CHAPTER 4
ANALYSIS AND DESIGN

4.1. Data Collection

The dataset consists of microplastics types in the first column and absorption bands obtained from FTIR in the next columns. Each data has a different number of absorption bands that can occur differently even when it is repeated. Here are some data from the existing dataset.

code	col_1	col_2	col_3	col_4	col_5	col_6	col_7	col_8	col_9	col_10	col_11	col_12	col_13	col_14	col_15	col_16
1	1193,94	1222,87	1273,02	1373,32	1465,9	1535,34	1633,71	1649,14	2852,72	2931,8	3300,2					
1	1193,94	1222,87	1271,09	1367,53	1465,9	1631,78	1651,07	2854,65	2927,94	3296,35						
1	1193,94	1222,87	1271,09	1369,46	1463,97	1631,78	1651,07	2856,58	2926,01	3296,35						
1	1192,01	1222,87	1273,02	1377,17	1465,9	1529,55	1631,78	1654,92	2858,51	2937,59	3300,2					
2	840,96	974,05	999,13	1170,79	1379,1	1456,26	2839,22	2922,16	2949,16							
2	840,96	974,05	997,2	1168,86	1377,17	1454,33	2839,22	2922,16	2953,02							
2	840,96	974,05	999,13	1168,86	1377,17	1454,33	2839,22	2920,23	2954,95							
2	840,96	974,05	999,13	1168,86	1377,17	1452,4	2839,22	2920,23	2956,87							
2	842,89	974,05	999,13	1168,86	1379,1	1462,04	2839,22									
2	840,96	974,05	999,13	1168,86	1377,17	1454,33	2839,22	2958,8								
2	840,96	974,05	999,13	1168,86	1379,1	1460,11	2839,22	2924,09	2953,02							
3	761,88	964,41	1028,06	1452,4	1492,9	1583,56	1600,92	2920,23	3028,24	3061,03						
3	763,81	964,41	1028,06	1452,4	1494,83	1525,69	1541,12	1583,56	1600,92	2850,79	2929,87	3003,17	3028,24	3061,03	3082,25	3101,54
3	759,95	964,41	1028,06	1452,4	1492,9	1541,12	1583,56	1600,92	2850,79	2926,01	3003,17	3026,31	3061,03	3082,25	3101,54	
3	756,1	964,41	1028,06	1452,4	1492,9	1539,2	1543,05	1583,56	1600,92	2850,79	2926,01	3003,17	3026,31	3061,03	3082,25	3101,54
4	968,27	1101,35	1253,73	1330,88	1427,32	1433,11										
4	966,34	1099,43	1257,59	1330,88	1427,32	1435,04										
4	702,09	966,34	1099,43	1257,59	1330,88	1435,04										
4	966,34	1099,43	1253,73	1328,95	1433,11											
4	968,27	1097,5	1255,66	1327,03	1427,32	1435,04										
5	889,18	910,4	1463,97	2848,86	2929,87											
5	719,45	731,02	1463,97	2854,65	2933,73											
5	721,38	729,09	1467,83	2852,72	2926,01											
5	721,38	729,09	1463,97	2931,8												
6	713,66	792,74	873,75	1049,28	1095,57	1128,36	1240,23	1346,31	1409,96	1448,54	1506,41	1579,7	1734,01	2347,37	2910,58	2966,52
6	717,52	968,27	1502,55	1573,91	1728,22	2351,23	3429,43									
6	794,67	848,68	873,75	972,12	1051,2	1344,38	1415,75	1448,54	1573,91	1957,75	2351,23	2910,58	2966,52	3049,46	3433,29	
6	715,59	794,67	873,75	972,12	1047,35	1413,82	1506,41	1573,91	1957,75	2349,3	3049,46	3429,43				
6	1047,35	1344,38	1415,75	1506,41	2351,23	3055,24	3431,36									

Figure 4.1 Example of dataset

In addition, there is also a dataset of reference absorption bands of a polymer that is used to convert absorption bands into polymers. The reference comes from several existing papers and is summarized by M. R. Jung et al. [6] in their paper. The reference only gives the value of a polymer without a range of values. In fact, the research using microplastics compounds and FTIR can change or will not always be the same depending on the conditions [9]. Here are some data about the reference of the absorption band on Figure 4.2 and Table 4.1.

Absorption Band	Polymer
694	Aromatic C-H out of plane bending
700	C-Cl stretching
757	Aromatic C-H stretching
840	C-CH3 stretching
966	CH2 rocking
967	C=C
972	C-CH3 rocking
997	C-CH3 rocking
1027	Aromatic C-H bending
1099	C-C stretching
1199	CH2 bending
1220	C-O-C
1255	C-H bending
1274	C-N stretching
1331	C-H bending
1372	CH2 bending
1375	C-CH3 symmetric
1377	CH3 groups
1547	Aromatic C-H stretching
1634	C=O stretching
2923	C-H stretching reflects
2932	C-H stretching
2952	CH3 symmetric
3055	Aromatic C-H stretching
3298	N-H stretching

Figure 4.2 Example of references dataset [6]

Table 4.1. Polyamide (PA) or Nylon Polymer Reference [6]

Absorption Bands	**Polymer**
3298	N-H stretching
2932; 2858	C-H stretching
1650; 1634	C=O stretching
1530	N-H bending
1274	C-N stretching
1464; 1372; 1199	CH2 bending
1220	C-O-C

4.2. Data Preprocessing

Data in Figure 4.1 are divided into 6 parts using K-fold. The K-fold is used due to the limited data (210). Splitting the data into train data, validation data, and test data is not possible with the amount of data. Reference data is also processed using Gaussian Mixture by considering the train data in each cross validator or K-Fold so as to produce a Gaussian Mixture Model (Normal Distribution) on each polymer.

To compare the performance of the model, data augmentation is used. In this research, crossover from genetic algorithm is used to create a new data. Crossover is a process from genetic algorithm that create a new genetic for the next generation by combining 2 DNA strain. The new genetic will be randomize and have least, partial, or major from the old generation [13].

Two data from the same type of microplastics will crossover each other to specific number. Each component from the data will be choose randomize to create a new data. After around 2000 data, the model is checked by its performance. For example, sample A (721.38, 729.09, 771.53, 1377.17, 1467.83) has A1 to A5 and sample B (721.38, 729.09, 775.38, 1377.17) has B1 to B4 are PVC microplastics. New data can be generate as 1 to 9 component long. To create new samples C with 5 component there are several result, such as (A1, B1, B3, A2, A5), (B1, B2, B3, A1, A2), and (A3, B1, B4, A4, A5). All of the component are chosen by random to create new data.

4.3. Experiment

4.3.1. Manual Matching

Manual Matching is done by matching the FTIR results against existing references. For example in Table 4.2 the absorption band column is one of the data with PA contamination. It can be seen that the data obtained is not exactly the same as the reference in Table 4.1. Therefore, the researcher needs to estimate without a definite reference to fill in the assignment and distance columns through Figure 4.2. Table 4.2 must be done against all microplastics references to get accurate results of the actual contamination. Therefore, Manual Matching requires longer time and high accuracy.

Table 4.2. Data Example of Polyamide (PA) Contamination

No	Absorption Band	Assignment	Distance (Point)	Part of PA (Yes/No)
1	1192.01	CH2 bending	6.99	Yes
2	1222.87	C-O-C	2.87	Yes
3	1273.02	C-N stretching	0.98	Yes
4	1377.17	CH3 groups	0.17	No
5	1465.90	CH2 bending	1.9	Yes
6	1529.55	N-H bending	0.45	Yes
7	1631.78	C=O stretching	2.22	Yes
8	1654.92	C=O stretching	4.92	Yes
9	2858.51	C-H stretching	0.51	Yes
10	2937.59	C-H stretching	5.59	Yes
11	3300.20	N-H stretching	2.2	Yes

From Table 4.2, it can be seen that the majority of absorption bands are part of microplastics PA. However, there is one data that is not part of microplastics PA, 1377.17, which is designated as CH3 groups.

4.3.2. Gaussian Mixture and Naive Bayes Matching

This matching method gives a probability that an absorption band belongs to a polymer. The probability is obtained from the results of preprocessing data obtained with Gaussian Mixture. In this research, the author uses Scikit-learn Gaussian Mixture which applies k-means to initialize the weights, the means and the precisions. In addition, the spherical covariance type is also used, in order to produce a variance value in the covariance variable.

Gaussian Mixture is one of the clustering methods that is soft clustering, it means it has a probability value for one or more clusters. To calculate a data into a cluster, the cluster mean data and the cluster variance are required.

By using Gaussian mixture sklearn, the first step to determine a cluster using k-means by entering the mean data from the existing reference absorption bands. Each absorption band will be calculated to get the cluster size. Furthermore, the cluster will be calculated continuously by recalculating the mean and variance of each cluster until there is no significant change in the data. This is the Expectation-Maximization (EM) algorithm, Expectation to calculate the data to a cluster and maximization to improve the cluster parameters.

As an illustration, if a and b are clusters and x is a data, then what needs to be calculated is the probability of cluster a if the data is x_i and the probability of x_i if it is cluster a through

function 1 and 2. This is done for all existing clusters. Symbol π represents a number of data, x_i the x value of a data minus the average of cluster a, σ_a^2 is the covariance of cluster a.

$$P(a|x_i) = \frac{P(x_i|a)\,P(a)}{P(x_i|a)\,P(a)+P(x_i|b)\,P(b)} \quad (1)$$

$$p(x_i|a) = \frac{1}{\sqrt{2\pi\sigma_a^2}} exp^{\left(-\frac{(x_i-\mu_a)^2}{2\sigma_a^2}\right)} \quad (2)$$

From these probabilities, each feature will be calculated towards class classification using Gaussian Naïve Bayes. Each group of data will go through preprocessing to convert absorption bands into polymer probabilities with the trained Gaussian Mixture Model. The same polymer with different absorption band values will be pooled together by taking the highest probabilities value. This set of probabilities is used as a one-hot-vector for classification using Gaussian Naïve Bayes. The result of Gaussian Naïve Bayes is the probability value of each class.

This method uses Numpy as numerical computing, and Pandas as csv data processing to numerical and vice versa.

4.4. Evaluation

Performance is measured with the Classification Report by Sklearn.metrics. This report contains precision, recall, f1-score, support, and accuracy for each class. This report is done for each cross-validator that has been set at the beginning, which is 6 times.

First, Precision (function 3) is the True Positive (TP) value compared to True Positive plus False Positive (FP). Second, recall (function 4) is the True Positive (TP) value compared to True Positive plus False Negative (FN). Third, f1-score (function 5) is twice the value of Precision multiplied by Recall compared to Precision plus Recall. The f1-score value is the harmony value of precision and recall.

$$P = \frac{T_p}{T_p+F_p} \quad (3)$$

$$R = \frac{T_p}{T_p+F_n} \quad (4)$$

$$F1 = 2\frac{P\ x\ R}{P+R} \quad (5)$$

Finally, K-means and Decision Tree are used to compare Gaussian Mixture and Naïve Bayes performance. Gaussian Mixture and K-means are 2 of clustering algorithms, but K-means works by grouping data based on the closest distance of data to the center of a cluster, so it only has 1 label. On the other hand, Decision Tree, is a classification algorithm that works by looking at the decision rules of the train data. The decisions are sorted from most definite to less definite so that an arrangement of these rules can be created. Because the two algorithms have the same function like Gaussian Mixture and Naive Bayes, K-Means and Decision Tree are used as a comparison.

4.5. Discussion

First, the proposed model should work well if the results of the Gaussian Mixture Model are close to the existing reference and the accuracy of Naïve Bayes is high enough. Secondly, the identification performance should also show how good the proposed preprocessing process is. With these two things, the supporting data from the Gaussian Mixture Model and the FTIR classification technique from Naïve Bayes should be shown good or not.