UCL Open Environment Preprints

Article Title:

Using hyperspectral imaging and machine learning to identify food contaminated compostable and recyclable plastics

Author(s):

Nutcha Taneepanichskul[1], Helen C Hailes[2], Mark Miodownik[3]

Affiliations:

Mechanical Engineering Dept, University College London [1] ; Chemistry Department, University College London [2] ; Mechanical Engineering Dept, University College London [3]

ORCID IDs:

0000-0001-5574-4742[2],

Corresponding Author:

Mark Miodownik (m.miodownik@ucl.ac.uk)

DOI:

https://doi.org/10.14324/ucloepreprints.282.v1

Preprint Statement:

This article is a preprint and has not been peer-reviewed. The article has been submitted to UCL Open Environment (ISSN: 2632-0886) and is under consideration for publication following peer review.

Preprint Version:

This is version of this article.

Data Availability Statement:

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Keywords:

food contamination, automatic sorting, machine learning, composting, recycling, hyperspectral imaging

Using hyperspectral imaging and machine learning to identify food contaminated compostable and recyclable plastics

Nutcha Taneepanichskul, Helen C. Hailes and Mark Miodownik

Abstract

With the increasing public legislation aimed at reducing plastic pollution, compostable plastics have emerged as an alternative to conventional plastics for some food packaging and food service items. However, the true value of compostable plastics can only be realized if they do not enter the environment as contaminants but instead are processed along with food and garden waste using industrial composting facilities. Distinguishing compostable plastics from conventional plastics in this waste stream is an outstanding problem. Currently, Near Infrared (NIR) technology is widely used to identify polymers, but it falls short in distinguishing plastics contaminated with food waste. This study investigates the application of hyperspectral imaging (HSI) to address this challenge, enhancing the detection and sorting of contaminated compostable plastics. By combining HSI with new machine learning algorithms we show it is possible to accurately identify and classify plastic packaging with food waste contamination, achieving up to 99% accuracy. The study also measures the impact of plastic features such as darkness, size, and level of contamination on model performance, with darkness having the most significant impact. Implementing HSI in waste management systems can significantly increase composting and recycling rates. This advanced deep learning approach supports the circular economy by ensuring that both compostable and recyclable plastics are effectively processed and recycled, minimizing environmental impact.

Keywords: food contamination, hyperspectral imaging, recycling, composting, machine learning, automatic sorting

1. Introduction

The increasing popularity of compostable and biodegradable plastics underscores the need for efficient sorting technologies to separate and collect them for waste processing. In 2023 they represented 52.1% of the global bioplastic production (EuropeanBioplastic, 2023). However, current waste management systems often fail to detect and separate compostable plastics especially when contaminated with food waste, leading to their improper disposal in landfills or incinerators for the majority of compostable plastics (Allison et al., 2022).

Near-Infrared (NIR) optical sorting is a widely used technology in recycling facilities for separating different types of plastics. This technology relies on the distinct spectral signatures of various plastic polymers to achieve accurate sorting (Taneepanichskul et al., 2022). However, food waste contamination, poses significant challenges to the efficiency and effectiveness of NIR optical sorting due to issues with spectral abortion reflection of NIR frequencies by the food residues on the plastic surfaces (Masoumi et al., 2012). Additionally, the presence of food waste introduces extra spectral signals in the NIR range, creating noise that makes it harder for the system to accurately identify the polymer type.

Hyperspectral imaging (HSI) coupled with machine learning algorithms offer an advanced solution for sorting plastics, surpassing traditional NIR optical sorting methods. HSI generates a hyperspectral cube, where each pixel contains a continuous spectrum, enabling detailed spectral analysis at each pixel in the image. This capability helps overcome the challenges posed by food contaminated plastics because uncontaminated pixels can be correctly identified rather than relying on the average signal from the whole sample as with NIR methods.

While numerous studies have explored the application of HSI for identifying various types of plastics, there remains a notable gap in research that specifically addresses the challenge of detecting food contaminated compostable plastics, which represents a significant issue within the context of current plastic waste management systems.

In 2013 Ulrici et al. used HSI and partial least squares discriminant analysis (PLS-DA) to distinguish PET and PLA achieving over 98% accuracy with just six variables on the reduced matrix (Ulrici et al., 2013). Subsequently, Bonifazi used HSI with machine learning to sort paper, cardboard, plastics, and multilayer packaging. A PLS-DA-based model achieved a 0.933

recognition and reliability rate, making HSI a reliable, low-cost solution for identifying impurities and composite materials in plastic waste streams (Bonifazi et al., 2021). Taneepanichskul et al then applied HSI together with PLS-DA to identify and classify compostable plastics (PLA and PBAT), compostable materials (sugarcane and palm leaf derived packaging) and conventional plastics (LDPE, PET and PP). PLS-DA achieved a perfect classification (100%) for virgin materials larger than 10 mm x 10 mm (Taneepanichskul et al., 2023). Taneepanichskul et al. also recently studied the impact of packaging properties such as darkness, colour, size, and contamination, showing how they all impacted identification. The accuracy of the system decreased when detecting plastics that were dark, thin, small, or had high levels of contamination (Taneepanichskul et al., 2024).

In this paper we present work developing new??? chemometric and machine learning algorithms combined with HSI and show data on their performance identifying compostable and recyclable plastics with varying types and levels of food contamination. Additionally, the study explores the impact of real-world food plastic packaging properties such as size, colour and darkness, on the performance of the system.

2. Materials and Methods

To develop the model to identify and classify food-contaminated compostable and recyclable plastic packaging samples were required for the development of three datasets: a calibration dataset, a cross-validation dataset, and a testing dataset. The training dataset is the initial set of data used to train a model (Wolff, 2020). The cross-validation dataset evaluates the model's predictive performance on new, unseen data, helping to identify issues like overfitting or selection bias and providing insight into the model's ability to generalize to an independent dataset (ScikitLearn, 2024). The testing dataset offers a final, real-world validation of the model's effectiveness on completely unseen data (Barkved, 2022). The details of the food contaminants, the plastic samples, the HSI system and the deep learning algorithms are described in the following sections.

2.1 Simulating Food Contamination

The contamination levels in this experimental setup were categorized into three levels: low (25%), medium (50%), and high (75%). Figure 1 illustrates the contamination process, depicting the simulation of 25%, 50%, and 75% contamination using tomato ketchup. Each sample was cut into 50 mm x 50 mm pieces with a thickness of 0.4 mm and divided into four equal sections.

Two sauces were used to simulate food contamination: tomato ketchup and mayonnaise. These were chosen due to their ability to be applied repeatably and consistently to the samples. The different compositions helps to create training data and cross-validation data. The compositions of these two sauces are shown in Table 1. These condiments are suitable proxies for food contamination because they can represent high water activity foods such as dips and sauces, prepared salads, and dairy products; acidic foods such as pickled products, fermented foods, and fruit-based sauces; emulsified foods such as salad dressings, processed meats, and butter and margarine; and fat-containing foods. Their compositional similarities to a wide range of other food products make them ideal for studying contamination and spoilage patterns across different food categories.

Table 1: The ingredients and components of HEINZ Tomato Ketchup and HEINZ Mayonnaise

HEINZ Tomato Ketchup Main Ingredient	**Component**
Tomatoes	Water
	Carbohydrate: Including sugars (glucose and fructose) and dietary fiber
	Acid: Citric acid and malic acid, contributing to the tartness
	Vitamin: Vitamin C, Vitamin A (from beta-carotene), and Vitamin K.
	Minerals: Potassium, magnesium, and iron
	Antioxidants: Lycopene, which gives tomatoes their red colour
Vinegar	Acetic Acid
Sugar	Glucose, fructose, sucrose
Salt	Sodium Chloride
Olive oil	Monounsaturated Fats: Predominantly oleic acid Antioxidants: Polyphenols and Vitamin E

	Vitamin Fat-Soluble Vitamins: Vitamin K and E
HEINZ Mayonnaise	**Component**
Oil	Triglycerides Fatty Acid
Egg Yolk	Water Proteins Fats Cholesterol Vitamin A,D,E and K
Vinegar	Acetic Acid
Water	-
Sugar	Glucose, fructose, sucrose
Starch	-
Salt	Sodium Chloride
Mustard	Water, Acids, Vitamin A and Vitamin C

Additionally, their viscosity and texture allow them to adhere well to surfaces, effectively simulating real-life conditions of food residue on plastics. This makes them ideal for testing cleaning and contamination processes. In this study, HEINZ tomato ketchup was used for both the training and cross-validation datasets, while HEINZ mayonnaise was used for the cross-validation dataset.

To achieve 25% contamination, tomato ketchup or mayonnaise was applied to one section; for 50% contamination, it was applied to two sections; and for 75% contamination, it was applied to three sections. The ketchup and mayonnaise were then spread to ensure they covered the entire plastic surface.

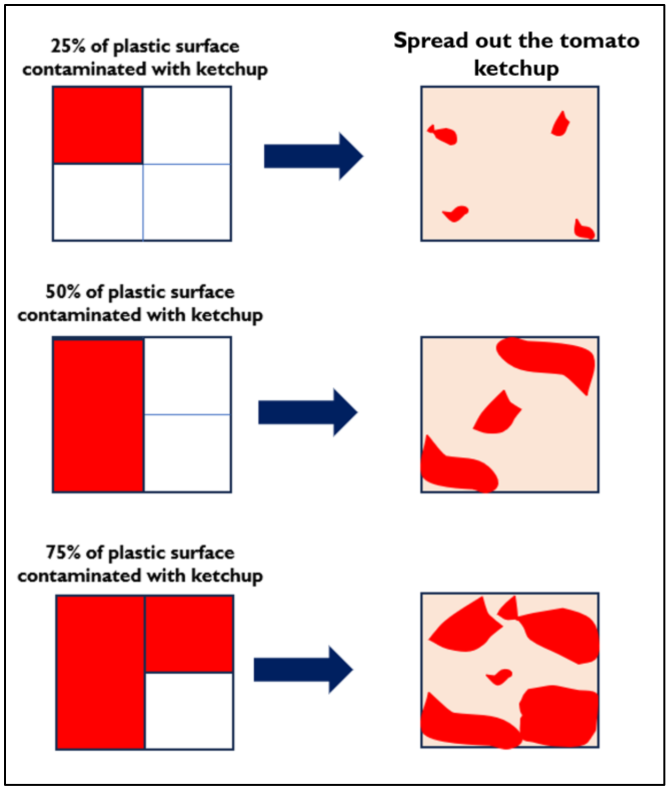

Figure 1: Simulated contamination levels of 25%, 50% and 75% sauces.

2.2 Sample Preparation

The experimental samples encompassed several size and contamination levels, with both conventional and compostable plastics. Within the category of conventional plastics, Low-Density Polyethylene (LDPE), High-Density Polyethylene (HDPE), Polyethylene Terephthalate (PET), and Polypropylene (PP) were represented. The compostable plastic category comprised Polylactic Acid (PLA), Polybutylene Adipate Terephthalate (PBAT), and Polyhydroxyalkanoate (PHA).
The materials were allocated into three datasets, namely calibration, cross-validation, and testing datasets as mentioned earlier. The training dataset encompassed both pristine plastics and plastics contaminated with low level of tomato ketchup (25%). The details of the materials within the training dataset are presented in Table 2.

Table 2: List of samples in training dataset

Material	**Material condition**	**Size**	**Number of replicates per plastic type**
LDPE, HDPE, PET, PP, PLA, PBAT and PHA	Pristine Plastic	50 mm x 50 mm	35
		40 mm x 40 mm	35
		30 mm x 30 mm	35
		20 mm x 20 mm	35
	Plastics with 25% level of tomato ketchup contamination	50 mm x 50 mm	21

In the cross-validation set, there were three replicates each with 50% and 75% tomato ketchup contamination of LDPE, HDPE, PET, PP, PLA, PBAT, and PHA. Additionally, for 25%, 50%, and 75% mayonnaise contamination, LDPE, HDPE, PET, PP, PLA, PBAT, and PHA, again each with three replicates.

In the testing dataset, 30 food waste contaminated plastic packaging items were collected from a various sources spanning across the city of London including tubs, trays, lids, plastic spoons. These sources were inclusive of both supermarkets, cafe and restaurants, resulting variety of packaging types, including take-away boxes, cutlery, lids, and more. We selected only the plastic packaging that had a label to show type of packaging on them in order to verify the model. Figure 2 provides examples of contaminated food packaging in testing dataset.

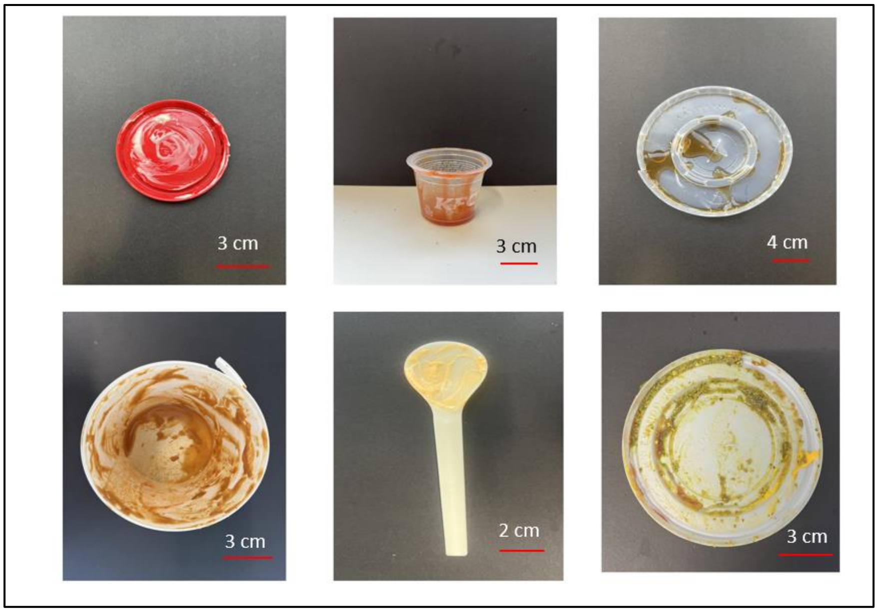

Figure 2: The example of real-world food contaminated plastic packaging used in the testing dataset.

2.3 Imaging methodology

2.3.1 Hyperspectral data acquisition and analysis schematic

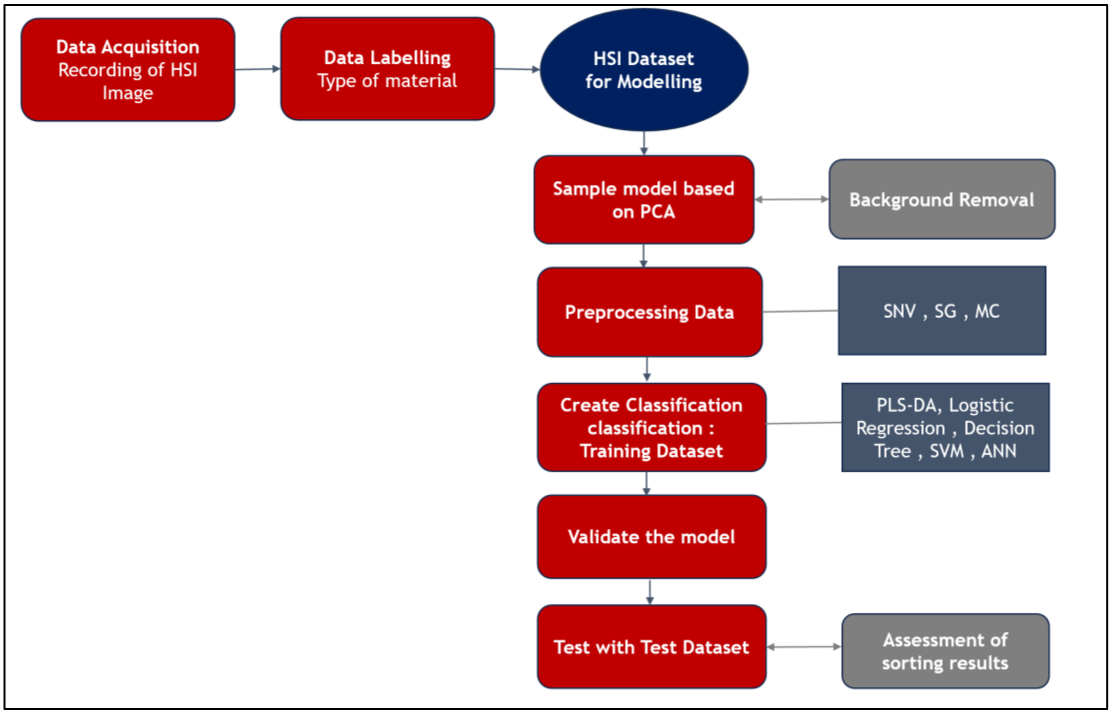

Figure 3: Schematic showing the hyperspectral data acquisition and analysis.

As shown in Figure 3 hyperspectral images were obtained from line scans of the samples on a conveyor belt passing under a HySpex Baldur S-640i N camera. The camera, positioned at a working distance of 1 metre with a 16° field of view, covered a spectral range of 950 to 1730 nm with a spectral resolution of 3.36 nm, resulting in a total of 232 spectral bands. The spatial pixel size of the images was 0.44 mm (Hyspex, 2023). The system's conveyor belt measured 700 mm in length, 215 mm in width, and 60 mm in height, with a maximum speed of 120 mm/s. The image capture background was the black conveyor belt. A halogen lamp, emitting light across the spectrum from 400 nm to 2500 nm, was employed as the light source. This experimental setup has been described in more detail in our previous work. (Taneepanichskul et al., 2024, Taneepanichskul et al., 2023). HyspexGround software facilitated the acquisition of the hyperspectral image. Subsequently, the Breeze software package was employed for PCA model development, spectrum preprocessing, application of diverse machine learning algorithms for classification, and production of classification results.

2.3.2 Principal component analysis (PCA) and spectrum pre-processing

PCA (Principal Component Analysis) was utilized to investigate the relationships between samples and measured variables, with the objective of unveiling patterns within the data. Its primary focus lies in identifying common features rather than distinguishing differences between classes (Castro-Díaz et al., 2023). PCA breaks down data into linear combinations of the original hyperspectral data, known as principal components (PCs). PC1 represents the greatest variability within the dataset, capturing the majority of the information. The subsequent principal components follow in descending order, representing the remaining variance. In our case, PCA was employed to eliminate background pixel and isolate objects (plastics) within the hyperspectral images.

Subsequently, spectral preprocessing was conducted using a combination of methods. This included applying a combination of Savitzky-Golay (SG) first derivative with a 2nd polynomial and a 15-point window, Standard Normal Variate (SNV) and mean centering. This technique was employed to eliminate insignificant baseline signals from the collected data and to rectify scatter data (Taneepanichskul et al., 2024).

2.3.3 Machine learning classification model

Various machine learning algorithms, including logistic regression, decision tree algorithms, support vector machines (SVM), artificial neural networks (ANN), and partial least squares discriminant analysis (PLS-DA), were applied to build classification models. The samples in the training dataset were used to develop these models.

2.3.3.1 Logistic regression

Logistic regression is a fundamental supervised learning method widely utilized for classification tasks, particularly in scenarios involving binary outcomes. Through the sigmoid function, it transforms spectral band values to produce probabilities for binary predictions, with coefficients assigned to each band indicating their predictive influence (Qian et al., 2012, Kabir et al., 2021).

Logistic regression can be extended to address multiclass classification problems through softmax regression. The softmax function normalizes the output into a probability distribution across multiple classes, ensuring that the sum of the predicted probabilities for all classes equals unity.

This way, the model can provide predictions for each class, and the class with the highest probability is considered as the final prediction (Tranmer and Elliot, 2008).

2.3.3.2 Decision tree (DT)

A decision tree (DT) is a non-parametric model structured as a tree, where each node contains a decision rule based on input data. This rule directs whether to move to the left or right sub-nodes, while the leaf nodes provide the final output. DTs are applicable to both classification and regression tasks and are particularly valued for their interpretability. One common method for building nodes in a DT is information gain, which uses entropy or the Gini index to measure the amount of information retained by each feature in the input data before making predictions. (Zhang et al., 2022).

2.3.3.3 Support vector machine (SVM)

A Support Vector Machine (SVM) was used as a supervised machine learning algorithm for classification, regression, and outlier detection. It was used to identify hyperplanes in the feature space that separate data points belonging to different classes. The hyperplane was positioned to maximize the margin, which is the distance between the hyperplane and the nearest data points of each class. SVM operates in the original feature space, but kernelized SVMs were also used, these transform data into higher-dimensional spaces through kernel functions. The algorithm requires labelled training data to learn and relies on support vectors, which are crucial points closest to the hyperplane. In a One-versus-One (OvO) approach, binary classifiers are created for each pair of classes. For N classes, this results in C(N,2) binary classifiers. In our scenario, where we sought to classify 7 types of plastics, we used 21 binary classifiers.

2.3.3.4 Artificial neural network (ANN)

The architecture of an artificial neural network (ANN) typically comprises three layers: input, hidden, and output. The input layer captures spectral information from hyperspectral imaging, where each input to the ANN is a vector representing the spectral signature of each sample. The hidden layer, containing numerous neurons, performs computations on the input data. Hidden layers enable ANNs to learn complex problems and nonlinear relationships. Each neuron in a hidden layer calculates a weighted sum of its inputs, applies an activation function, and produces an output that becomes the input for the next layer. Various activation functions, such as linear, sigmoid, tanh, and ReLU, can be employed based on the task.

The input to the ANN was represented by a vector that encapsulates the spectral information, with its length determined by the number of spectral bands or channels in the hyperspectral data. Each element of the vector corresponded to the intensity or reflectance value of the pixel in a specific spectral band. The hidden layer, with 100 neurons, utilized the ReLU activation function to process the hyperspectral data and extract relevant features for classifying the types of plastics (MicrosoftBuild, 2021). The output layer produced the final classification results, with each neuron corresponding to a different type of plastic, typically using a softmax activation function to provide probabilities for each class.

2.3.3.5 Partial least squares discriminant analysis (PLS-DA)

PLS-DA, a blend of partial least squares regression (PLS-R) and discriminant analysis (DA), is a supervised ML method for dimensionality reduction and material class prediction. It necessitates an X matrix with calibration spectra and a corresponding Y matrix denoting class identity (types of plastic). In binary cases, Y is a single column; for multiclass scenarios, it's a dummy matrix with 1's and 0's indicating class membership. The model's output isn't strictly binary, requiring a threshold establishment during prediction. Setting thresholds employs various methods, with Bayes' Theorem being a prevalent choice. Alternatively, a 0.5 cut-off point is often employed for binary classification tasks (Amigo et al., 2015). In our PLS-DA, the linear equation was modelled with around 5 latent variables, enabling graphical visualization and understanding through LV scores and loadings.

2.4 Classification model performance (model validation)

Model validation is a crucial step in machine learning, particularly for assessing the performance of classification models. Various metrics are utilized for evaluation, including sensitivity (Equation 1), specificity (Equation 2), F1 score (Equation 3), and accuracy (Equation 4). The formulas for these metrics are based on the following definitions: True Positive (TP) represents instances where the model correctly predicts the positive class, while True Negative (TN) indicates instances where the model correctly predicts the negative class. False Positive (FP) refers to instances where the model incorrectly predicts the positive class, and False Negative (FN) denotes instances where the model incorrectly predicts the negative class.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \text{(Equation 1)}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad \text{(Equation 2)}$$

$$F1 - Score = \frac{True\ positive}{True\ positive + \frac{1}{2}(False\ Positive + False\ Negative)} \quad \text{(Equation 3)}$$

$$Accuracy = \frac{True\ Negative + True\ Positive}{True\ Negative + True\ Positive + False\ Negative + False\ Positive} \quad \text{(Equation 4)}$$

2.5 The evaluation of plastic features in testing dataset

To measure the impact of plastic features on the performance of classification models, the properties of plastics in the testing dataset, including darkness, level of contamination, and size, were evaluated using image processing algorithms to ensure precise evaluation (Taneepanichskul et al., 2024).

2.5.1 Size

The plastic packaging images in the testing dataset were resized to 10 cm x 15 cm and converted to greyscale. Otsu's thresholding method was then applied to remove the background and convert the greyscale images to binary format. In this process, pixels with values below the threshold were set to 0, while those above the threshold were set to 255. Following this, the percentages of foreground and background areas were calculated. These percentages were then multiplied by 150 cm^2 (the total area of the frame) to determine the area occupied by the plastic packaging. The size was classified into 3 categories: small ($< 20\ cm^2$), medium ($20\ cm^2 \leq$ area $< 80\ cm^2$) and large ($\geq 80\ cm^2$) (Taneepanichskul et al., 2024).

2.5.2 Level of contamination

K-means clustering was applied to assess the level of contamination in plastic packaging within the testing dataset. The images were loaded and converted to greyscale, with each pixel represented as a vector based on its greyscale intensity. We selected the number of clusters to be 3. The centroids of each cluster were initialized, and for each pixel in the image, a similarity measure was calculated to determine its proximity to each cluster centroid using a distance metric, such as Euclidean distance. Based on this calculation, the pixel was assigned to the cluster with the closest centroid, forming the initial clusters. Upon convergence, the algorithm produced the final clustering results. At this stage, each pixel was firmly assigned to a specific cluster, and the cluster centroids represented the average greyscale intensities of the pixels within their respective clusters. The number of pixels in each cluster was counted, and their ratios were calculated to determine the percentage of contamination. The level of contamination in the plastic packaging was classified into four categories: low contamination (< 25%), medium contamination (25% ≤ contamination < 60%), high contamination (≥ 60%), and indeterminate due to multicoloured packaging or oily contamination (Taneepanichskul et al., 2024).

2.5.3 Darkness

The images in the testing dataset were loaded and converted into greyscale. Otsu's threshold theory was applied to separate foreground and background. The average pixel of foreground was calculated to determine the darkness level. The darkness level was classified into three categories: bright (≥157), dark (<157) and transparent (Taneepanichskul et al., 2024).

3. Results

3.1 Average raw absorbance spectrum and pre-processed spectrum

Samples of seven types of plastics including conventional plastic (PP, LDPE, HDPE and PET) and compostable plastics (PLA, PBAT and PHA) were passed underneath the HSI camera by a conveyor belt. The data obtained was used to develop an identification and classification model of plastics with tomato ketchup contamination using machine learning algorithms. Raw absorbance spectrum of pristine plastic samples and plastic samples with 25% of surface covered with tomato ketchup were shown in Figure 4(a) and 4(b) respectively. Raw absorbance of these materials in training dataset was pre-processed using Savitzky-Golay (1st derivative, 2nd polynomial and 15 points window) method to identify spectral signatures. The pre-processed absorbance spectra are shown in Figure 4(c).

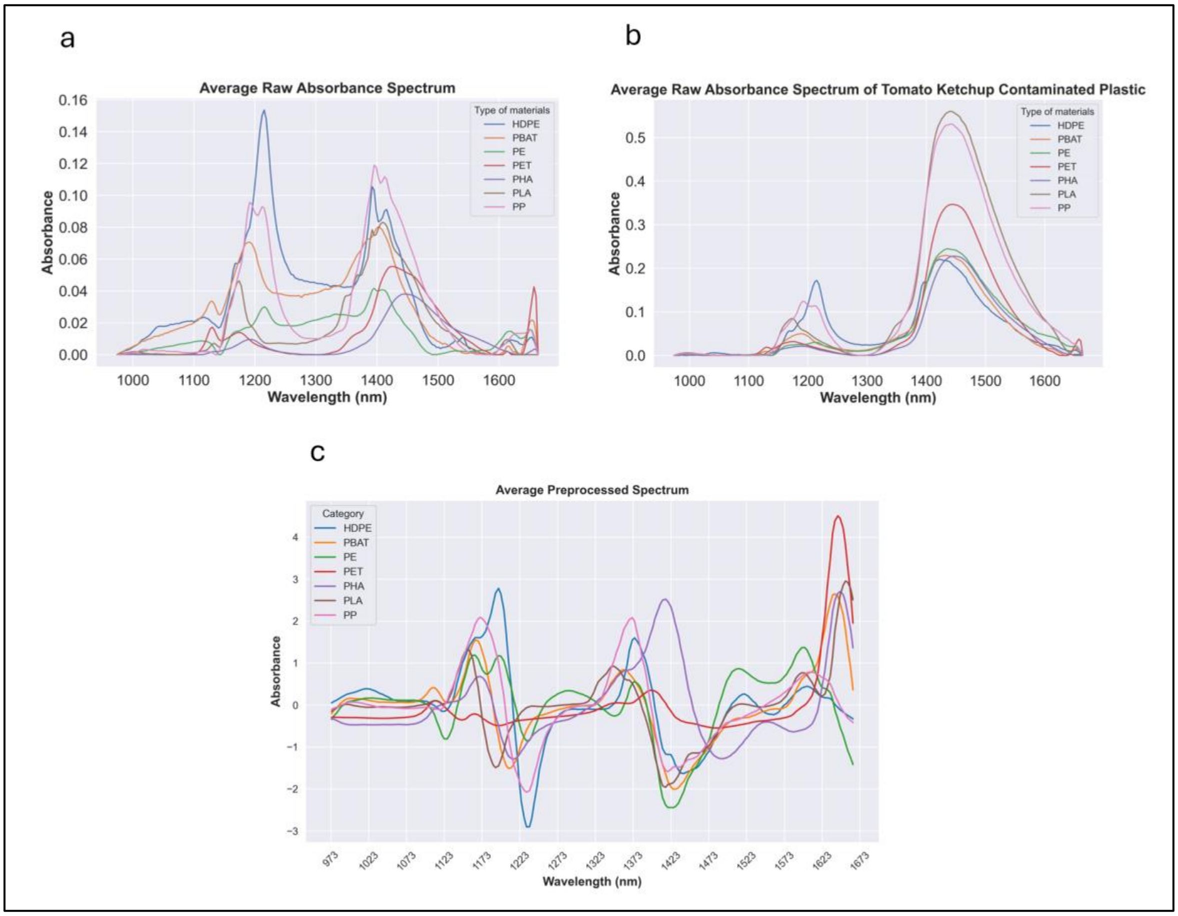

Figure 4: Raw absorbance spectrum of (a) pristine plastics PP, PET, LDPE, HDPE, PLA, PBAT and PHA; (b) the same plastics with 25% of plastic surface contaminated with tomato ketchup; (c) pre-processed absorbance spectrum of plastics in training dataset (pristine and contaminated with tomato ketchup)

3.2 Principal Component Analysis (PCA)

Following the preprocessing of the absorbance spectrum, a principal component analysis (PCA) was carried out to achieve dimensional reduction. The spectra of pristine and tomato ketchup contaminated plastics from the training dataset were then utilized to generate a PCA score plot, as depicted in Figure 5(a). The results indicate that a substantial portion of the variance is effectively captured by the first principal component (PC1), which accounts for 46%, and the second principal component (PC2), which contributes 20%. Pristine plastics showed a high level of separability.

Specifically, pristine HDPE and PP are situated in the second quadrant, while LDPE, PBAT, and PLA are in the third quadrant, and PET and PHA are in the fourth quadrant. Plastics contaminated with tomato ketchup, which are indicated by the red box, are all located in the first quadrant but show some overlap with each other.

3.3 Performance of classification models

3.3.1 Performance of classification models on calibration dataset

The calibration dataset consists of pristine plastics and plastics with the low level of tomato ketchup (25%). We have applied five machine learning techniques to build classification model on training dataset. Figure 5(b), (c) and (d) illustrates the decision boundary of logistic regression, decision trees, and SVM classification respectively, providing a visual representation of how these algorithms partition the feature space to classify different plastic samples in training dataset. The decision boundary delineates the regions where each class is predicted, offering insights into the complexity and separability of the dataset. This visualization aids understanding of the underlying behaviour of the models and their ability to discriminate between different classes of plastics based on the provided features. However, ANN and PLS-DA do not have a straightforward decision boundary. ANN operates through complex transformations of the input data. The decision-making process in an ANN involves a series of interconnected neurons with weighted connections. PLS-DA works by finding linear combinations of features that best separate the classes in the data. Unlike traditional classifiers, PLS-DA does not directly define a decision boundary. Instead, it projects the data into a new space where the classes are maximally separated along latent variables. Consequently, it is not as intuitive to visualize the decision boundary in the original feature space.

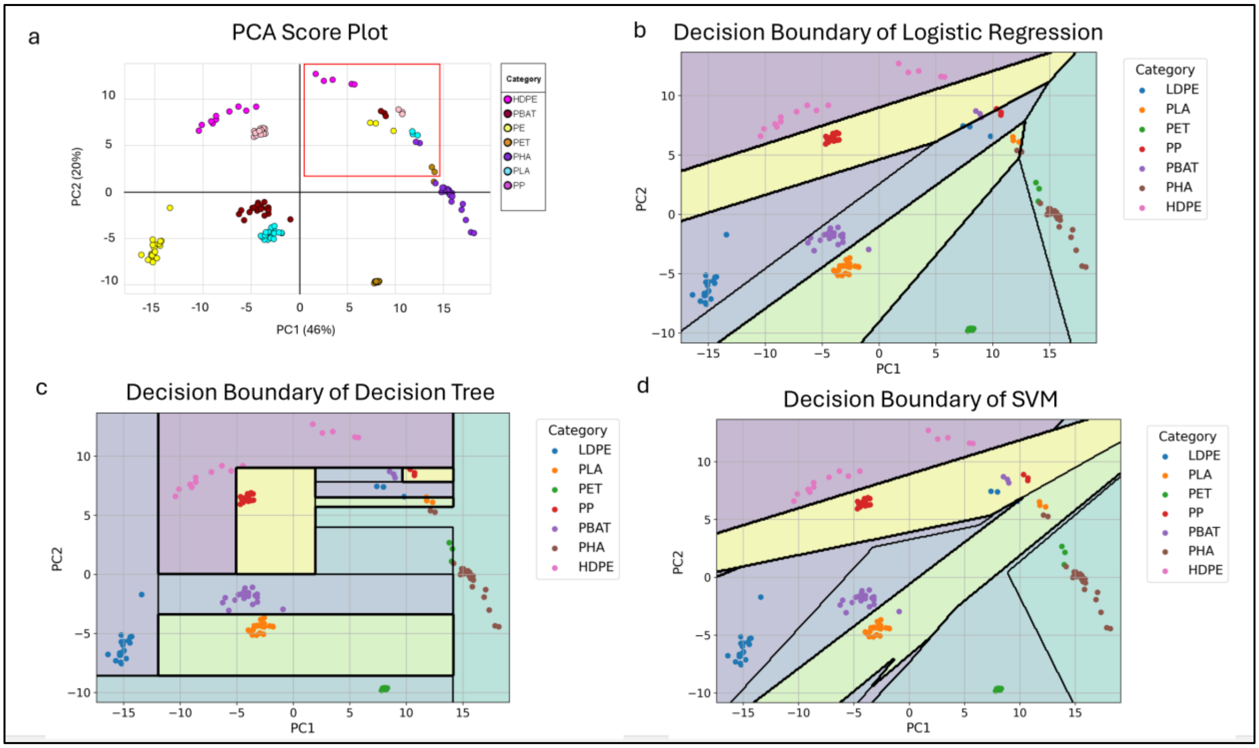

Figure 5: (a) PCA score plot of training dataset, contaminated plastics identified within red box; (b) Logistic regression; (c) Decision tree; (d) SVM for the training dataset.

Table 3 shows the performance of each classification model. For logistic regression, SVM, decision trees, and ANN, sensitivity, specificity, and F1 score all reached 1, resulting in an overall accuracy of 100%. Conversely, other models achieved 100% accuracy. However, PLS-DA exhibited slightly lower accuracy (90.6%) due to its increased sensitivity to outliers, particularly noticeable when identifying plastics contaminated with tomato ketchup.

Table 3: the performance of various machine learning algorithms in identifying plastics within the training dataset.

Machine Learning Methods	Polymer	Sensitivity	Specificity	F1 Score	Overall Accuracy
Logistic Regression, SVM, Decision Tree, ANN	PLA	1	1	1	100%
	PBAT	1	1	1	
	PHA	1	1	1	
	PET	1	1	1	
	PP	1	1	1	
	HDPE	1	1	1	
	LDPE	1	1	1	
PLS-DA	PLA	0.91	1	0.95	90.6%
	PBAT	1	1	1	
	PHA	0.87	1	0.93	
	PET	0.87	1	0.93	
	PP	0.87	1	0.93	
	HDPE	1	1	1	
	LDPE	0.91	1	0.95	

3.3.2 Performance of classification models on cross-validation dataset

Before testing the model with real world contaminated food packaging, we applied these models to classify types of materials in cross validation dataset to assess generalization of data which included a new type of contamination (mayonnaise). The results are summarised in Table 4.

The logistic regression model performed well on datasets 95% accuracy. For PLA, PBAT, PET, HDPE, and LDPE it achieved perfect scores of 1 for sensitivity, specificity, and F1 score. However, it encountered challenges in accurately detecting PHA due to a new type of contamination and the presence of thin film. Consequently, instances of PHA were misclassified as PP, resulting in a decrease in sensitivity for PHA to 0.67 and a decrease in specificity for PP to 0.94.

The SVM model achieved 94% accuracy. For PLA, HDPE, and LDPE it achieved perfect scores of 1 for sensitivity and specificity. However, like logistic regression, its performance declined when classifying PHA and PBAT. The sensitivity of PHA and PBAT was 0.67 and 0.93 respectively. Misclassifications of PHA (66.7%) and PBAT (6.7%) as PET led to decreases in specificity for PP and PET, resulting in values of 0.94 and 0.99 respectively. Consequently, the F1 scores for PBAT, PHA, PP, and PET decreased to 0.96, 0.8, 0.85, and 0.97 respectively.

The decision tree model achieved 88% accuracy, encountering difficulties in accurately identifying PBAT, PET, and PHA. Specifically, the sensitivity for PBAT, PHA, and PET was 0.8, 0.6, and 0.7 respectively, while other types of plastics achieved a sensitivity of 1. PBAT was often misclassified as PP (13.3%) and LDPE (6.7%), while PET was misclassified as PP (26.7%). PHA suffered misclassifications as LDPE (6.7%) and PP (33.3%). Regarding specificity, PP and LDPE exhibited lower values compared to other plastic types, with scores of 0.88 and 0.98 respectively. Consequently, the F1 scores for PBAT, PHA, PP, and PET were 0.89, 0.75, 0.97, and 0.85 respectively.

The ANN model demonstrated strong overall performance with an accuracy of 90%. In the cross-validation dataset, it achieved excellent sensitivity, specificity, and F1 scores for all types of plastic except for PHA and LDPE, where sensitivity dropped to 0.73 and 0.6 respectively. Furthermore, the model exhibited misclassifications, 40% of LDPE being incorrectly labelled as PP, while 26.7% of PHA samples were misclassified as PP. Additionally, the specificity of PP was low at 0.89. Consequently, the F1 scores for PHA, PP, and LDPE were computed as 0.84, 0.75, and 0.75 respectively. Overall, while the model achieved impressive accuracy and performance for most plastic types, there are evidently areas for improvement, particularly in accurately distinguishing PHA and LDPE, as well as reducing misclassifications, especially between LDPE and PP.

The performance of PLS-DA fell short compared to other machine learning algorithms, achieving an overall accuracy of only 75%. Due to the introduction of a new type of contamination (mayonnaise), misclassifications occurred across various plastic types: 6.7% of PBAT, 20% of PET, 13.3% of PLA, 20% of PP, and 46.7% of LDPE could not be identified. Additionally, 13.3% of PBAT samples were misclassified as LDPE. Misclassifications were observed between various plastic types as well, with 20.3% of LDPE and 26.7% of PLA incorrectly labelled as PBAT, while

6.7% of PHA samples were misclassified as PLA. Consequently, the sensitivity of PHA was the lowest at 0.47, followed by PBAT, LDPE, PET, and PP at 0.8, while PLA had a sensitivity of 0.6. For specificity, all polymers in the cross-validation dataset achieved values greater than 0.9, indicating strong performance in correctly identifying true negatives. However, PLA, PBAT, and PP exhibited slightly lower specificity compared to others, with values of 0.99, 0.92, and 0.98 respectively. Additionally, the F1 score for PHA was the lowest at 0.63, followed by PBAT, PLA, LDPE, PET, and PP, which achieved scores of 0.7, 0.72, 0.8, and 0.89 respectively.

Table 4: The performance of classification models on cross validation dataset

Machine Learning Methods	Polymer	Sensitivity	Specificity	F1 Score	Overall Accuracy
Logistic regression	PLA	1	1	1	95%
	PBAT	1	1	1	
	PHA	0.67	1	0.8	
	PET	1	1	1	
	PP	1	0.94	0.85	
	HDPE	1	1	1	
	LDPE	1	1	1	
SVM	PLA	1	1	1	94%
	PBAT	0.93	1	0.96	
	PHA	0.67	1	0.8	
	PET	1	0.99	0.97	
	PP	1	0.94	0.85	
	HDPE	1	1	1	
	LDPE	1	1	1	
Decision tree	PLA	1	1	1	88%
	PBAT	0.8	1	0.89	
	PHA	0.6	1	0.75	
	PET	0.7	1	0.85	

	PP	1	0.88	0.73	
	HDPE	1	1	1	
	LDPE	1	0.98	0.93	
ANN	PLA	1	1	1	90%
	PBAT	1	1	1	
	PHA	0.73	1	0.84	
	PET	1	1	1	
	PP	1	0.89	0.75	
	HDPE	1	1	1	
	LDPE	0.6	1	0.75	
PLS-DA	PLA	0.6	0.99	0.72	75%
	PBAT	0.8	0.92	0.7	
	PHA	0.47	1	0.63	
	PET	0.8	1	0.89	
	PP	0.8	1	0.8	
	HDPE	1	1	1	
	LDPE	0.8	0.98	0.82	

Figure 6 demonstrates the impact of contamination levels on the accuracy of various classification models. For plastic with a low level of contamination, logistic regression, SVM, ANN, and PLS-DA achieved 100% accuracy, while the decision tree model reached 95% accuracy. As contamination levels increased to a medium level (50%), the accuracy of all models decreased: logistic regression and SVM dropped to 95%, while ANN and PLS-DA fell to 88%. At high contamination levels (75%), the accuracy further declined to 93% for logistic regression, 90% for SVM, 76% for the decision tree, 86% for ANN, and 50% for PLS-DA.

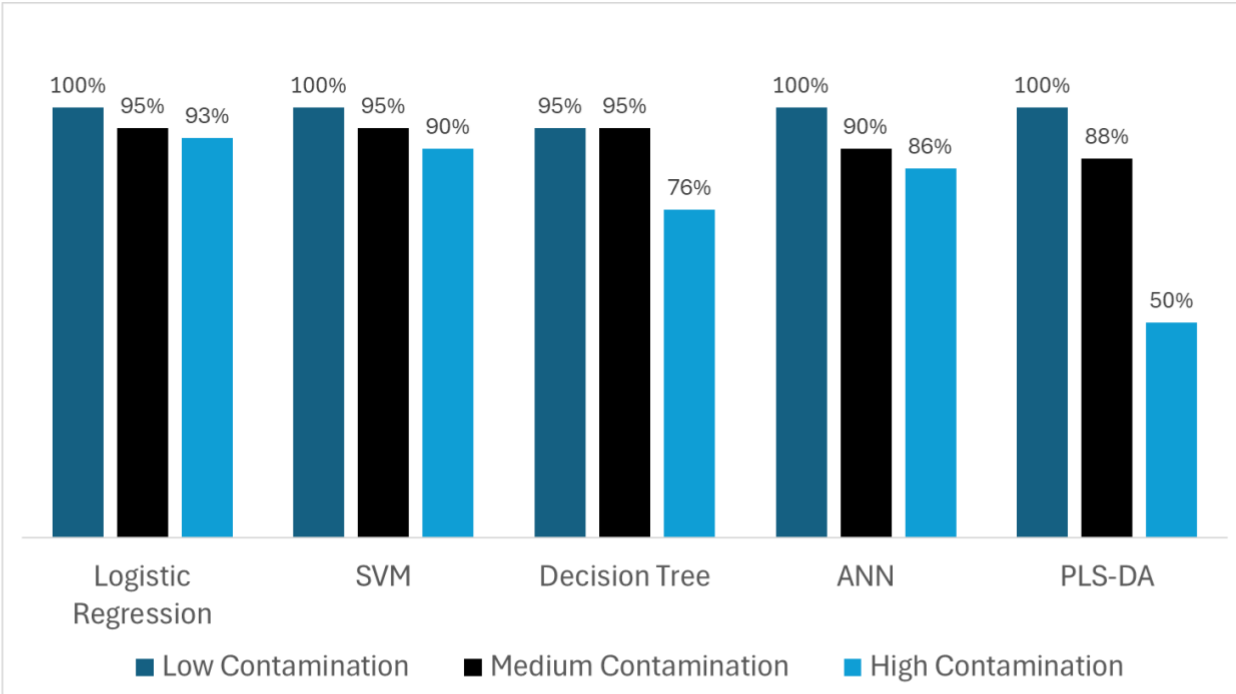

Figure 6: The impact of contamination level on the accuracy of the model

3.3.2 Performance of classification models on testing dataset

The classification models were employed to categorize 30 real-world packaging samples with various types and level of contamination. The performance of each classification model showed in table 5.

Table 5: The prediction accuracy of SVM, logistic regression, decision tree, ANN and PLSDA on testing dataset

Machine Learning Methods	Polymer	Sensitivity	Specificity	F1 Score	Overall Accuracy
SVM	HDPE	1	1	1	99%
	PET	0.86	1	0.92	
	PHA	1	1	1	
	PLA	1	1	1	
	PP	1	0.94	0.97	

	Overall	0.97	0.99	0.98	
Logistic regression, decision tree, ANN	HDPE	1	1	1	98%
	PET	1	1	1	
	PHA	1	1	1	
	PLA	1	0.96	0.8	
	PP	0.86	1	0.92	
	Overall	0.81	0.99	0.94	
PLS-DA	HDPE	1	1	1	96%
	PET	1	1	1	
	PHA	1	1	1	
	PLA	1	0.89	0.57	
	PP	0.79	1	0.88	
	Overall	0.96	0.98	0.89	

From Table 5, it is evident that the overall accuracy of SVM surpasses that of other machine learning algorithms. However, SVM exhibits lower sensitivity in detecting PET packaging compared to other types of plastic. This is mainly due to the limited reflectance demonstrated by PET, resulting in a weak Short-Wave Infrared (SWIR) signal. Consequently, identifying thin or transparent materials like PET becomes inherently challenging. Additionally, the model tends to classify contamination as PP, leading to a lower specificity for PP compared to other material types see Figure 7.

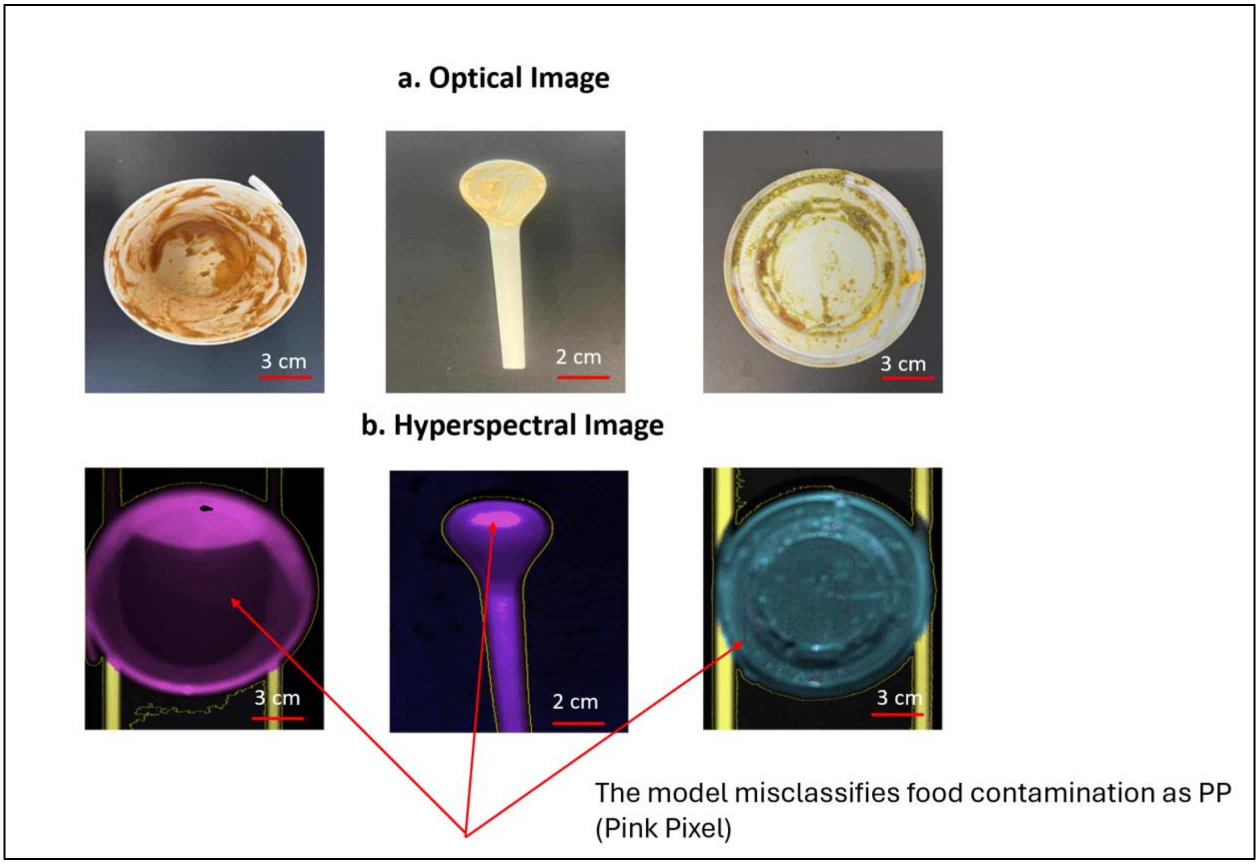

Figure 7: The example of (a) optical Images and (b) hyperspectral images of testing dataset (PP: pink, PHA, purple, PLA: blue)

For logistic regression, decision tree and ANN, the sensitivity in detecting PP is the lowest at 0.86. PP was misclassified as PLA and LDPE. Thus, the specificity and F1 score of PLA is 0.96 and 0.8 respectively while other types of plastics are 1. The overall accuracy of these models is 98%. The misclassified samples have translucent colour and dark colour. Logistic regression, decision tree, ANN classified translucent PP lid and red dark colour Japanese rice bowl was misclassified as PLA and LDPE respectively. Some pixels were misclassified by each model, leading to the same final classification outcome. For example, in Figure 8, we used various classification models to identify the type of food-contaminated spoon. The majority of pixels were classified as PHA. However, for spicy mayo contamination, SVM classified it as PP, while Decision Tree, Logistic Regression, and ANN each classified spicy mayo as a combination of PP (pink) and PLA (blue) but in different positions. PLS-DA classified it as a mix of unidentified pixels (Red) and PP (pink).

For PLS-DA, the overall accuracy is the lowest at 96%. 21.4% of PP were classified incorrectly as PLA. Therefore, the sensitivity of PP drops to 0.79 and the specificity of PLA decreases to 0.89. For F1-score, PLA achieves only 0.57 and PP achieves 0.88 while others are 1.

Figure 8 shows a sample (spoon) made from PHA (purple) with some contaminated areas (spicy mayo) incorrectly classified as PLA (blue) or PP (pink). This highlights a limitation of PLS-DA in accurately identifying specific materials in contaminated regions compared to other algorithms used in the study.

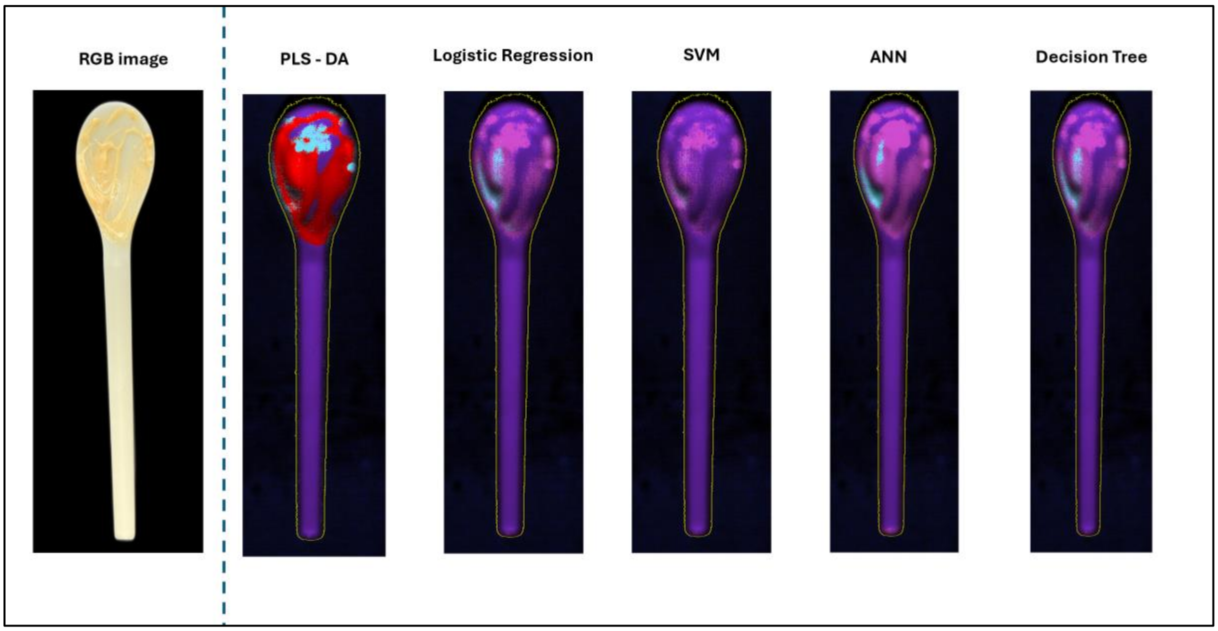

Figure 8: Material Classification of PHA Spoon with Spicy Mayo Using Various Machine Learning Models (Purple Pixel: PHA, Blue Pixel: PLA, Pink Pixel: PP and Red Pixel : Unidentified pixel)

3.4 Material Properties of Contaminated Plastic Packaging in Testing Dataset

In this set of experiments, we investigated which properties of contaminated plastic have impact on the accuracy of selected classification models. Specifically, we measured the size of packaging, the level of contamination, and darkness of the packaging.

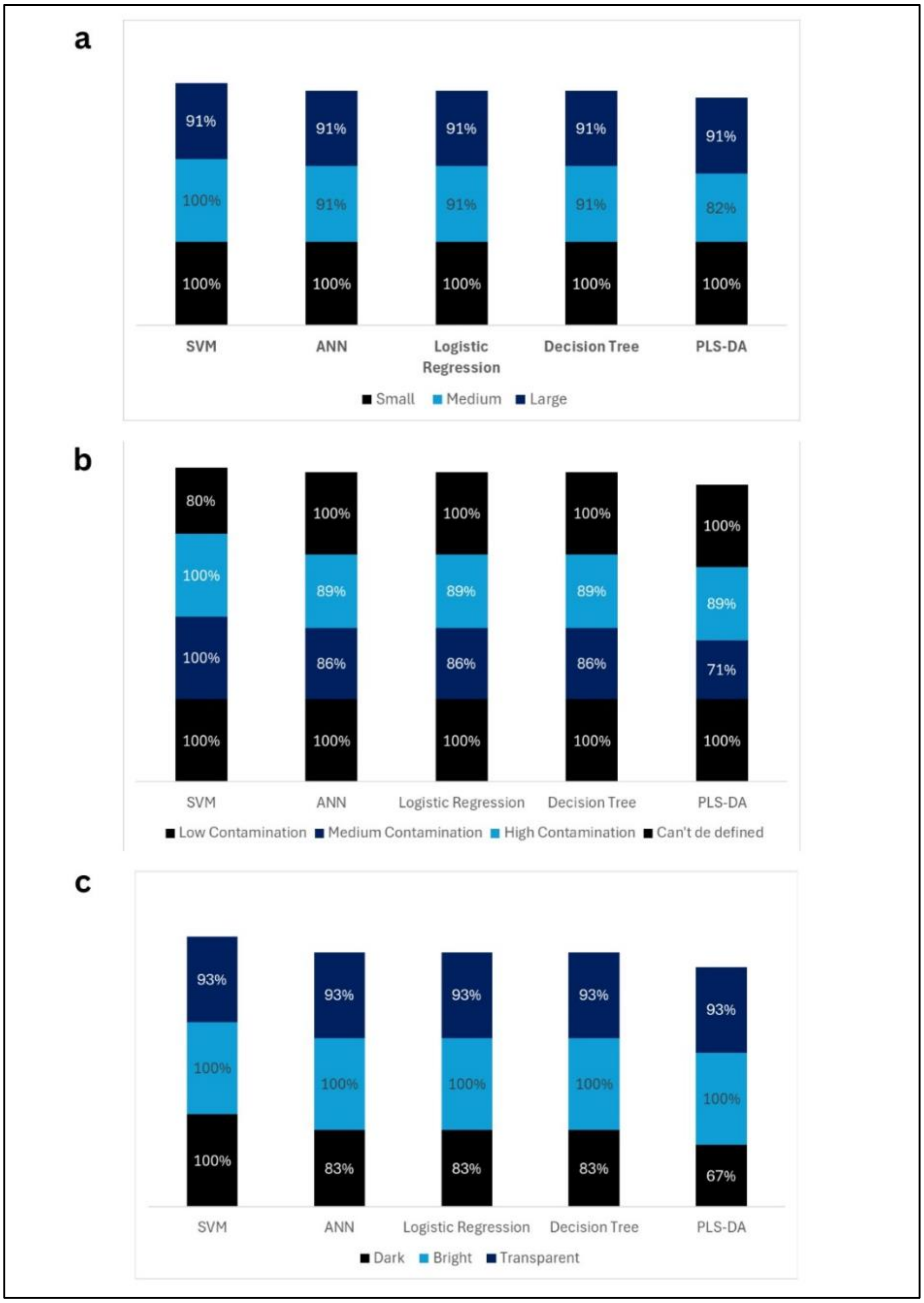

Figure 9: Accuracy of models for identifying plastic samples from the testing data set; (a) small, medium, and large plastics; (b) low, medium and high level of contaminated plastic packaging; (c) transparent, bright and dark plastic.

3.4.1 Size

The size of packaging was determined through surface area estimation algorithm. The average size of plastic packaging in the testing dataset (30 real world plastic packaging) was 63.94 cm^2. The number of small, medium and large packaging was 8,11 and 11 respectively.

The results are shown in Figure 9 (a). For SVM, the system achieved 100% accuracy for small and medium plastic packaging but experienced a drop to 91% accuracy for large plastic packaging . Similarly, for ANN, logistic regression, and Decision Tree models, the accuracy in detecting small plastic packaging was 100%. For PLS-DA, the accuracy in identifying small plastic packaging was 100%, but it decreased to 82% and 91% when detecting medium and large sizes, respectively. The accuracy of the models dropped when detecting large and medium plastics, as some samples in these categories have opaque colours. However, the accuracy for detecting brightly coloured plastics, regardless of size, is 100%. Thus, sizes larger than 8 cm^2, which is the size of the smallest plastic packaging, have no impact on the model's accuracy.

3.4.2 Level of contamination

The average level of contamination of real-world plastic packaging was measured at 37%. Figure 9(b) illustrates the accuracy of the system in identifying types of polymers with several degrees of contamination. In the testing dataset, there were 9 plastic packaging samples with a low level of contamination, 10 with medium contamination, and 6 with a high contamination level. 5 pieces of the plastic packaging could not have their contamination level measured due to the presence of labels and transparent oily contamination.

The SVM models performed best, the level of contamination had a low impact on the accuracy of the system. Even with the highest level of contamination reaching 83%, the model still correctly identified the plastic. For ANN, logistic regression, and decision tree models, the accuracy of the model in identifying low-level contaminated plastic was 100%, but it decreased to 86% and 89% when identifying medium and highly contaminated packaging, respectively. Similarly, for PLS-DA, the accuracy of the model in identifying plastic with a low level of contamination was 100%, but it dropped to 71% and 89% when identifying medium and high levels of contamination in plastic packaging, respectively.

3.4.3 Darkness Level

The darkness level was identified using average pixel value of greyscale image of samples in testing dataset. The testing dataset consisted of 15 transparent plastic, 6 dark coloured plastic and 9 brightly coloured plastics. The average darkness level was found to be a greyscale value of 157. The impact of darkness on successful identification is shown in Figure 9(c).

The accuracy of SVM to detect dark coloured and bright coloured plastic was 100% but it dropped to 93% when identifying transparent plastic. For ANN, Logistic Regression and Decision Tree, the accuracy of models in identifying dark plastic and transparent was 83% and 93% respectively, the accuracy increased to 100% when identifying brightly coloured plastics. For PLS-DA, the accuracy to identify dark plastic (67%) was much lower than identifying brightly coloured plastic (100%). However, the accuracy in identifying transparent plastic was 93%.

3.4.4 Food Contaminant Colour

The colour of the contaminant exerted a significant impact on the accuracy of the system. This effect is attributed to its influence on the darkness of the material. The interplay between colour and darkness proved to be a crucial factor affecting the model's performance. Figure 10(a) displays a PLA lid surface with applied food contamination indicators in black, yellow, and green colours.

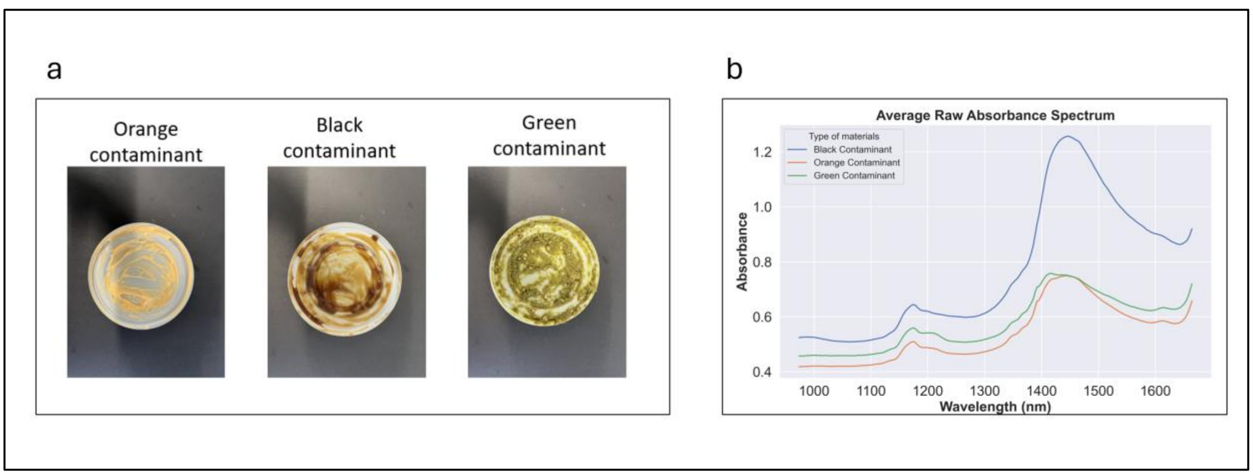

Figure 10: Effect of colour of contamination. (a). PLA lid with orange, black and green contaminant; (b) the raw absorbance spectrum of PLA packaging with various colours of contaminant (black, orange and green).

Black contaminants absorb lighter than contaminants of other colours, see figure 10(b). However, our classification models are robust due the use of HSI; it can correctly identify the type of plastic even when the contaminant is black. This result aligns with our previous research (Taneepanichskul et al., 2024).

4 Discussion and Conclusions

4.1 Classification model performance comparison

The combination of HSI and machine learning has been applied to identify and classify types of plastics with various types of contamination and contamination level. The samples in this experiment included conventional plastics (PP, PET, LDPE and HDPE) and compostable plastics (PLA, PBAT and PHA). In the training dataset, machine learning including SVM, decision tree, logistic regression, ANN achieved 100% accuracy even when classifying the plastic with low level of tomato ketchup, the models still have impressive performance. On the other hand, PLS-DA demonstrated the lowest accuracy among the models, registering a rate of 90.6% in identifying plastic samples within the training dataset.

To enhance robustness of model and mitigate overfitting, the utilization of cross validation dataset is crucial. Table 4 explains accuracy of algorithms on the cross-validation dataset, revealing that Logistic regression, SVM exhibit superior performance. These models perform well on the cross-validation dataset, indicating that they can handle new, unseen data effectively including various levels mayonnaise contamination on sample surface.

ANN and Decision Tree models exhibited accuracy rates 90% and 88% respectively in identifying contaminated plastic samples within the cross-validation dataset. These models encountered challenges, particularly in misclassifying PHA samples with high mayonnaise contamination.

The obstacles faced by models in detecting contaminated PHA samples are multi-faceted. Firstly, the inherent characteristics of thin and transparent films pose difficulties, given their low absorption in the Short-Wave Infrared (SWIR) range. HSI relies on the absorption of light by molecular vibrations, and when dealing with thin and transparent films, the limited absorption features become a hurdle for the sensor to effectively detect and differentiate materials. Secondly, the presence of thin films introduces scattering effects, causing alterations in the direction of incident light. This scattering effect can introduce variability in the measured spectra, creating challenges in maintaining the consistency required for reliable classification. Thirdly, contamination on the surface of PHA induces shifts in the absorbance spectrum, further complicating the classification process. The introduction of contaminants alters the characteristic molecular vibrations, making it challenging for the models to accurately identify and categorize the material. PLS-DA may face challenges when dealing with intricate relationships within the data, especially in scenarios where the underlying patterns are highly complex. Moreover, the dataset is small so other machine learnings performed better than PLS-DA.

4.2 Influence of material properties on the performance of the models

Our focus extended to three material properties: size, level of contamination, and darkness. An analysis reveals no discernible correlation between the size of the material (particularly when exceeding 8 cm^2) and the accuracy of the model. Surprisingly, the level of contamination demonstrated minimal influence on the system's accuracy. Darkness showed significant impact on the accuracy of the system. Opaque plastic is more difficult to be classified due to high light absorbance of SWIR region. Transparent plastic is also difficult to be identified due to the scattering of light.

Black contaminants absorb lighter than contaminants of other colours, see figure 10(b). However, our classification models are robust; it can correctly identify the type of plastic even when the contaminant is black. This result aligns with our previous research (Taneepanichskul et al., 2024).

Our SVM model for identifying polymer contamination performed comparably to the PLS-DA model developed by Bonifazi, both achieving a sensitivity of 0.99. However, our model was applied to packaging with a higher level of contamination (Bonifazi et al., 2021). Additionally, Cucuzza's Hierarchical PLS-DA model demonstrated impressive accuracy, reaching up to 1.0. These findings highlight that the integration of hyperspectral imaging with machine learning significantly enhances the recycling rate by accurately identifying polymer contamination (Cucuzza et al., 2021). Krasniewski applied various machine learning techniques to identify 11 types of polymers, finding that PET had the lowest accuracy due to its transparency, which aligns with our results (Kraśniewski et al., 2021). Importantly, our SVM model developed here enhances the performance of our previous PLS-DA model (Taneepanichskul et al., 2023). Even with highly contaminated packaging, the SVM model can identify polymers with very high accuracy, whereas our previous PLS-DA model could only accurately identify pristine plastics with accuracy dropping to 75% for highly contaminated plastics.

4.3 Application of HSI in anaerobic digestion, in-vessel composting and recycling plant for detecting food contaminated compostable plastics.

In anaerobic digestion (AD) and in-vessel composting (IVC), the first step involves sorting the waste. Pre-consumer waste, often referred to as source-separated, includes a wide range of organic materials and other contaminants. The primary task is to remove all packaging and separate organic matter from non-organic materials such as metals, minerals, dirt, and various unexpected objects. This ensures that only appropriate organic materials are processed further, improving efficiency and output quality (AnaerobicDigestion, 2023).

Depackaging and separation are carried out using machines called depackagers. The reject stream from these machines consists of packaging materials, including contaminated plastics, cardboard, glass, and metal. After the separation process, IVC primarily relies on manual sorting combined with visual inspection to identify compostable plastics, which is labor-intensive and costly (WRAP, 2009). Contaminated plastics that cannot be composted are sent to landfill or incineration. Similarly, in AD, all contaminated plastics are directed to landfill or incineration (Taneepanichskul et al., 2022).

The integration of hyperspectral imaging (HSI) with machine learning methods can enhance the system by reducing the landfill and incineration of plastics and increasing recycling and composting rates. With a detection system in place, compostable plastics can be reintroduced into the system, and recyclable plastics can be detected and sent to recycling plants.

If recycling plants were employing this detection system, a high percentage of food contaminated compostable plastics can be identified and redirected to composting facilities for proper processing. Additionally, the system can help identify the 17% of recyclable plastics are rendered non-recyclable due to food contamination (Biffa, 2022). Implementing this system would require an automatic separation system to act on identification and characterisation provided by the HSI system. These already exist in modern waste recycling facilities. They are less common in AD and

IVCs. Investment in these facilities would have to be driven by a return on investment which is incentivised by lower number of plastics being sent to landfill and incineration.

5. References

ALLISON, A. L., PURKISS, D., LORENCATTO, F., MIODOWNIK, M. & MICHIE, S. 2022. Improving compostable plastic disposal: An application of the Behaviour Change Wheel intervention development method. *Frontiers in Sustainability,* 3**,** 92.

AMIGO, J. M., BABAMORADI, H. & ELCOROARISTIZABAL, S. 2015. Hyperspectral image analysis. A tutorial. *Analytica chimica acta,* 896**,** 34-51.

ANAEROBICDIGESTION. 2023. *Anaerobic Digestion of Pre-Consumer Food Waste & 3 Generations of Depackaging Machines* [Online]. Available: https://blog.anaerobic-digestion.com/anaerobic-digestion-of-pre-consumer-food-waste/ [Accessed 11 July 2024].

BARKVED, K. 2022. *The Difference Between Training Data vs. Test Data in Machine Learning* [Online]. Available: https://www.obviously.ai/post/the-difference-between-training-data-vs-test-data-in-machine-learning [Accessed 6 Jan 2024].

BIFFA. 2022. *Nearly 17% of recycling is wasted due to contamination* [Online]. Available: https://www.recycling-magazine.com/2022/10/26/nearly-17-of-recycling-is-wasted-due-to-contamination/ [Accessed 3 July 2024].

BIOFERTILISER. 2022. *Annual Report 2022 Compost / Biofertiliser Certification Schemes* [Online]. Available: https://www.biofertiliser.org.uk/images/upload/news_1183_BCS-Annual-Report-2022Digitalv2-single-pages.pdf [Accessed 5 June 2024].

BONIFAZI, G., GASBARRONE, R. & SERRANTI, S. 2021. Detecting contaminants in post-consumer plastic packaging waste by a NIR hyperspectral imaging-based cascade detection approach. *Detritus,* 15**,** 94-106.

BRAUN, S. 2023. *Why most plastic can't be recycled* [Online]. Available: https://www.dw.com/en/why-most-plastic-cant-be-recycled/a-64978847 [Accessed 31 May 2024].

CASTRO-DÍAZ, M., OSMANI, M., CAVALARO, S., CACHO, Í., URIA, I., NEEDHAM, P., THOMPSON, J., PARKER, B. & LOVATO, T. 2023. Hyperspectral Imaging Sorting of Refurbishment Plasterboard Waste. *Applied Sciences,* 13**,** 2413.

CUCUZZA, P., SERRANTI, S., BONIFAZI, G. & CAPOBIANCO, G. 2021. Effective recycling solutions for the production of high-quality PET flakes based on hyperspectral imaging and variable selection. *Journal of imaging,* 7**,** 181.

DABO, M. 2023. *Compostable plastic: a growing trend with limited impact* [Online]. Available: https://www.packaging-gateway.com/comment/compostable-plastic-a-growing-trend-with-limited-impact/ [Accessed 31 May 2024].

DESAI, S. 2023. *Compostable Plastic Is Garbage* [Online]. Available: https://www.beyondplastics.org/news-stories/compostable-plastic-trash [Accessed 3 July 2024].

EUROPEANBIOPLASTIC. 2019. *PLA in the waste stream* [Online]. Available: https://www.european-bioplastics.org/pla-in-the-waste-stream/ [Accessed 13 May 2024].

EUROPEANBIOPLASTIC. 2023. *BIOPLASTICS MARKET DEVELOPMENT UPDATE 2023* [Online]. Available: https://www.european-bioplastics.org/bioplastics-market-development-update-2023-2/ [Accessed 1 April 2024].

FRANKLIN-WALLIS, O. 2019. *'Plastic recycling is a myth': what really happens to your rubbish?* [Online]. Available: https://www.theguardian.com/environment/2019/aug/17/plastic-recycling-myth-what-really-happens-your-rubbish [Accessed 31 May 2024].

HYSPEX. 2023. *Hyspex Baldur S-640iN* [Online]. Available: https://www.hyspex.com/hyspex-products/hyspex-baldur/hyspex-baldur-s-640i-n/ [Accessed].

KABIR, M. H., GUINDO, M. L., CHEN, R. & LIU, F. 2021. Geographic origin discrimination of millet using Vis-NIR spectroscopy combined with machine learning techniques. *Foods,* 10**,** 2767.

KRAŚNIEWSKI, J., DĄBAŁA, Ł. & LEWANDOWSKI, M. Hyperspectral imaging for analysis and classification of plastic waste. 2020 25th International Conference on Pattern Recognition (ICPR), 2021. IEEE, 4805-4812.

MASOUMI, H., SAFAVI, S. M. & KHANI, Z. 2012. Identification and classification of plastic resins using near infrared reflectance. *International Journal of Mechanical and Industrial Engineering,* 6**,** 213-20.

MICROSOFTBUILD. 2021. *Multiclass Neural Network component* [Online]. Available: https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/multiclass-neural-network?view=azureml-api-2 [Accessed 13 May 2024].

QIAN, Y., YE, M. & ZHOU, J. 2012. Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features. *IEEE Transactions on Geoscience and Remote Sensing,* 51**,** 2276-2291.

SCIKITLEARN. 2024. *Cross-validation: evaluating estimator performance* [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html [Accessed 5 Jan 2024].

TANEEPANICHSKUL, N., HAILES, H. & MIODOWNIK, M. 2024. Using Hyperspectral Imaging to Identify and Classify Large Microplastic Contamination in Industrial Composting Processes. *Frontiers in Sustainability,* 5**,** 1332163.

TANEEPANICHSKUL, N., HAILES, H. C. & MIODOWNIK, M. 2023. Automatic identification and classification of compostable and biodegradable plastics using hyperspectral imaging. *Frontiers in Sustainability,* 4**,** 1125954.

TANEEPANICHSKUL, N., PURKISS, D. & MIODOWNIK, M. 2022. A review of sorting and separating technologies suitable for compostable and biodegradable plastic packaging. *Frontiers in Sustainability,* 3**,** 901885.

TRANMER, M. & ELLIOT, M. 2008. Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR),* 5**,** 1-5.

ULRICI, A., SERRANTI, S., FERRARI, C., CESARE, D., FOCA, G. & BONIFAZI, G. 2013. Efficient chemometric strategies for PET–PLA discrimination in recycling plants using hyperspectral imaging. *Chemometrics and Intelligent Laboratory Systems,* 122**,** 31-39.

WOLFF, R. 2020. *What Is Training Data in Machine Learning?* [Online]. Available: https://monkeylearn.com/blog/training-data/ [Accessed 5 Jan 2024].

WRAP. 2009. *Review of Food Waste Depackaging Equipment* [Online]. Available: http://www.organicsrecycling.org.uk/uploads/article1762/Wrap%20Report%20on%20Food%20Waste%20Depackaging.pdf [Accessed 11 July].

ZHANG, W., KASUN, L. C., WANG, Q. J., ZHENG, Y. & LIN, Z. 2022. A review of machine learning for near-infrared spectroscopy. *Sensors,* 22**,** 9764.