of samples [39,40]. Some of the advantages offered by this technology offers are following: 1) There are no additional criteria for sample size, water content, shape and sample preparation is straightforward and non-destructive. 2) The robust detector is sensitive, the measurement zone is modest, and the precision of detection can reach several microns. 3) It is convenient to search the infrared spectrum database and analyse the chemical functional groups to determine the types and properties of substances. 4) ATR instrument can be installed on conventional FTIR to provide a cost effective and simple measurement solution.

3.2. The microplastics sample dataset

This work adopts 2 datasets from previous works. One dataset is from the Kedzierski et al.[12] containing 970 spectra. The MP of this spectral dataset mainly originate from the Mediterranean coast. All MP wavelengths data was recorded in absorbance mode ranging 4000–600 cm^{-1} with a 4 cm^{-1} resolution and 16 scans. The other dataset is from Jung et al. [41] containing 798 spectra for 5 types of MP ingested by sea turtles. The spectra are collected from 4000 cm^{-1} to 450 cm^{-1} with a data interval of 1 cm^{-1} with a 4 cm^{-1} resolution. The 'Unknown' type is defined by the datasets, which represents a small amount of MP samples that are not labelled. The Kedzierski dataset was collected from 120 sites along the Mediterranean coast and the organic materials were removed through dissecting microscope. The Jung dataset was collected from the digestive tract of turtles. MP samples were cleaned with a cleanroom wiper after being washed with nano porous deionized water.

This work has modified Kedzierski dataset (by deleting the 6 types of MP that only contain 1 sample) and Jung dataset (by deleting 3 types of MPs that only contain l sample and moving Polystyrene that only has 5 samples to the Unknown type) for training ML models. The modified Kedzierski dataset comprises 12 categories of MPs (polypropylene (PP), polyethylene, polyamide, PEVA, etc). The modified Jung dataset comprises 5 categories of MPs (HDPE, LDPE, PP, etc.). The details of dataset are depicted in Figure A.4.

The Jung dataset only contains 5 types of MPs that have sufficient

samples for training. Consequently, it is not suitable for comprehensive evaluation of the ML approach performance. Hence, it is only used for evaluating the proposed ensemble learning algorithm. The Kedzierski dataset is used for both the comparisons of the ML algorithm and the evaluation of the proposed ensemble learning algorithm.

3.3. Tools for programming

In this work, all the data analytical programs are written in the Python language. The ML methods of SVM, KNN, MLP, LDA, RF are implemented by scikit-learn (Sklearn) library [42]. PCA, PLSDA, SIMCA are implemented by Numpy and other libraries of Python. The functions for metrics are also from the Sklearn library. For each simulation, all source codes are available in a supplementary file (Appendix B) for reproducing and improved by peers.

3.4. Procedure for Comparison of MP Identification Algorithms

Fig. 1 illustrates the overall process of the comparison workflow for comparing the ML algorithms that have been used in the existing works. This workflow consists of three major steps, i.e. data pre-processing (i.e., smoothing, baseline correction, normalization, and PCA processing), hyperparameter adjustment and ML model training (training and testing the ML models), and systematic comparison of the ML models (generating the evaluation metrics, e.g. radar charts, confusion matrix, and class reports).

3.4.1. Dataset pre-processing

The smoothing method may cause information loss in spectral data [43]. Since the dataset adopted in this paper is not noisy, the smoothing method is not employed in the pre-processing procedure. For baseline correction methods, the airPLS method can retain the original information and the peaks of spectral data [20]. Consequently, this work adopts the airPLS method to correct the baseline effect of the dataset. For normalizing the data samples for ML approaches, MinMaxScaler, StandardScaler, and Normalize are commonly used. This work utilizes the Normalize method with parameter max as performing better than MinMaxScaler, StandardScaler. In this paper, PCA is adopted to reduce the dimension of the spectral data and to retain relevant information. To explore the impacts of data dimension reduction on the classification results, this work simultaneously analyses the dataset processed with PCA and without PCA while evaluating the ML models.

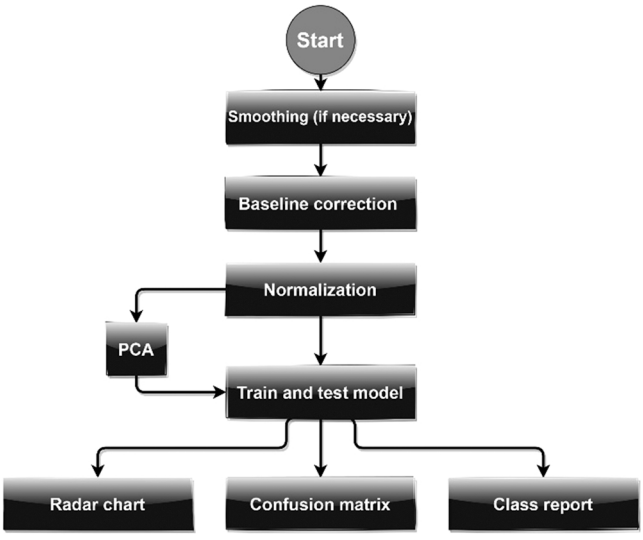

Fig. 1. The workflow of comparison for MP identification methods.

3.4.2. Hyperparameter adjustment and ML model training

In this work the hyperparameters of each ML based approaches are adjusted using the GridSearchCV function of Sklearn, e.g., finding the optimal K value of the KNN model. In previous work, Back et al. [16] adopt the Monte Carlo Cross-Validation (MCCV) to randomly split the FTIR dataset, train the model with the dataset, and then average the resulting metrics of multiple runs [44]. Similar to their method, this work randomly divides the dataset into a training set and a validation set by a train_test_split function from Sklearn, and then train and validate the ML model. This process is repeated 200 times. The collected metrics are averaged.

3.4.3. Systematic evaluation of the ML models

This paper employs metrics that can depict the performance of the ML models from both macroscopic and microscopic perspectives in terms of 5 metrics, i.e. recall, precision, F1 score, accuracy, and Kappa score. For the microscopic analysis, the result for each metric, each MP and each ML method is calculated. In addition, the confusion matrix (CM) for each ML method is presented, which displays whether multiple categories are confused, i.e., one MP type is predicted as another MP type. For the macroscopic analysis, the average metric of each ML method for classifying all the MPs are calculated and displayed in a radar chart. For simplicity purposes, only a part of microscopic results are presented in the paper, i.e. recall, precision and F1 score.

3.5. Metrices for benchmarking ML based approaches

This work adopts the following metrics to benchmark the performance of the ML based MP identification algorithms, as shown in Eqs. (1)–(5), where TP is True Positives, FP is False Positives, TN is True Negatives, and FN is False Negatives [45].

Recall (also called Sensitivity) can be considered as the proportion correctly identified by the model for the same type of MP [46]. Its definition is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

Similar to Recall, Precision is the percentage of the positive results that are correctly identified as positive by the model. For MP identification, Precision is the ratio of the correctly classified MP to the number of MP that are identified as this MP kind. Its definition is as follows:

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

The F1 score is the harmonic mean of the Recall and Precision [47]. When the Recall and Precision metrics are contradictory, the F1 score should be adopted to compare the two metrics above comprehensively. Its definition is as follows:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

Accuracy is another measurement of the robustness of the model, which is the percentage of all the tested data that is correctly identified (i.e., TP and TN). Accuracy is the simplest and most intuitive metric for benchmarking ML methods [30]. Its definition is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Confusion matrix (CM). In addition to the metrices above, CM is used to depict the relationship between all the identified classes [48]. In a CM, each row represents that predicted class for each actual class. The diagonal element contained in a row represents the correct prediction result for the actual class (i.e. the recall score), whereas the other elements show the misclassification results. The greater diagonal value in the matrix represents the better classification performance for the correspondent MP (class). The CM can also show which MP is prone to

be misclassified to which MP.

Kappa score is utilized for consistency testing and used to assess classification accuracy considering the effects of sample imbalance in the dataset [30]. Its definition is as follows:

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{5}$$

Kappa score can be calculated based on CM. Where p_0 is accuracy score of predicted results, i.e. the sum of the diagonal elements of the CM divided by the sum of all the elements in the CM. Assuming that the sums of the elements for each row in the CM are A_1, A_2., A_i respectively, and the sums of elements for each column in the CM are B_1, B_2., B_i respectively, the total number of samples are N, then $pe = (A_1 * B_1 + A_2 * B_2 + . A_i * B_i) / (N*N)$.

4. Results and discussion for systematic evaluation of ML based methods

4.1. Macroscopic performance of ML based approaches

To calculate the appropriate reduced dimensional data (PCs) by PCA, this work calculates the accuracy of each method for various PCs. When the number of PCs is set as 200, each approach achieves the best results. The average values of the Kappa score, F1 score, accuracy score and recall score and precision score of different ML based algorithms are calculated and depicted in Figure A.5.

SVM and PCA_SVM based identification method performs the best amongst all the ML methods (Kappa score, F1 score, accuracy, recall, precision are 92.24 %, 90.77 %, 93.53 %, 92.77 %, and 90.08 %

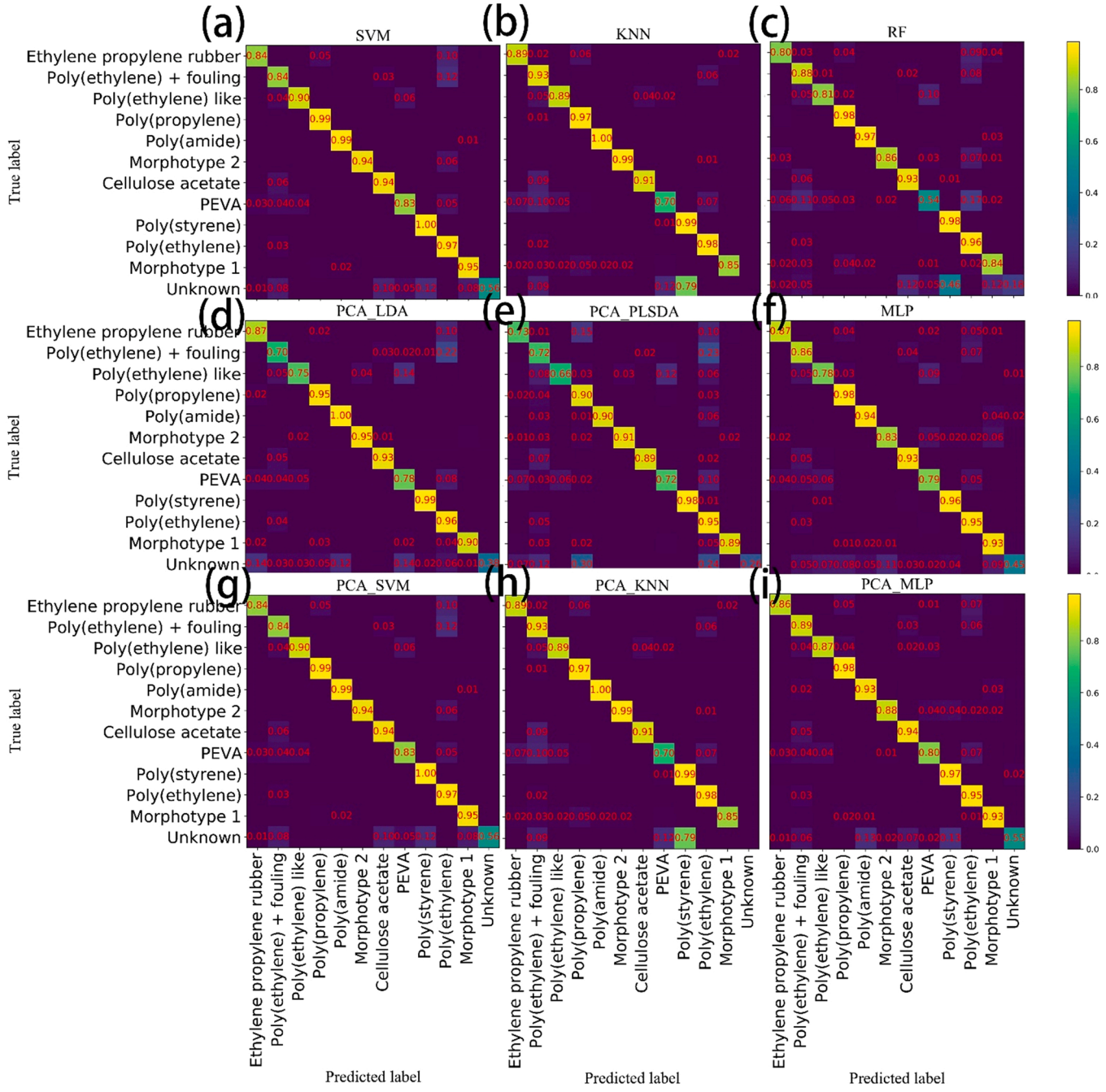

Fig. 2. the confusion matrices of each identification approach. (a) the CM of SVM (b) the CM of KNN (c) the CM of RF (d) the CM of LDA (e) the CM of PCA_PLSDA (f) the CM of MLP (g) the CM of PCA_SVM (h) the CM of PCA_KNN (i) the CM of PCA_MLP.

respectively). However, KNN, PCA_KNN and MLP also perform well and their five metrics in the radar chart are near to those of SVM. More importantly, each metric calculated in the radar chart is an average score for all types of MP. If one type of MP occupies a large portion of the dataset, it will have a significant impact on the average metrics. To avoid this bias, the next section will provide a zoomed in analysis to these ML algorithms, focusing on their performance for each individual MP.

4.2. CMs of ML based approaches for each MP

Fig. 2 shows the 9 CMs of the ML based approaches that perform well in the tests. The results for other approaches are omitted (PCA_RF, LDA, SIMCA, PCA_SIMCA). For each sub-figure (a–i) in Fig. 2, the horizontal coordinate represents the predicated label of each MP and the vertical coordinate represents the True label of MP. For each ML based identification method, the percentage results of each MP classification are provided. It also exhibits which types of MPs are commonly confused and misidentified as others for each ML approach. It also shows the impacts of PCA on the performance of each ML method while identifying each MP.

The results show that Poly(propylene), Poly(amide) Poly(styrene) are identified correctly by SVM (99–100 %). Ethylene propylene rubber (EPR), Poly(ethylene) fouling and PEVA are not well identified (83–84 %) and confused with other MP (which have been depicted in the CM). KNN obtains high scores in identifying Poly(amide), Morphotype2, Poly (styrene) and Poly(ethylene) (98–99 %), but it does not recognize the MP well when the amount of training samples is not sufficient (such as Poly(ethylene) like, PEVA). PCA_LDA has high identification rates in Poly(amide), Poly(styrene), Poly(ethylene) (96–100 %). PCA_SVM, PCA_KNN and PCA_MLP all perform well and are similar to the respective ML methods without using PCA. All the ML based methods can identify the pure Poly(propylene), Poly(amide), Poly(styrene), and Poly (ethylene) very well. However, certain mixtures are often misclassified, e.g., EPR, Poly(ethylene) fouling, Poly(ethylene), and Morphotype2.

4.3. Class report of ML based approaches for each MP

To obtain additional in-depth analysis to the ML algorithms while identifying each MP, and display the specific performance of different methods in different types of MP identification, this work adopts Precision, Recall, F1 score to compare different ML approaches for each MP (CM just demonstrates the Recall score).

SVM and PCA_SVM both perform the best in identifying the EPR, Poly(propylene), PEVA, Morphotype 1, which demonstrates SVM is quite effective in identifying the mixed MPs. However, for identifying Poly(ethylene) fouling, Poly(ethylene), and Cellulose acetate, KNN and PCA_KNN present the best result. This probably means that KNN could identify the MPs with noise (e.g. Poly(ethylene) fouling) with sufficient samples. The other ML methods for identifying each MP based on the highest value of the sum of the Precision, Recall and F1 score are summarized in Table A.1. (All the class reports are shown in Table A.2-A.11 and Figure A.6). The reasons of the different performance are various ML techniques extract features in different ways e.g., SVM classifies the MPs by calculating the hyperplane among MPs.(as mentioned in Section 2.2). In high-dimensional FTIR data, these methods will be affected by additional characteristics (-OH, -CO) produced by contaminated or oxidized MPs.

Although SVM is superior to the other ML methods in the average metrices, the CMs and the 3 metrices show that the ML method should be selected based on the sample types of MPs. In next section, this work intends to build an ensemble learning to exploit advantages of different ML based methods according to the characteristics of dataset.

5. Methods and materials for the ensemble learning algorithm

The CM and class report show that, for different MPs, each ML method exhibits different performances. The results (Fig. 2 and Table A.1) show that SVM, PCA_SVM, KNN, PCA_KNN and PCA_LDA can cover different scenarios for identifying MPs. To explore a general method to utilise the strength of each of the above ML algorithms, this work designs an ensemble learning algorithm to aggregate the output of the 5 ML algorithms with a neural network.

The algorithm architecture is shown in Fig. 3. It includes 3 key steps: 1) inputs the dataset (either training dataset or test dataset) into the 5 chosen ML methods. Each of the ML methods will generate the classification probabilities, i.e. probability that the input MP should be classified as each MP; 2) sends the output probabilities of each ML method to the input layer of the neural network. 3) the neural network outputs 12 classification probabilities which correspond to the 12 MP types. In Fig. 3, $A^{[0]}$ represents the input layer of the neural network, which contains 60 neurons (5 ML methods generate 5*12 probabilities). A [1], A [2], and A [3] represent the hidden layers that contain 64 neurons in each layer. These 3 layers all adopt ReLU (Rectified Linear Unit) function as the activation function. A [4] represents the output layers containing 12 neurons. This output layer utilizes softmax as the activation function. Cross entropy (categorical_crossentropy) is used as the loss function during the training.

6. Results and discussion for the ensemble learning algorithm

6.1. Kedzierski dataset

The 5 average metrics (Kappa score, F1 score, accuracy, recall, precision) of this ensemble learning algorithm are 93.04 %, 90,13 %, 94.19 %, 91.74 %, 89.78 % separately based on the Kedzierski dataset. Although the recall score is 0.1 % less than that of SVM, the average accuracy is 0.7 % higher than that of SVM, which means this method improves the successful identification rate.

Fig. 4 shows the CM of the proposed ensemble learning algorithm. 10 % of ERP (containing ethylene and propylene) samples are misidentified as Poly(propylene) and Poly(ethylene) fouling, PEVA, Morphotype 1 (4 %, 2 %, 2 % and 2 % respectively). The 7 % of Poly(ethylene) fouling and PEVA are confused with Poly(ethylene). The most important reason of these misclassification is that the MP samples contain the identical chemical compound (e.g., ethylene and/or propylene) that generates the same peaks in the FTIR spectra. It is very difficult for sub-model (SVM, KNN, LDA, etc) to extract sufficient features with limited samples to identify the MPs. However, compared with the SVM CM (which performs the best amongst the ML models) in the Fig. 2, the recall scores for EPR and poly(ethylene) fouling are both improved from 84 % to 89 %, and the recall scores for the other MPs are similar to those of SVM.

In class report, the average value across all the scores (precision, F1 score and recall) is 91.5 % (shown in Table A.12 and specific scores shown in Figure A.7). The average score of SVM is 91.1 % (shown in Table A.2). Although the ensemble learning algorithm does not outperform SVM in all MP types, the sum of precision, recall, F1 score of Poly(ethylene) fouling, Poly(ethylene) like, Cellulose acetate and Poly (ethylene) are improved. Other MPs are similar to SVM results.

6.2. Jung dataset

Back et al. test the SVM algorithm with linear kernel on Jung dataset [16]. The accuracy reaches 94.0 %. This proposed ensemble learning algorithm has been evaluated with the same dataset. The average values of the five metrics (Kappa score, F1 score, accuracy, recall, precision) are 90.4 %, 78.9 %, 94.5 %, 84.7 %, 76.0 %. The accuracy (94.5 %) is better than Back's method. The CM of the proposed algorithm while using Jung dataset shows that 50 % of LDPE and 19 % of Mixture are misclassified (depicted in Figure A.8). The class report results demonstrate that this method performs excellent in HDPE and PP (shown in Table A.13). 50 % of LDPE and 19 % of Mixture are misclassified because they contain identical chemical characteristics, and the samples are insufficient (both

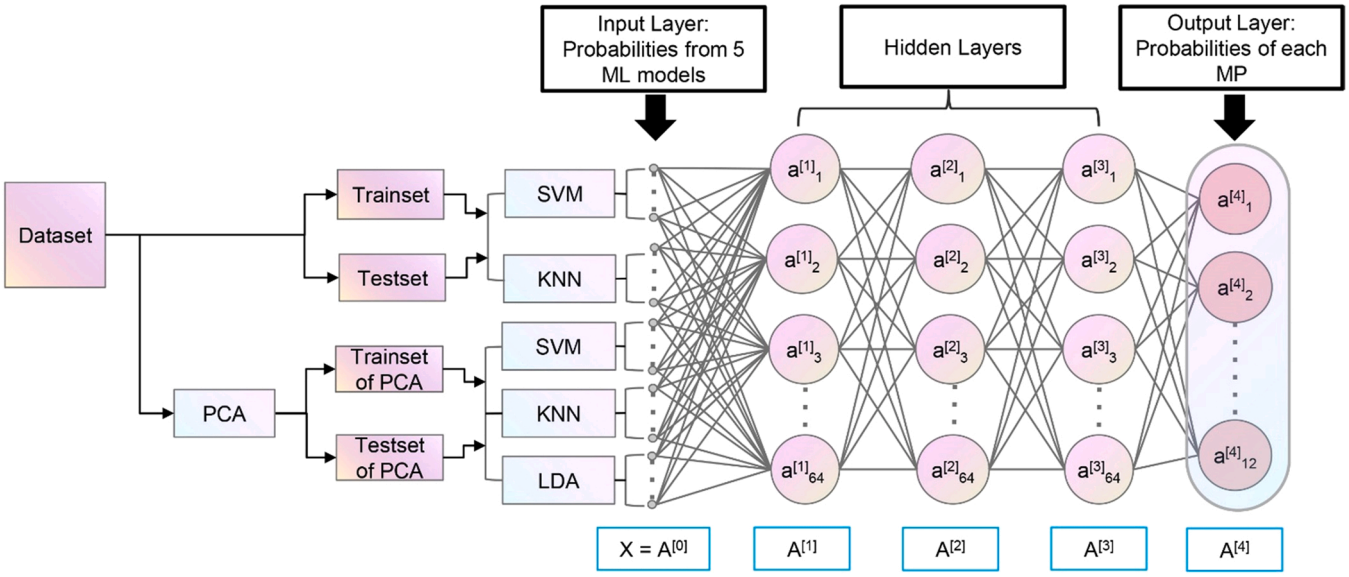

Fig. 3. The design of ensemble learning algorithm.

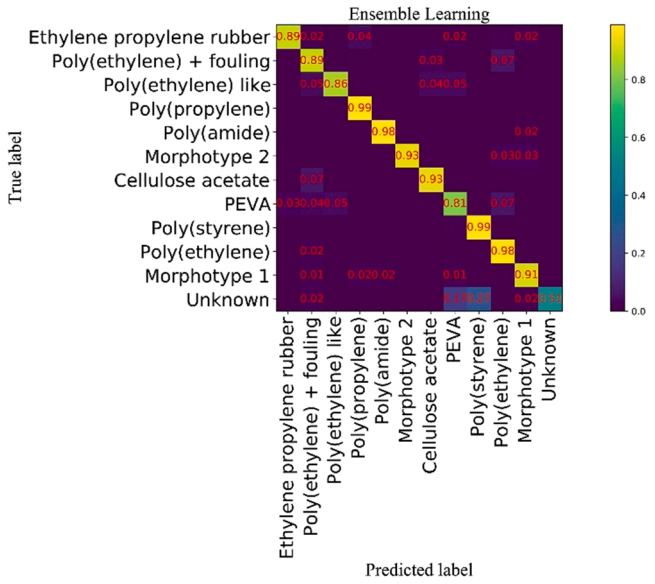

Fig. 4. CM of the Ensemble Learning Algorithm.

occupy 5 %). In Back's work, the two MP are also not classified well due to the identical reasons (their recall scores are 65 % and 82 % because they transfer some spectrums into Unknown type).

7. Conclusion & future work

This work adopts a systematic workflow to evaluate the performance of 7 existing machine learning based approaches in identifying and characterising MPs using one standard FTIR spectral dataset. The results show that the imbalanced sample size in the dataset and the MP fouling effects have significant impacts on the performance of the ML algorithms while identifying each MP type. A ML algorithm may perform differently while identifying different types of MPs.

In practice, a user would not know the MP type before identification. Therefore, it is difficult to choose an appropriate ML model for the specific MP identification task. Hence, this work proposes an ensemble learning based method to aggregate the best performing ML models identified in the systematic evaluation to avoid the poor performance scenarios of each individual ML model. The proposed ensemble learning method has been evaluated with 2 datasets, i.e. the Kedzierski and Jung Dataset. The results show that this method obtains higher performance than the state-of-the-art works in terms of 5 metrics (Kappa score, F1 score, accuracy, recall, precision). It obtains a clearer confusion matrix (i.e. less confused classifications), and a higher class report for each MP.

The future work includes expanding the application domain from aquatic systems to other areas, e.g. applying this method on MP in soils and plants. Another planned extension of this work is to integrate other MP characterisation technologies, e.g. DSC and TGA, into the ensemble learning algorithm.

CRediT authorship contribution statement

Xinyu Yan: Algorithm design and implementation. **Zhi Cao:** Algorithm design. **Alan Murphy**: Algorithm design. **Yuansong Qiao**: Algorithm design.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This publication has emanated from research conducted with the financial support of the Technological University of the Shannon (TUS), Ireland under President's Doctoral Scholarship 2020, and Science Foundation Ireland (SFI) under Grant Number SFI 16/RC/3918, co-funded by the European Regional Development Fund.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jece.2022.108130.

References

[1] M. Haward, Plastic pollution of the world's seas and oceans as a contemporary challenge in ocean governance, Nat. Commun. 9 (2018) 9–11, https://doi.org/10.1038/s41467-018-03104-3.

[2] S. Gündoğdu, in: M.Z. Hashmi (Ed.), Polymer Types of Microplastic in Coastal Areas BT - Microplastic Pollution: Environmental Occurrence and Treatment Technologies, Springer International Publishing, Cham, 2022, pp. 77–88, https://doi.org/10.1007/978-3-030-89220-3_4.

[3] W. Leal Filho, U. Saari, M. Fedoruk, A. Iital, H. Moora, M. Klöga, V. Voronova, An overview of the problems posed by plastic products and the role of extended producer responsibility in Europe, J. Clean. Prod. 214 (2019) 550–558, https://doi.org/10.1016/j.jclepro.2018.12.256.

[4] M. Atamanalp, M. Köktürk, V. Parlak, A. Ucar, G. Arslan, G. Alak, A new record for the presence of microplastics in dominant fish species of the Karasu River Erzurum, Turkey, Environ. Sci. Pollut. Res. 29 (2022) 7866–7876, https://doi.org/10.1007/s11356-021-16243-w.

[5] V.H. Da Silva, F. Murphy, J.M. Amigo, C. Stedmon, J. Strand, Classification and quantification of microplastics (<100 μm) using a focal plane array-fourier transform infrared imaging system and machine learning, Anal. Chem. 92 (2020) 13724–13733, https://doi.org/10.1021/acs.analchem.0c01324.

[6] J.G.B. Neto, F.L. Rodrigues, I. Ortega, L. dos, S. Rodrigues, A.L. d F. Lacerda, J. L. Coletto, F. Kessler, L.G. Cardoso, L. Madureira, M.C. Proietti, Ingestion of plastic debris by commercially important marine fish in southeast-south Brazil, Environ. Pollut. 267 (2020), https://doi.org/10.1016/j.envpol.2020.115508.

[7] M.R. Jung, F.D. Horgen, S.V. Orski, V. Rodriguez C, K.L. Beers, G.H. Balazs, T. T. Jones, T.M. Work, K.C. Brignac, S.J. Royer, K.D. Hyrenbach, B.A. Jensen, J. M. Lynch, Validation of ATR FT-IR to identify polymers of plastic marine debris, including those ingested by marine organisms, Mar. Pollut. Bull. 127 (2018) 704–716, https://doi.org/10.1016/j.marpolbul.2017.12.061.

[8] F. Julienne, N. Delorme, F. Lagarde, From macroplastics to microplastics: role of water in the fragmentation of polyethylene, Chemosphere 236 (2019), 124409, https://doi.org/10.1016/j.chemosphere.2019.124409.

[9] S. Gündoğdu, C. Çevik, S. Karaca, Fouling assemblage of benthic plastic debris collected from Mersin Bay, NE Levantine coast of Turkey, Mar. Pollut. Bull. 124 (2017) 147–154, https://doi.org/10.1016/j.marpolbul.2017.07.023.

[10] N. Cebi, M.T. Yilmaz, O. Sagdic, A rapid ATR-FTIR spectroscopic method for detection of sibutramine adulteration in tea and coffee based on hierarchical cluster and principal component analyses, Food Chem. 229 (2017) 517–526, https://doi.org/10.1016/j.foodchem.2017.02.072.

[11] G. Renner, P. Sauerbier, T.C. Schmidt, J. Schram, Robust automatic identification of microplastics in environmental samples using FTIR microscopy, Anal. Chem. 91 (2019) 9656–9664, https://doi.org/10.1021/acs.analchem.9b01095.

[12] M. Kedzierski, M. Falcou-Préfol, M.E. Kerros, M. Henry, M.L. Pedrotti, S. Bruzaud, A machine learning algorithm for high throughput identification of FTIR spectra: Application on microplastics collected in the Mediterranean Sea, Chemosphere 234 (2019) 242–251, https://doi.org/10.1016/j.chemosphere.2019.05.113.

[13] A.P.M. Michel, A.E. Morrison, V.L. Preston, C.T. Marx, B.C. Colson, H.K. White, Rapid identification of marine plastic debris via spectroscopic techniques and machine learning classifiers, Environ. Sci. Technol. 54 (2020) 10630–10637, https://doi.org/10.1021/acs.est.0c02099.

[14] B. Hufnagl, D. Steiner, E. Renner, M.G.J. Löder, C. Laforsch, H. Lohninger, A methodology for the fast identification and monitoring of microplastics in environmental samples using random decision forest classifiers, Anal. Methods 11 (2019) 2277–2285, https://doi.org/10.1039/c9ay00252a.

[15] S. Primpke, C. Lorenz, R. Rascher-Friesenhausen, G. Gerdts, An automated approach for microplastics analysis using focal plane array (FPA) FTIR microscopy and image analysis, Anal. Methods 9 (2017) 1499–1511, https://doi.org/10.1039/c6ay02476a.

[16] H. de, M. Back, E.C. Vargas Junior, O.E. Alarcon, D. Pottmaier, Training and evaluating machine learning algorithms for ocean microplastics classification through vibrational spectroscopy, Chemosphere 287 (2022), 131903, https://doi.org/10.1016/j.chemosphere.2021.131903.

[17] H. Zhu, M. Nyström, Cleaning results characterized by flux, streaming potential and FTIR measurements, Colloids Surf. A Physicochem. Eng. Asp. 138 (1998) 309–321, https://doi.org/10.1016/S0927-7757(97)00072-1.

[18] G. Renner, A. Nellessen, A. Schwiers, M. Wenzel, T.C. Schmidt, J. Schram, Data preprocessing & evaluation used in the microplastics identification process: a critical review & practical guide, TrAC - Trends Anal. Chem. 111 (2019) 229–238, https://doi.org/10.1016/j.trac.2018.12.004.

[19] B. Zimmermann, A. Kohler, Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy, Appl. Spectrosc. 67 (2013) 892–902, https://doi.org/10.1366/12-06723.

[20] Z.M. Zhang, S. Chen, Y.Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 135 (2010) 1138–1146, https://doi.org/10.1039/b922045c.

[21] K.H. Liland, 4S Peak Filling - Baseline estimation by iterative mean suppression, MethodsX 2 (2015) 135–140, https://doi.org/10.1016/j.mex.2015.02.009.

[22] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine Learning and Materials Informatics: Recent Applications and Prospects, 2017. ⟨http://arxiv.org/abs/1707.07294⟩.

[23] S. Primpke, M. Wirth, C. Lorenz, G. Gerdts, Reference database design for the automated analysis of microplastic samples based on Fourier transform infrared (FTIR) spectroscopy, Anal. Bioanal. Chem. 410 (2018) 5131–5141, https://doi.org/10.1007/s00216-018-1156-x.

[24] L.C. Lee, C.Y. Liong, A.A. Jemain, A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum, Chemom. Intell. Lab. Syst. 163 (2017) 64–75, https://doi.org/10.1016/j.chemolab.2017.02.008.

[25] P. Saganowska, M. Wesolowski, Principal component and cluster analyses as supporting tools for co-crystals detection, J. Therm. Anal. Calorim. 130 (2017) 45–55, https://doi.org/10.1007/s10973-017-6436-8.

[26] J.S. Bae, S.K. Oh, W. Pedrycz, Z. Fu, Design of fuzzy radial basis function neural network classifier based on information data preprocessing for recycling black plastic wastes: comparative studies of ATR FT-IR and Raman spectroscopy, Appl. Intell. 49 (2019) 929–949, https://doi.org/10.1007/s10489-018-1300-5.

[27] J. Gal, C. Bailleux, D. Chardin, T. Pourcher, J. Gilhodes, L. Jing, J.M. Guigonis, J. M. Ferrero, G. Milano, B. Mograbi, P. Brest, Y. Chateau, O. Humbert, E. Chamorey, Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer, Comput. Struct. Biotechnol. J. 18 (2020) 1509–1524, https://doi.org/10.1016/j.csbj.2020.05.021.

[28] G. Bonifazi, G. Capobianco, S. Serranti, A hierarchical classification approach for recognition of low-density (LDPE) and high-density polyethylene (HDPE) in mixed plastic waste based on short-wave infrared (SWIR) hyperspectral imaging, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 198 (2018) 115–122, https://doi.org/10.1016/j.saa.2018.03.006.

[29] J.A. Fine, A.A. Rajasekar, K.P. Jethava, G. Chopra, Spectral deep learning for prediction and prospective validation of functional groups, Chem. Sci. 11 (2020) 4618–4630, https://doi.org/10.1039/c9sc06240h.

[30] F.Y. Osisanwo, J.E.T. Akinsola, O. Awodele, J.O. Hinmikaiye, O. Olakanmi, J. Akinjobi, Supervised machine learning algorithms: classification and comparison, Int. J. Comput. Trends Technol. 48 (2017) 128–138, https://doi.org/10.14445/22312803/ijctt-v48p126.

[31] Ö. Akar, O. Güngör, Classification of multispectral images using Random Forest algorithm, J. Geod. Geoinf. 1 (2012) 105–112, https://doi.org/10.9733/jgg.241212.1.

[32] Y.T. Wang, B. Li, X.J. Xu, H. Bin Ren, J.Y. Yin, H. Zhu, Y.H. Zhang, FTIR spectroscopy coupled with machine learning approaches as a rapid tool for identification and quantification of artificial sweeteners, Food Chem. 303 (2020), https://doi.org/10.1016/j.foodchem.2019.125404.

[33] P. Thanh Noi, M. Kappas, Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 Imagery, Sensors 18 (2017), https://doi.org/10.3390/s18010018.

[34] M. Murugappan, Electromyogram signal based human emotion classification using KNN and LDA, Proc. - 2011 IEEE Int, Conf. Syst. Eng. Technol. ICSET 2011 (2011) 106–110, https://doi.org/10.1109/ICSEngT.2011.5993430.

[35] M. Imani, H. Ghassemian, Band clustering-based feature extraction for classification of hyperspectral images using limited training samples, IEEE Geosci. Remote Sens. Lett. 11 (2014) 1325–1329, https://doi.org/10.1109/LGRS.2013.2292892.

[36] R. Calvini, G. Orlandi, G. Foca, A. Ulrici, Developmentof a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging, J. Spectr. Imaging 7 (2018) 1–15, https://doi.org/10.1255/jsi.2018.a13.

[37] A.M. Jiménez-Carvelo, M.T. Osorio, A. Koidis, A. González-Casado, L. Cuadros-Rodríguez, Chemometric classification and quantification of olive oil in blends with any edible vegetable oils using FTIR-ATR and Raman spectroscopy, LWT - Food Sci. Technol. 86 (2017) 174–184, https://doi.org/10.1016/j.lwt.2017.07.050.

[38] F.B. de Santana, W. Borges Neto, R.J. Poppi, Random forest as one-class classifier and infrared spectroscopy for food adulteration detection, Food Chem. 293 (2019) 323–332, https://doi.org/10.1016/j.foodchem.2019.04.073.

[39] J.E. Halstead, J.A. Smith, E.A. Carter, P.A. Lay, E.L. Johnston, Assessment tools for microplastics and natural fibres ingested by fish in an urbanised estuary, Environ. Pollut. 234 (2018) 552–561, https://doi.org/10.1016/j.envpol.2017.11.085.

[40] B.H. Stuart, Infrared Spectroscopy: Fundamentals and Applications, 2005. ⟨https://doi.org/10.1002/0470011149⟩.

[41] M.R. Jung, G.H. Balazs, T.M. Work, T.T. Jones, S.V. Orski, V. Rodriguez C, K. L. Beers, K.C. Brignac, K.D. Hyrenbach, B.A. Jensen, J.M. Lynch, Polymer identification of plastic debris ingested by pelagic-phase sea turtles in the Central Pacific, Environ. Sci. Technol. 52 (2018) 11535–11544, https://doi.org/10.1021/acs.est.8b03118.

[42] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, 2013, pp. 1–15. ⟨http://arxiv.org/abs/1309.0238⟩.

[43] R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Trans. Speech Audio Process 9 (2001) 504–512, https://doi.org/10.1109/89.928915.

[44] Q.S. Xu, Y.Z. Liang, Monte Carlo cross validation, Chemom. Intell. Lab. Syst. 56 (2001) 1–11, https://doi.org/10.1016/S0169-7439(00)00122-2.

[45] K. Shah, H. Patel, D. Sanghvi, M. Shah, A comparative analysis of logistic regression, random forest and KNN Models for the Text classification, Augment. Hum. Res. 5 (2020), https://doi.org/10.1007/s41133-020-00032-0.

[46] V.K. Unnikrishnan, K.S. Choudhari, S.D. Kulkarni, R. Nayak, V.B. Kartha, C. Santhosh, Analytical predictive capabilities of laser induced breakdown spectroscopy (LIBS) with principal component analysis (PCA) for plastic classification, RSC Adv. 3 (2013) 25872–25880, https://doi.org/10.1039/c3ra44946g.

[47] A.R. Parsons, J.C. Pober, J.E. Aguirre, C.L. Carilli, D.C. Jacobs, D.F. Moore, A per-baseline, delay-spectrum technique for accessing the 21cm cosmic reionization signature, Astrophys. J. 756 (2012), https://doi.org/10.1088/0004-637X/756/2/165.

[48] A. Luque, A. Carrasco, A. Martín, A. de las Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix, Pattern Recognit. 91 (2019) 216–231, https://doi.org/10.1016/j.patcog.2019.02.023.