THE IDENTIFICATION AND CLASSIFICATION OF MICROPLASTICS BY FTIR USING GAUSSIAN MIXTURE AND NAIVE BAYES

[1]Tan, Yudistira Suryakencana Adisatria, [2]Yonathan Purbo Santosa
[1,2]Department of Informatics Engineering Faculty of Computer Science,
Soegijapranata Catholic University
[2]yonathansantosa@unika.ac.id

ABSTRACT

Microplastics has become more widely discussed recently. Detecting microplastics can be done using Fourier Transform Infrared Spectroscopy (FTIR). The results provide an absorption band that must be translated into a polymer. However, these results have different sizes of data, varied data, and take a long time to translate if done manually. This can be solved using Gaussian Mixture and Naïve Bayes by modifying the preprocessing to create same-sized data. The results are preprocessing which succeed in equalizing the length of the data, having good performance in the means value which is likely the same as the reference and having high accuracy, also being able to be used as supporting data when manual matching is done.

Keywords: microplastics, FTIR, classification, gaussian mixture, naïve bayes

INTRODUCTION

Background

Microplastics have been known since 2004 after a lot of sedimentation appeared in European waters. Research on microplastics began to develop from year to year to define what microplastics are. In 2019, Frias and Nash, defined microplastics as any synthetic solid particle or polymeric matrix which are insoluble in water. The size is from 1 micrometre to 5 millimetre and its shape can be regular or irregular of either primary or secondary manufacturing origin [1]. The consequences of microplastics include, affecting pregnant women and their babies, imbalances the hormones, carcinogenicity, and many more [2].

Fourier Transform Infrared Spectroscopy (FTIR) is a tool that shoots infrared radiation through an existing sample. The radiation will be fully absorbed, partially, or not even absorbed by part of the existing sample. From these data, a special spectrum will emerge that represents the characteristics of the existing sample [3]. The result of FTIR is a spectrum wave that changes from the one that was fired. Each spectrum will be a marker of the chemical group characteristics of the microplastics [4].

Spectrum data processing as a characteristic of chemical groups can use gaussian mixture and naive bayes. Gaussian mixture is a statistical distribution model used to measure the distribution of a category to its members. The heterogeneity in the population can cause a parametric family fails to model the data properly [5]. In this case, the gaussian mixture is used to determine the chemical groups of the spectrum present.

Naive Bayes is one of the probabilistic models to determine the class of a thing with Bayes theory [6]. In this paper, the Naive Bayes method used is Gaussian because it is used to calculate the microplastics probability of many plastic classes.

In this study, the author suspects that the use of gaussian mixture and naive bayes to determine the microplastics content can improve accuracy compared to manual matching (one by one) against existing references. This is because, the current reference is the single absorption bands number, not the absorption bands range.

Problem Formulation

1. How to answer the ambiguity of which polymer to classify at manual matching?
2. How to solve the difference in component data size of a classification?
3. How precise is the use of Gaussian Mixture and Gaussian Naïve Bayes?

Scope

The data used is the data that retrieved from the Faculty of Agricultural Technology, Soegijapranata Catholic University, Semarang. All the data that I use are assumed as ground truth for each microplastics. The data obtained was tested in the lab by deliberately creating microplastics contamination. The experiment was carried out in sterile conditions using 96% ethanol, not using plastic equipment, and covered with aluminum foil so as to minimize contamination.

Objective

The main objectives of this research are to get the scope range of absorption bands and create a model that can identify the spectrum of absorption bands and classify the group of absorption bands into what kind of microplastics pollution.

LITERATURE STUDY

Frias and Nash [1], summarized the history of microplastics definition from so many references. They state that there have been many changes from 1907 to 2019. It is also convinced by Gago et al. [7] who stated the latest definition of microplastics and comparison. This article [1] just states briefly and was explained in the second article [7]. These two articles are good for finding the definition and history of microplastics, especially for common people.

Sharma and Kaushik [8], described the source, health risk, and ways to reduce microplastics. They also stated that current events such as Covid-19, contribute to the increased number of microplastics. This article is good for those who want to know more about how microplastics are related to human behavior. This also became support information for the history of microplastics that have been stated by Frias and Nash [1] as well as Gago et al. [7].

Dutta [9] has shown that Fourier Transform Infrared Spectrum (FTIR) can detect contents inside a sample. Dutta explained how it can be detected in very detail but easy to understand for common people. On the other hand, Song et al. [10] stated about why the result of FTIR can be

missed, one of the causes is the weathered and contaminated surfaces of plastics. Therefore, Jung et al. [6] gave a list of references due to type of microplastics and the absorption bands based on FTIR result. This reference is used for the calculation of absorption bands into types of microplastics.

Zhang and Chen [11] and Weinberger and Bresler [5] explained the concept of Gaussian Mixture from its roots. The history, related study, and mathematical proof of the equation are explained in their articles. These articles help the author to learn the concept and its usage for this paper. These two articles are useful for those who need the detail of a Gaussian Mixture but full of mathematical problems.

Abbas et al. [12] and Narayanan et al. [13] explained about Naïve Bayes Theorem and how it can be a classifier. These 2 paper are good for those who want to learn the concept of Naïve Bayes. Furthermore, the multinomial Naive Bayes classification [12] was useful for the author to consider using this algorithm in this paper.

Jahromi and M. Taheri [14] explained the concept of Gaussian Naive Bayes. They also compared the Gaussian Naïve Bayes to the other 6 classifiers. As a result, Gaussian Naive Bayes can be compared to the others. This paper is useful to understand more about Gaussian Naive Bayes. Kamel et al. [4] propose cancer classification that is using Gaussian Naive Bayes in 2019. The dataset has 9 features with 1 to 10 values. This paper supports the knowledge of using Gaussian Naive Bayes for microplastics classification that have multiple features and values.

Liu et al. [15] explained the concept of genethic algorithm and its steps. They gave detail especially the crossover process. This paper is useful to create crossover process for data augmentation due to the limited data.

RESEARCH METHODOLOGY

Data Collection

The dataset used in this project are from the experiments of Agricultural Technology, Soegijapranata Catholic University students, Alice Septiana Dewi and Irmadella Rana Nathania, in their research. This dataset has 6 types of microplastics with references to absorption bands from several literature and the absorption bands themselves. The types of microplastics are PA/Nylon, PP, PS, PVC, PE/LDPE, and PET/PETE.

The dataset has their references for each microplastics. They gathered it from several references, one of them is M. R. Jung et al. [6] The paper has summarized many microplastics and its absorption band polymer.

DATA PREPROCESSING

There are several steps for preprocessing the dataset. First, split Data Using K-fold. Dataset from Faculty of Agricultural Technology are splitted using K-fold. K-fold is used because of the limited data for the dataset. Second, calculate the mean of gaussian mixture. References from

literature become predefined to gaussian mixture. Fit the gaussian mixture model with the data train that has been splitted with K-fold. Lastly, identify the polymers. After the gaussian mixture model is created, the absorption band data are converted into a polymer with probability. The maximum probability of the polymer is selected into an array.

For experiment, data augmentation is used to compare the performance of the model. Crossover become the base idea for the augmentation. The crossover of 2 data from the same microplastic will generate several data.

Experiment

Experiments were conducted with 2 different methods, Manual Matching and Gaussian Mixture Naïve Bayes matching.

Manual Matching

Each absorption band from the FTIR result is matched to the reference of polymer absorption bands. The user must estimate the value of the absorption band into polymer reference.

Gaussian Mixture and Naive Bayes Matching

All data are converted into a csv file separated from the references and the dataset. The data were processed in gaussian mixture model to identify the polymer and gaussian naive bayes to classify the microplastics types (PA/Nylon, PP, PS, PVC, PE/LDPE, PET/PETE).

Evaluation

For the evaluation, the author used K-fold to check the performance of the Gaussian Naive Bayes Model. The dataset was split into 6 folds and iterated the process The performance was measured with Classification Report by Sklearn.metrics. The report contains precision, recall, f1-score, support, and accuracy. As an addition, the performance was compared to another algorithm.

Discussion

The results of this paper will be the range of absorption band scope based on gaussian mixture model. This will be used as a reference to know the standard deviation of absorption bands polymer around the other reference. Furthermore, it can be used for any identification of polymer absorption band as long as has the reference absorption band. This also can be used to classify the other microplastics with adjustment on the reference.

ANALYSIS AND DESIGN

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 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 [10]. Here are some data about the reference of the absorption band on Figure 2 and Table 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 2. Example of references dataset [6]

Table 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

Data Preprocessing

Data in Figure 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 [15].

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.

Experiment

Manual Matching

Manual Matching is done by matching the FTIR results against existing references. For example in Table 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 1. Therefore, the researcher needs to estimate without a definite reference to fill in the assignment and distance columns through Figure 2. Table 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 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 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.

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.

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.

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.

IMPLEMENTATION AND RESULTS

Implementation

This experiment was conducted at Google Colab with a time frame of September 19, 2022-October 28, 2022. Before preprocessing the Gaussian Mixture Model, the library needs to be prepared by importing it.

```
1. import numpy as np
2. import pandas as pd
3. from sklearn.mixture import GaussianMixture
4. from sklearn.metrics import classification_report
5. from sklearn.naive_bayes import GaussianNB
6.
7. random = 777
8. np.random.seed(random)
9.
10. np.set_printoptions(suppress=True)
```

Lines 1-5 import the library used by the author, numpy as numerical computing, and pandas as csv data processing to numerical and vice versa. Lines 7-10 initiate random values in order to get the same results every time you run the program. Line 10 is used to create a fixed numeric value because if it is in exponential form (for example 1e+03) it is considered a string/object so it cannot run properly.

```
11. loc = "/content/GroundTruthLabelFull.csv"
12. ref = "/content/ReferencesFull.csv"
13. 
14. # Import CSV to Numpy
15. data_raw = pd.read_csv(loc).to_numpy()
16. 
17. from sklearn.model_selection import KFold
18. 
19. kf = KFold(n_splits=6, random_state=random, shuffle=True)
20. 
21. train = []
22. test = []
23. for train_index, test_index in kf.split(data_raw):
24.     train.append(train_index)
25.     test.append(test_index)
```

Lines 11-15, the dataset in csv format is loaded into the program via pandas and converted to numpy for processing. Next, lines 17-25 are the process of separating into 6 parts for the cross-validator.

```
26. data_ref = pd.read_csv(ref).to_numpy()
27. means, Feature = np.hsplit(data_ref, 2)
```

Lines 26-27 load the reference and separate the absorption band with its polymer as shown in Figure 2. The separation is done because the means will be used in the Gaussian Mixture Model as the initial and feature as the list of polymers to look for.

```
28. def processing(p_data):
29.     # Bag of Probabilities / X
30.     bop = []
31. 
32.     # Bag of Target
33.     bot = []
34. 
35.     # Identification
36.     bodata = bagOfData(p_data)
37.     for dt in bodata:
38.         temp = np.zeros(len(bof))
39.         for x in dt[1:]:
40.             pred = gm.predict([[x]])[0]
41.             ind = bof.index(Feature[pred])
42.             prob = gm.predict_proba([[x]])[0][pred]
43.             if temp[ind] is 0:
44.                 temp[ind] = prob
45.             else:
46.                 if temp[ind] < prob:
47.                     temp[ind] = prob
48. 
```

```
49.          # probability one hot vector of feature, number label
50.          bot.append(dt[0])
51.          bop.append(temp)
52.
53.      return bop, bot, bodata
```

Lines 28-53 are the function to process the data into a Gaussian Mixture Model that produces a bag of data that is the same size as the existing feature. Lines 40-42 are the process to convert absorption bands into polymers with probability values. Lines 43-47 look for the highest value of a polymer probability, as explained in 0. Lines 50-53 enter the variable as a separator for each data and return it.

```
54. def microplasticsProba(p_bop):
55.      result = []
56.      for i in range(len(p_bop)):
57.          temp = []
58.          pred = clf.predict_proba(p_bop[i].reshape(1, -1))
59.          max_value = np.amax(pred)
60.          max_index = np.argmax(pred)
61.          temp.append(pred[0])
62.          temp.append(convertClassIdx(max_index))
63.          result.append(temp)
64.
65.      return result
```

Lines 54-65 is a function to convert the polymer opportunity group into microplastics classification with probability. Line 58 is the classification process using Gaussian Naïve Bayes. Line 61-63 is a process to make it easier for readers to see the classification results.

Next for the main process,

```
66. for j in range(len(train)):
67.      data_train = [data_raw[z] for z in train[j]]
68.      data_test = [data_raw[z] for z in test[j]]
69.
70.      data_train = np.array(data_train)
71.      data_test = np.array(data_test)
72.
73.      data = preprocessingTrain(data_train)
```

Line 66 is the function for the cross-validator loop. Lines 67-68 are used to retrieve the data that has been separated from the k-fold index result. Lines 70-71 convert the data into ndarray type, so that the preprocessing function can process the train data and test data.

```
74. bof = []
75.      for fea in Feature:
76.          if fea[0] not in bof:
77.              bof.append(fea[0])
78.
79. x = data_x(data)
```

Lines 74-79 are used to unify polymers of various absorption bands.

```
80.      # Spherical = covariances between its own
```

```
81.     gm = GaussianMixture(n_components=means.shape[0], random_state=ran
   dom, means_init=means,covariance_type="spherical")
82.     gm.fit(data)
```

Lines 80-82 is the initiation of the Gaussian Mixture Model, which uses the size of the dataset means and covariance_type spherical. Spherical is used in order to produce a variance value in the covariance.

```
83.     bop, bot, bod = processing(data)
84.
85.     X = bop
86.     Y = bot
87.
88.     clf = GaussianNB()
89.     clf.fit(X, Y)
```

Line 83 is used to obtain the probability gaussian mixture, the target of the dataset, and the whole data. Lines 85-86 separate X and Y and are processed using Gaussian Naïve Bayes on lines 88-89.

```
90. def testing(p_bop, p_bot):
91.     result = []
92.     for i in range(len(p_bop)):
93.         temp = []
94.         pred = clf.predict(p_bop[i].reshape(1, -1))
95.         target = p_bot[i]
96.         temp.append(pred[0])
97.         temp.append(target)
98.         result.append(temp)
99.
100.    return result
```

Lines 90-100 is a function to perform classification and enter it into an array containing targets and predictions. This prediction uses Gaussian Naïve Bayes that has been set on line 88.

```
101.data = preprocessingTrain(data_test)
102.bop, bot, bod = processing(data)
103.result = np.array(testing(bop, bot))
104.print("K-Fold = ",    j,    "\n",    classification_report(result[:,1],
   result[:,0]))
```

Lines 101-104 perform the same preprocessing, identification, and classification processes as the train data against the test data. We print the results using the help of sklearn classification_report by entering the target and also the prediction of the result variable on line 104 to produce the classification report.

Results

From this research, several results were obtained. First, that the center value of a polymer from the Gaussian Mixture Model does not differ much from the existing reference. The range of absorption band values can also be obtained. The results of this range can be used as supporting data for manual matching which can be seen in Table 3. The polymer column is the name of the polymer in a microplastics. The reference column is a reference value from previous research,

namely Jung et al. [6]. Calculated Means column is the center value of absorption band of a polymer from Gaussian Mixture Model. Calculated Variance column is the variance of the value of the polymer.

Table 3. Result of Gaussian Mixture Model (First five, complete data in appendix)

No	Polymer	Reference	Calculated Means	Difference	Calculated Variance
0	C-Cl stretching	700	707.7361	7.7361	18.11433
1	Polar ester groups and benzene ring interaction	712	711.73	0.27	1E-06
2	CH2 rocking	720	721.1141	1.1141	2.381933
3	CH2 rocking	730	729.7961	0.2039	0.864328
4	Aromatic C-H stretching	757	759.6795	2.6795	12.19483
5	Ethyl branching	775	773.7793	1.2207	6.10176

Table 3 shows that 59 out of 65 polymers, have a difference in value with the reference of less than 10 with an average of 2.50. Differences of more than 10 and less than 20 is only found in 1 data, namely for polymer C-CH3 symmetric (reference value 1375). In addition, there are 5 polymer data that have a very far distance with the reference, namely, Aromatic rings 1,2,4,5; Tetra replaced (872), Methylene group and ester C-O bond vibrations (1050), C-C stretching (1099), Terephthalate Group (OOCC6H4-COO) (1124), and Symmetric CH stretch (2969). Therefore, the Gaussian Mixture Model is not much different from the reference.

Too large a deviation can be caused by the absence of data in the vicinity of the polymer from the experimental material. It can be seen from numbers 17 and 60 which have no absorption band value, as well as number 19 with a value of 3.236. In addition to the zero value, outliers can be a factor that makes the absorption band value deviate. Outlier data will make the average value of data in a polymer shift.

Second, variance shows the distribution of data in the polymer. In simple terms, we know what range of absorption band values are included in a polymer. For example, Table 3 number 5, the absorption band value for Ethyl branching is in the range of 771.3091 to 776.2495 with the highest probability at 773.7793. This shows agreement with the reference, which is 775. From this result, we can answer the ambiguity of which polymer to classify at manual matching.

Third, the different data lengths in the dataset (e.g. Figure 3) can be equalized through the preprocessing applied. The different data lengths are unified into an array that has a probability value for all polymers as shown in Figure 4. The probability value points to the polymers in Table 3 which has been grouped by polymer name into Table 4.

```
Example A:  10 data
[1.0, 1193.94, 1222.87, 1271.09, 1367.53, 1633.71, 1656.85, 2856.58,
2937.59, 3296.35]

Example B:  17 data
[6.0, 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]
```

Figure 3. Example of different size data before preprocessing

```
Example A :  41 data
 [0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.99990443 0.99990585
 0.         0.99989775 0.         0.         0.         0.
 0.         0.         0.         0.         0.99955471 0.
 0.         0.         0.         0.88620758 0.         0.
 0.         0.         0.         1.         0.        ]

 Example B :  41 data
 [0.85610124 0.         0.         0.         0.         0.98673432
 0.         0.         0.98130207 0.         0.         0.
 0.         1.         0.99274475 0.         0.98173239 0.
 0.         0.         0.99994436 0.         0.         0.99932807
 0.         0.         0.99995317 0.         0.         0.99995791
 1.         0.         0.         0.         0.99224994 0.
 0.         0.         0.         0.         0.        ]
```

Figure 4. Example of same size data after preprocessing

Finally, although the Gaussian Mixture Model has some polymers that are far from the reference, the performance of Gaussian Naïve Bayes obtained from Classification Report by Scikit-learn shows a value of 1.0 which indicates that this model can do its job very well. These results can be seen in 0.

Table 4. Polymer Grouping (from 1 to 5, the complete data are in appendix)

No	Polymer
1	C-Cl stretching
2	Polar ester groups and benzene ring interaction
3	CH2 rocking
4	Aromatic C-H stretching
5	Ethyl branching

Table 5. K-Fold Classification Report

Parameter	*Gaussian Mixture + Naïve Bayes*	*Gaussian Mixture + Decision Tree*	*Kmeans + Naïve Bayes*	*Kmeans + Decision Tree*
Accuracy	1	0.96572	1	0.97619
Precision	1	0.9615	1	0.981996
Recall	1	0.95858	1	0.966534
F1-score	1	0.95806	1	0.976207
Accuracy Sdtev	0	0.03724	0	0.011664
Precision Sdtev	0	0.047546	0	0.005977
Recall Sdtev	0	0.047112	0	0.015035
F1-score Sdtev	0	0.049158	0	0.014254

In Table 5, K-means is used to compare the performance of Gaussian Mixture because both can be used for identification. However, K-means is a hard clustering which means it has no probability. In addition, Decision Tree is one of the classification methods which in this case is also used as a comparator for Naïve Bayes. To simplify, there are also the graphic of the report on Figure 5 and Figure 6.

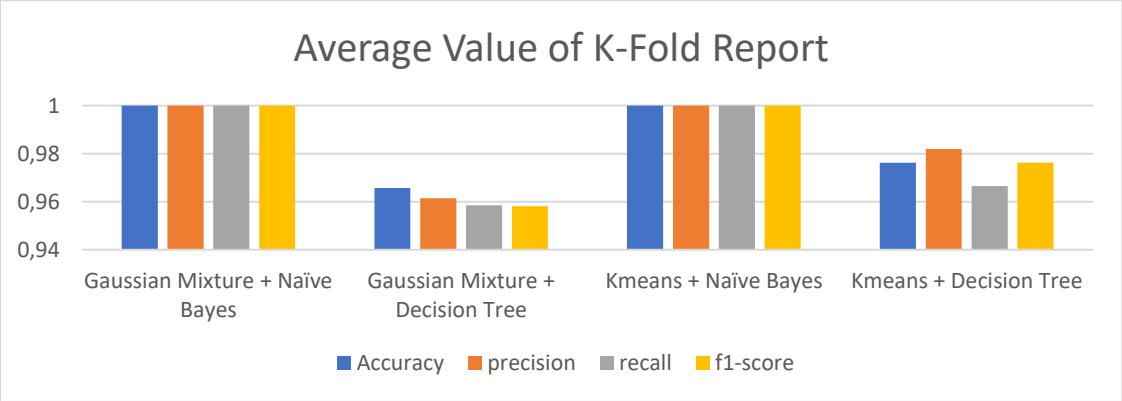

Figure 5. Chart of Average Value K-fold report\

From the comparison, it can be seen that the use of the Gaussian Mixture and K-Means is not much different. If we look at Figure 5.2, we can get the probability value of a polymer up to 100%, while K-means forces the data into one polymer type. This will be a problem if the dataset used is not clean or has a lot of noise. Since the dataset in this case was done in a laboratory with low contamination, this problem does not arise. In the use of Gaussian Mixture, more information is obtained such as, polymer variance and chance value to support the manual matching process so that it is more accurate.

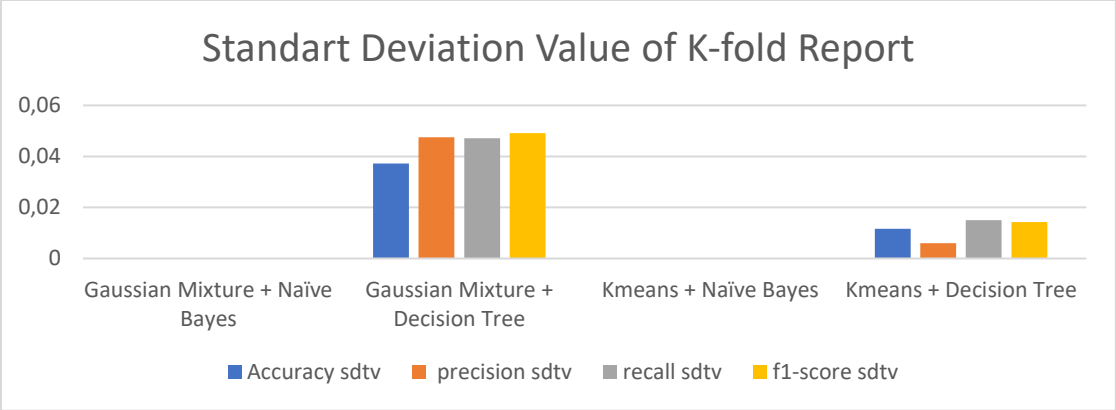

Figure 6. Chart of Standart Deviation Value K-fold report

On the other hand, the use of Naïve Bayes is better than Decision Tree because it manages to perform the classification process better even though the values are not much different.

After generate 9 new data for each data for the data augmentation, the accuary of two model did not change differently and it can be seen on below, 0. For the full report of the model on number of augmentation 9 can be seen on **Error! Reference source not found.**.

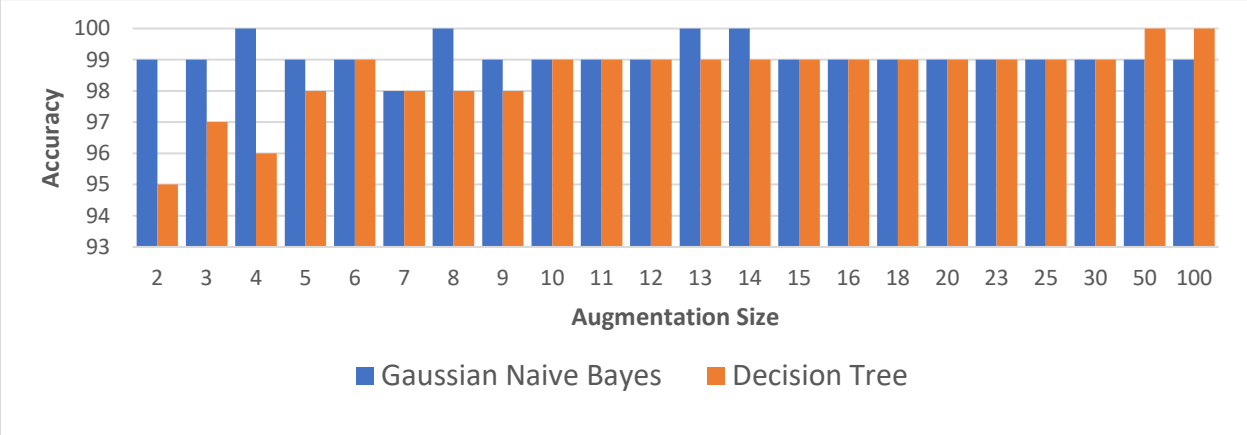

Figure 7. Number of augmentation and the model accuracy

Gaussian Naive Bayes

	precision	recall	f1-score	support
1.0	1.00	0.99	1.00	104
2.0	1.00	1.00	1.00	108
3.0	0.97	1.00	0.98	111
4.0	0.96	1.00	0.98	105
5.0	1.00	0.96	0.98	102
6.0	1.00	0.97	0.99	118
accuracy			0.99	648
macro avg	0.99	0.99	0.99	648
weighted avg	0.99	0.99	0.99	648

Figure 8. Classification Report of Gaussian Naive Bayes using 9 Augmentation

Dessicion Tree

	precision	recall	f1-score	support
1.0	0.95	0.98	0.97	104
2.0	0.97	0.97	0.97	108
3.0	0.98	0.96	0.97	111
4.0	0.99	1.00	1.00	105
5.0	0.99	1.00	1.00	102
6.0	0.97	0.94	0.95	118
accuracy			0.98	648
macro avg	0.98	0.98	0.98	648
weighted avg	0.98	0.98	0.98	648

Figure 9. Classification Report of Decision Tree using 9 Augmentation

From the data above, it produces good results because the mean value of the Gaussian Mixture Model is close to the reference [6] by 59 out of 65. In addition, the process of equalizing the component data size is achieved by changing the absorption band data to the highest probability of a polymer. This achievement also makes the performance of Naïve Bayes perfect with an accuracy value of 100% for 6 K-fold.

However, this perfect result can be different if done in different places and times. Because the absorption band value can be significantly different from the existing reference depending on the climate and weather as well as the contamination contained in the sample according to Song et al. [10] in their research.

CONCLUSION

In this study, the author used Gaussian Mixture and Gaussian Naïve Bayes as absorption band identification and microplastics classification, this is due to their accuracy and information detail given by Gaussian Mixture and Naive Bayes compared to K-Means and Decision Tree. Unequal component data size for each aspect value is no longer a problem for machine learning identification and classification. Preprocessing by separating each group of data and converting them into polymer groups can make the component data length equal.

Gaussian Mixture works very well with a difference of 2.50 points against the reference. However, there are 6 data with large differences. In addition, Gaussian Mixture can generate a range of absorption band values of a polymer as supporting data.

The use of Gaussian Naïve Bayes is considered appropriate based on the results of precision, recall, and f1-score. Each cross-validator that was tried produced a value of 1.0 for each aspect.

Therefore, the combination of Gaussian Mixture and Naïve Bayes can solve the existing problems. The model and method proposed by the authors can also be useful for identification and classification of various things that both use Fourier Transform Infrared Spectroscopy (FTIR) as a polymer detection tool.

The limitation of this project is that the author uses data conducted by Agricultural Technology, Soegijapranata Catholic University students, where the data obtained is very limited, only 210 data for 6 classes of microplastics. The recommendation for future research is to add more data. In addition, further research can compare this research in areas which have different contamination and weather, to get a more varied absorption band scope.

REFERENCES

[1] J. P. G. L. Frias and R. Nash, "Microplastics: Finding a consensus on the definition," *Marine Pollution Bulletin*, vol. 138, pp. 145–147, Jan. 2019, doi: 10.1016/j.marpolbul.2018.11.022.

[2] "A Detailed Review Study on Potential Effects of Microplastics and Additives of Concern on Human Health - PMC." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7068600/ (accessed Apr. 18, 2019).

[3] T. Liu, Z. Chen, H. Liu, and Z. Zhang, "FTIR spectral imaging enhancement for teacher's facial expressions recognition in the intelligent learning environment," *Infrared Physics & Technology*, vol. 93, pp. 213–222, Sep. 2018, doi: 10.1016/j.infrared.2018.07.035.

[4] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," in *2019 International Engineering Conference (IEC)*, Erbil, Iraq: IEEE, Jun. 2019, pp. 165–170. doi: 10.1109/IEC47844.2019.8950650.

[5] N. Weinberger and G. Bresler, "The EM algorithm is adaptively-optimal for unbalanced symmetric Gaussian mixtures," *J. Mach. Learn. Res*, 2021, [Online]. Available: https://www.jmlr.org/papers/volume23/21-0186/21-0186.pdf

[6] M. R. Jung *et al.*, "Validation of ATR FT-IR to identify polymers of plastic marine debris, including those ingested by marine organisms," *Marine Pollution Bulletin*, vol. 127, pp. 704–716, Feb. 2018, doi: 10.1016/j.marpolbul.2017.12.061.

[7] J. Gago, A. Filgueiras, M. L. Pedrotti, M. Caetano, and J. Frias, "Standardised protocol for monitoring microplastics in seawater. Deliverable 4.1.," 2019, doi: 10.25607/OBP-605.
[8] R. Sharma and H. Kaushik, "Micro-plastics: An invisible danger to human health," *CGCIJCTR*, vol. 3, no. 2, pp. 182–186, Jul. 2021, doi: 10.46860/cgcijctr.2021.06.31.182.
[9] A. Dutta, "Fourier Transform Infrared Spectroscopy," in *Spectroscopic Methods for Nanomaterials Characterization*, Elsevier, 2017, pp. 73–93. doi: 10.1016/B978-0-323-46140-5.00004-2.
[10] Y. K. Song *et al.*, "A comparison of microscopic and spectroscopic identification methods for analysis of microplastics in environmental samples," *Marine Pollution Bulletin*, vol. 93, no. 1–2, pp. 202–209, Apr. 2015, doi: 10.1016/j.marpolbul.2015.01.015.
[11] Q. Zhang and J. Chen, "Distributed Learning of Finite Gaussian Mixtures," *Journal of Machine Learning Research*, vol. 23, no. 99, pp. 1–40, 2022.
[12] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed, "Multinomial Naive Bayes classification model for sentiment analysis," *IJCSNS Int. J. Comput. Sci. Netw. Secur*, vol. 19, no. 3, p. 62, 2019.
[13] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model," 2013, pp. 194–201. doi: 10.1007/978-3-642-41278-3_24.
[14] A. H. Jahromi and M. Taheri, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features," in *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, Shiraz: IEEE, Oct. 2017, pp. 209–212. doi: 10.1109/AISP.2017.8324083.
[15] H. Liu, X. Chen, and R. Li, "The Optimization of Finishing Train Based on Improved Genetic Algorithm," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 12, no. 5, pp. 3555–3559, 2014.

APPENDIX

Polymer Grouping

No	Polymer
1	C-Cl stretching
2	Polar ester groups and benzene ring interaction
3	CH2 rocking
4	Aromatic C-H stretching
5	Ethyl branching
6	Adjacent two aromatic H vibration and aromatic bands
7	C-CH3 stretching
8	Aromatic rings 1,2,4,5; Tetra replaced
9	Vinylidene group
10	Terminal vinyl group
11	C=C
12	C-CH3 rocking
13	Aromatic C-H bending
14	Methylene group and ester C-O bond vibrations
15	C-C stretching
16	Terephthalate Group (OOCC6H4-COO)
17	CH2 bending

No	Polymer
18	C-O-C
19	C-H bending
20	C-N stretching
21	C-O group stretching of the O-H group deformation and ethylene glycol bending and wagging vibration
22	C-CH3 symmetric
23	CH3 groups
24	CH2 scissors
25	CH2 symmetric
26	CH2 scissors vibration
27	C=C aromatic stretch
28	N-H bending
29	C=O stretching
30	C=O stretch
31	CO2 axial symmetric deformation
32	C-H stretching reflects
33	Symmetric CH2 stretch
34	C-H stretching
35	Symmetric C-H stretch
36	CH2 asymmetric
37	Asymmetric CH2 stretch
38	CH3 symmetric
39	Symmetric CH stretch
40	N-H stretching
41	OH group (hydroxyl)

Result of Gaussian Mixture Model

No	Polymer	Reference	Calculated Means	Difference	Calculated Variance
0	C-Cl stretching	700	707.7361	7.7361	18.11433
1	Polar ester groups and benzene ring interaction	712	711.73	0.27	1E-06
2	CH2 rocking	720	721.1141	1.1141	2.381933
3	CH2 rocking	730	729.7961	0.2039	0.864328
4	Aromatic C-H stretching	757	759.6795	2.6795	12.19483
5	Ethyl branching	775	773.7793	1.2207	6.10176
6	Adjacent two aromatic H vibration and aromatic bands	795	794.0669	0.9331	2.662722
7	C-CH3 stretching	840	840.96	0.96	1E-06
8	Aromatic rings 1,2,4,5; Tetra replaced	848	848.8536	0.8536	11.82362
9	Aromatic rings 1,2,4,5; Tetra replaced	872	0	872	0.000001
10	Vinylidene group	890	880.1908	9.8092	59.83227
11	Terminal vinyl group	910	910.0427	0.0427	6.614106

No	Polymer	Reference	Calculated Means	Difference	Calculated Variance
12	CH2 rocking	966	964.41	1.59	1E-06
13	C=C	967	967.3387	0.3387	2.046454
14	C-CH3 rocking	972	974.1955	2.1955	1.673468
15	C-CH3 rocking	997	998.7791	1.7791	0.554114
16	Aromatic C-H bending	1027	1031.516	4.5155	54.2538
17	Methylene group and ester C-O bond vibrations	1050	0	1050	0.000001
18	Methylene group and ester C-O bond vibrations	1096	1099.213	3.2133	3.116134
19	C-C stretching	1099	3.236379	1095.764	3.220798
20	Terephthalate Group (OOCC6H4-COO)	1124	5	1119	1E-06
21	C-CH3 rocking	1165	1169.299	4.2986	0.654167
22	CH2 bending	1199	1194.009	4.9911	0.660411
23	C-O-C	1220	1223.146	3.1457	0.456111
24	Terephthalate Group (OOCC6H4-COO)	1240	1238.3	1.7	1E-06
25	C-H bending	1255	1254.534	0.4658	4.009442
26	C-N stretching	1274	1271.504	2.4964	1.425345
27	C-H bending	1331	1329.915	1.0846	1.788595
28	C-O group stretching of the O-H group deformation and ethylene glycol bending and wagging vibration	1342	1345.948	3.9481	0.567466
29	CH2 bending	1372	1369.003	2.9973	3.395884
30	C-CH3 symmetric	1375	1386.82	11.82	1E-06
31	CH3 groups	1377	1377.723	0.7227	0.909211
32	C-O group stretching of the O-H group deformation and ethylene glycol bending and wagging vibration	1410	1414.34	4.3396	1.30592
33	CH2 scissors	1427	1427.321	0.3211	0.823466
34	CH2 scissors	1435	1434.536	0.4636	1.041862
35	CH2 bending	1451	1452.4	1.4	1E-06
36	C-O group stretching of the O-H group deformation and ethylene glycol bending and wagging vibration	1453	1459.371	6.3710	33.15763
37	CH2 symmetric	1455	1454.33	0.67	0.000001
38	CH2 scissors vibration	1463	1463.97	0.97	9.98E-07
39	CH2 bending	1464	1466.752	2.7518	0.91891
40	CH2 scissors vibration	1475	1475.54	0.54	1.01E-06
41	C=C aromatic stretch	1504	1500.182	3.8180	38.95998
42	N-H bending	1530	1533.916	3.9159	36.45554
43	Aromatic C-H stretching	1547	1540.728	6.2716	2.351811

No	Polymer	Reference	Calculated Means	Difference	Calculated Variance
44	C=C aromatic stretch	1577	1586.251	9.2509	121.0465
45	C=O stretching	1634	1632.578	1.4222	1.849629
46	C=O stretching	1650	1648.816	1.1837	21.4591
47	C=O stretch	1730	1731.115	1.115	8.381026
48	Adjacent two aromatic H vibration and aromatic bands	1960	1959.906	0.0935	4.324141
49	CO2 axial symmetric deformation	2350	2350.907	0.9067	8.367101
50	CH2 symmetric	2838	2840.099	2.0986	7.077286
51	C-H stretching reflects	2850	2850.79	0.79	1.01E-06
52	Symmetric CH2 stretch	2852	2853.42	1.4202	0.889716
53	C-H stretching	2858	2856.546	1.4536	2.162724
54	Symmetric C-H stretch	2908	2910.483	2.4829	1.664519
55	CH2 asymmetric	2917	2922.911	5.9113	1.700227
56	C-H stretching reflects	2923	2922.503	0.4969	2.11109
57	Asymmetric CH2 stretch	2927	2926.01	0.99	1.01E-06
58	C-H stretching	2932	2932.071	0.0712	8.37503
59	CH3 symmetric	2952	2960.43	8.4302	67.7309
60	Symmetric C-H stretch	2969	0	2969	0.000001
61	Symmetric CH stretch	3054	3044.725	9.2747	1905.845
62	Aromatic C-H stretching	3055	3057.607	2.6069	15.43262
63	N-H stretching	3298	3297.725	0.275	9.51709
64	OH group (hydroxyl)	3432	3431.623	0.3768	2.131814