



UCL Open Environment Preprints

Article Title:

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

Author(s):

Nutchana Taneepanichskul¹, Helen C Hailes², Mark Miodownik³

Affiliations:

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

ORCID IDs:

0000-0001-5574-4742²,

Corresponding Author:

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

DOI:

<https://doi.org/10.14324/ucloepreprints.282.v2>

Licence information:

This is an open access preprint distributed under the terms of the Creative Commons Attribution License (CC BY) 4.0. <https://creativecommons.org/licenses/by/4.0/>.

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This article is a preprint and has not been peer-reviewed.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 2 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.

Preprint first published online:

18-02-2025

Keywords:

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

1 **Using Hyperspectral Imaging and Machine Learning to Identify**
2 **Food-Contaminated Compostable and Recyclable Plastics**

3 Nutchana Taneepanichskul, Helen C. Hailes and Mark Miodownik

4 **Abstract**

5 With the increasing public legislation aimed at reducing plastic pollution, compostable plastics
6 have emerged as an alternative to conventional plastics for some food packaging and food service
7 items. However, the true value of compostable plastics can only be realized if they do not enter the
8 environment as contaminants but instead are processed along with food waste using industrial
9 composting facilities. Distinguishing compostable plastics from other plastics in this waste stream
10 is an outstanding problem. Currently, near-infrared (NIR) technology is widely used to identify
11 polymers, but it falls short in distinguishing plastics contaminated with food waste. This study
12 investigates the application of hyperspectral imaging (HSI) to address this challenge, enhancing
13 the detection and sorting of contaminated compostable plastics. By combining HSI with various
14 machine learning algorithms we show it is possible to accurately identify and classify plastic
15 packaging with food waste contamination, achieving up to 99% accuracy. The study also measures
16 the impact of plastic features such as darkness, size, and level of contamination on model
17 performance, with darkness having the most significant impact. The developed machine learning
18 model can detect plastic with higher levels of contamination more accurately compared to our
19 previous study. Implementing HSI in waste management systems can significantly increase
20 composting and recycling rates, improve the quality of recycled products. This advanced approach
21 supports the circular economy by ensuring that both compostable and recyclable plastics are
22 effectively processed and recycled, minimizing environmental impact.

23
24 Keywords: food contaminated plastics, hyperspectral imaging (HSI), recycling, composting,
25 machine learning, automatic sorting

26

27

28

29 1. Introduction

30 The increasing popularity of compostable and biodegradable plastics underscores the need for
31 efficient sorting technologies to separate and collect them for waste processing. In 2023 they
32 represented 52.1% of the global bioplastic production (EuropeanBioplastic, 2023). In the current
33 UK waste management system, organic waste is managed through two primary methods: In-Vessel
34 Composting (IVC) and Anaerobic Digestion (AD). Compostable and biodegradable plastics
35 should be directed to IVC, where they can break down into compost. In contrast, AD is not suitable
36 for compostable plastics, as they can clog the system. However, the systems often fail to detect
37 and separate compostable plastics, especially when contaminated with food waste, leading to their
38 improper disposal in landfills, incinerators, or misdirection to AD or recycling plants. This
39 significantly contributes to low recycling and composting rates (Allison et al., 2022).

40 NIR optical sorting is a widely used technology in recycling facilities for separating different types
41 of plastics (Taneepanichskul et al., 2022). NIR offers significant advantages over RGB-based
42 imaging methods, as it relies on the distinct spectral signatures of various plastic polymers to
43 achieve accurate sorting. In contrast, RGB-based methods often struggle to identify plastics with
44 varying transparency, colour, and quality (Tamin et al., 2023b). Combining RGB with near-infrared
45 imaging can enhance the model's overall performance (Tamin et al., 2023a, Tamin et al., 2022b,
46 Tamin et al., 2022a). However, food waste contamination, poses significant challenges to the
47 efficiency and effectiveness of NIR optical sorting due to issues with spectral absorption reflection
48 of NIR frequencies by the food residues on the plastic surfaces (Masoumi et al., 2012).
49 Additionally, the presence of food waste introduces extra spectral signals in the NIR range, creating
50 noise that makes it harder for the system to accurately identify the polymer type.

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

57 While numerous studies have explored the application of HSI for identifying various types of
58 plastics, there remains a notable gap in research that specifically addresses the challenge of

59 detecting food contaminated compostable plastics, which represents a significant issue within the
60 context of current plastic waste management systems.

61 In 2013 Ulrici et al. used HSI and partial least squares discriminant analysis (PLS-DA) to
62 distinguish PET and PLA achieving over 98% accuracy with just six variables on the reduced
63 matrix (Ulrici et al., 2013). Subsequently, Bonifazi used HSI with machine learning to sort paper,
64 cardboard, plastics, and multilayer packaging. A PLS-DA-based model achieved a 0.933
65 recognition and reliability rate, making HSI a reliable, low-cost solution for identifying impurities
66 and composite materials in plastic waste streams (Bonifazi et al., 2021). Taneepanichskul et al then
67 applied HSI (Taneepanichskul et al., 2024) together with PLS-DA to identify and classify
68 compostable plastics (PLA and PBAT), compostable materials (sugarcane and palm leaf derived
69 packaging) and conventional plastics (LDPE, PET and PP). PLS-DA achieved a perfect
70 classification (100%) for virgin materials larger than 10 mm x 10 mm (Taneepanichskul et al.,
71 2023). Taneepanichskul et al. also recently studied the impact of packaging properties such as
72 darkness, colour, size, and contamination, showing how they all impacted identification. The
73 accuracy of the system decreased when detecting plastics that were dark, thin, small, or had high
74 levels of contamination (Taneepanichskul et al., 2024).

75 Currently, no research has been conducted on identifying plastics contaminated with food waste,
76 which constitutes a significant portion of plastic waste in the UK waste management system. In
77 this paper, we present the development of chemometric and machine learning algorithms
78 integrated with HSI, demonstrating their effectiveness in identifying compostable and recyclable
79 plastics with varying types and levels of food contamination. This study improves upon our
80 previous works (Taneepanichskul et al., 2023, Taneepanichskul et al., 2024), which focused solely
81 on virgin plastics and those contaminated with compost from IVC. The enhanced models
82 demonstrate excellent performance, effectively detecting plastics with diverse types and varying
83 levels of food contamination. The overall model accuracy for identifying contaminated plastics
84 has improved from 85% to 97%. Additionally, the study explores the impact of real-world food
85 plastic packaging properties such as size, colour and darkness, on the performance of the system.

86 The remainder of this manuscript is structured as follows: there is a Materials and Methods section
87 which details the simulation of food contamination, the materials used to construct the training,
88 cross-validation, and testing datasets, and the process of hyperspectral image acquisition and

89 analysis. Additionally, the performance of the classification models is analysed, and the algorithm
90 for evaluating plastic features, including size, darkness, contamination levels and food
91 contaminant colours is described. In the Results section, the average raw and pre-processed spectra
92 are presented, and the performance of each classification model is calculated and compared. The
93 impact of plastic features (size, darkness, contamination levels and food contaminant colours) on
94 the accuracy of the developed classification model is analysed. In the Discussion and Conclusion
95 sections, the results are compared with findings from our previous studies as well as other related
96 studies to provide valuable insight. Additionally, the use of HSI for increasing the efficiency of
97 detecting food-contaminated compostable plastics in the waste streams of AD, IVC, and recycling
98 plants is discussed.

99 **2. Materials and Methods**

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

110 **2.1 Simulating Food Contamination**

111 The contamination levels in this experimental setup were categorized into three levels: low (25%),
112 medium (50%), and high (75%). Figure 1 illustrates the contamination process, depicting the
113 simulation of 25%, 50%, and 75% contamination using tomato ketchup. Each sample was cut into
114 50 mm x 50 mm pieces with a thickness of 0.4 mm and divided into four equal sections.

115 Two sauces were used to simulate food contamination: tomato ketchup and mayonnaise. These
116 were chosen due to their ability to be applied repeatably and consistently to the samples. The
117 different compositions help to create training data and cross-validation data. The compositions of
118 these two sauces are shown in Table 1. These condiments are suitable proxies for food

119 contamination because they can represent high water activity foods such as dips and sauces,
 120 prepared salads, and dairy products; acidic foods such as pickled products, fermented foods, and
 121 fruit-based sauces; emulsified foods such as salad dressings, processed meats, and butter and
 122 margarine; and fat-containing foods. Their compositional similarities to a wide range of other food
 123 products make them ideal for studying contamination and spoilage patterns across different food
 124 categories.

125 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

126

127 Additionally, their viscosity and texture allow them to adhere well to surfaces, effectively

128 simulating real-life conditions of food residue on plastics. This makes them ideal for testing

129 cleaning and contamination processes. In this study, HEINZ tomato ketchup was used for both the

130 training and cross-validation datasets, while HEINZ mayonnaise was used for the cross-validation

131 dataset.

132 To achieve 25% contamination, tomato ketchup or mayonnaise was applied to one section; for

133 50% contamination, it was applied to two sections; and for 75% contamination, it was applied to

134 three sections. The ketchup and mayonnaise were then spread to ensure they covered the entire

135 plastic surface.

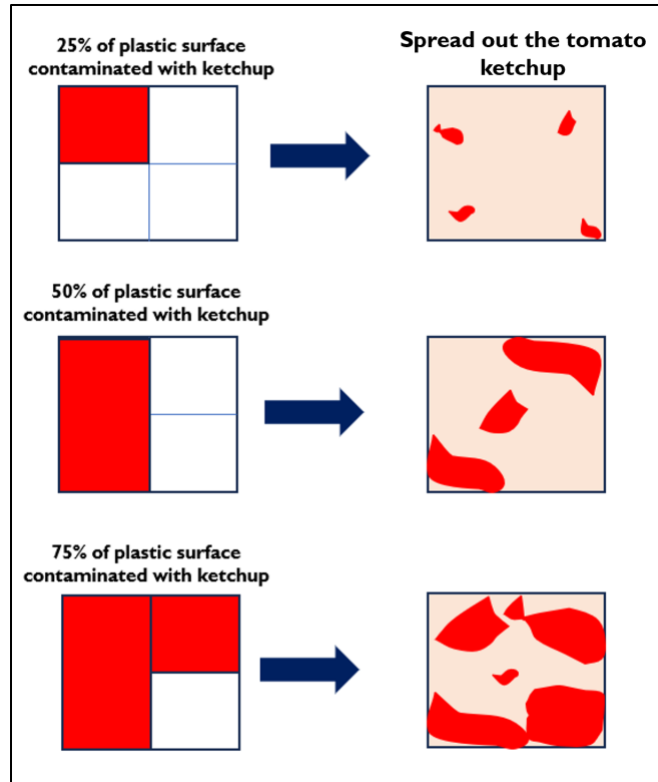


Figure 1: Simulated contamination levels of 25%, 50% and 75% sauces.

136

137

138

139 2.2 Sample Preparation

140 The experimental samples encompassed several size and contamination levels, with both
 141 conventional and compostable plastics. Within the category of conventional plastics, Low-Density
 142 Polyethylene (LDPE), High-Density Polyethylene (HDPE), Polyethylene Terephthalate (PET),
 143 and Polypropylene (PP) were represented. The compostable plastic category comprised Polylactic
 144 Acid (PLA), Polybutylene Adipate Terephthalate (PBAT), and Polyhydroxyalkanoate (PHA).

145 The materials were allocated into three datasets, namely calibration, cross-validation, and testing
 146 datasets as mentioned earlier. The training dataset encompassed both pristine plastics and plastics
 147 contaminated with low level of tomato ketchup (25%). This approach allowed the models to learn
 148 patterns associated with both clean and contaminated plastics. It also helped the models identify
 149 the essential features of plastics from pristine samples while adapting to variations introduced by
 150 specific type of contamination (tomato ketchup). The details of the materials within the training
 151 dataset are presented in Table 2.

152

153

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

155

156 In the cross-validation set, there were three replicates each with 50% and 75% tomato ketchup
 157 contamination of the plastic samples and 25%, 50%, and 75% mayonnaise contamination of the
 158 plastic samples as shown in Table 3. The models were validated using high levels of contamination
 159 from tomato ketchup and new type of contamination (mayonnaise) to assess their robustness. This
 160 approach ensured that the models could effectively handle complex scenarios and accurately
 161 identify plastic types, even under higher levels of tomato ketchup contamination and new
 162 contamination conditions (mayonnaise).

163

164

165

166

167

168

169

170

171

172

173

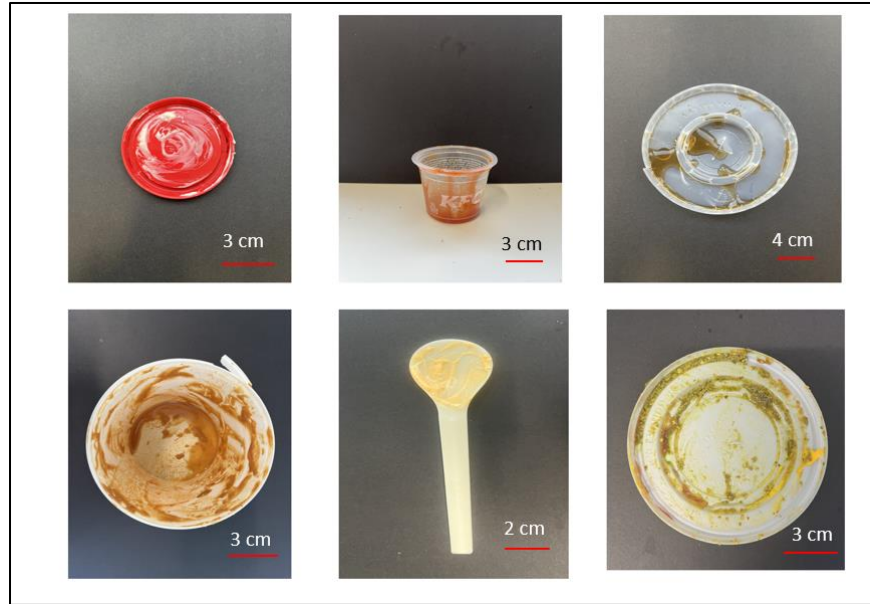
Table 3: List of samples in cross validation dataset

Material	Material condition	Size	Number of replicates per plastic type
LDPE, HDPE, PET, PP, PLA, PBAT and PHA	Plastics with 50% level of tomato ketchup contamination	50 mm x 50 mm	21
	Plastics with 75% level of tomato ketchup contamination		21
	Plastics with 25% level of mayonnaise		21
	Plastics with 50% level of mayonnaise		21
	Plastics with 75% level of mayonnaise		21

175

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

182



183
184
185
186

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

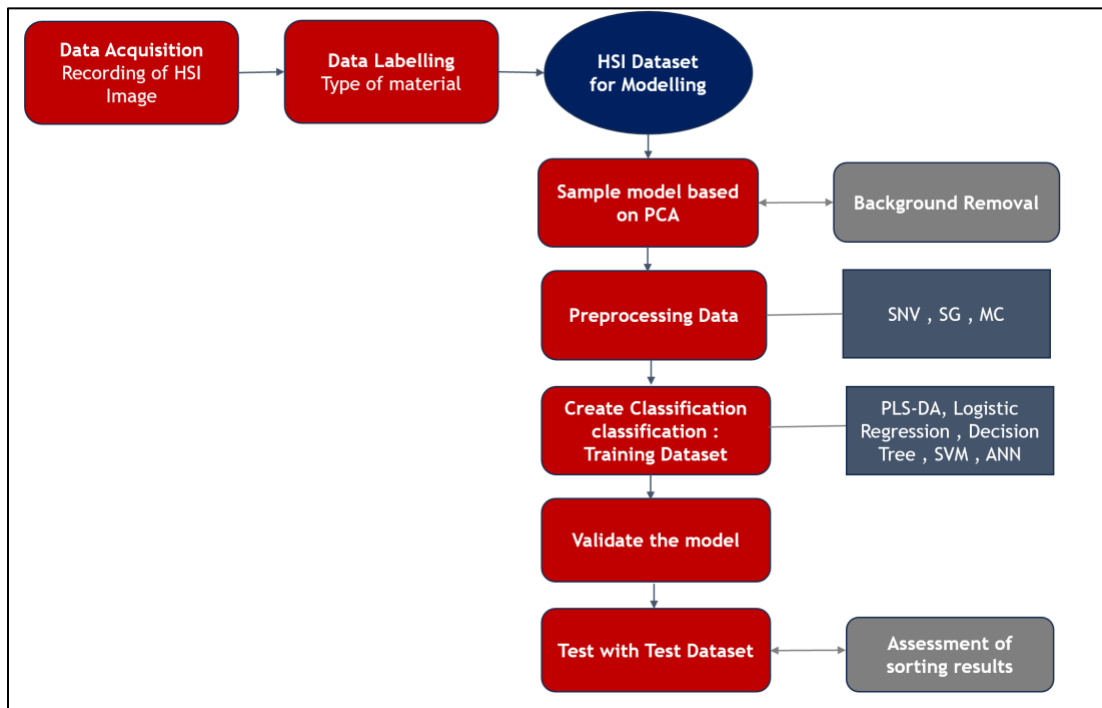
187 The distribution of training, cross validation and testing dataset was 54%, 36 % and 10%. This
188 distribution reflected a strong focus on training the model, supported by a well-balanced validation
189 dataset and a thoughtfully allocated testing set to ensure comprehensive evaluation.

190
191
192
193
194
195
196
197
198
199

200 **2.3 Imaging methodology**

201 HSI captures a continuous spectrum for each pixel of an object, enabling precise material
202 identification based on unique spectral signatures. Unlike NIR, HSI generates spatial maps that
203 visualize material distribution allowing multi-materials to be detected in a single piece of packaging.
204 Unlike NIR it can detect heterogeneity of contamination, and analyse complex mixtures with
205 unparalleled precision (Bhargava et al., 2024, Taneepanichskul et al., 2022).

206 **2.3.1 Hyperspectral data acquisition and analysis schematic**



207

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

209 As shown in Figure 3 hyperspectral images were obtained from line scans of the samples on a
210 conveyor belt passing under a HySpex Baldur S-640i N camera. The camera, positioned at a
211 working distance of 1 metre with a 16° field of view, covered a spectral range of 950 to 1730 nm
212 with a spectral resolution of 3.36 nm, resulting in a total of 232 spectral bands. The spatial pixel
213 size of the images was 0.44 mm (Hypex, 2023). The system's conveyor belt measured 700 mm in
214 length, 215 mm in width, and 60 mm in height, with speed of 120 mm/s. The image capture
215 background was the black conveyor belt. A halogen lamp, emitting light across the spectrum from
216 400 nm to 2500 nm, was employed as the light source. This experimental setup has been described

217 in more detail in our previous work. (Taneepanichskul et al., 2024, Taneepanichskul et al., 2023).
218 HypexGround software facilitated the acquisition of the hyperspectral image. Subsequently, the
219 Breeze software package was employed for PCA model development, spectrum preprocessing,
220 application of diverse machine learning algorithms for classification, and production of
221 classification results.

222 **2.3.2 Principal component analysis (PCA) and spectrum pre-processing**

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

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

236 **2.3.3 Machine learning classification model**

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

241 **2.3.3.1 Logistic regression**

242 Logistic regression is a fundamental supervised learning method widely utilized for classification
243 tasks, particularly in scenarios involving binary outcomes. Through the sigmoid function, it

244 transforms spectral band values to produce probabilities for binary predictions, with coefficients
245 assigned to each band indicating their predictive influence (Qian et al., 2012, Kabir et al., 2021).

246 Logistic regression can be extended to address multiclass classification problems through softmax
247 regression. The softmax function normalizes the output into a probability distribution across
248 multiple classes, ensuring that the sum of the predicted probabilities for all classes equals unity.
249 This way, the model can provide predictions for each class, and the class with the highest
250 probability is considered as the final prediction (Tranmer and Elliot, 2008).

251 **2.3.3.2 Decision tree (DT)**

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

259 **2.3.3.3 Support vector machine (SVM)**

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

270 **2.3.3.4 Artificial neural network (ANN)**

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

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

287 **2.3.3.5 Partial least squares discriminant analysis (PLS-DA)**

288 PLS-DA, a blend of partial least squares regression (PLS-R) and discriminant analysis (DA), is a
289 supervised ML method for dimensionality reduction and material class prediction. It necessitates
290 an X matrix with calibration spectra and a corresponding Y matrix denoting class identity (types
291 of plastic). In binary cases, Y is a single column; for multiclass scenarios, it's a dummy matrix with
292 1's and 0's indicating class membership. The model's output isn't strictly binary, requiring a
293 threshold establishment during prediction. Setting thresholds employs various methods, with
294 Bayes' Theorem being a prevalent choice. Alternatively, a 0.5 cut-off point is often employed for
295 binary classification tasks (Amigo et al., 2015). In our PLS-DA, the linear equation was modelled
296 with around 5 latent variables, enabling graphical visualization and understanding through LV
297 scores and loadings.

298 **2.4 Classification model performance (model validation)**

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

307
$$\text{Sensitivity (Recall)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \text{ (Equation 1)}$$

308

309
$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \text{ (Equation 2)}$$

310

311
$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \text{ (Equation 3)}$$

312

313
$$\text{F1 - Score} = \frac{\text{True positive}}{\text{True positive} + \frac{1}{2}(\text{False Positive} + \text{False Negative})} \text{ (Equation 4)}$$

314

315
$$\text{Accuracy} = \frac{\text{True Negative} + \text{True Positive}}{\text{True Negative} + \text{True Positive} + \text{False Negative} + \text{False Positive}} \text{ (Equation 5)}$$

316

317 **2.5 The evaluation of plastic features in testing dataset**

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

322 **2.5.1 Size**

323 The plastic packaging images in the testing dataset were resized to 10 cm x 15 cm and converted
324 to greyscale. Otsu's thresholding method was then applied to remove the background and convert

325 the greyscale images to binary format. In this process, pixels with values below the threshold were
326 set to 0, while those above the threshold were set to 255. Following this, the percentages of
327 foreground and background areas were calculated. These percentages were then multiplied by 150
328 cm² (the total area of the frame) to determine the area occupied by the plastic packaging. The size
329 was classified into 3 categories: small ($< 20 \text{ cm}^2$), medium ($20 \text{ cm}^2 \leq \text{area} < 80 \text{ cm}^2$) and large
330 ($\geq 80 \text{ cm}^2$) (Taneepanichskul et al., 2024).

331 **2.5.2 Level of contamination**

332 K-means clustering was applied to assess the level of contamination in plastic packaging within
333 the testing dataset. The images were loaded and converted to greyscale, with each pixel represented
334 as a vector based on its greyscale intensity. We selected the number of clusters to be 3. The
335 centroids of each cluster were initialized, and for each pixel in the image, a similarity measure was
336 calculated to determine its proximity to each cluster centroid using a distance metric, such as
337 Euclidean distance. Based on this calculation, the pixel was assigned to the cluster with the closest
338 centroid, forming the initial clusters. Upon convergence, the algorithm produced the final
339 clustering results. At this stage, each pixel was firmly assigned to a specific cluster, and the cluster
340 centroids represented the average greyscale intensities of the pixels within their respective clusters.
341 The number of pixels in each cluster was counted, and their ratios were calculated to determine
342 the percentage of contamination. The level of contamination in the plastic packaging was classified
343 into four categories: low contamination ($< 25\%$), medium contamination ($25\% \leq \text{contamination} <$
344 60%), high contamination ($\geq 60\%$), and indeterminate due to multicoloured packaging or oily
345 contamination (Taneepanichskul et al., 2024).

346 **2.5.3 Darkness**

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

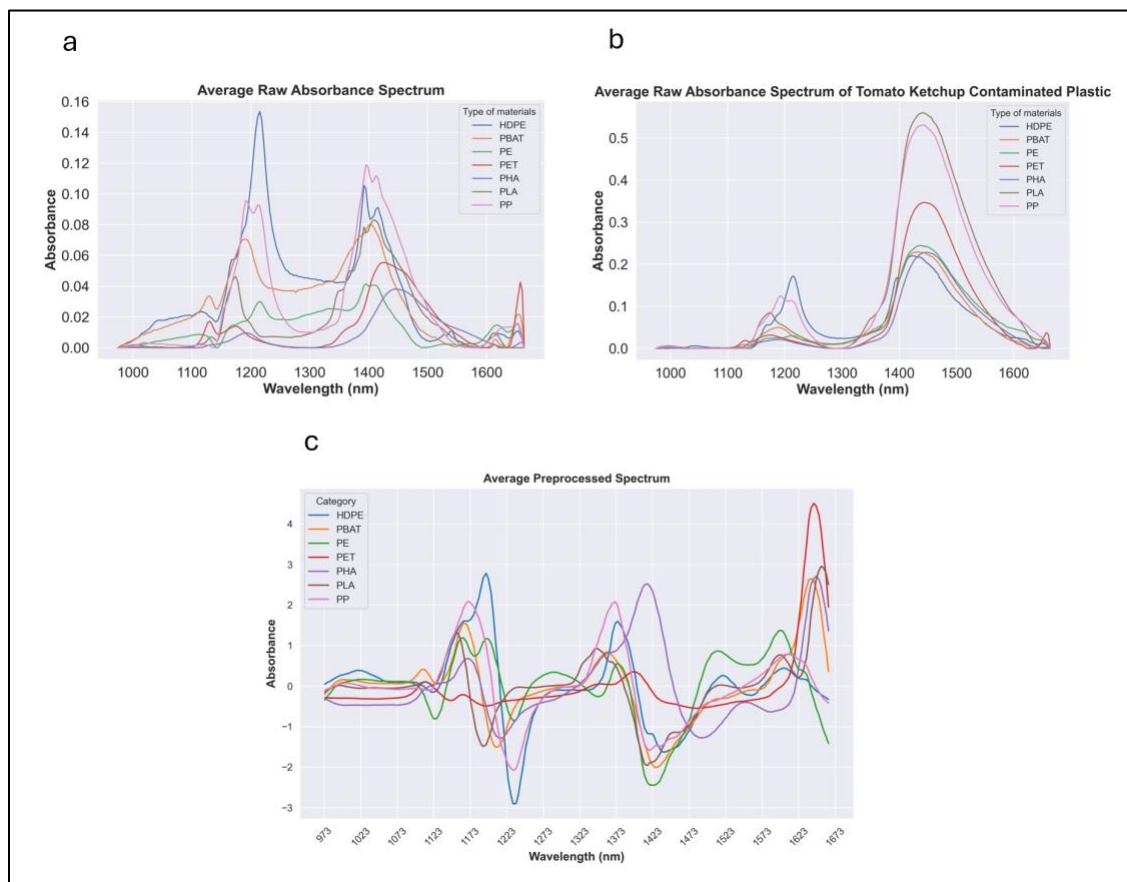
351 **3. Results**

352 The results of this study are presented to evaluate the effectiveness of the developed classification
353 models and the impact of plastic attributes, such as size, darkness, and contamination levels, on

354 accuracy. The section begins with an analysis of the raw and pre-processed spectra, followed by a
355 comparison of model performances. Finally, the influence of plastic packaging attributes on
356 classification accuracy is demonstrated.

357 3.1 Average raw absorbance spectrum and pre-processed spectrum

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



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

372 **3.2 Principal Component Analysis (PCA)**

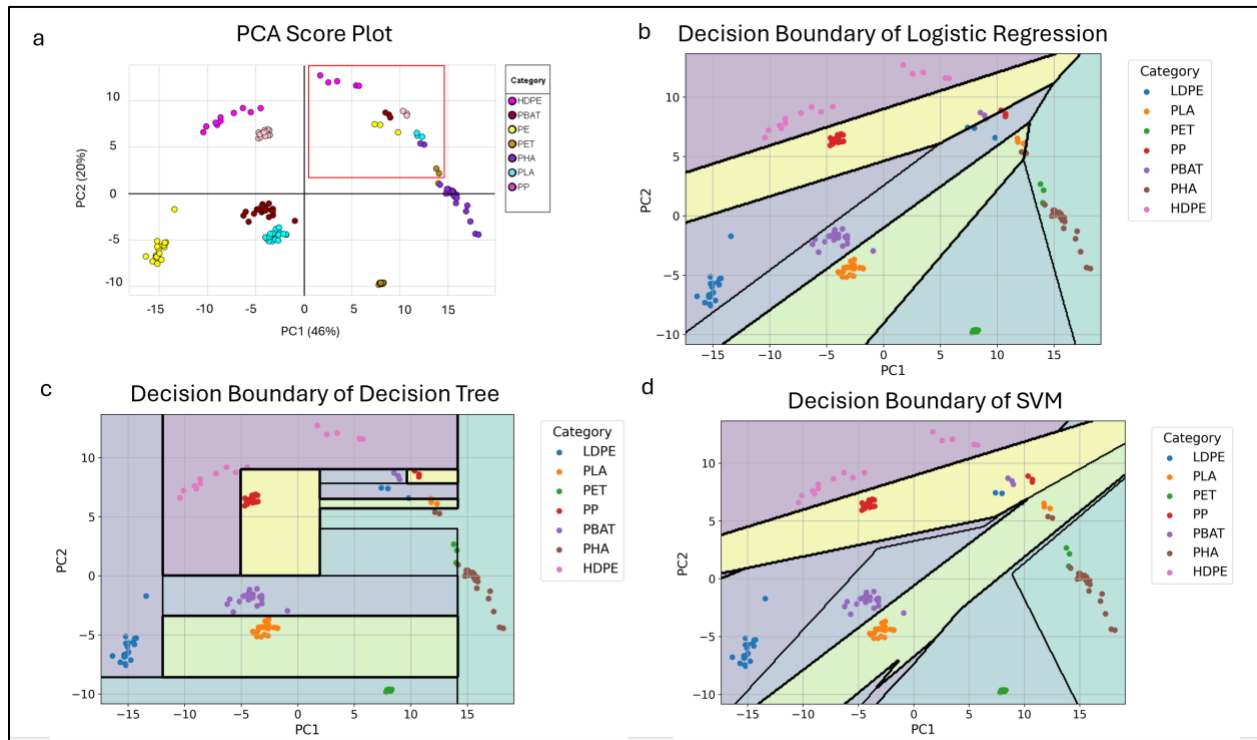
373 Following the preprocessing of the absorbance spectrum, a principal component analysis (PCA)
374 was carried out to achieve dimensional reduction. The spectra of pristine and tomato ketchup
375 contaminated plastics from the training dataset were then utilized to generate a PCA score plot, as
376 depicted in Figure 5(a). The results indicate that a substantial portion of the variance is effectively
377 captured by the first principal component (PC1), which accounts for 46%, and the second principal
378 component (PC2), which contributes 20%. Pristine plastics showed a high level of separability.
379 Specifically, pristine HDPE and PP are situated in the second quadrant, while LDPE, PBAT, and
380 PLA are in the third quadrant, and PET and PHA are in the fourth quadrant. Plastics contaminated
381 with tomato ketchup, which are indicated by the red box, are all located in the first quadrant but
382 show some overlap with each other.

383 **3.3 Performance of classification models**

384 **3.3.1 Performance of classification models on calibration dataset**

385 The calibration dataset consists of pristine plastics and plastics with the low level of tomato
386 ketchup (25%). We have applied five machine learning techniques to build classification model on
387 training dataset. Figure 5(b), (c) and (d) illustrates the decision boundary of logistic regression,
388 decision trees, and SVM classification respectively, providing a visual representation of how these
389 algorithms partition the feature space to classify different plastic samples in training dataset. The
390 decision boundary delineates the regions where each class is predicted, offering insights into the
391 complexity and separability of the dataset. This visualization aids understanding of the underlying
392 behaviour of the models and their ability to discriminate between different classes of plastics based
393 on the provided features. However, ANN and PLS-DA do not have a straightforward decision
394 boundary. ANN operates through complex transformations of the input data. The decision-making
395 process in an ANN involves a series of interconnected neurons with weighted connections. PLS-
396 DA works by finding linear combinations of features that best separate the classes in the data.

397 Unlike traditional classifiers, PLS-DA does not directly define a decision boundary. Instead, it
398 projects the data into a new space where the classes are maximally separated along latent variables.
399 Consequently, it is not as intuitive to visualize the decision boundary in the original feature space.



400
401 Figure 5: (a) PCA score plot of training dataset, contaminated plastics identified within red box;
402 (b) Logistic regression; (c) Decision tree; (d) SVM for the training dataset.

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

408

409

410

411

412

413

414 Table 4: The performance of various machine learning algorithms in identifying plastics within
415 the training dataset.

Machine Learning Methods	Polymer	Sensitivity (Recall)	Specificity	Precision	F1 Score	Overall Accuracy
Logistic Regression, SVM, Decision Tree, ANN	PLA	1	1	1	1	100%
	PBAT	1	1	1	1	
	PHA	1	1	1	1	
	PET	1	1	1	1	
	PP	1	1	1	1	
	HDPE	1	1	1	1	
	LDPE	1	1	1	1	
PLS-DA	PLA	0.91	1	0.99	0.95	90.6%
	PBAT	1	1	1	1	
	PHA	0.87	1	1	0.93	
	PET	0.87	1	1	0.93	
	PP	0.87	1	1	0.93	
	HDPE	1	1	1	1	
	LDPE	0.91	1	0.99	0.95	

416

417 3.3.2 Performance of classification models on cross-validation dataset

418 Before testing the model with real world contaminated food packaging, we applied these models
419 to classify types of materials in cross validation dataset to assess generalization of data which
420 included a new type of contamination (mayonnaise). The results are summarised in Table 5.

421

422 The logistic regression model performed well on datasets 95% accuracy. For PLA, PBAT, PET,
423 HDPE, and LDPE it achieved perfect scores of 1 for sensitivity, specificity, precision and F1 score.

424 However, it encountered challenges in accurately detecting PHA due to a new type of
425 contamination and the presence of thin film. Consequently, instances of PHA were misclassified

426 as PP, resulting in a decrease in sensitivity for PHA to 0.67 and a decrease in specificity for PP to
427 0.94.

428 The SVM model achieved 94% accuracy. For PLA, HDPE, and LDPE it achieved perfect scores
429 of 1 for sensitivity and specificity. However, like logistic regression, its performance declined
430 when classifying PHA and PBAT. The sensitivity of PHA and PBAT was 0.67 and 0.93
431 respectively. Misclassifications of PHA (66.7%) and PBAT (6.7%) as PET led to decreases in
432 specificity for PP and PET, resulting in values of 0.94 and 0.99 respectively. Consequently, the
433 precision for PBAT, PHA, PP, and PET dropped to 0.99, 0.99, 0.74 and 0.94 respectively and the
434 F1 scores for PBAT, PHA, PP, and PET decreased to 0.96, 0.8, 0.85, and 0.97 respectively.

435
436 The decision tree model achieved 88% accuracy, encountering difficulties in accurately identifying
437 PBAT, PET, and PHA. Specifically, the sensitivity for PBAT, PHA, and PET was 0.89, 0.6, and
438 0.85 respectively, while other types of plastics achieved a sensitivity of 1. PBAT was often
439 misclassified as PP (13.3%) and LDPE (6.7%), while PET was misclassified as PP (26.7%). PHA
440 suffered misclassifications as LDPE (6.7%) and PP (33.3%). Regarding specificity, PP and LDPE
441 exhibited lower values compared to other plastic types, with scores of 0.88 and 0.98 respectively.
442 Consequently, the precision for PBAT, PP, and PET decreased to 0.73, 0.57 and 0.6 respectively.
443 Additionally, the F1 scores for PBAT, PHA, PP, and PET were 0.89, 0.75, 0.97, and 0.85
444 respectively.

445 The ANN model demonstrated strong overall performance with an accuracy of 90%. In the cross-
446 validation dataset, it achieved excellent sensitivity, specificity, and F1 scores for all types of plastic
447 except for PHA and LDPE, where sensitivity dropped to 0.73 and 0.6 respectively. Furthermore,
448 the model exhibited misclassifications, 40% of LDPE being incorrectly labelled as PP, while
449 26.7% of PHA samples were misclassified as PP. Additionally, the specificity of PP was low at
450 0.89. Consequently, the precision for PHA and PP dropped to 0.99 and 0.6 and the F1 scores for
451 PHA, PP, and LDPE were computed as 0.84, 0.75, and 0.75 respectively. Overall, while the model
452 achieved impressive accuracy and performance for most plastic types, there are evidently areas for
453 improvement, particularly in accurately distinguishing PHA and LDPE, as well as reducing
454 misclassifications, especially between LDPE and PP.

455 The performance of PLS-DA fell short compared to other machine learning algorithms, achieving
 456 an overall accuracy of only 75%. Due to the introduction of a new type of contamination
 457 (mayonnaise), misclassifications occurred across various plastic types: 6.7% of PBAT, 20% of
 458 PET, 13.3% of PLA, 20% of PP, and 46.7% of LDPE could not be identified. Additionally, 13.3%
 459 of PBAT samples were misclassified as LDPE. Misclassifications were observed between various
 460 plastic types as well, with 20.3% of LDPE and 26.7% of PLA incorrectly labelled as PBAT, while
 461 6.7% of PHA samples were misclassified as PLA. Consequently, the sensitivity of PHA was the
 462 lowest at 0.47, followed by PBAT, LDPE, PET, and PP at 0.8, while PLA had a sensitivity of 0.6.
 463 For specificity, all polymers in the cross-validation dataset achieved values greater than 0.9,
 464 indicating strong performance in correctly identifying true negatives. However, PLA, PBAT, and
 465 PP exhibited slightly lower specificity compared to others, with values of 0.99, 0.92, and 0.98
 466 respectively. Additionally, the precision for PLA, PBAT, PHA, PP was 0.9, 0.62, 0.96 and 0.8
 467 respectively the F1 score for PHA was the lowest at 0.63, followed by PBAT, PLA, LDPE, PET,
 468 and PP, which achieved scores of 0.7, 0.72, 0.8, and 0.89 respectively.

469 Table 5: The performance of classification models on cross validation dataset

Machine Learning Methods	Polymer	Sensitivity (Recall)	Specificity	Precision	F1 Score	Overall Accuracy
Logistic regression	PLA	1	1	1	1	95%
	PBAT	1	1	1	1	
	PHA	0.67	1	0.99	0.8	
	PET	1	1	1	1	
	PP	1	0.94	0.74	0.85	
	HDPE	1	1	1	1	
	LDPE	1	1	1	1	
SVM	PLA	1	1	1	1	94%
	PBAT	0.93	1	0.99	0.96	
	PHA	0.67	1	0.99	0.8	
	PET	1	0.99	0.94	0.97	

	PP	1	0.94	0.74	0.85	
	HDPE	1	1	1	1	
	LDPE	1	1	1	1	
Decision tree	PLA	1	1	1	1	88%
	PBAT	0.89	1	0.73	0.8	
	PHA	0.6	1	1	0.75	
	PET	0.85	1	0.60	0.7	
	PP	1	0.88	0.57	0.73	
	HDPE	1	1	1	1	
	LDPE	1	0.98	0.87	0.93	
ANN	PLA	1	1	1	1	90%
	PBAT	1	1	1	1	
	PHA	0.73	1	0.99	0.84	
	PET	1	1	1	1	
	PP	1	0.89	0.60	0.75	
	HDPE	1	1	1	1	
	LDPE	0.6	1	1	0.75	
PLS-DA	PLA	0.6	0.99	0.9	0.72	75%
	PBAT	0.8	0.92	0.62	0.7	
	PHA	0.47	1	0.96	0.63	
	PET	0.8	1	1	0.89	
	PP	0.8	1	0.8	0.8	
	HDPE	1	1	1	1	
	LDPE	0.8	0.98	0.84	0.82	

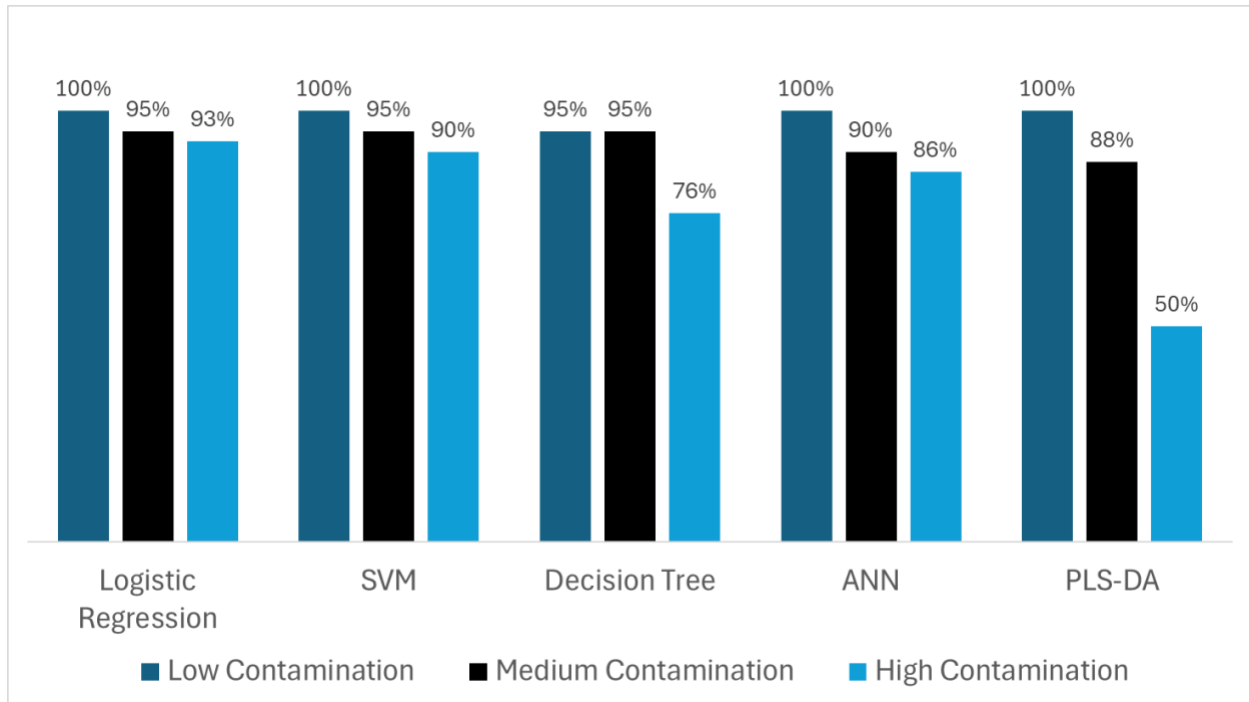
470

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

476 contamination levels (75%), the accuracy further declined to 93% for logistic regression, 90% for
477 SVM, 76% for the decision tree, 86% for ANN, and 50% for PLS-DA.

478

479



480

481

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

482

483 3.3.2 Performance of classification models on testing dataset

484 The classification models were employed to categorize 30 real-world packaging samples with
485 various types and level of contamination. The performance of each classification model showed in
486 table 6.

487

488

489

490

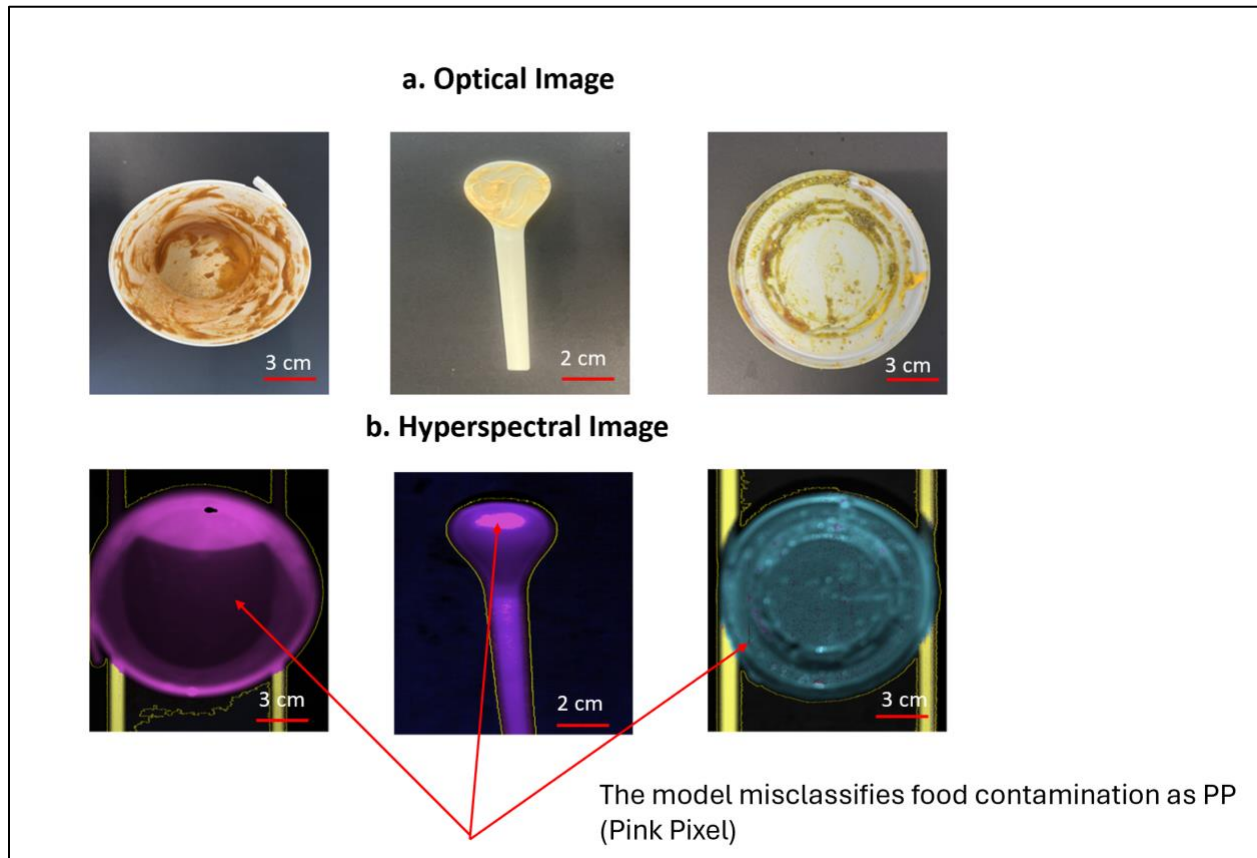
491

492 Table 6: The prediction accuracy of SVM, logistic regression, decision tree, ANN and PLSDA on
 493 testing dataset

Machine Learning Methods	Polymer	Sensitivity (Recall)	Specificity	Precision	F1 Score	Overall Accuracy
SVM	HDPE	1	1	1	1	97%
	PET	0.86	1	0.98	0.92	
	PHA	1	1	1	1	
	PLA	1	1	1	1	
	PP	1	0.94	0.94	0.97	
	Overall	0.97	0.99	0.99	0.98	
Logistic regression, decision tree, ANN	HDPE	1	1	1	1	93%
	PET	1	1	1	1	
	PHA	1	1	1	1	
	PLA	1	0.96	0.67	0.8	
	PP	0.86	1	0.99	0.92	
	Overall	0.81	0.99	0.93	0.94	
PLS-DA	HDPE	1	1	1	1	90%
	PET	1	1	1	1	
	PHA	1	1	1	1	
	PLA	1	0.89	0.40	0.57	
	PP	0.79	1	0.99	0.88	
	Overall	0.96	0.98	0.87	0.89	

494
 495 From Table 6, it is evident that the overall accuracy of SVM surpasses that of other machine
 496 learning algorithms. However, SVM exhibits lower sensitivity in detecting PET packaging
 497 compared to other types of plastic. This is mainly due to the limited reflectance demonstrated by
 498 PET, resulting in a weak Short-Wave Infrared (SWIR) signal. Consequently, identifying thin or
 499 transparent materials like PET becomes inherently challenging. Additionally, the model tends to

500 classify contamination as PP, leading to a lower specificity for PP compared to other material
501 types see Figure 7.



502

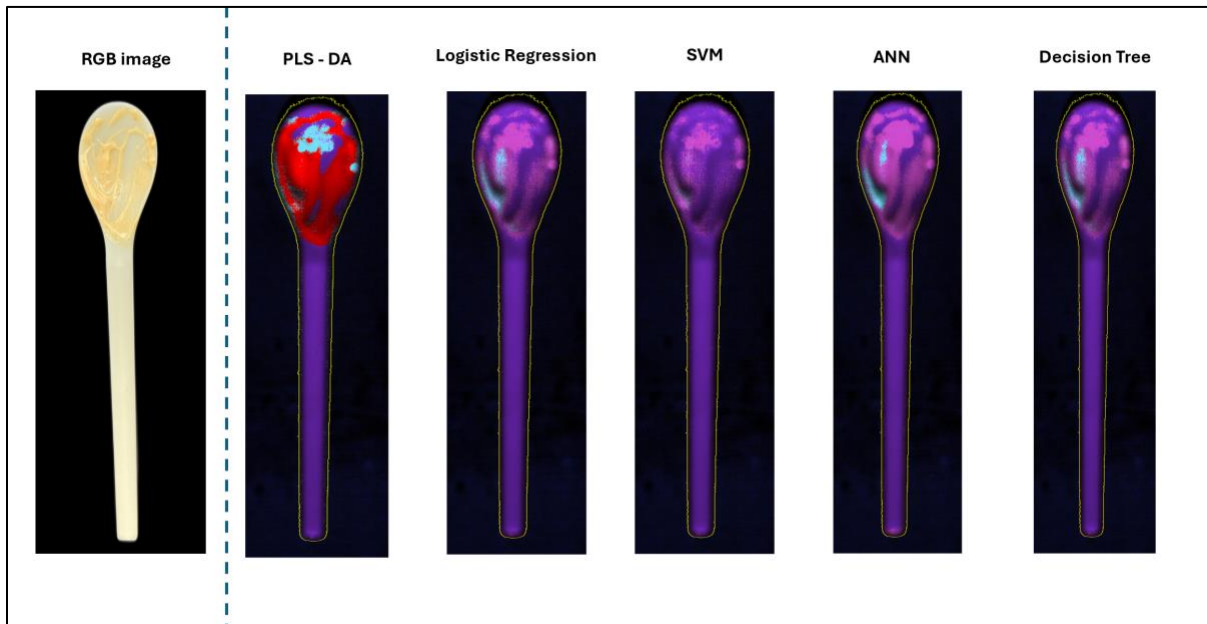
503 Figure 7: The example of (a) optical Images and (b) hyperspectral images of testing dataset (PP:
504 pink, PHA, purple, PLA: blue)

505 For logistic regression, decision tree and ANN, the sensitivity in detecting PP is the lowest at
506 0.86. PP was misclassified as PLA and LDPE. Thus, the specificity and precision and F1 score of
507 PLA is 0.96, 0.67 and 0.8 respectively while other types of plastics are 1. The overall accuracy of
508 these models is 93%. The misclassified samples have translucent colour and dark colour. Logistic
509 regression, decision tree, ANN classified translucent PP lid and red dark colour Japanese rice
510 bowl was misclassified as PLA and LDPE respectively. Some pixels were misclassified by each
511 model, leading to the same final classification outcome. For example, in Figure 8, we used
512 various classification models to identify the type of food-contaminated spoon. The majority of
513 pixels were classified as PHA. However, for spicy mayo contamination, SVM classified it as PP,
514 while Decision Tree, Logistic Regression, and ANN each classified spicy mayo as a combination

515 of PP (pink) and PLA (blue) but in different positions. PLS-DA classified it as a mix of
516 unidentified pixels (Red) and PP (pink).

517 For PLS-DA, the overall accuracy is the lowest at 90%. 21.4% of PP were classified incorrectly
518 as PLA. Therefore, the sensitivity of PP drops to 0.79 and the specificity of PLA decreases to
519 0.89. For precision and F1-score, PLA achieves only 0.4 and 0.57 and PP achieves 0.99 and 0.88
520 while others are 1.

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

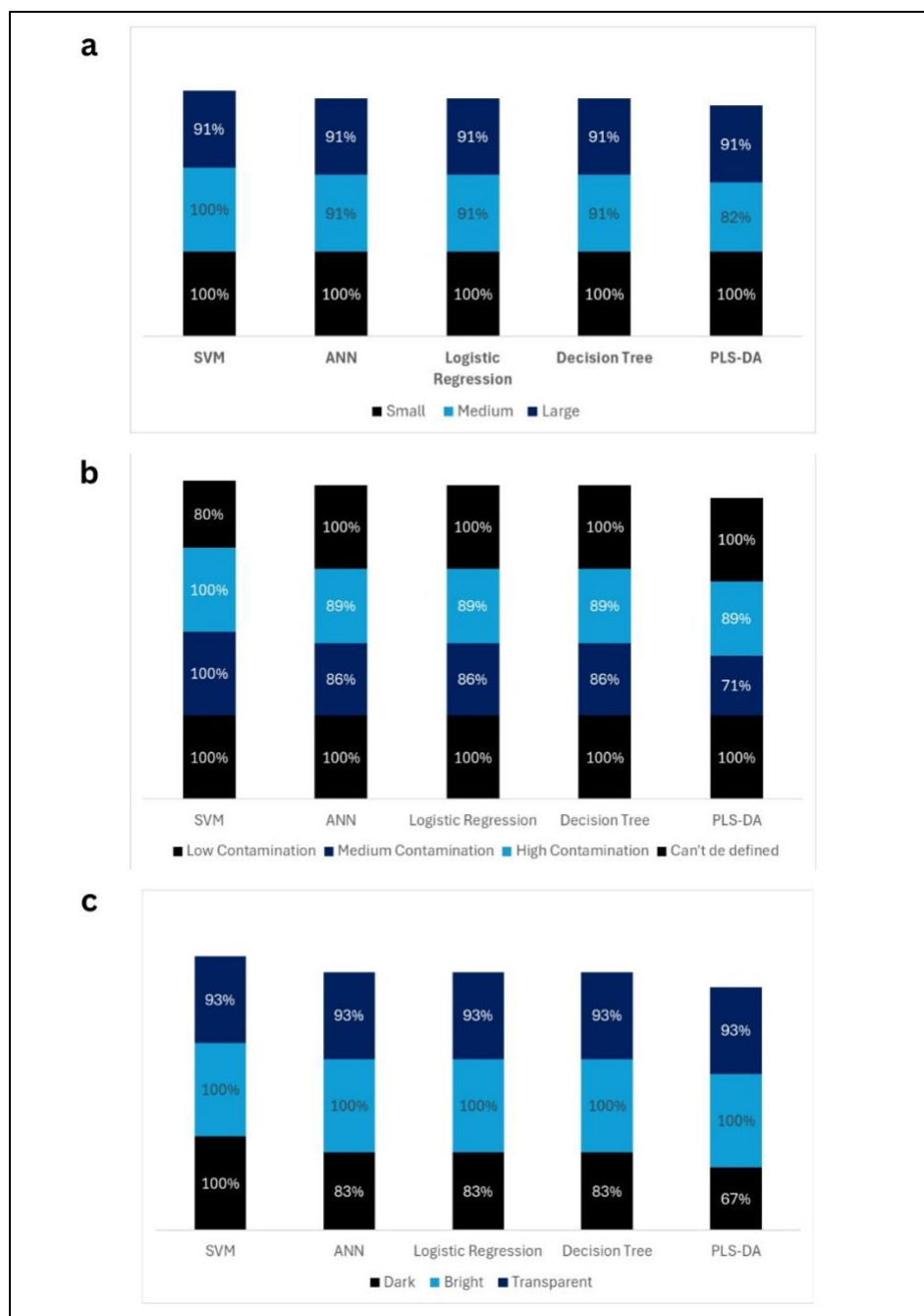


525

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

529 3.4 Material Properties of Contaminated Plastic Packaging in Testing Dataset

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



533

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

537

538

539 3.4.1 Size

540 The size of packaging was determined through surface area estimation algorithm. The average size
541 of plastic packaging in the testing dataset (30 real world plastic packaging) was 63.94 cm². The
542 number of small, medium and large packaging was 8,11 and 11 respectively.

543 The results are shown in Figure 9 (a). For SVM, the system achieved 100% accuracy for small and
544 medium plastic packaging but experienced a drop to 91% accuracy for large plastic packaging .
545 Similarly, for ANN, logistic regression, and Decision Tree models, the accuracy in detecting small
546 plastic packaging was 100%. For PLS-DA, the accuracy in identifying small plastic packaging was
547 100%, but it decreased to 82% and 91% when detecting medium and large sizes, respectively. The
548 accuracy of the models dropped when detecting large and medium plastics, as some samples in
549 these categories have opaque colours. However, the accuracy for detecting brightly coloured
550 plastics, regardless of size, is 100%. Thus, sizes larger than 8 cm², which is the size of the smallest
551 plastic packaging, have no impact on the model's accuracy.

552 3.4.2 Level of contamination

553 The average level of contamination of real-world plastic packaging was measured at 37%. Figure
554 9(b) illustrates the accuracy of the system in identifying types of polymers with several degrees of
555 contamination. In the testing dataset, there were 9 plastic packaging samples with a low level of
556 contamination, 10 with medium contamination, and 6 with a high contamination level. 5 pieces of
557 the plastic packaging could not have their contamination level measured due to the presence of
558 labels and transparent oily contamination.

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

568 darkness in the samples. Most medium contamination samples were dark in colour, while the
569 majority of high contamination samples were bright in colour.

570 **3.4.3 Darkness Level**

571 The darkness level was identified using average pixel value of greyscale image of samples in
572 testing dataset. The testing dataset consisted of 15 transparent plastic, 6 dark coloured plastic and
573 9 brightly coloured plastics. The average darkness level was found to be a greyscale value of 157.
574 The impact of darkness on successful identification is shown in Figure 9(c).

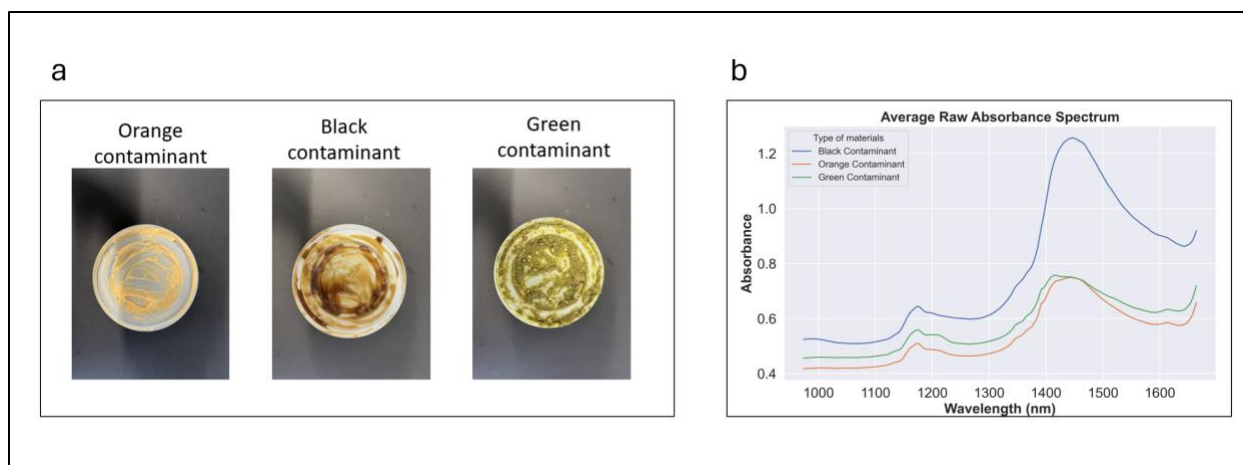
575 The accuracy of SVM to detect dark coloured and bright coloured plastic was 100% but it dropped
576 to 93% when identifying transparent plastic. For ANN, Logistic Regression and Decision Tree, the
577 accuracy of models in identifying dark plastic and transparent was 83% and 93% respectively, the
578 accuracy increased to 100% when identifying brightly coloured plastics. For PLS-DA, the
579 accuracy to identify dark plastic (67%) was much lower than identifying brightly coloured plastic
580 (100%). However, the accuracy in identifying transparent plastic was 93%.

581 **3.4.4 Food Contaminant Colour**

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

586

587



588

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

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

596 4. Discussion

597 4.1 Classification model performance comparison

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

606 To enhance robustness of model and mitigate overfitting, the utilization of cross validation dataset
 607 is crucial. Table 5 explains accuracy of algorithms on the cross-validation dataset, revealing that
 608 Logistic regression, SVM exhibit superior performance. These models perform well on the cross-

609 validation dataset, indicating that they can handle new, unseen data effectively including various
610 levels mayonnaise contamination on sample surface.

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

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

628 **4.2 Influence of material properties on the performance of the models**

629 Our focus extended to three material properties: size, level of contamination, and darkness. An
630 analysis reveals no discernible correlation between the size of the material (particularly when
631 exceeding 8 cm²) and the accuracy of the model. Surprisingly, the level of contamination
632 demonstrated minimal influence on the system's accuracy. Darkness showed significant impact on
633 the accuracy of the system. Opaque plastic is more difficult to be classified due to high light
634 absorbance of SWIR region. Transparent plastic is also difficult to be identified due to the
635 scattering of light.

636 Black contaminants absorb lighter than contaminants of other colours, see figure 10(b). However,
637 our classification models are robust; it can correctly identify the type of plastic even when the
638 contaminant is black. This result aligns with our previous research (Taneepanichskul et al., 2024).
639 Our SVM model for identifying polymer contamination performed comparably to the PLS-DA
640 model developed by Bonifazi, both achieving a sensitivity of 0.97. However, our model was
641 applied to packaging with a higher level of contamination (Bonifazi et al., 2021). Additionally,
642 Cucuzza's Hierarchical PLS-DA model demonstrated impressive accuracy, reaching up to 1.0.
643 These findings highlight that the integration of HSI with machine learning significantly enhances
644 the recycling rate by accurately identifying polymer contamination (Cucuzza et al., 2021).
645 Krasniewski applied various machine learning techniques to identify 11 types of polymers, finding
646 that PET had the lowest accuracy due to its transparency, which aligns with our results
647 (Kraśniewski et al., 2021). Importantly, our SVM model developed here enhances the performance
648 of our previous PLS-DA model (Taneepanichskul et al., 2023). Even with highly contaminated
649 packaging, the SVM model can identify polymers with 100% accuracy. In contrast, our previous
650 PLS-DA model could only accurately identify pristine plastics, with accuracy dropping to 75% for
651 highly contaminated plastics.

652 **4.3 Application of HSI in anaerobic digestion, in-vessel composting and recycling plant for** 653 **detecting food contaminated compostable plastics.**

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

660 Depackaging and separation are carried out using machines called depackagers. The reject stream
661 from these machines consists of packaging materials, including contaminated plastics, cardboard,
662 glass, and metal. After the separation process, IVC primarily relies on manual sorting combined
663 with visual inspection to identify compostable plastics, which is labour-intensive and costly
664 (WRAP, 2009). Contaminated plastics that cannot be composted are sent to landfill or incineration.

665 Similarly, in AD, all contaminated plastics are directed to landfill or incineration (Taneepanichskul
666 et al., 2022).

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

671 If recycling plants were employing this detection system, a high percentage of food contaminated
672 compostable plastics can be identified and redirected to composting facilities for proper
673 processing. Additionally, the system can help identify the 17% of recyclable plastics are rendered
674 non-recyclable due to food contamination (Biffa, 2022). Implementing this system would require
675 an automatic separation system to act on identification and characterisation provided by the HSI
676 system. These already exist in modern waste recycling facilities. They are less common in AD and
677 IVCs. Investment in these facilities would have to be driven by a return on investment which is
678 incentivised by lower number of plastics being sent to landfill and incineration. The next steps in
679 development would be to trial this system in a commercial AD or IVC setting where the effect of
680 the speed of the conveyor belt on accuracy, and other real-time variables could be tested.

681 **5. Conclusion**

682 In this study, we demonstrated that the combination of hyperspectral imaging and machine learning
683 algorithms can effectively identify different types of plastic packaging, even when highly
684 contaminated with food waste. Five machine learning algorithms including SVM, logistic
685 regression, decision tree, ANN, and PLS-DA—were applied to classify real-world contaminated
686 plastic packaging. Among these, SVM achieved the highest accuracy at 97%, followed by logistic
687 regression, ANN and decision tree, achieving 93%. PLS-DA showed the lowest accuracy at 90%.
688 However, the model faced challenges in identifying PET due to its transparency.

689 Furthermore, the influence of material attributes, including size, contamination level, and
690 darkness, was evaluated. Darkness significantly impacted the system's accuracy, as opaque plastics
691 were more challenging to classify due to more light absorption in the SWIR region. In contrast,
692 size and contamination level had a lesser effect on the models' accuracy.

693 This technology has the potential to be implemented in real waste management facilities,
694 addressing current challenges and significantly improving the efficiency of waste sorting systems.
695 As a result, recycling and composting rates could increase, contributing to the creation of a circular
696 economy for plastic waste.

697 6. References

- 698 ALLISON, A. L., PURKISS, D., LORENCATTO, F., MIODOWNIK, M. & MICHIE, S. 2022. Improving
699 compostable plastic disposal: An application of the Behaviour Change Wheel intervention
700 development method. *Frontiers in Sustainability*, 3, 92.
- 701 AMIGO, J. M., BABAMORADI, H. & ELCOROARISTIZABAL, S. 2015. Hyperspectral image analysis. A
702 tutorial. *Analytica chimica acta*, 896, 34-51.
- 703 ANAEROBICDIGESTION. 2023. *Anaerobic Digestion of Pre-Consumer Food Waste & 3 Generations*
704 *of Depackaging Machines* [Online]. Available: [https://blog.anaerobic-](https://blog.anaerobic-digestion.com/anaerobic-digestion-of-pre-consumer-food-waste/)
705 [digestion.com/anaerobic-digestion-of-pre-consumer-food-waste/](https://blog.anaerobic-digestion.com/anaerobic-digestion-of-pre-consumer-food-waste/) [Accessed 11 July 2024].
- 706 BARKVED, K. 2022. *The Difference Between Training Data vs. Test Data in Machine Learning*
707 [Online]. Available: [https://www.obviously.ai/post/the-difference-between-training-data-](https://www.obviously.ai/post/the-difference-between-training-data-vs-test-data-in-machine-learning)
708 [vs-test-data-in-machine-learning](https://www.obviously.ai/post/the-difference-between-training-data-vs-test-data-in-machine-learning) [Accessed 6 Jan 2024].
- 709 BHARGAVA, A., SACHDEVA, A., SHARMA, K., ALSHARIF, M. H., UTHANSAKUL, P. & UTHANSAKUL,
710 M. 2024. Hyperspectral Imaging and Its Applications: A Review. *Heliyon*, 10.
- 711 BONIFAZI, G., GASBARRONE, R. & SERRANTI, S. 2021. Detecting contaminants in post-consumer
712 plastic packaging waste by a NIR hyperspectral imaging-based cascade detection
713 approach. *Detritus*, 15, 94-106.
- 714 CASTRO-DÍAZ, M., OSMANI, M., CAVALARO, S., CACHO, Í., URÍA, I., NEEDHAM, P., THOMPSON, J.,
715 PARKER, B. & LOVATO, T. 2023. Hyperspectral Imaging Sorting of Refurbishment
716 Plasterboard Waste. *Applied Sciences*, 13, 2413.
- 717 CUCUZZA, P., SERRANTI, S., BONIFAZI, G. & CAPOBIANCO, G. 2021. Effective recycling solutions
718 for the production of high-quality PET flakes based on hyperspectral imaging and variable
719 selection. *Journal of imaging*, 7, 181.
- 720 EUROPEANBIOPLASTIC. 2023. *BIOPLASTICS MARKET DEVELOPMENT UPDATE 2023* [Online].
721 Available: [https://www.european-bioplastics.org/bioplastics-market-development-update-](https://www.european-bioplastics.org/bioplastics-market-development-update-2023-2/)
722 [2023-2/](https://www.european-bioplastics.org/bioplastics-market-development-update-2023-2/) [Accessed 1 April 2024].
- 723 HYSPEX. 2023. *Hyspex Baldur S-640iN* [Online]. Available: [https://www.hyspex.com/hyspex-](https://www.hyspex.com/hyspex-products/hyspex-baldur/hyspex-baldur-s-640i-n/)
724 [products/hyspex-baldur/hyspex-baldur-s-640i-n/](https://www.hyspex.com/hyspex-products/hyspex-baldur/hyspex-baldur-s-640i-n/) [Accessed].
- 725 KABIR, M. H., GUINDO, M. L., CHEN, R. & LIU, F. 2021. Geographic origin discrimination of millet
726 using Vis-NIR spectroscopy combined with machine learning techniques. *Foods*, 10, 2767.
- 727 KRAŚNIEWSKI, J., DĄBAŁA, Ł. & LEWANDOWSKI, M. Hyperspectral imaging for analysis and
728 classification of plastic waste. 2020 25th International Conference on Pattern Recognition
729 (ICPR), 2021. IEEE, 4805-4812.
- 730 MASOUMI, H., SAFAVI, S. M. & KHANI, Z. 2012. Identification and classification of plastic resins
731 using near infrared reflectance. *International Journal of Mechanical and Industrial*
732 *Engineering*, 6, 213-20.
- 733 MICROSOFTBUILD. 2021. *Multiclass Neural Network component* [Online]. Available:
734 [https://learn.microsoft.com/en-us/azure/machine-learning/component-](https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/multiclass-neural-network?view=azureml-api-2)
735 [reference/multiclass-neural-network?view=azureml-api-2](https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/multiclass-neural-network?view=azureml-api-2) [Accessed 13 May 2024].

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

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

741 TAMIN, O., MOUNG, E. G., DARGHAM, J. A., YAHYA, F., FARZAMNIA, A., SIA, F., NAIM, N. F. M. &
742 ANGELINE, L. 2023a. On-Shore Plastic Waste Detection with YOLOv5 and RGB-Near-
743 Infrared Fusion: A State-of-the-Art Solution for Accurate and Efficient Environmental
744 Monitoring. *Big Data and Cognitive Computing*, 7, 103.

745 TAMIN, O., MOUNG, E. G., DARGHAM, J. A., YAHYA, F. & OMATU, S. 2023b. A review of hyperspectral
746 imaging-based plastic waste detection state-of-the-arts. *Int. J. Electr. Comput. Eng.(IJECE)*,
747 13, 3407-3419.

748 TAMIN, O., MOUNG, E. G., DARGHAM, J. A., YAHYA, F., OMATU, S. & ANGELINE, L. A comparison of
749 RGB and RGNIR color spaces for plastic waste detection using the YOLOv5 architecture.
750 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology
751 (IICAET), 2022a. IEEE, 1-6.

752 TAMIN, O., MOUNG, E. G., DARGHAM, J. A., YAHYA, F., OMATU, S. & ANGELINE, L. Machine learning
753 for plastic waste detection: State-of-the-art, challenges, and solutions. 2022 International
754 Conference on Communications, Information, Electronic and Energy Systems (CIEES),
755 2022b. IEEE, 1-6.

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

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

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

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

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

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

772 WRAP. 2009. *Review of Food Waste Depackaging Equipment* [Online]. Available:
773 [http://www.organics-](http://www.organics-recycling.org.uk/uploads/article1762/Wrap%20Report%20on%20Food%20Waste%20Depackaging.pdf)
774 [recycling.org.uk/uploads/article1762/](http://www.organics-recycling.org.uk/uploads/article1762/Wrap%20Report%20on%20Food%20Waste%20Depackaging.pdf)
775 [Wrap%20Report%20on%20Food%20Waste%20Depa](http://www.organics-recycling.org.uk/uploads/article1762/Wrap%20Report%20on%20Food%20Waste%20Depackaging.pdf)
[ckaging.pdf](http://www.organics-recycling.org.uk/uploads/article1762/Wrap%20Report%20on%20Food%20Waste%20Depackaging.pdf) [Accessed 11 July].

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

778