Automated Classification of Undegraded and Aged Polyethylene Terephthalate Microplastics from ATR-FTIR Spectroscopy using Machine Learning Algorithms

Christian Ebere Enyoh (cenyoh@gmail.com)
Saitama University, Japan https://orcid.org/0000-0003-4132-8988
Qingyue Wang (seiyo@mail.saitama-u.ac.jp)
Saitama University, Japan https://orcid.org/0000-0002-7673-2836

Research Article

Keywords: Automated classification, PET microplastics, ATR-FTIR spectroscopy, Machine learning algorithms, Environmental analysis, Method standardization

DOI: https://doi.org/

Automated Classification of Undegraded and Aged Polyethylene Terephthalate Microplastics from ATR-FTIR Spectroscopy using Machine Learning Algorithms

Christian Ebere Enyoh and Qingyue Wang

Graduate School of Science and Engineering, Saitama University, 255 Shimo-Okubo, Sakura-ku, Saitama City, Saitama 338-8570, Japan.

Correspondence: cenyoh@gmail.com, ORCID ID: 0000-0003-4132-8988(C.E.E); seiyo@mail.saitama-u.ac.jp, ORCID ID: 0000-0002-7673-2836 (Q.W.)

Abstract

Automated analysis of microplastics is essential due to the labor-intensive, time-consuming, and error-prone nature of manual methods. Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) spectroscopy offers valuable molecular information about microplastic composition. However, efficient data analysis tools are required to effectively differentiate between various types of microplastics due to the large volume of spectral data generated by ATR-FTIR. In this study, we propose a machine learning (ML) approach utilizing ATR-FTIR spectroscopy data for accurate and efficient classification of undegraded and aged polyethylene terephthalate (PET) microplastics (MPs). We evaluate seven ML algorithms, including Random Forest (RF), Gradient Boosting (GB), Decision Tree (DT), k-Nearest Neighbors (k-NN), Logistic Regression (LR), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP), to assess their performance. The models were optimized using 5-fold cross-validation and evaluated using multiple metrics such as confusion matrix, accuracy, precision, recall (sensitivity), and F1-score. The experimental results demonstrate exceptional performance by RF, GB, DT, and k-NN models, achieving an accuracy of 99% in correctly classifying undegraded and aged PET MPs. The proposed approach capitalizes on the potential of ATR-FTIR spectra to discern distinct chemical signatures of undegraded and aged PET particles, enabling precise and reliable classification. Furthermore, the method offers the benefit of automating the classification process, streamlining the analysis of environmental samples. It also presents the advantage of providing an effective means for method standardization, facilitating more automated and optimized extraction of information from spectral data. The method's versatility and potential for large-scale application make it a valuable contribution to the field of MP environmental research.

Keywords: *Automated classification; PET microplastics; ATR-FTIR spectroscopy; Machine learning algorithms; Environmental analysis; Method standardization*

1. Introduction

Microplastics (MPs), defined as plastic particles smaller than 5mm in size, have become a pervasive environmental concern due to their widespread distribution and potential ecological impacts (Verla et al, 2019; Enyoh et al, 2021). Among the various types of MPs, polyethylene terephthalate (PET) MPs are particularly prevalent, originating from commonly used items such as single-use plastic bottles and polyester textiles (Enyoh et al, 2023; Chowdury et al, 2022). While in the environment PET MPs can degrade and thus can serve as an efficient vector for toxic pollutants to ecosystems (Verla et al, 2019a). Therefore, understanding the abundance and distribution of both undegraded and aged PET microplastics is crucial for assessing their environmental risks and formulating effective mitigation strategies.

Analyzing microplastics manually is labor-intensive, time-consuming, and error-prone, necessitating the development of automated approaches to streamline the process. One promising technique for microplastic analysis is

Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) spectroscopy, which provides valuable molecular information about the composition of microplastics (Ioakeimidis et al, 2016; Chowdhury et al, 2022). However, the vast amount of spectral data generated by ATR-FTIR necessitates the utilization of powerful data analysis tools to accurately distinguish between different MPs types.

Recently researchers are now developing automated analytical method for the characterization of MPs. Hufnagl et al. (2019) proposed a Radom Forest Classifier method utilizing micro Fourier Transform Infrared (μ-FTIR) hyperspectral images to identify various types of microplastics (MPs), including polyethylene, polypropylene, poly(methyl methacrylate), polyacrylonitrile, and polystyrene, in environmental samples. They developed a model for four plastic types using spectral descriptors determined by spectroscopy experts for polymer characterization. Kedzierski et al. (2019) developed automated methods for identifying the chemical nature of microplastics (MPs) through FTIR-ATR spectra, using k-nearest neighbors' classification. The spectra were collected during the Tara Expedition in the Mediterranean Sea, and a learning database containing 969 microplastic spectra was created for testing. The results demonstrated the effectiveness of machine learning in identifying spectra of common polymers, such as poly(ethylene). However, it was noted that the learning database would benefit from enhancement with less common microplastic spectra. The method was further applied to over 4,000 spectra of unidentified

microplastics. The verification protocol revealed less than a 10% difference in the results between the proposed automated method and human expertise. Notably, 75% of the discrepancies could be easily corrected with minimal intervention, indicating the reliability and efficiency of the automated approach in identifying the chemical nature of microplastics. These findings highlight the potential of machine learning in large-scale microplastic characterization studies and underscore the importance of continuously updating the learning database for enhanced performance.

On the other hand, Wander et al. (2020) conducted an exploratory analysis of µ-FTIR imaging by employing Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) to reduce data dimensionality and visualize particle similarity. Although this strategy significantly reduced the analyzed data and removed background information from the images, further analysis was necessary for spectra characterization. Da Silva et al. (2020) presented an automated analytical method for characterizing small microplastics (< 100 µm) using µ-FTIR hyperspectral imaging and machine learning (ML) tools. They evaluated Partial Least Squares Discriminant Analysis (PLS-DA) and Soft Independent Modelling of Class Analogy (SIMCA) models with different data pre-processing strategies for classifying nine of the most common polymers worldwide. Additionally, they analyzed the hyperspectral images to automatically quantify particle abundance and size. PLS-DA demonstrated superior analytical performance, exhibiting higher sensitivity, sensibility, and lower misclassification error

compared to SIMCA models. Moreover, PLS-DA was less sensitive to edge effects on spectra and poorly focused regions of particles. The approach was successfully tested on a seabed sediment sample from Roskilde Fjord, Denmark, showcasing the method's efficiency. This proposed method offers an automated and efficient approach for microplastic polymer characterization, abundance enumeration, and size distribution, thereby contributing to methods standardization with significant benefits.

In a recent study, Yan et al (2022) developed an ensemble model comprising of 7 ML models including support vector machine, K nearest neighbor, least discriminant analysis etc to identify MPs types from ATR-FTIR spectra. The Kedzierski and Jung Datasets have been used to assess the suggested ensemble learning approach. The findings demonstrate that, in terms of 5 metrics (Kappa score, F1 score, accuracy, recall, and precision). Their technique showed excellent performance with Each MP receiving a higher-class report and a clearer confusion matrix (i.e., less muddled categories). Moses et al. (2023) utilized a focal plane array (FPA) based micro-FTIR (μFTIR) to compare two widely used data analysis algorithms concerning the abundance, polymer composition, and size distributions of MPs derived from selected environmental water samples in the size range of 11–500 μm. The two algorithms under investigation were: (a) the siMPle analysis tool (systematic identification of MicroPlastics in the environment) in combination with MPAPP (MicroPlastic Automated Particle/fibre analysis Pipeline), and (b) the BPF (Bayreuth Particle Finder). The findings of the study revealed a

generally good agreement between the two algorithms, but certain discrepancies were observed, particularly concerning specific polymer types/clusters and the smallest MP size classes. This highlights the importance of conducting a detailed comparison of MP algorithms, as it is crucial for ensuring better comparability of MP data. By addressing these differences and potential limitations, researchers can enhance the accuracy and reliability of MP characterization, thus advancing the understanding of MP pollution in aquatic environments.

These studies have demonstrated the use of ML models for MPs identification, however, there are needs to develop systems that could identify MPs at different states (undegraded or aged). As a first step, in this research, we present a novel approach that harnesses the potential of unsupervised machine learning algorithms to automate the classification of undegraded and aged PET MPs using ATR-FTIR spectroscopic data. By employing unsupervised methods, our approach reduces the need for manual labeling of data, making it both cost-effective and adaptable to a wide range of microplastic samples. The ultimate goal of this study is to develop a robust and efficient tool that can accurately differentiate between undegraded and aged PET MPs, facilitating a comprehensive understanding of their prevalence in various environmental matrices.

2. Methodology

2.1. Samples preparation

Commercial plastic products such disposable water bottles made of PET, was used to obtain the MPs tested in this work (excluding the caps). The PET MPs preparation has been described in detailed in previous studies (Enyoh et al, 2022; 2023; Enyoh and Wang, 2022; 2023). The plastic waste was first cut into small pieces using stainless-steel scissors. Subsequently, the resulting pieces were further ground into fine particles using a High-Speed Blender. The grinding process involved cycles of blending with cooling intervals. The number of cycles varied based on the amount of plastic to be ground, with more than 10 cycles required for PET. After grinding, the particles were sieved to obtain PET MPs within a specific size range. The prepared PET MPs (500 µm) were then cleaned by soaking in methanol overnight, followed by drying in an oven and rinsing with ultrapure water.

The cleaned MPs were labeled as undegraded PET MPs. Further undegraded PET MPs were prepared by just putting the PET MPs in deionized water for 24 hrs at room temperature and using a pristine PET MPs. To age the PET MPs, they were treated with sulfuric acid (H_2SO_4) at an elevated temperature (60 ^oC). Additionally, artificial thermal aging was performed on PET MPs, which involved exposing the PET MPs to hydrogen peroxide (H_2O_2) and elevated temperature (60 ^oC) in an oven. The resulting aged MPs were washed and dried, and they were categorized as Aged PET MPs. To confirm the presence of undegraded and aged PET microplastics (MPs), a scanning electron microscopy (SEM) technique was employed using a Variable Pressure Scanning Electron Microscope (VP-SEM) SU-1510 with an

accelerating voltage of 15kV (Hitachi Ltd, Tokyo, Japan). The samples underwent preparation and were placed on a stud before undergoing sputter coating. Sputter coating involved applying a film approximately 15 nm thick using argon gas at a pressure of around 4 psi and a current of approximately 16 mA for a duration of 4 minutes. The sputter coating process was carried out using the E102 Ion Sputter (Hitachi Ltd, Tokyo, Japan). By employing SEM and sputter coating, a high-resolution images and surface information was obtained, allowing the verification of the presence and characteristics of undegraded and aged PET MPs with a greater level of detail.

2.2. Sample analysis- ATR-FTIR spectral acquisition

Attenuated Total Reflectance-Fourier-Transform Infrared (ATR-FTIR) spectroscopy is a widely employed method to assess changes in the functional groups of polyethylene terephthalate (PET) during environmental degradation (Chowdhury et al., 2022). For the analysis, the functional groups on the surface of the prepared microplastic particles (MPs) were determined using an ATR-FTIR system (JASCO FTIR-6100). Before analyzing the MPs, the instrument was blanked to ensure accurate measurements. The MP samples were securely attached to a KBr disc and placed in the FTIR instrument for measurement. The infrared spectrum was recorded in the range of 400–4000 cm^{-1} by averaging 64 scans at a resolution of 4 cm^{-1}. This process provided valuable molecular information about the composition of the PET microplastics, aiding in the investigation of environmental deterioration effects.

2.3. Machine learning (ML) development

The ML sequence of workflow for the classification of undegraded and aged PET MPs is illustrated in Figure 1. The entire process, including data processing, model training, and classification, was executed using the Python programming language within a Jupyter Notebook environment. The computations were conducted on a system equipped with a 64-bit Intel Core i5 vPro processor and 4 GB of RAM.

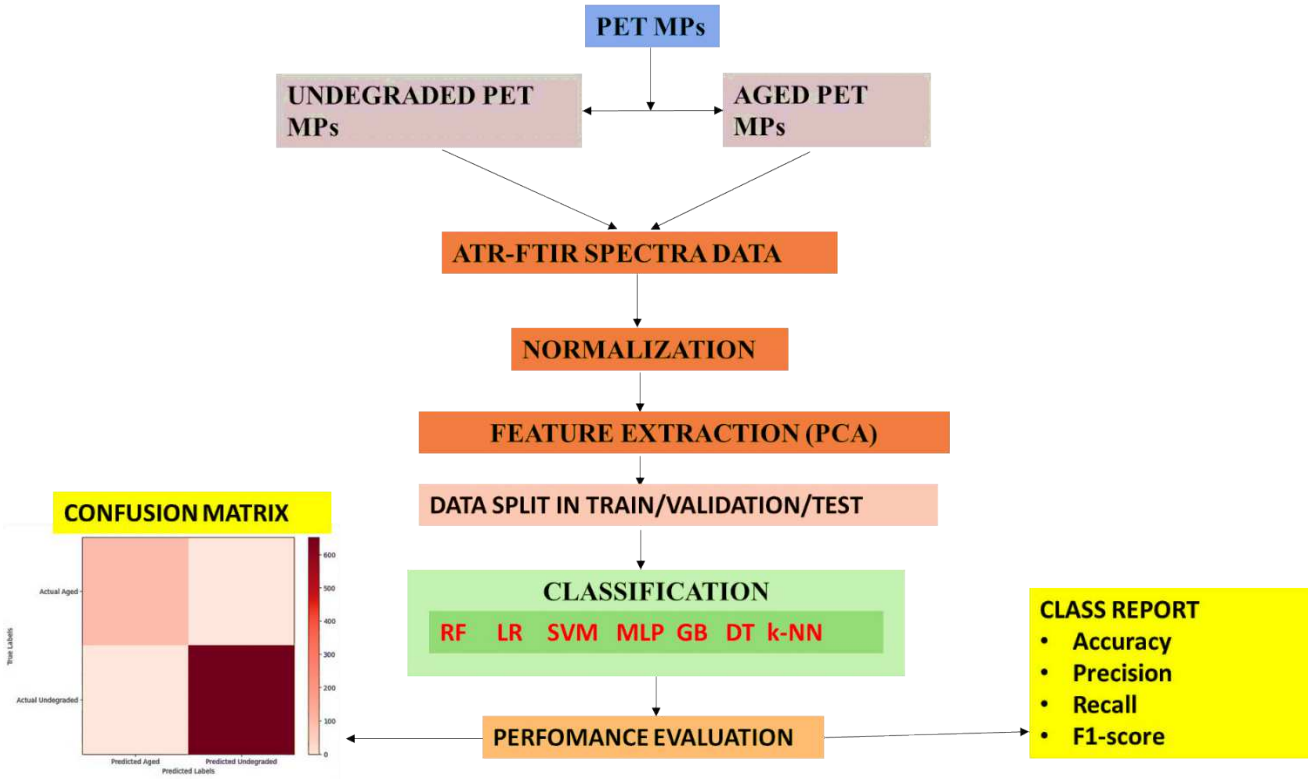


Figure 1: ML sequence of workflow for the classification of PET MPs

2.3.1. Data normalization

The ATR-FTIR spectra data typically consist of multiple columns representing different features, such as undegraded and aged spectra. However, these features may have different scales, which could potentially introduce bias

towards columns with larger values during the modeling process (Yan et al, 2022). To address this issue, the data was normalized using min-max scaling, also known as feature scaling or data normalization. Min-max scaling transforms each feature column in the dataset to a common range, typically between 0 and 1. The formula for min-max scaling is represented as follows:

$$\text{Scaled\_value} = \frac{\text{original\_value} - \text{min\_value}}{\text{max\_value} - \text{min\_value}} \quad (1)$$

Where: original_value is the value of a data point in the feature column; min_value is the minimum value of the feature column; max_value is the maximum value of the feature column and scaled_value is the resulting normalized value for the data point, which lies between 0 and 1.

By applying this transformation, all the feature columns, including undegraded (0) and aged spectra (1), were scaled to the same range, and further splitting the data into features X and target labels y. The X DataFrame contains the numerical features (0 and 1), and the y Series contains the corresponding numerical class labels (undegraded and aged). This makes them comparable and preventing any bias due to different scales. Normalizing the data in this manner ensures that each feature contributes equally to the machine learning model and improves the model's performance and convergence during training.

2.3.2. Feature Selection

For the feature selection step, the principal component analysis (PCA) an unsupervised ML technique was employed on the previously scaled data. PCA

is a widely used dimensionality reduction method that aims to transform the original features in a multivariate dataset (in this case, ATR-FTIR spectral data) into a new set of uncorrelated variables called principal components (PCs) (Enyoh et al, 2023b). These PCs are ordered in such a way that the first component captures the most significant variance in the data, the second component captures the second most significant variance, and so on. The primary objective of using PCA for feature selection is to reduce the number of dimensions (features) while preserving the most important information in the data (Da Silva et al., 2020). The mathematical model for PCA decomposition is shown in Equation (2):

$$X = TP^T + E \quad (2)$$

In this equation, X represents the measured spectral data (sample by wavenumber), T is the score matrix (sample by component), P^T is the loading matrix (component by wavenumber), and E is the residuals (unexplained data, sample by wavenumber).

By applying PCA to the scaled data, the number of dimensions is reduced to a specified number of principal components (in this case, 2 components). The selection of the number of components is based on the analyzed explained variance ratios provided by PCA. Once the PCs are obtained, they serve as the new feature set for constructing ML classification models.

2.3.4. Data Splitting

After performing data normalization and feature selection using PCA, the next step in the modeling process involved splitting the spectral dataset into two separate parts by using a train_test_split function from Sklearn: the training data and the testing data. This division was done to create a clear distinction between the data used to train the machine learning model and the data used to evaluate its performance. The split was done in a specific ratio to ensure an effective and reliable evaluation of the model's generalization capabilities.

The dataset was divided into two subsets:

1. Training Data: This subset comprised 80% of the original spectral dataset. The training data was used to train the ML model, allowing it to learn the underlying patterns and relationships present in the data. During training, the model adjusted its internal parameters to minimize the prediction errors and optimize its performance on the training data.
2. Testing Data: The remaining 20% of the spectral dataset formed the testing data. This independent subset served as a previously unseen sample for the model. After training, the model was evaluated on this testing data to assess its performance in predicting the target labels (e.g., undegraded or aged spectra) for new, unseen samples. The testing data acted as a simulation of real-world scenarios where the model encounters new observations that it has not seen during training.

By splitting the dataset into training and testing subsets, the model's ability to generalize and make accurate predictions on unseen data was assessed.

This process is crucial to determine if the model has learned patterns that can be applied to new, unseen data without overfitting or underfitting.

2.3.5. ML data training and testing Methodology for PET MPs classification

Seven (7) ML models Random Forest classifier (RF), Logistic Regression (LR), Support Vector Machines classifier (SVM), Neural Networks based on multilayer perceptron classifier (MLP), Gradient Boost (GB), Decision Trees (DT) and k-Nearest Neighbor (k-NN) were evaluated for classifying the undegraded and aged PET MPs in this study.

RF

The RF classifier is composed of multiple DTs. When making a new classification, each DT independently provides a classification for the input data. The RF algorithm then evaluates these classifications and selects the final prediction based on the class that receives the most votes from the individual trees (Mao and Wang, 2012; Cinar and Koklu, 2019). RF is particularly efficient in handling datasets with a large number of variables (Enyoh et al., 2023a). The simplified equation for the RF, as represented by equation 3, is as follows:

$$RF_{(x)} = \text{mode}(DT_1(x), DT_2(x), \ldots, DT_n(x)) \quad (3)$$

Here, RF(x) represents the class prediction made by the RF for a given input instance x. The mode function selects the most frequently occurring class prediction from the individual decision trees DT_1, DT_2, ..., DT_n, where n is the number of trees in the forest. Based on a randomly selected portion of the

training data, each decision tree in the RF is built individually. A random selection of predictor variables is also taken into account for partitioning the data at each node of the tree.

LR

LR primary purpose is to elucidate the relationship between these dependent and independent variables. To achieve this, LR fits the weights of the input variables to the training data, aiming to minimize the discrepancy between the predicted probabilities and the actual class labels (Cruyff et al., 2016). The simplified equation for logistic regression, represented as equation (4), is as follows:

$$y = \frac{1}{(1 + e^{(-z)})} \quad (4)$$

where the variable "y" denotes the predicted output or the probability of a specific class. This probability is obtained by passing the linear combination of the input variables and their respective weights, represented by "z," through the sigmoid function. The sigmoid function transforms any real-valued number to a value within the range of 0 to 1, enabling the interpretation of the output as a probability. This property makes logistic regression suitable for tasks where the prediction is associated with a probability score, allowing for a more nuanced understanding of the model's predictions.

SVM

SVM is a fundamental technique used for both classification and regression tasks. It creates a hyperplane that aids in distinguishing between different

classes or predicting numerical values. In two-dimensional space, SVM achieves linear separation, while in three-dimensional space, it uses a planar separation. In multidimensional space, it relies on a hyperplane for effective separation of data points (Schölkopf et al., 2001). The classification process in SVM involves identifying the optimal hyperplane that maximizes the margin between different classes. The larger the margin, the better the separation and generalization of the model (Cinar and Koklu, 2019). The simplified form for the predicted output from SVM, represented by equation (5), is as follows (Enyoh et al, 2023):

$$y(x)_{pre} = \sum_{i=1}^{n} a_i.K(x_i,x_j) + b \qquad (5)$$

where K(xi ,xj) is the radial basis function kernel. αi and b denote Lagrange multiplier and threshold parameter, respectively.

MLP

In this study, we further utilized a popular artificial neural network (ANN) known as the Multilayer Perceptron (MLP). The MLP learns through a technique called backpropagation, where weights are adjusted either after analyzing the entire dataset or after each individual data point. The architecture of the MLP involves organizing neurons into layers, with a hidden layer situated between the input and output layers. Depending on the complexity of the problem, an MLP can consist of multiple hidden layers. The input layer captures information about the problem to be addressed, while the output layer produces the final results or predictions. The study's findings

and data processing within the network are conveyed through this output layer (Enyoh et al., 2023). The equation (6), in which f is the activation function, N is the number of inputs per neuron, and k is the layer (hidden, output), may be used to represent the ANN system in its simplest form (Enyoh et al, 2023a).

$$Y_j^{k+1} = f(\sum_{i=1}^{N} X_i^k w_{ij}^k + b_i^k) \quad (6)$$

In this research, the model was configured with 100 hidden layers, and the activation function used was the Rectified Linear Unit (ReLU). ReLU, represented by the function f(x) = max(0, x), introduces non-linearity to the model and effectively addresses the issue of vanishing gradients. It is a widely adopted activation function in deep learning due to its popularity and effectiveness. For optimizing the training process, the Adam optimization algorithm was employed, and a random state of 42 was set. Adam is known for its adaptive learning rate strategy, which dynamically adjusts the learning rate during training. This adaptive approach scales the gradients based on their estimated first and second moments, resulting in faster convergence and improved performance when compared to traditional gradient descent algorithms. By using Adam, the model achieves faster convergence, allowing for better generalization to unseen data.

GB

The Gradient Boosting (GB) classifier is an ensemble learning technique that combines multiple weak learners, represented by Decision Trees (DTs), to

create a robust and accurate model (Hastie et al, 2009). The algorithm follows an iterative process, where it gradually adds new DTs to the ensemble. Each subsequent tree focuses on reducing the errors made by the previous trees (Hastie et al, 2009). During each iteration, the algorithm calculates the gradient of the loss function concerning the predicted values and constructs a new tree to minimize this gradient (Piryonesi et al, 2021). The predictions from all the trees in the ensemble are then combined to make the final prediction. The simplified equation for the Gradient Boosting classifier is the sum of weak learners, where each weak learner compensates for the errors made by the preceding learner. It can be expressed as shown in equation (7).

$$y(x) = y0(x) + \eta * g1(x) + \eta * g2(x) + \ldots + \eta * gn(x) \quad (7)$$

Where y(x) is the predicted output (whether undegraded or aged), y0(x) is the initial prediction, g1(x), g2(x), …, gn(x) are the weak learners (usually decision trees), and η is the learning rate (in this case = 0.1). At the beginning, y(x) is initialized with y0(x), which is the mean or median value of the target variable.

DT

DT is often visualized as a tree diagram, where each branch and node represent a classification query. The root node stands for an attribute, and the inner nodes indicate tests or evaluations of properties. The branches depict the outcomes of these evaluations, leading to the final decision represented by the leaf nodes, which correspond to the classes (Enyoh et al., 2023; Rokach and Maimon, 2005). DT offers several advantages, making it

well-suited for handling complex problems and providing inferences in the form of logical classification rules (Cinar and Koklu, 2019). Its distinct advantages include ease of implementation, seamless integration into databases, and high reliability (Wu et al., 2008). In its simplified form, a Decision Tree can be expressed as shown in equation (8).

$$y_{(x)} = (x_1, x_2, x_3 \ldots\ldots\ldots x_n y) \quad (8)$$

Where y is the target variable for classifying (undegraded or aged). The vector x is composed of the features, $x_1, x_2, \ldots$ etc., that are used for that task.

k-NN

k-NN is a popular and widely used machine learning model, especially for large-scale training datasets. It operates based on a distance metric to identify the most similar data points in the training set (Ibeto et al., 2021). In the k-NN algorithm, each data point is conceptually plotted in a multi-dimensional space, where each axis represents a different variable or feature. When a new data point needs to be classified (the test data), the algorithm compares it with all the available data points in the training set. The test data will have several neighbors that are close to it in terms of all the measured characteristics. To determine the class of the test data, the algorithm selects the k nearest data points based on the distance metric. The class with the majority of data points among these selected neighbors is assigned to the test data (Richman, 2011).

In this specific study, the k value, representing the number of nearest neighbors to consider, was chosen as 5. This means that when classifying new

data points, the algorithm will look at the class labels of the 5 nearest neighbors to make the final prediction.

2.4. Model Optimization

The ML model was optimized using cross-validation to avoid overfitting and improve its performance. In this specific case, the k-fold cross-validation process is performed with 5 folds, meaning the dataset is divided into 5 equal parts, and the model is trained and tested 5 times, each time using a different fold as the test set and the remaining four folds as the training set (Figure 2).

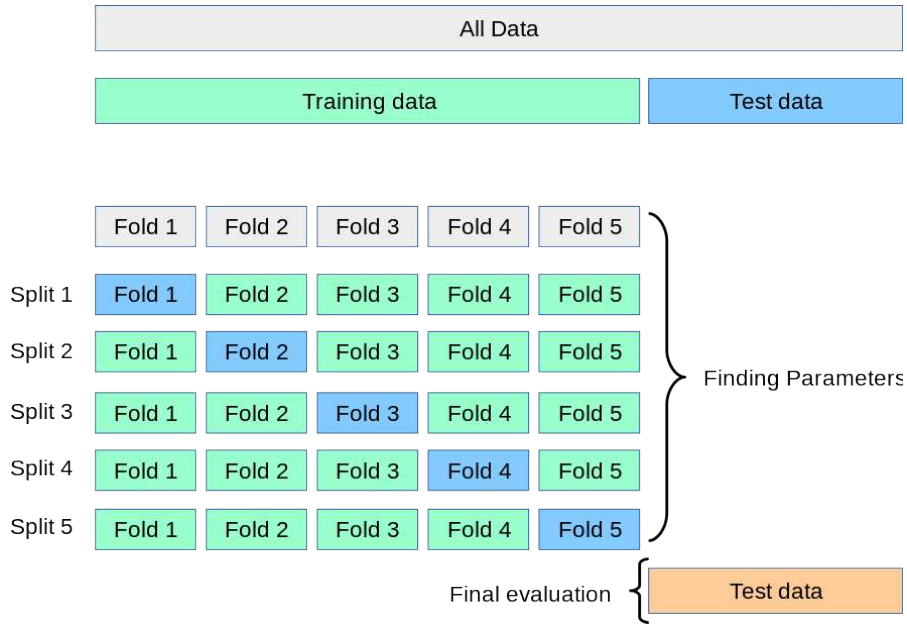


Figure 2. 5-fold cross-validation applied in this study (Adapted from https://scikit-learn.org/stable/modules/cross_validation.html, assessed 28/07/2023)

The accuracy values range from approximately 93.33 % to 100 % (Table 1). The accuracy measures the proportion of correctly classified samples over the total number of samples in each fold. The cross-validation process helps

to assess the model's performance on different subsets of the data, providing an estimate of how well the model generalizes to unseen data. In this case, the model's performance is consistently high in Folds 2, 3, and 4, achieving perfect accuracy (100 %), meaning it correctly classified all samples in these folds. In Folds 1 and 5, the accuracy is slightly lower (93.33 % and 99.47 %, respectively), but still relatively high, suggesting that the model is performing well on different subsets of the data.

Table 1. The reported cross-validation (CV) scores obtained in each of the 5 folds for all ML algorithms

Folds	**Accuracy (%)**						
	RF	**LR**	**SVM**	**MLP**	**GB**	**DT**	**KNN**
1	93.33	100	97.19	99.47	93.45	93.45	97.99
2	100	100	100	100	100	100	100
3	100	100	100	100	100	100	100
4	100	100	100	100	100	100	100
5	94.81	94.78	94.91	96.52	94.64	94.64	94.78

2.5. Model Evaluation metrics

To evaluate the model's performance on the testing data to see how well it generalizes to unseen samples, common evaluation metrics for classification

tasks such as confusion matrix, accuracy, precision, recall, and F1-score was computed. By utilizing these metrics, a comprehensive understanding of a model's strengths and weaknesses in classifying undegraded and aged PET MPs accurately is obtained. Additionally, a learning curve of the different ML were also evaluated.

2.5.1. Confusion Matrix (CM)

A confusion matrix (CM) is a table that summarizes the performance of a classification model. It presents the actual class labels (undegraded and aged) against the predicted class labels (undegraded and aged). The confusion matrix includes four key terms:

- True Positives (TP): The number of positive instances correctly classified as positive.
- False Positives (FP): The number of negative instances incorrectly classified as positive.
- True Negatives (TN): The number of negative instances correctly classified as negative.
- False Negatives (FN): The number of positive instances incorrectly classified as negative.

The confusion matrix helps to visualize the model's performance across different classes and serves as the foundation for calculating accuracy, precision, recall, and F1-score.

2.5.2. Accuracy

Accuracy measures the proportion of correctly classified instances over the total number of instances in the dataset. Mathematically, accuracy is defined as in equation (10):

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (10)$$

2.5.3. Precision

Precision quantifies the ability of the model to correctly identify positive instances among the instances predicted as positive. It focuses on minimizing false positives. The precision is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (11)$$

2.5.4. *Recall (Sensitivity or True Positive Rate)*

Recall evaluates the model's ability to correctly identify positive instances out of all the actual positive instances in the dataset. It focuses on minimizing false negatives. The recall is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (12)$$

2.5.5. F1-score

The F1-score is the harmonic mean of precision and recall and provides a balanced assessment of the model's performance. It takes into account both false positives and false negatives. The F1-score is calculated as:

$$\text{F1 - score} = \frac{2 \times \text{Precision} + \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \quad (13)$$

3. Results and discussion

3.1. Description of the ATR-FTIR spectral and SEM

Figure 3 presents the ATR-FTIR spectra for both undegraded and aged PET microplastics (MPs). The spectra indicate the principal bands corresponding to various functional groups in each material. According to standard PET spectroscopy, the bands at 1000 cm^{-1}, 1099 cm^{-1}, 1701 cm^{-1}, 2925 cm^{-1}, and 3400 cm^{-1} in the undegraded PET MPs correspond to aromatic -CH vibrations, O-C-C, C = O (carbonyl), -C-C- (alkyl), and -OH groups in the PET structure. These bands provide evidence of the chemical structure specific to undegraded PET (Chowdhury et al, 2022). However, it is noteworthy that the undegraded 2 (pristine PET MPs) showed no -OH group in its spectra. The presence of broad -OH peaks in the spectra of other undegraded PET MPs could be attributed to treatment processes, such as using methanol to remove additives or exposure to water. After aging, the spectra indicate that the major functional groups in the PET are retained, but there is a slight change in the spectrum for both undegraded and aged PET MPs. The band strength at 1099 cm^{-1} decreased in the aged PET MPs due to the thermal aging process that the pristine PET MPs underwent. The thermal treatment dispersed the PET's long chain backbone into smaller fragments, leading to the formation of an interactive cross-linked reactive PET product, primarily associated with the -O-C-C- group present in undegraded PET MPs. The cross-linking process generates radicals, and when these radicals' peroxide, peroxy radicals are formed. During the termination stage, the generated peroxy radical interacts with additional free radical PET monomers. As a result, the band at 1099 cm^{-1},

related to pristine PET MPs, is nearly lost. This observation confirms that the FTIR-ATR spectra support the alteration and aging of the PET MPs.

To provide additional confirmation of the aging process in comparison to undegraded PET MPs, SEM was performed, as depicted in Figure 3c. The SEM images distinctly display a noticeable difference in the surface characteristics of the undegraded and aged PET MPs. In the SEM images, the gaps between the strands of the aged PET MPs appear expanded when compared to the undegraded PET MPs. This observation serves as strong evidence confirming the occurrence of aging in the PET MPs. The SEM analysis visually confirms the structural changes that have taken place on the surface of the aged PET MPs, further corroborating the findings of the aging process.

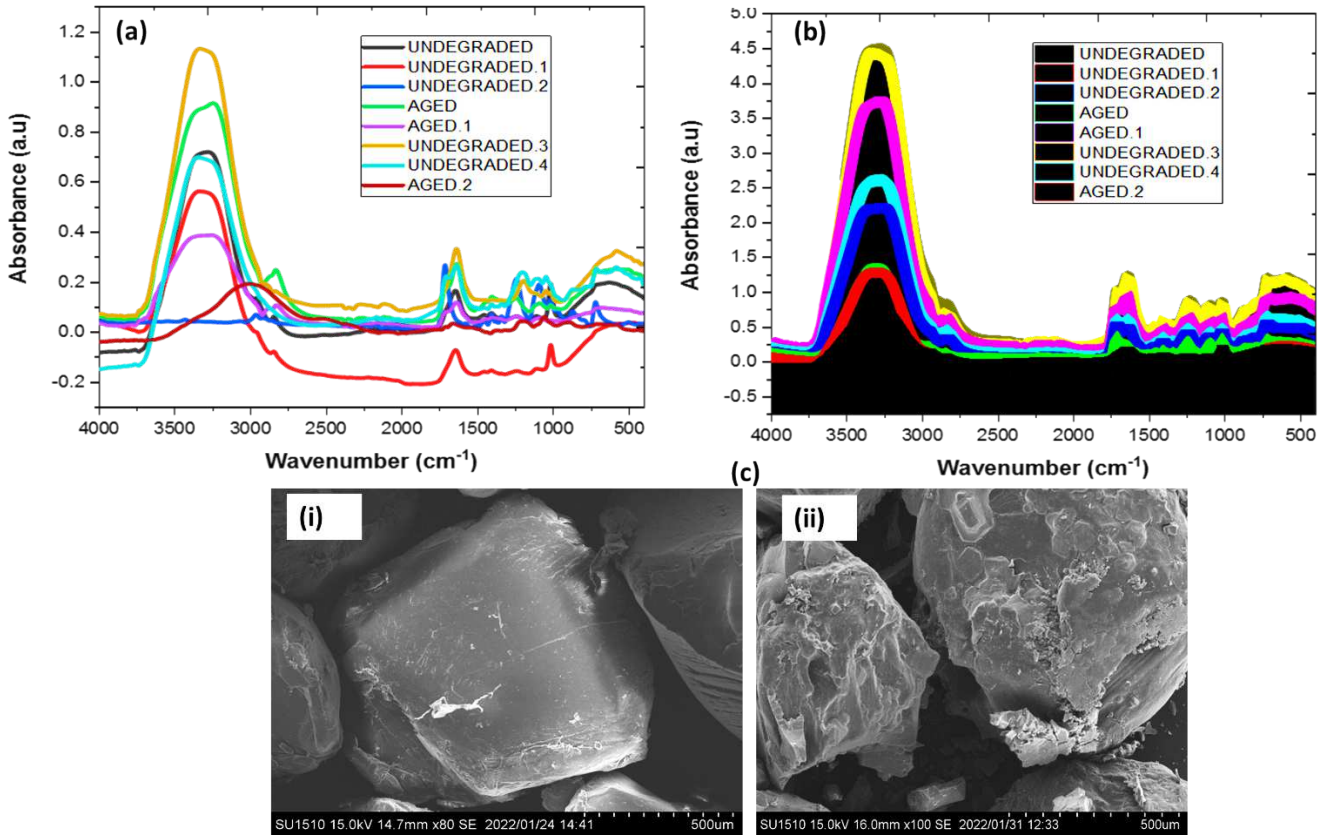
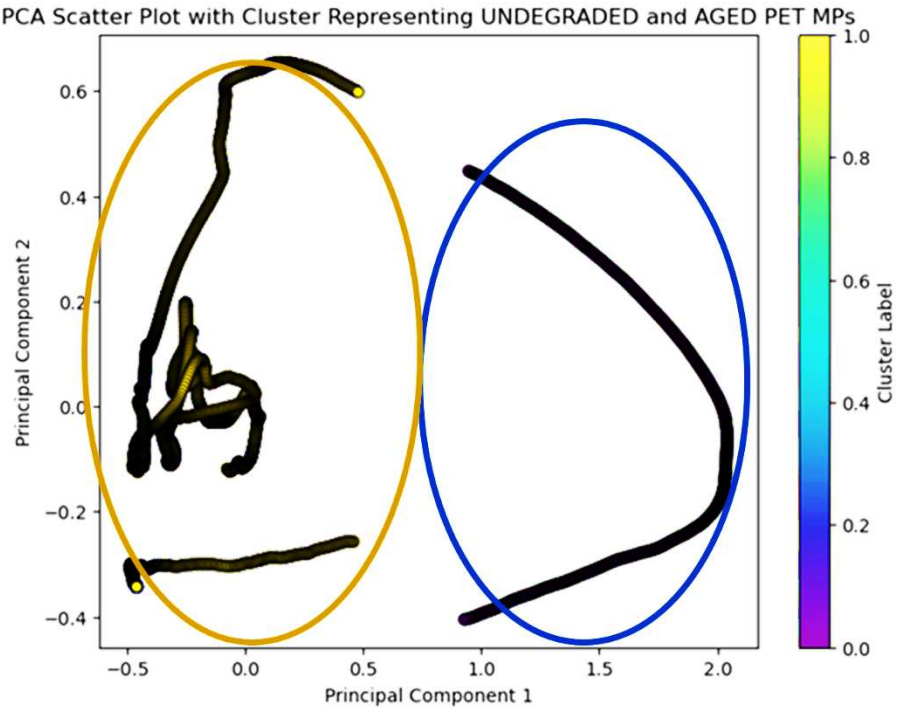


Figure 3. (a) Pre-processed spectra ATR-FTIR spectra of the undegraded and aged PET MPs, (b) the dataset used for PCA and (c) surface morphology of the degraded and aged PET MPs

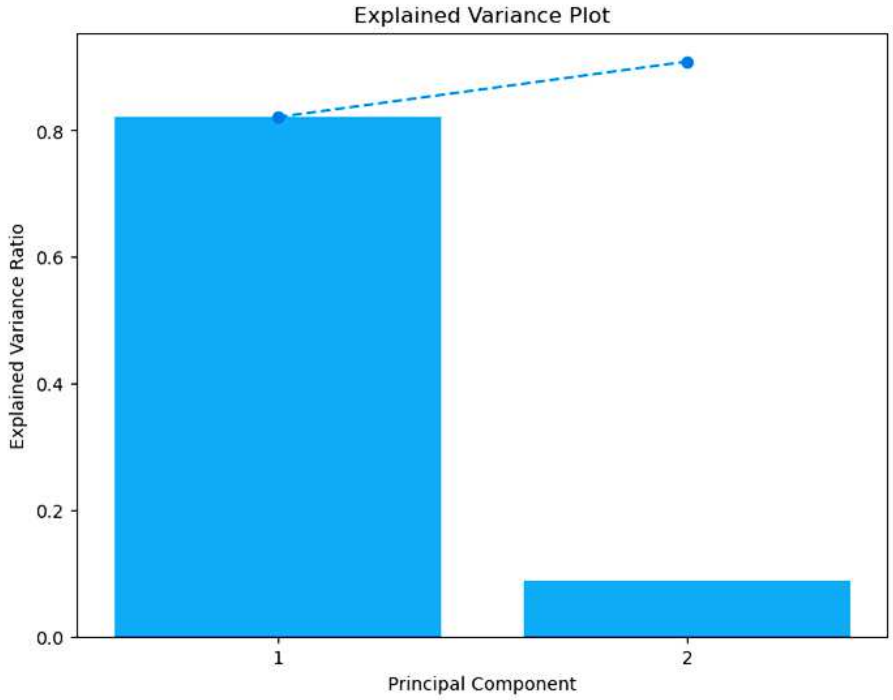
3.2. Exploratory analysis (PCA) of the normalized ATR-FTIR spectra

The application of Principal Component Analysis (PCA) in this study served as a feature selection technique to reduce the dimensionality of the dataset. As a result, two principal components (PC 1 and PC 2) were extracted from the original data. These principal components captured the most important information from the dataset and allowed for a more concise representation of the samples. In the PCA plot (Figure 4), the samples were visualized based on their scores in the PC 1 and PC 2 axes. It was observed that the undegraded PET MPs were predominantly clustered together in one group, while the aged PET MPs formed a separate cluster. This segregation of samples based on their respective clusters indicates that the PCA successfully identified underlying patterns and distinctive characteristics of each group. Furthermore, the explained variance plot revealed that 80% of the data's variability was captured in PC 1, while PC 2 accounted for 10% of the variability (Figure 4b). This indicates that PC 1 carries a substantial amount of information, making it the most significant component for discriminating between undegraded and aged PET MPs. PC 2, although explaining less variance, still contributes valuable information for differentiation between the two groups. In general, the PCA analysis with the two extracted principal components demonstrated its efficacy in distinguishing undegraded and aged

PET MPs, providing an informative and concise representation of the dataset with a substantial proportion of the data's variability retained.



(a)



(b)

Figure 4. (a) The PCA plots in space from the normalized spectral for undegraded and aged PET MPs and (b) the explained variance plot for the principal components extracted

3.3. ML classification models evaluation

3.3.1 Confusion matrix (CM)

The confusion matrix (CM) is a matrix of numbers that provides valuable insights into how a ML model performs in classifying undegraded and aged PET MPs dataset. In Figure 5 and summarized in Table 2, the CM shows the predicted class labels on the x-axis and the true class labels on the y-axis for each ML algorithm.

By analyzing the CM, we can observe the percentage of correctly classified undegraded and aged PET MPs for each ML algorithm. It also reveals

instances where the model gets confused and misidentifies certain PET MPs as others. These misclassifications can be crucial in understanding the strengths and limitations of each ML approach in accurately classifying the MPs.

Among the ML models such as Random Forest (RF), Gradient Boosting (GB), Decision Tree (DT), and k-Nearest Neighbors (k-NN), most of the samples are correctly classified with no false positives (FP) expected (Table 2). However, Logistic Regression (LR), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) had a few samples i.e. 22, 7 and 4 respectively classified as false positives (FP). Additionally, these models had some samples i.e. 10, 6 and 2 respectively, that were expected to be of the positive class but were classified as the negative class, resulting in false negatives (FN).

Understanding the FP and FN rates is crucial as it helps to identify the areas where the ML models may have challenges in classification. By examining these misclassifications and the impacts of Principal Component Analysis (PCA) on the ML methods, informed decisions on model improvement and identify strategies to enhance the classification accuracy of undegraded and aged PET MPs can be made.

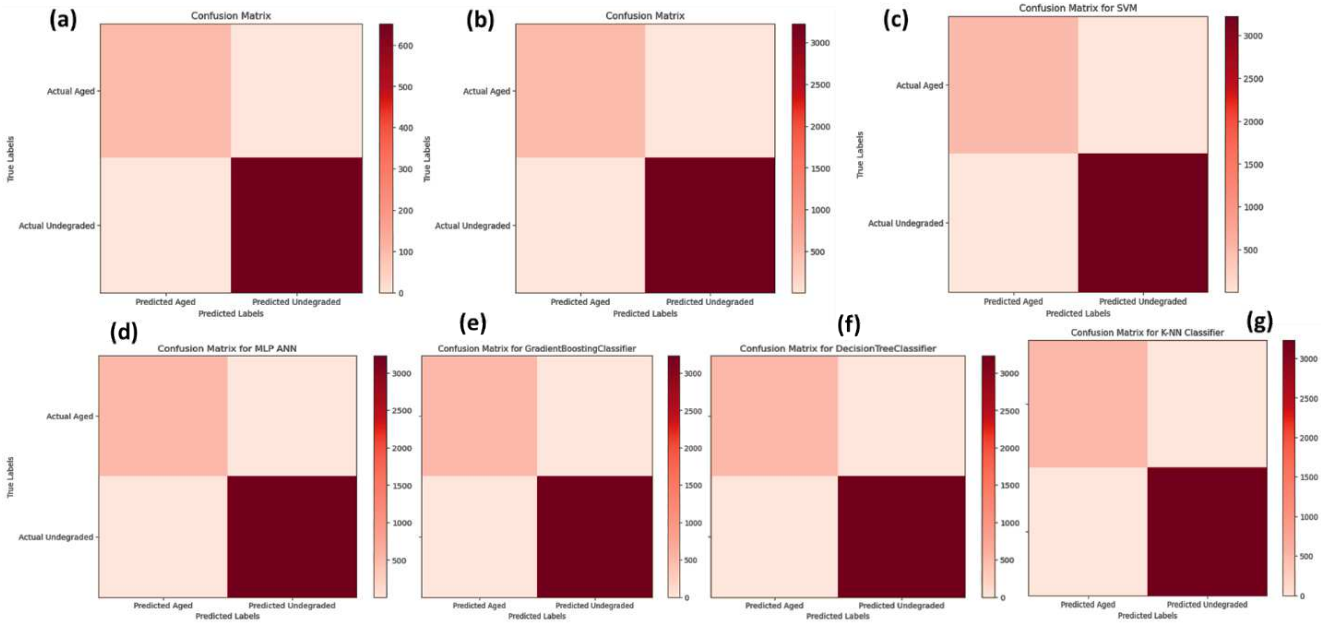


Figure 5. Confusion matrix plot for all ML classifiers (a) RF (b) LR (c) SVM (d) MLP (e) GB (f) DT (g) K-NN

Table 2. Complexity matrix of algorithms used

Algorithm	Confusion Matrix		
	Positive	Negative	
RF	92	0	Positive
	0	656	Negative
LR	481	22	Positive
	10	3223	Negative
SVM	496	7	Positive
	6	3227	Negative
MLP	499	4	Positive

	2	3231	Negative
GB	503	0	Positive
	0	3233	Negative
DT	503	0	Positive
	0	3233	Negative
k-NN	503	0	Positive
	0	3233	Negative

3.3.2. Class reports for all ML algorithms

In this study, a more comprehensive analysis of the ML algorithms' performance in classifying PET microplastics (MPs) is conducted. To evaluate the specific performance of different methods for undegraded and aged PET MPs classification, accuracy, precision, recall, and F1 score are used as performance metrics. The results of these classification performance measurements are presented in Table 3, providing valuable insights into the strengths and weaknesses of each ML approach in accurately classifying undegraded and aged PET MPs. Based on the evaluation of various performance metrics, it is evident that the ML models generally performed exceptionally well in classifying undegraded and aged PET microplastics (MPs). The accuracy of the models, ranging from 0.98 to 0.99 following 5-fold cross-validation, indicates their high ability to predict the correct classes.

Among the models, LR, MLP, and k-NN demonstrated the highest accuracy, achieving a remarkable 0.99. Additionally, RF, GB, DT, and k-NN showed

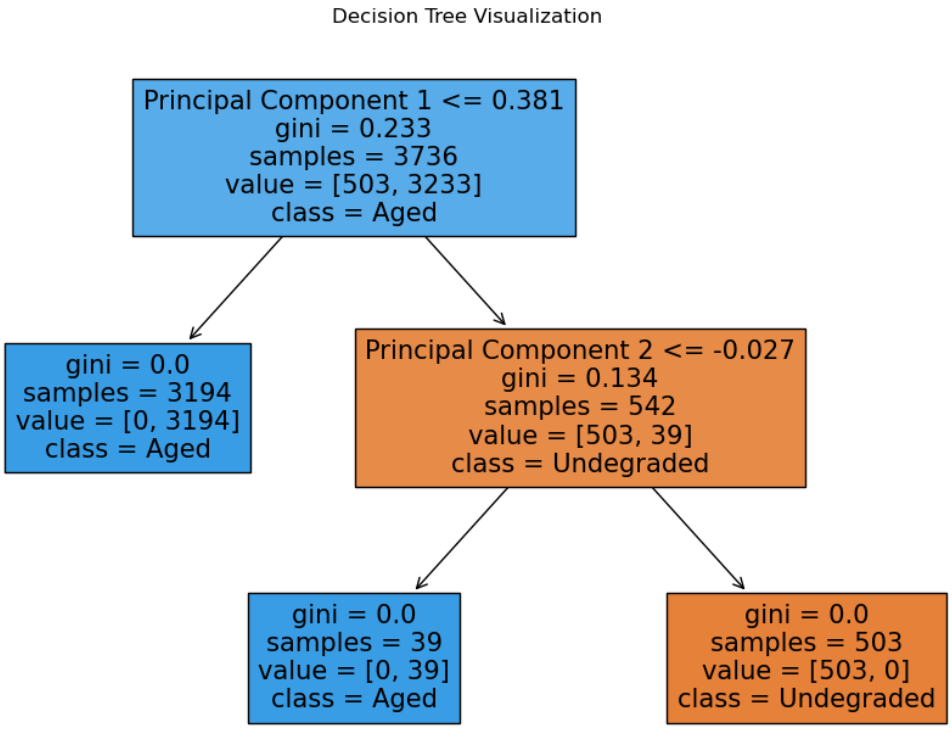
perfect precision, recall, and F1 scores (all equal to 1.00) for correctly classifying both undegraded and aged PET MPs. These models showcased excellent performance and robustness in distinguishing between the two classes, resulting in no false positives or false negatives. However, other ML models such as LR, SVM, and MLP, while still achieving high accuracy, showed metrics slightly below 1.00, especially when classifying aged PET MPs. This indicates that these models may have some challenges in accurately identifying aged PET MPs, as they had some misclassifications.

Based on the overall performance and considering the precision, recall, and F1 scores, RF, GB, DT (Figure 6a), and k-NN (Figure 6b) are the most favorable models for the classification of undegraded and aged PET MPs. These models demonstrated superior accuracy and robustness in correctly identifying both classes with no misclassifications. However, further examination and potential fine-tuning of LR, SVM, and MLP may be needed to improve their performance, particularly in classifying aged PET MPs.

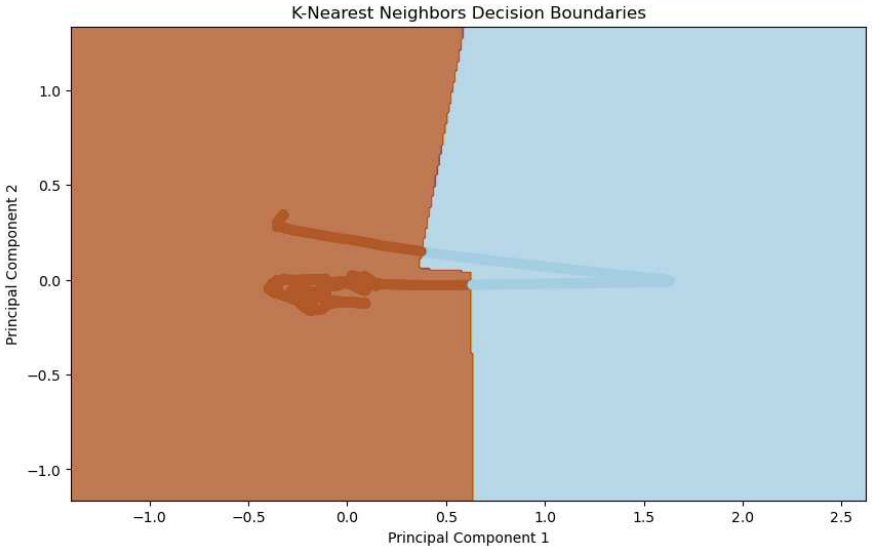
Table 2. Classification report for the different ML algorithms

Measure	**Class (0 =undegraded and 1 = aged)**	**RF**	**LR**	**SVM**	**MLP**	**GB**	**DT**	**k-NN**
Accuracy		0.98	0.99	0.98	0.99	0.98	0.98	0.99
Precision	0	1.00	0.98	0.99	1.00	1.00	1.00	1.00

	1	1.00	0.99	1.00	1.00	1.00	1.00	1.00
Recall	0	1.00	0.96	0.99	0.99	1.00	1.00	1.00
(Sensitivity)	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F1-Score	0	1.00	0.97	0.99	0.99	1.00	1.00	1.00
	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00



(a)



(b)

Figure 6. Visualization of (a) DT and (b) k-NN for classifying undegraded and aged PET MPs using ATR-FTIR spectral

3.3.3. Learning curves for ML algorithms

Interpreting learning curves is crucial in assessing the performance and behavior of a machine learning model during the training process. Learning curves depict the model's training and validation (or cross-validation) performance as a function of the training set size or the number of training iterations (Perlich, 2011).

The learning curves for best ML models is presented in Figure 7, which typically show the training set loss(accuracy) and cross-validation set loss on the y-axis and the number of training samples on the x-axis. The training set loss measures how well the model is fitting the training data, while the validation set loss evaluates the model's generalization performance on unseen data. The learning curves play a vital role in assessing the

performance and behavior of a machine learning model as the training set size increases. In this case, the learning curves clearly demonstrated positive signs of a well-behaved model. The decreasing and converging nature of both the training and cross-validation curves indicates that the model is learning from the data and improving its performance as more training samples are added. The fact that both curves decrease implies that the model is fitting the training data well, capturing the underlying patterns and trends. The convergence of the training and cross-validation curves is a positive indicator, as it suggests that the model is not suffering from significant overfitting. Overfitting occurs when a model memorizes the training data too closely and fails to generalize to new, unseen data (Ying, 2019). A small gap between the training and cross-validation curves implies that the model's performance on new data is similar to its performance on the training data, which is a desirable outcome. This finding indicates that the model is exhibiting good generalization capabilities, meaning it can make accurate predictions on data it has never seen before. The convergence of the learning curves suggests that the model is not just memorizing the training data but is learning to capture the underlying patterns that can be applied to new data. These ML models are likely to make accurate predictions on new, unseen data, making it a reliable and effective tool for the task at hand.

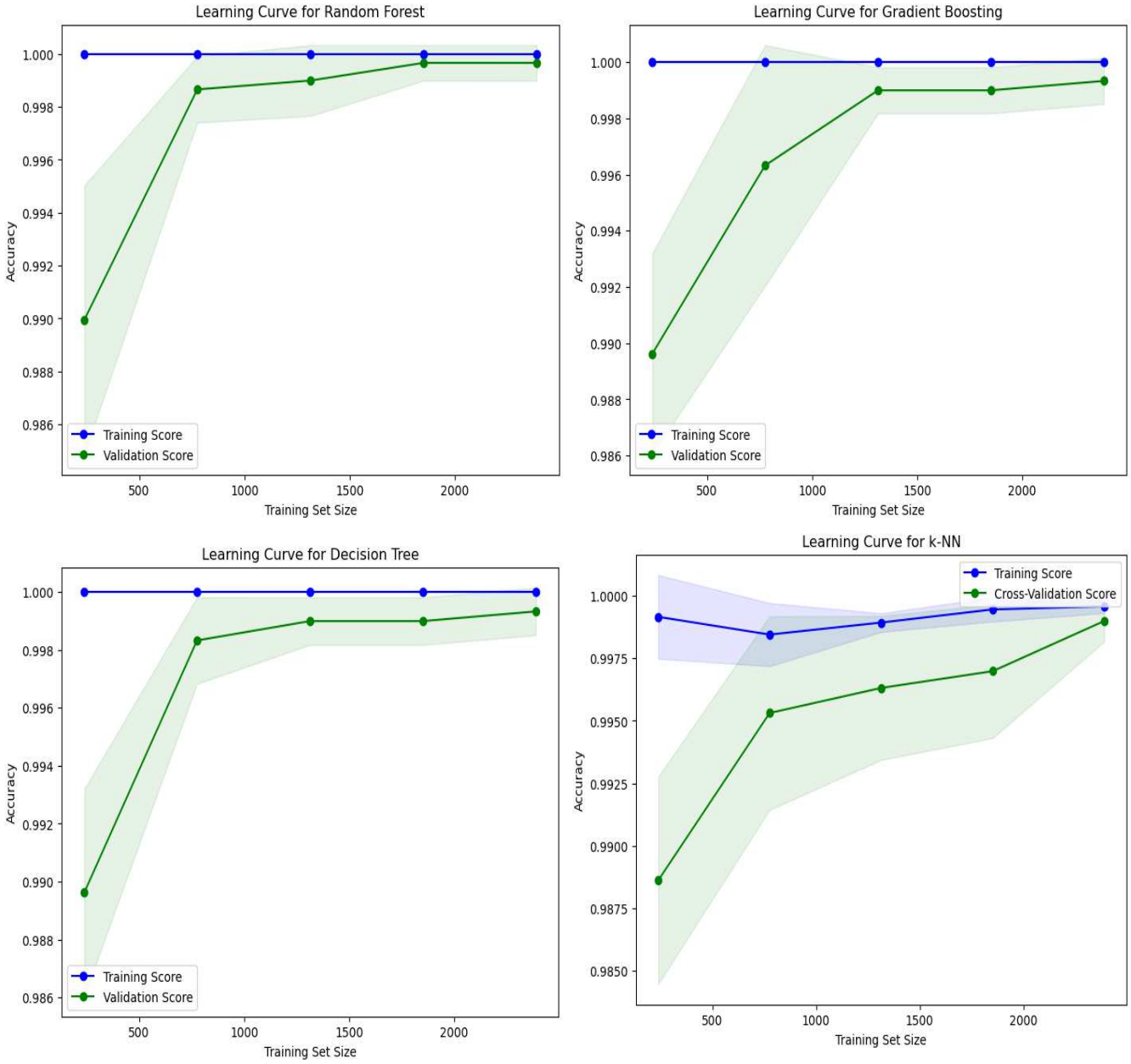


Figure 7. Learning curves for the best ML algorithms for the classification of undegraded and aged PET MPs.

4. Conclusion

In this study, we developed a machine learning approach utilizing ATR-FTIR spectral data to classify undegraded and aged PET microplastics (MPs) particles. Among the seven ML models evaluated, Random Forest (RF),

Gradient Boosting (GB), Decision Tree (DT), and k-Nearest Neighbors (k-NN) demonstrated the best performance with an impressive accuracy of 99% in classifying undegraded and aged PET MPs. These results showcase the significant potential of ATR-FTIR spectra in accurately distinguishing between undegraded and aged PET MPs particles. The proposed strategy not only enables effective classification but can also be adapted for use with various environmental samples. Furthermore, our method offers the advantage of automating the sorting process for microplastic particles, making it a valuable tool for standardizing methods. By optimizing spectra and extracting essential information from the data, our approach streamlines and enhances the classification process, providing more reliable and efficient results. Finally, the method's versatility and potential for method standardization make it a valuable contribution to the field of microplastic analysis in environmental research.

Contributions

C.E.E.: Conceptualization, Methodology, Software, Formal analysis, Validation, Visualization, Investigation, Data curation, Project administration, Writing- Original draft preparation, Writing- Reviewing and Editing. **W.Q***:* Supervision, Project administration, Funding acquisition, Resources, Writing- Reviewing and Editing.

Funding

This study was partially supported by the Special Funds for Basic Research (B) (No.22H03747, FY2022-FY2024) of Grant-in-Aid for Scientific Research of Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT).

References

Chowdhury, T., Wang, Q. & Enyoh, C.E. Degradation of Polyethylene Terephthalate Microplastics by Mineral Acids: Experimental, Molecular Modelling and Optimization Studies. J Polym Environ (2022). https://doi.org/10.1007/s10924-022-02578-z

Cinar, I, and Koklu, M. (2019). Classification of Rice Varieties Using Artificial Intelligence Methods. International Journal of Intelligent Systems and Applications in Engineering IJISAE, 2019, 7(3), 188–194.

Cruyff, M.; Böckenholt, U.; van der Heijden, P.G.M.; Frank, L.E.; Chaudhuri, A.; Christofides, C.T.; Rao, C.R. (2016). Handbook of Statistics, Volume 34. Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits, pp. 287 - 315, Elsevier. p. 287-315.

Da Silva, V. H., Murphy, F., Amigo, J. M., Stedmon, C., & Strand, J. (2020). Classification and Quantification of Microplastics (<100 µm) Using a Focal Plane Array–Fourier Transform Infrared Imaging System and Machine Learning. Analytical Chemistry, 92(20), 13724–13733. doi:10.1021/acs.analchem.0c01324

Enyoh C.E. and Wan, Q. (2022). Combined experimental and molecular dynamics removal processes of contaminant phenol from simulated

wastewater by polyethylene terephthalate microplastics, Environmental Technology, DOI: 10.1080/09593330.2022.2139636

Enyoh C.E., A.W. Verla, F.O. Ohiagu & E.C. Enyoh (2021). Progress and future perspectives of microplastic research in Nigeria. International Journal of Environmental Analytical Chemistry, https://doi.org/10.1080/03067319.2021.1887161

Enyoh C.E., Duru C.E., Prosper E., Wang Q. (2023). Evaluation of Nanoplastics Toxicity to the Human Placenta in Systems. Journal of Hazardous Materials 446:130600. DOI: 10.1016/j.jhazmat.2022.130600

Enyoh, C.E., Qingyue Wang , Prosper O. (2022). Response Surface Methodology for modeling the Adsorptive uptake of Phenol from Aqueous solution Using Adsorbent Polyethylene Terephthalate Microplastics. Chemical Engineering Journal Advances. DOI: 10.1016/j.ceja.2022.100370

Enyoh, C.E.; Wang, Q.; (2023) Adsorption and toxicity characteristics of ciprofloxacin on differently prepared polyethylene terephthalate microplastics from both experimental and theoretical perspectives. Journal of Water Process Engineering, 53, 103909. DOI: https://doi.org/10.1016/j.jwpe.2023.103909

Enyoh, C.E.; Wang, Q.; Momimul, R.H.; Senlin, L.; (2023b) Preliminary characterization and probabilistic risk assessment of microplastics and potentially toxic elements (PTEs) in garri (cassava flake), a common staple food consumed in West Africa Environmental Analysis Health and Toxicology 2023; 38(1): e2023005. https://doi.org/10.5620/eaht.2023005

Enyoh, C.E.; Wang, Q.; Senlin, L.(2023) Optimizing the Efficient Removal of Ciprofloxacin from Aqueous Solutions by Polyethylene Terephthalate Microplastics using Multivariate Statistical Approach. Chemical Engineering Science 278(12):118917:. DOI: 10.1016/j.ces.2023.118917

Hastie, T.; Tibshirani, R.; Friedman, J. H. (2009). 10. Boosting and Additive Trees. The Elements of Statistical Learning (2nd ed.). New York: Springer. pp. 337–384.

Hufnagl, B.; Steiner, D.; Renner, E.; Löder, M. G. J.; Laforsch, C.; Lohninger, H. A Methodology for the Fast Identification and Monitoring of Microplastics in Environmental Samples Using Random Decision Forest Classifiers. Anal. Methods. 2019, 11 (17), 2277–2285.

Ibeto C.N., C.E. Enyoh, A.C. Ofomatah, L.A. Oguejiofor, T. Okafocha & V. Okanya (2021): Microplastics pollution indices of bottled water from South

Eastern Nigeria, International Journal of Environmental Analytical Chemistry, DOI: 10.1080/03067319.2021.1982926

Ioakeimidis, C., Fotopoulou, K. N., Karapanagioti, H. K., Geraga, M., Zeri, C., Papathanassiou, E., … Papatheodorou, G. (2016). The degradation potential of PET bottles in the marine environment: An ATR-FTIR based approach. Scientific Reports, 6(1). doi:10.1038/srep23501

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

Mao, W. and F. Wang, *New advances in intelligence and security informatics*. 2012: Academic Press.

Moses, S.R., Roscher, L., Primpke, S. et al. Comparison of two rapid automated analysis tools for large FTIR microplastic datasets. Anal Bioanal Chem 415, 2975–2987 (2023). https://doi.org/10.1007/s00216-023-04630-w

Perlich, C. (2011). Learning Curves in Machine Learning. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_452

Piryonesi, S. Madeh; El-Diraby, Tamer E. (2021). Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling. Journal of Infrastructure Systems. 27 (2): 04021005. doi:10.1061/(ASCE)IS.1943-555X.0000602.

Richman, J.S. (2011), Multivariate neighborhood sample entropy: a method for data reduction and prediction of complex data, in Methods in enzymology. Elsevier. p. 397-408.

Rokach, L.; Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*. **35** (4): 476–487.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural computation, 13(7), 1443–1471. https://doi.org/10.1162/089976601750264965

Verla AW, Enyoh CE, Verla EN (2019). Microplastics, an emerging concern: a review of analytical techniques for detecting and quantifying microplatics. Anal Methods Environ Chem J. l2:15–32.

Verla, A.W., Enyoh, C.E., Verla, E.N. et al. (2019a). Microplastic–toxic chemical interaction: a review study on quantified levels, mechanism and implication. SN Appl. Sci. 1, 1400. https://doi.org/10.1007/s42452-019-1352-0

Wander, L.; Vianello, A.; Vollertsen, J.; Westad, F.; Braun, U.; Paul, A. Exploratory Analysis of Hyperspectral FTIR Data Obtained from Environmental Microplastics Samples. Anal. Methods. 2020, 12 (6), 781–791.

Wu, Xindong; Kumar, Vipin; Ross Quinlan, J.; Ghosh, Joydeep; Yang, Qiang; Motoda, Hiroshi; McLachlan, Geoffrey J.; Ng, Angus; Liu, Bing; Yu, Philip S.; Zhou, Zhi-Hua (2008). Top 10 algorithms in data mining. Knowledge and Information Systems. 14 (1): 1–37. doi:10.1007/s10115-007-0114-2

Yan X., Zhi, C., Alan, M., Yuansong, Q. (2022). An ensemble machine learning method for microplastics identification with FTIR spectrum. Journal of Environmental Chemical Engineering, 10, 4, 108130. https://doi.org/10.1016/j.jece.2022.108130

Ying X. (2019). An Overview of Overfitting and its Solutions. *J. Phys.: Conf. Ser.* **1168** 022022. **DOI** 10.1088/1742-6596/1168/2/022022

Figures

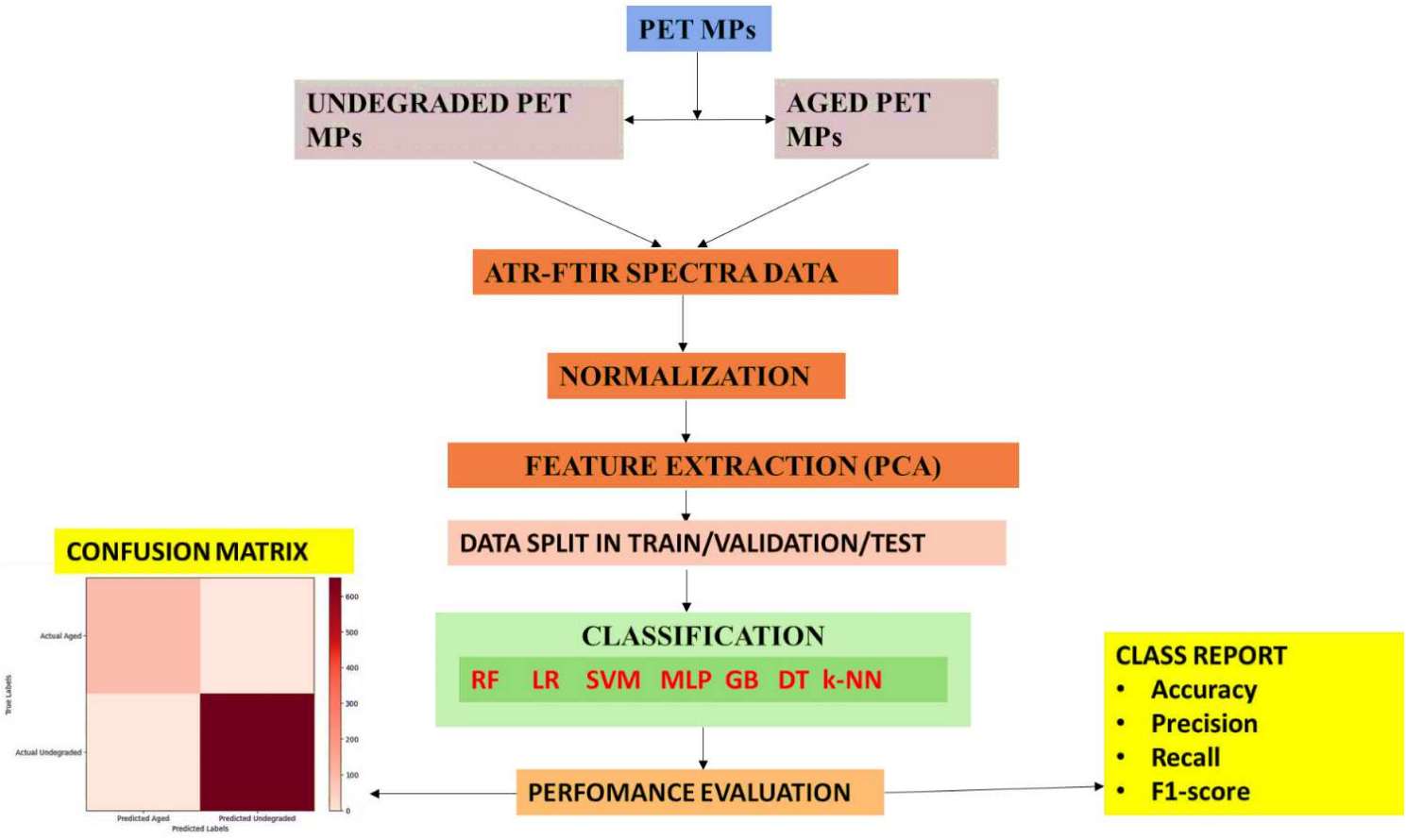


Figure 1

ML sequence of workflow for the classification of PET MPs

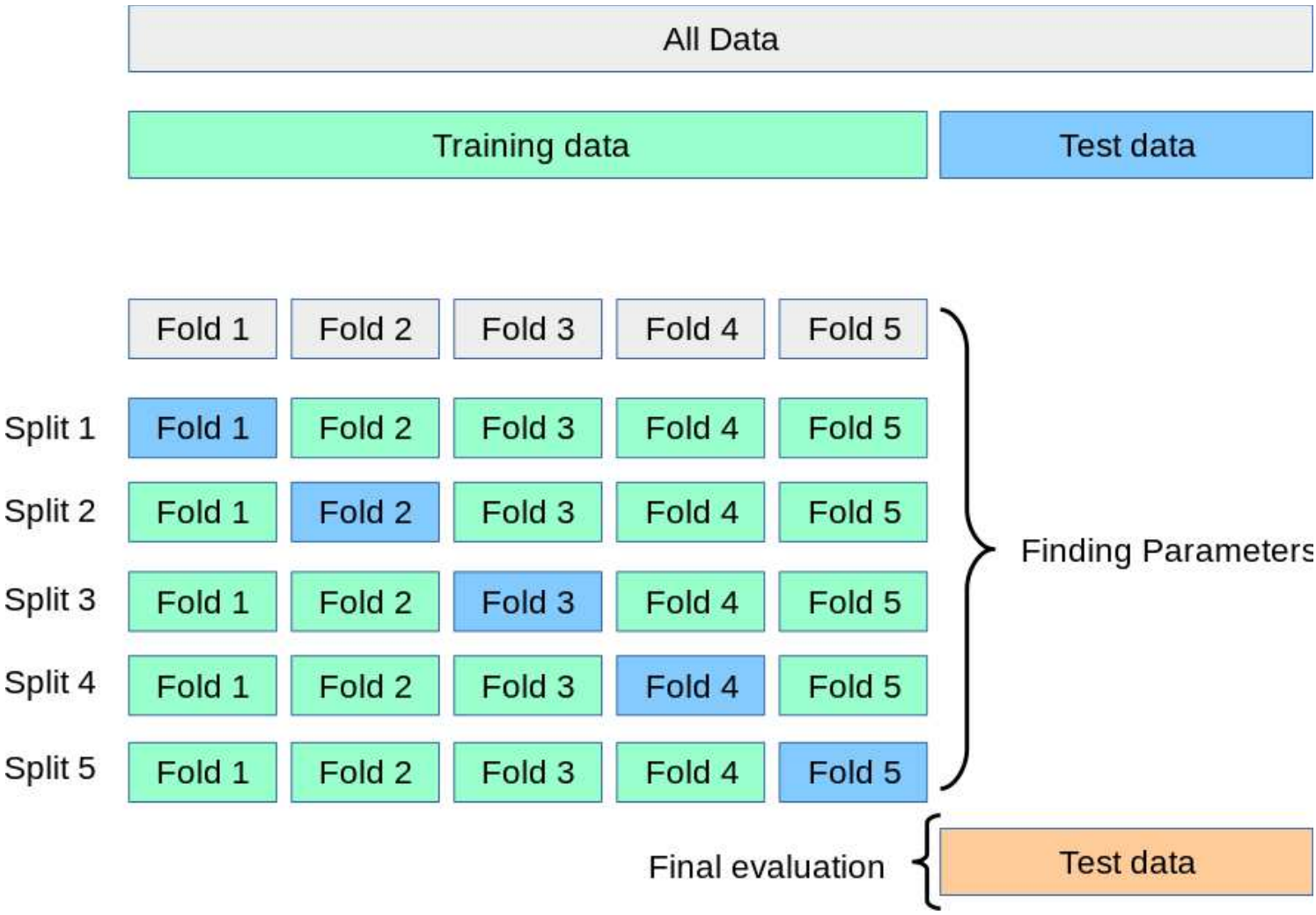


Figure 2

5-fold cross-validation applied in this study (Adapted from https://scikit-learn.org/stable/modules/cross_validation.html, assessed 28/07/2023)

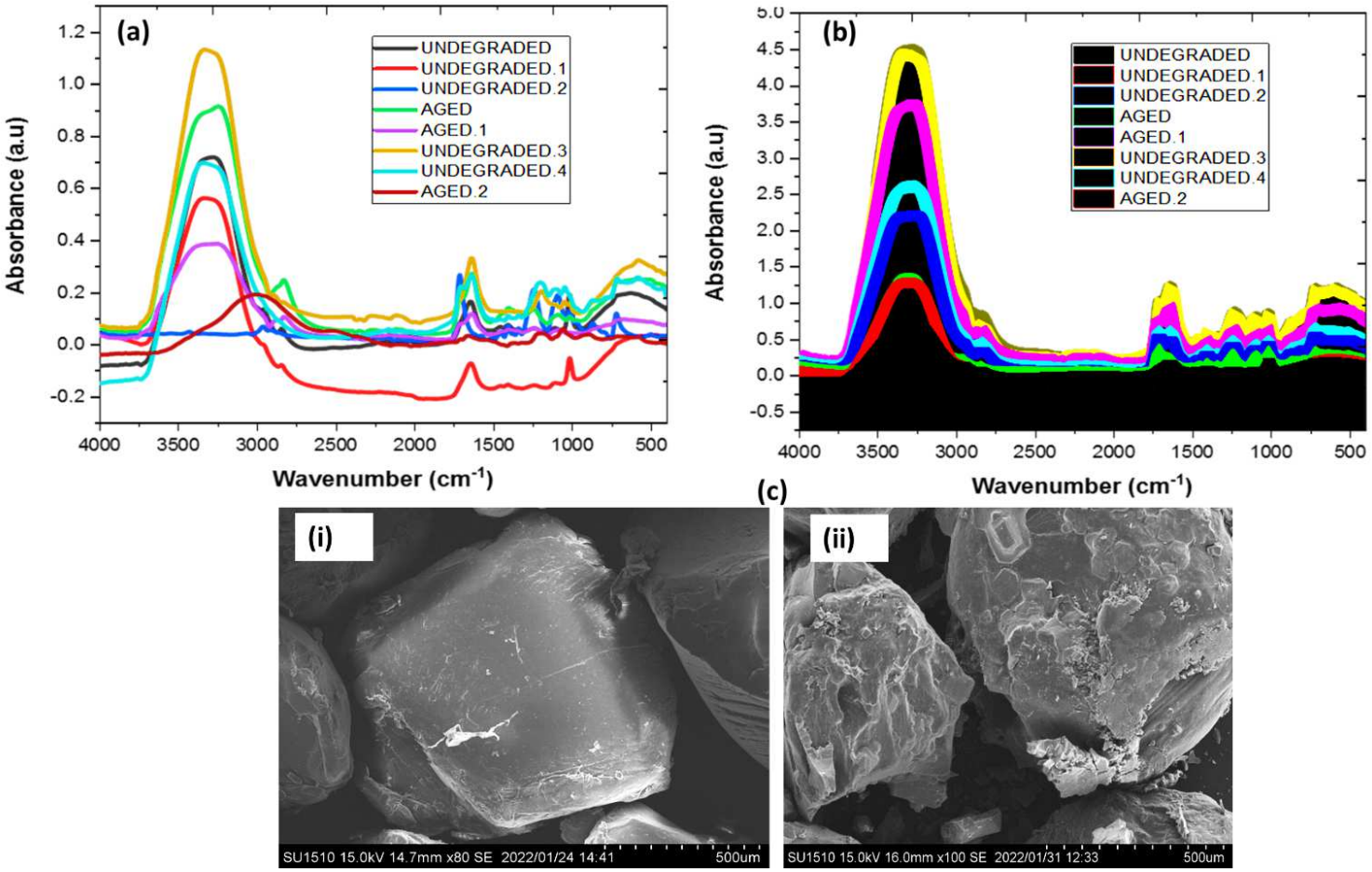
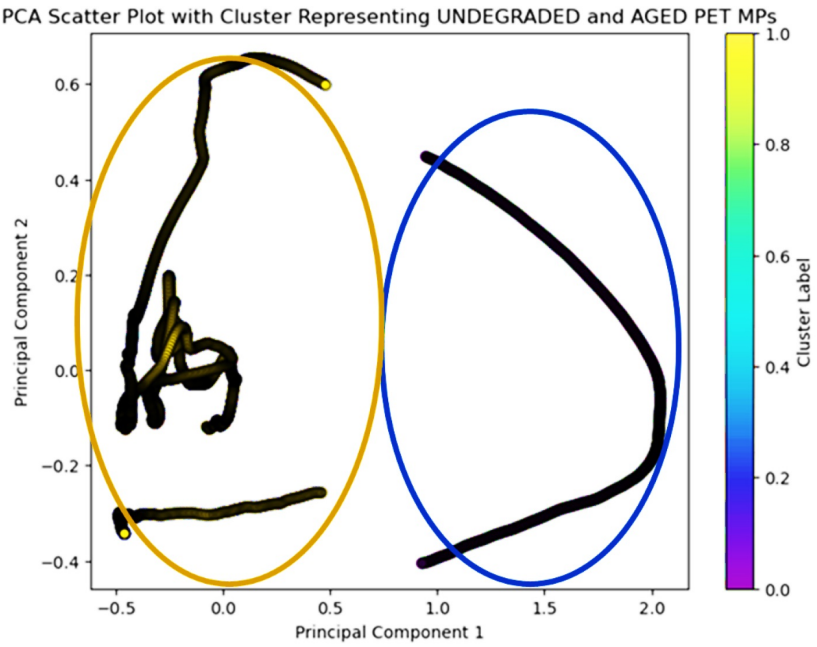
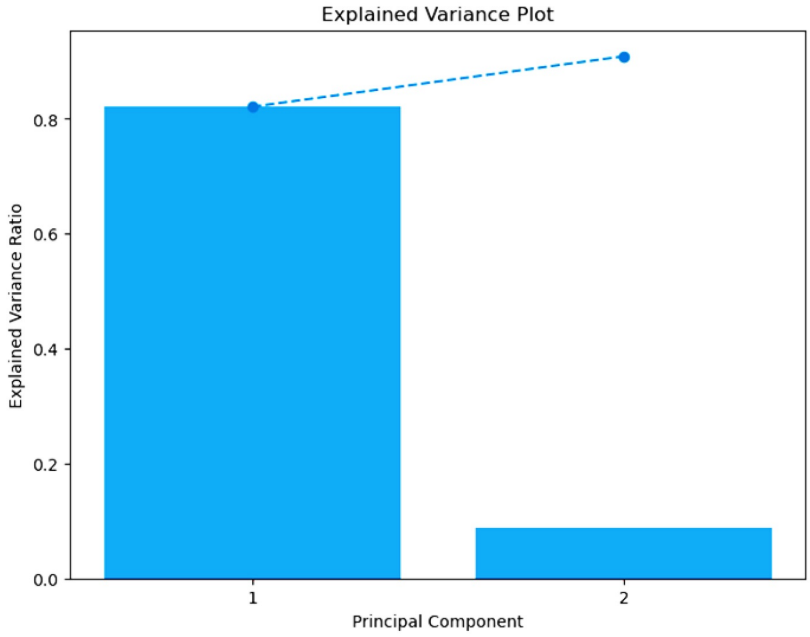


Figure 3

(a) Pre-processed spectra ATR-FTIR spectra of the undegraded and aged PET MPs, (b) the dataset used for PCA and (c) surface morphology of the degraded and aged PET MPs



(a)



(b)

Figure 4

(a) The PCA plots in space from the normalized spectral for undegraded and aged PET MPs and (b) the explained variance plot for the principal components extracted

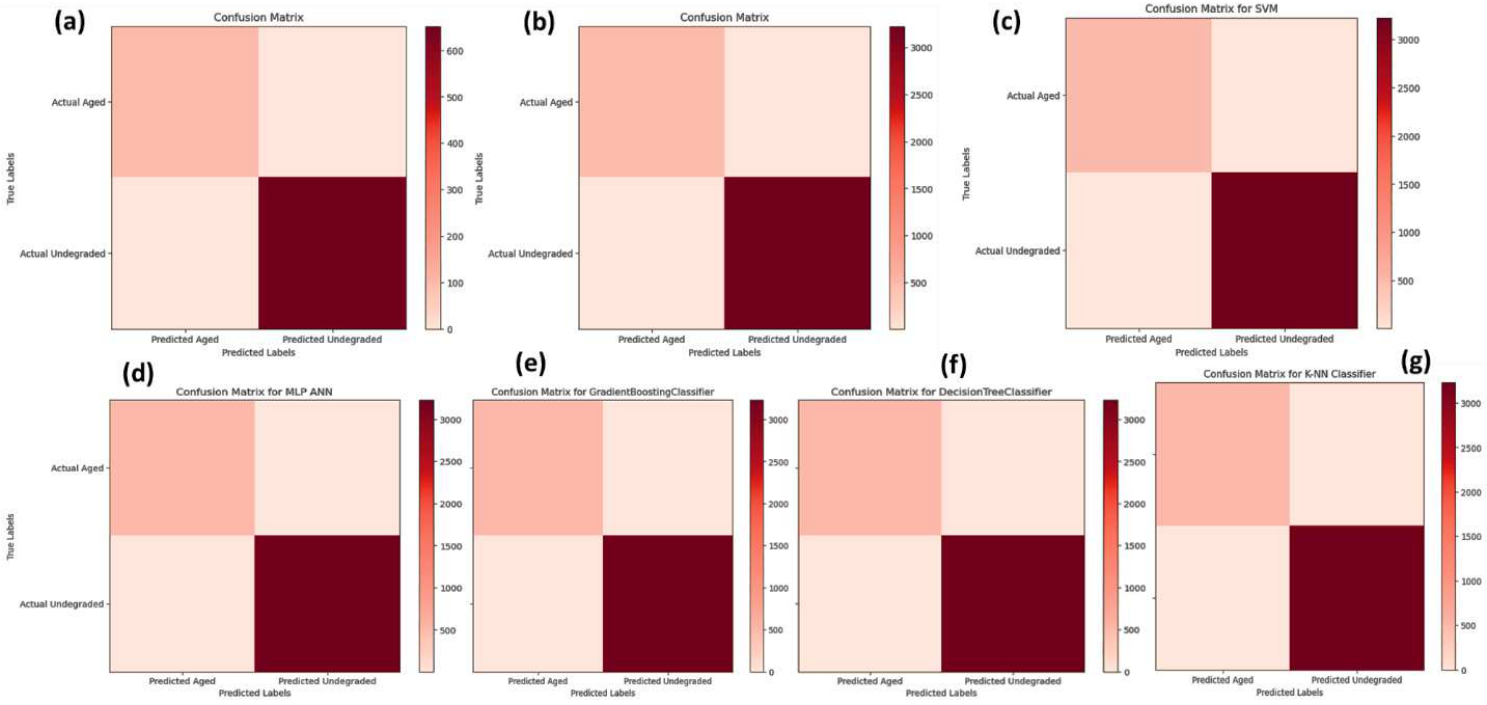
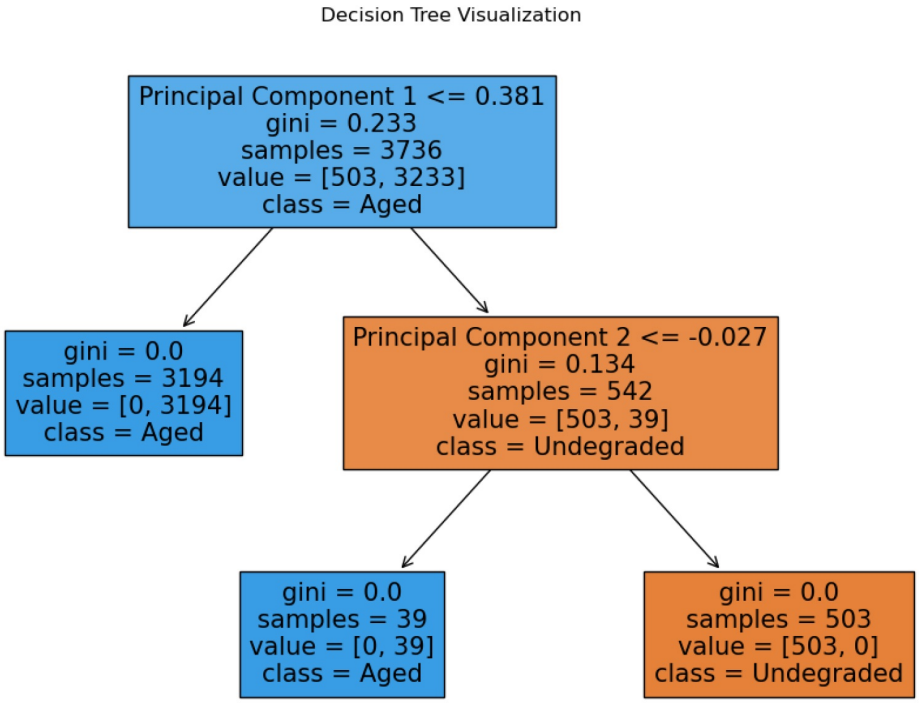
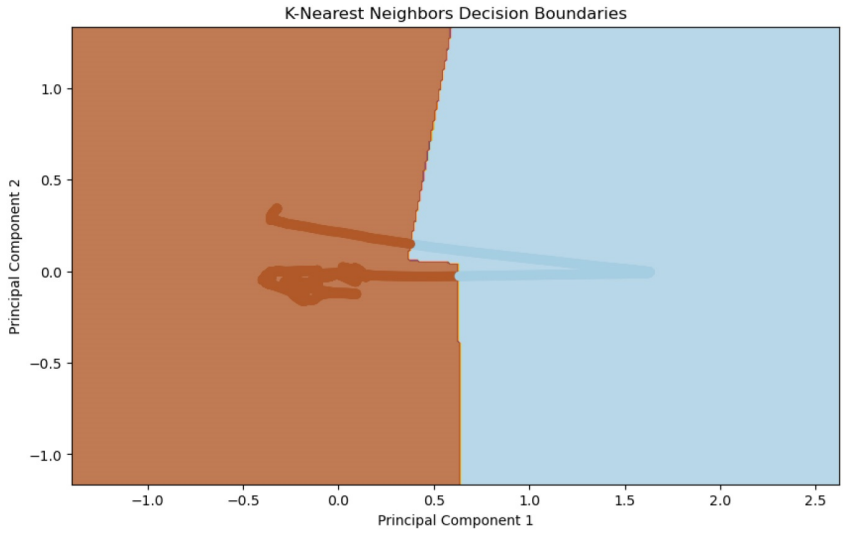


Figure 5

Confusion matrix plot for all ML classifiers (a) RF (b) LR (c) SVM (d) MLP (e) GB (f) DT (g) K-NN



(a)



(b)

Figure 6

Visualization of (a) DT and (b) k-NN for classifying undegraded and aged PET MPs using ATR-FTIR spectral

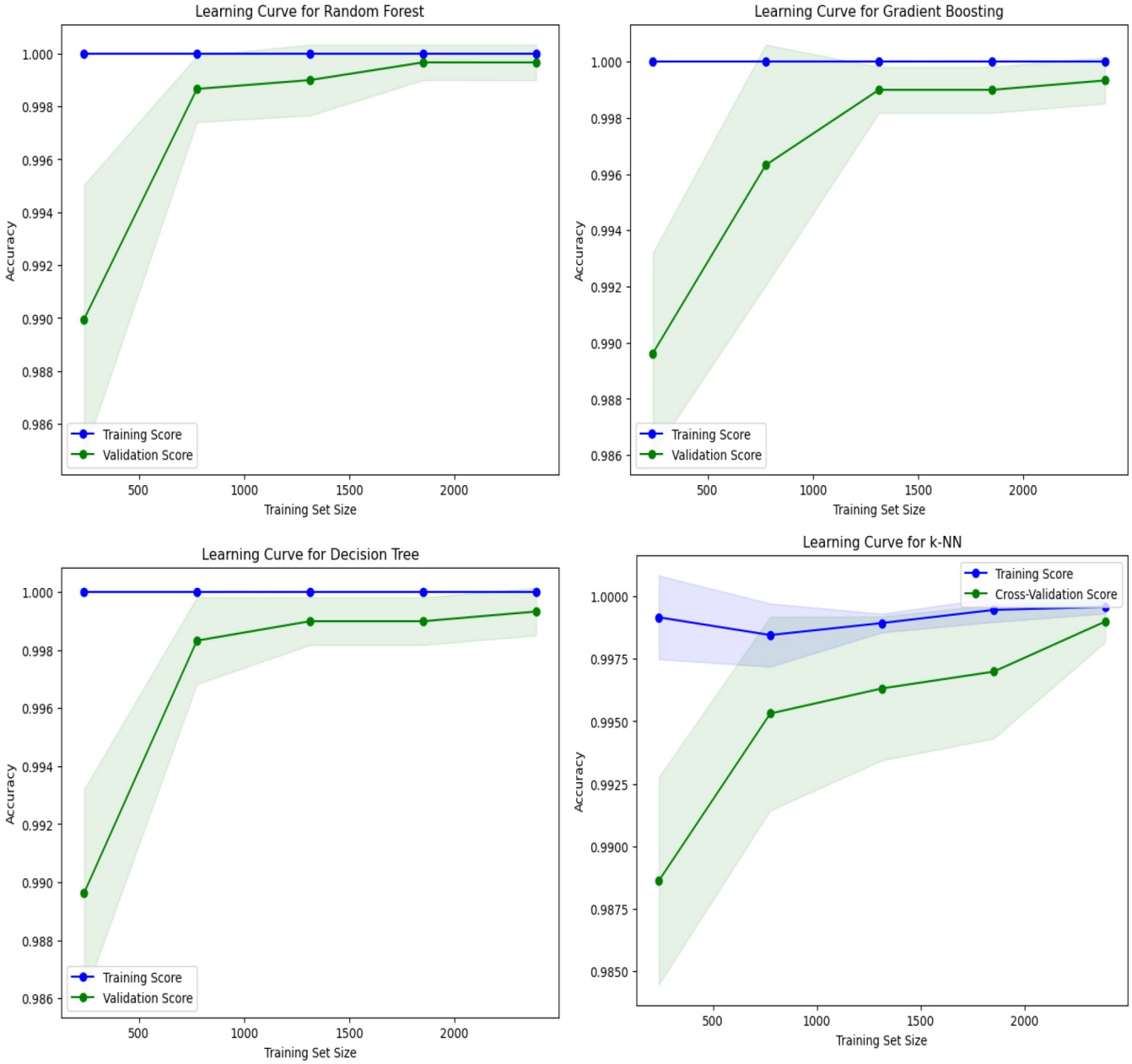


Figure 7

Learning curves for the best ML algorithms for the classification of undegraded and aged PET MPs.