# PCA-K-means Based Clustering Algorithm for High Dimensional and Overlapping Spectra Signals

Nian Zhang, Keenan Leatham
Dept. of Electrical and Computer Eng.
Univ. of the District of Columbia
Washington, D.C. 20008 USA
nzhang@udc.edu, keenan.leatham@udc.edu

Jiang Xiong, Jing Zhong
College of Computer Science and Eng.
Chongqing Three Gorges University
Chongqing, 404000, China
xjcq123@sohu.com, zhongandy@sohu.com

***Abstract*—This paper applied a PCA-K-means method to exploit photo-thermal infrared imaging spectroscopy based trace explosives with overlapping spectral absorption bands. We intend to explore the underlying patterns that affect the clustering performance using top principal components. We also strive to investigate the effectiveness of the clustering algorithm on different analytes and substrates. We reduced the dimensions by applying the principal component analysis (PCA) on the data to transform the original data to the top principal components' feature space. The data were revealed in the feature space and formed into clusters. Then we used the K-mean based clustering algorithm to classify them into six classes including RDX, PC, Copper/Steel, TNT, DNT, and PE. After that, we conducted the performance evaluation. We found that the F1 score of the classification of RDX, PC, Copper/Steel, TNT, DNT, and PE is 85%, 39%, 71%, 99%, 92%, and 86%, respectively. The results demonstrated that the proposed algorithm can effectively reduce dimension and accurately determined the classes of those analytes and substrates.**

***Keywords—classification; clustering; principal component analysis; k-mean clustering; big data; high dimensional; overlapped spectra***

## I. INTRODUCTION

Trace analyte detection has become an emergent goal in the fields of military, homeland security, and law enforcement. It provides an early warning of concealed threats and therefore can save people's lives and protect the public facilities. This technology includes remote detection systems capable of detecting explosives and other hazardous materials from a standoff distance. As the demand from military and security markets has increased. promote research and development for efficient detection systems to face the problems of hidden explosives at public places, such as suicide, leave-behind, and vehicle-borne explosives in airports, railway, ship, bus, truck, container, bridge, tunnel, tower, terminal environments, and coach stations. Also, the development of analytical tools can identify explosive remains of tremendous importance in the forensic field for crime-scene reconstruction. In addition, the detection of explosives is also used in many peaceful applications. For example, it can be relevant in environmental areas to monitor the quality of soil, water, and groundwater suspected of being contaminated by explosives and their degradation products, in order to prevent poisoning of populations of humans and animals.

There is an emerging thrust to potentially replace the traditional point detection of trace residues (which requires physical contact) with standoff sensing capability from a safe distance. Improving survivability and situational awareness has spurred a wide variety of recent development of sensor technology. Most of them are laser-based trace detection technologies, such as laser-induced breakdown spectroscopy (LIBS), Raman spectroscopy, laser-induced-fluorescence spectroscopy, and Fourier transform IR spectroscopy. The resonant infrared (IR) photothermal technology that can remotely detect trace explosives material on relevant substrate surfaces from significant standoff distances using photo-thermal infrared (IR) imaging spectroscopy (PT-IRIS). In this technique, light from a specific infrared wavelength is directed to the surface of interest and the thermal response is viewed with an infrared camera [1]. Comparing the thermal image as a function of incident wavelength with the absorption spectrum of explosives reveals the presence and location of trace residues. By varying the incident wavelength, other analytes of interest (e.g., drugs and chemical agents) could also be imaged. Compared to other trace detection techniques, this technology has the potential to generate chemical images of the chemical composition of surfaces and bulk materials with a spatial resolution of ~1um.

However the ability to detect small amounts of analytes across large relevant substrates is complicated by the optical and thermal analyte/substrate interactions. The key challenge of remote detection techniques is to distinguish materials such as explosives from the substrate materials on which they lay, such as glasses, paint, or clothes. While substrate materials are chemically distinct from explosives, they nonetheless have complicated and overlapping IR features with explosives. A

complication using polymeric materials, they tend to absorb throughout the IR. These universal considerations introduced by real-world surfaces complicate the detection and identification of explosive materials.

The advancements in infrared (IR) and Raman spectroscopy have led to an explosive growth in stored or transient data and have generated an urgent need for new and automated methods of spectral data analysis. The emerging photo-thermal infrared imaging spectroscopy (PT-IRIS) technique which allows for further increase of the spatial resolution from the current ~10 microns to ~1 micron makes this data analysis demand more critical. This paper will focus on the PT-IRIS data analysis which was used for the application of trace analyte detection. The aim of trace analyte detection is to distinguish illicit analytes such as explosives from the substrates on which they rest. To date, there have been insignificant efforts to analyze the photo-thermal infrared data sets using computational intelligence techniques.

The rest of the paper is organized as follows. In Section II, the high dimensional and overlapped data set is described. In Section III, the proposed methodology is presented. A combined principal component analysis (PCA)-K-means clustering algorithm is described. In Section IV, the analysis and results are presented. In Section V, the conclusions are given.

## II. Photo-Thermal Infrared Imaging Spectroscopy

This section will use a PCA-K-means method to exploit PT-IRIS based trace explosives with overlapping spectral absorption bands. We intend to explore the underlying patterns that affect the clustering performance using top principal components. We also strive to investigate the effectiveness of the clustering algorithm on different analytes and substrates.

### *A. Data Set*

The advanced photo-thermal infrared imaging spectroscopy (PT-IRIS) technique that can be used for standoff detection application [1]. The two fundamental components are infrared (IR) quantum cascade lasers (QCL) and IR focal plane array detectors. Specifically, IR QCL is used to illuminate a surface potentially contain residues of interest. If the excitation wavelength of the light is resonant with collection wavelength of surface residues, the residues of interest will heat up by (~1ºC). The IR focal plane array detectors are used for imaging.

The temperature increase, Tmax is measured as a function of excitation and collection wavelengths at the end of a laser pulse. Tmax normalized to the average power of the laser pulse will then be used as feature vectors, as shown in Fig. 1. Simulated samples include 5 analytes (TNT, DNT, RDX, Polyethylene, and Polycarbonate) on 4 substrates (Copper, Steel, Polyethylene, and Polycarbonate) using 28 excitation wavelengths (6.0 μm to 6.6 μm and 7.0 μm to 7.7 μm) and 26 collection wavelengths (8.0 μm to 10.5 μm).

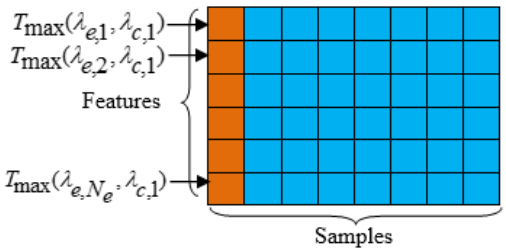


Fig. 1 Data matrix with feature vectors. *Tmax* values for each excitation and collection wavelength pair were arranged into a feature vector (column).

The fundamental spectroscopic characteristics for the PT-IRIS is shown in Fig. 2. It shows the IR absorbance spectra of the residues of interest. Peaks in the curves reflect unique "signatures" for each analyte. According to Kirchhoff's Law, the emissivity of a material and its absorptivity are equivalent at thermal equilibrium. Thus, the measured absorption spectrum of a material obtained at ambient temperature can be used to accurately predict its emission spectrum. In another word, if we can determine the most important features (i.e. Tmax values) among the 728 features, which are correlated with the absorption spectrum of a material, we can use the absorption spectrum to predict its emission spectrum. Since the thermal emission from analyte of interest and substrate materials have different spectral signatures, the unique thermal emission spectrum can ultimately determine the type of this material.

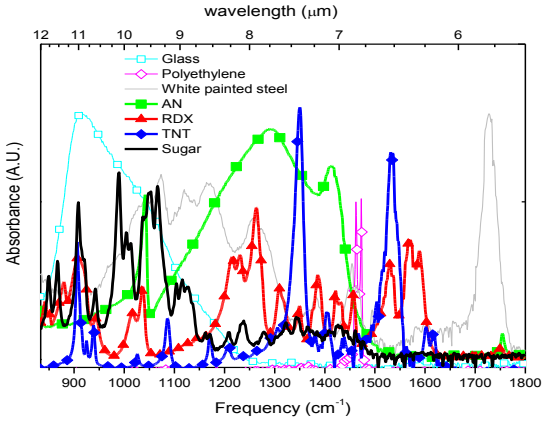


Fig. 2 IR absorbance spectra of glass, polyethylene, white painted steel, ammonium nitrate, RDX, TNT and sucrose (sugar).

Each pixel in the camera frame was a column in this data matrix, which includes 728 features. For each feature, it is a function of Tmax in terms of different excitation wavelengths and collection wavelengths. With 28 excitation wavelengths and 26 collection wavelengths, they would generate 728 features. Therefore, we may demonstrate the photo-thermal signal matrix for all the 468 samples. This can be seen by display ,the data set in false color plot ,which will show visible or non-visible parts of the electromagnetic spectrum, shown in Fig. 3. From left to right, the particle sizes are 8 μm, 12 μm, 20 μm, 3 μm, 1.5 μm, and 5 μm. Each loop takes 76 columns per particle size. Each column contains 728 features. In Fig. 3,

the color of the data point is proportional to signal strength, i.e. red represents high, and blue represents low.

For each analyte including TNT, DNT, and RDX, they will be made of all 6 possible particle sizes, and two pixels in the camera frame (i.e. columns) are on the particle, the two pixels will rest on all 4 substrates (i.e. copper, steel, PC, and PE). Thus, there will be a maximum of 48 samples for each analyte. In addition, for each substrate including copper, steel, PC, and PE, they will spread over the 6 particle sizes with two pixels off each particle, and they interact with 5 analytes (TNT, DNT, RDX, PC, and PE), so the maximum of substrate samples is 60.

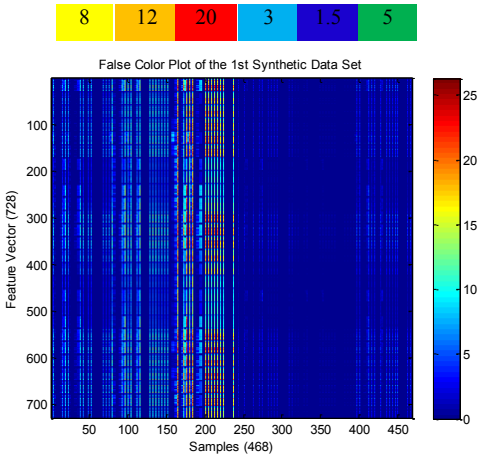


Fig. 3 Six clusters consisting of the four analytes and two substrates were formed using the K-means clustering method.

PC can be used for both analyte and substrate, so the maximum sample amount would be the total of possible analyte and substrate samples, which is therefore 108. PC has a strong signature. PE can also be used for both analyte and substrate. There will be 108 samples at maximum. PE is known as a poor thermal conductor, as a result the temperature increase with illumination is nearly zero. Thus it does not have its own signature.

Since the samples comprise both "on" and "off" particle pixels, some samples may have analyte, some may have substrate, but some complicated and overlapping sample can happen by optically "through" a particle. The mixing IR absorption/emission features reflects the primary challenge to a useful detection technique in the real-world application.

## III. Research Methodology

Feature selection is a very important pre-processing technique for large scale pattern recognition problems, especially when the number of samples is relatively small [2]-[4]. Feature selection techniques are designed to find a subset of relevant feature subset of the original features which can facilitate clustering, classification and retrieval [5][6]. A data set contains relevant, irrelevant, and redundant features. However, irrelevant and redundant features are not useful for classification, and they may even reduce the classification performance due to the large search space known as "the curse of dimensionality". By eliminating irrelevant and redundant features, feature selection could helps in understanding data, reducing computation requirement, reducing the effect of curse of dimensionality and improving the predictor performance [7]-[13].

### *A. Principle Component Analysis (PCA)*

Principle component analysis (PCA) is quantitatively rigorous method for achieving dimensional reduction before applying the feature selection methods. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences [14]. The method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. Several top ranking principal components will be selected to form a new feature space. The original samples will be transformed to this new feature space in the directions of the principal components. Although the PCA can effectively reduce the number of dimensions by selecting the top ranking principle components, PCA method is not able to select a subset of features which are important to distinguish the classes. It only guarantees that when you project each observation on an axis (along a principle component) in a new space, the variance of the new variable is the maximum among all possible choices of that axis. This means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance.

In the application of trace analyte detection, we consider the feature selection problem in unsupervised learning scenario, which is particularly difficult due to the absence of class labels that would guide the search for relevant information. The feature selection problem is essentially a combinatorial optimization problem which is computationally expensive. The existing and most powerful unsupervised feature selection technique is principle component analysis (PCA). It is often useful to map data onto their principal components rather than on the original x-y axis. In this way the underlying structure in the data can be revealed. We applied the PCA technique to the data set to reveal the patterns in data, as well as reduce the dimension of feature vectors (i.e. vectors containing the principle components). First we deconstruct the set into eigenvectors and eigenvalues. An eigenvector is a direction, and an eigenvalue is a number, telling you how much variance there is in the data in that direction. The amount of eigenvectors/values that exist equals the number of dimensions the data set has. The reason for this is that eigenvectors put the data into a new set of dimensions, and these new dimensions have to be equal to the original amount of dimensions. It is worthwhile to investigate the PCA algorithm because it allows us to exploit the correlation of most significant eigenvectors and analyte types.

### B. *PCA-K-means Clustering Algorithm*

PCA algorithm will be applying to reduce the dimensions, and then the k-means clustering algorithm will be applied.

| Steps | **K-means Clustering Algorithm** |
|---|---|
| 1 | k initial "means" (k is an estimated value) are randomly generated within the data domain. |
| 2 | k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram [8] generated by the means. |
| 3 | The centroid of each of the k clusters becomes the new mean. |
| 4 | Steps 2 and 3 are repeated until convergence has been reached. |

## IV. ANALYSIS AND RESULTS

### A. *PCA-K-means Clustering Results*

We presented the principal component analysis (PCA) results using a combination of top PCs, i.e. PC1 and PC2. We displayed the data in PC1-PC2 axes, as shown in Fig. 4.

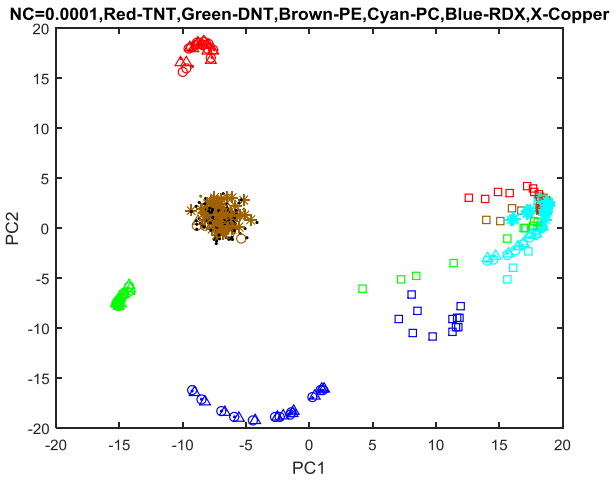


Fig. 4 All the data projected onto PC1-PC2 space after using the PCA method.

Then we applied the K-mean clustering algorithm to make them into 6 clusters. The number of analytes and substrates in each of the six classes are shown in Table I. Letter C represents the cluster. The rows represent TNT, DNT, PE, PC, RDX, and Copper, respectively. The columns represent Cluster 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5, and Cluster 6.

Table I. Number of Samples in Classifiers after Using the PCA-K-means Clustering Algorithm

| | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|
| TNT | 0 | 10 | 0 | 36 | 0 | 0 |
| DNT | 0 | 6 | 0 | 0 | 36 | 0 |
| PE | 0 | 12 | 96 | 0 | 0 | 4 |
| PC | 0 | 104 | 0 | 0 | 0 | 0 |
| RDX | 34 | 0 | 0 | 0 | 0 | 12 |
| Copper | 0 | 0 | 118 | 0 | 0 | 0 |

By counting the largest number of analytes/substrate in each class, we can find the analyte/substrate dominating that class. For each cluster, the majority analytes will determine the class that this cluster belongs to. Therefore, we find Cluster 1 represents RDX, Cluster 2 represents PC, Cluster 3 represents Copper, Cluster 4 represents TNT, Cluster 5 represents DNT, and Cluster 6 represents PE. These labels are shown in the tables in the following Clustering Analysis section.

The classes are demonstrated in PC1-PC2 axes, as shown in Fig. 5. To enhance the visualization efficiency, the same color code is used in the k-means algorithm as in the PCA algorithm. Red dots represent TNT, Green dots represent DNT, Yellow dots represents PE, Cyan dots represent PC, Blue dots represents RDX, and Black dots represent Copper/Steel. The centroids of the classes are indicated by black crosses. Thus, it is easier to compare Fig. 4 and Fig. 5.

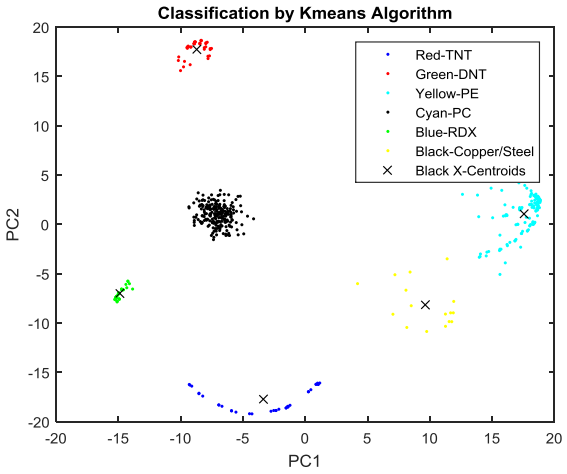


Fig. 5 Six clusters consisting of the four analytes and two substrates were formed using the K-means clustering method.

### B. *Clustering Analysis*

We conducted the popular clustering performance evaluation which consists of probability of detection (POD), false alarm rate (FAR), accuracy, precision, recall, and F1 score for each residue of interest. F1 score is usually more useful than accuracy, especially if there exists an uneven class distribution. The procedure involves the determination of true positive (TP), false negative (FN), false positive (FP), and true negative (TN) of residues of interest. In the pre-processing

stage, based on Table I, six tables are generated for the six analytes in Table II-VII. Different colors are chosen for these metrics, as shown below.


Table II. Labeling for TNT

| Analytes | Class 1 (RDX) | Class 2 (PC) | Class 3 (Copper) | **Class 4 (TNT)** | Class 5 (DNT) | Class 6 (PE) |
|---|---|---|---|---|---|---|
| **TNT** | 0 FN | 10 FN | 0 FN | 36 TP | 0 FN | 0 FN |
| DNT | 0 TN | 6 TN | 0 TN | 0 FP | 36 TN | 0 TN |
| PE | 0 TN | 12 TN | 96 TN | 0 FP | 0 TN | 4 TN |
| PC | 0 TN | 104 TN | 0 TN | 0 FP | 0 TN | 0 TN |
| RDX | 34 TN | 0 TN | 0 TN | 0 FP | 0 TN | 12 TN |
| Copper | 0 TN | 0 TN | 118 TN | 0 FP | 0 TN | 0 TN |

Table III. Labeling for DNT

| Analytes | Class 1 (RDX) | Class 2 (PC) | Class 3 (Copper) | Class 4 (TNT) | **Class 5 (DNT)** | Class 6 (PE) |
|---|---|---|---|---|---|---|
| TNT | 0 TN | 10 TN | 0 TN | 36 TN | 0 FP | 0 TN |
| **DNT** | 0 FN | 6 FN | 0 FN | 0 FN | 36 TP | 0 FN |
| PE | 0 TN | 12 TN | 96 TN | 0 TN | 0 FP | 4 TN |
| PC | 0 TN | 104 TN | 0 TN | 0 TN | 0 FP | 0 TN |
| RDX | 34 TN | 0 TN | 0 TN | 0 TN | 0 FP | 12 TN |
| Copper | 0 TN | 0 TN | 118 TN | 0 TN | 0 FP | 0 TN |

Table IV. Labeling for PE

| Analytes | Class 1 (RDX) | Class 2 (PC) | Class 3 (Copper) | Class 4 (TNT) | Class 5 (DNT) | **Class 6 (PE)** |
|---|---|---|---|---|---|---|
| TNT | 0 TN | 10 TN | 0 TN | 36 TN | 0 TN | 0 FP |
| DNT | 0 TN | 6 TN | 0 TN | 0 TN | 36 TN | 0 FP |
| **PE** | 0 FN | 12 FN | 96 FN | 0 FN | 0 FN | 4 TP |
| PC | 0 TN | 104 TN | 0 TN | 0 TN | 0 TN | 0 FP |
| RDX | 34 TN | 0 TN | 0 TN | 0 TN | 0 TN | 12 FP |
| Copper | 0 TN | 0 TN | 118 TN | 0 TN | 0 TN | 0 FP |

Table V. Labeling for PC

| Analyes | Class 1 (RDX) | **Class 2 (PC)** | Class 3 (Copper) | Class 4 (TNT) | Class 5 (DNT) | Class 6 (PE) |
|---|---|---|---|---|---|---|
| TNT | 0 TN | 10 FP | 0 TN | 36 TN | 0 TN | 0 TN |
| DNT | 0 TN | 6 FP | 0 TN | 0 TN | 36 TN | 0 TN |
| PE | 0 TN | 12 FP | 96 TN | 0 TN | 0 TN | 4 FN |
| **PC** | 0 FN | 104 TP | 0 FN | 0 FN | 0 FN | 0 FN |
| RDX | 34 TN | 0 FP | 0 TN | 0 TN | 0 TN | 12 TN |
| Copper | 0 TN | 0 FP | 118 TN | 0 TN | 0 TN | 0 TN |

Table VI. Labeling for RDX

| Analytes | **Class 1 (RDX)** | Class 2 (PC) | Class 3 (Copper) | Class 4 (TNT) | Class 5 (DNT) | Class 6 (PE) |
|---|---|---|---|---|---|---|
| TNT | 0 FP | 10 TN | 0 TN | 36 TN | 0 TN | 0 TN |
| DNT | 0 FP | 6 TN | 0 TN | 0 TN | 36 TN | 0 TN |
| PE | 0 FP | 12 TN | 96 TN | 0 TN | 0 TN | 4 FN |
| PC | 0 FP | 104 TN | 0 TN | 0 TN | 0 TN | 0 TN |
| **RDX** | 34 TP | 0 FN | 0 FN | 0 FN | 0 FN | 12 FN |
| Copper | 0 FP | 0 TN | 118 TN | 0 TN | 0 TN | 0 TN |

Table VII. Labeling for Copper/Steel

| Analytes | Class 1 (RDX) | Class 2 (PC) | **Class 3 (Copper)** | Class 4 (TNT) | Class 5 (DNT) | Class 6 (PE) |
|---|---|---|---|---|---|---|
| TNT | 0 TN | 10 TN | 0 FP | 36 TN | 0 TN | 0 TN |
| DNT | 0 TN | 6 TN | 0 FP | 0 TN | 36 TN | 0 TN |
| PE | 0 TN | 12 TN | 96 FP | 0 TN | 0 TN | 4 TN |
| PC | 0 TN | 104 TN | 0 FP | 0 TN | 0 TN | 0 TN |
| RDX | 34 TP | 0 TN | 0 FP | 0 TN | 0 TN | 12 TN |
| **Copper** | 0 FN | 0 FN | 118 TP | 0 FN | 0 FN | 0 FN |

### C. *Clustering Performance Evaluation*

We then conducted the clustering performance evaluation by calculating the evaluation metrics, including the probability of detection (POD), false alarm rate (FAR), accuracy, precision, recall, and F1 score (the higher the better) for each residue of interest. The above evaluation metrics are defined as follows [9]:

*Probability of Detection*: POD = TP/(TP+FN)
*False Alarm Rate*: FAR = FP/(FP+TN)
*Precision*: P = TP/(TP+FP)
*Recall*: R = TP/(TP+FN)
*Accuracy*: Accuracy = (TP+TN)/(TP+FP+FN+TN)
*F1 Score*: F1 Score = 2 * Precision * Recall/(Precision + Recall)

From the equations, we see that precision measures how often an instance was predicted as positive that is actually positive, while recall measures how often a positive class instance in the data set was predicted as a positive class instance by the classifier. In imbalanced data set, the goal is to improve recall without hurting precision. These goals, however, are often conflicting, since in order to increase the TP for the minority class, the number of FP is also often increased, resulting in reduced precision.

The k-means clustering algorithm on the data on PC1 and PC2 was performed. The clustering performance including the probability of detection (POD), false alarm rate (FAR), accuracy, precision, recall and F1 score is conducted. The performance results are shown in Table VIII. The six clusters were determined by the majority analyte type in each cluster.

Accuracy can be significantly affected by the number of true negatives which in the application of trace analyte detection, are not as critical indicators as false negative and false positive. Therefore, F1 score is usually a better measure to evaluate if we need to seek a balance between precision and recall and the data has an uneven class distribution.

Compared to polyethylene (PE), copper, and steel, only polycarbonate (PC) has its own “spectrum”. Spectral mixing

problem becomes worst when the trace analytes rest on such active substrate. It will result in poor F1 score.

TABLE VIII. PCA-K-Means Clustering Performance on PC1 and PC2

| | Class1 (RDX) | Class2 (PC) | Class3 (Cop) | Class4 (TNT) | Class5 (DNT) | Class6 (PE) |
|---|---|---|---|---|---|---|
| POD | 74% | 100% | 100% | 78% | 86% | 10% |
| FAR | 7% | 8% | 27% | 0 | 0 | 1% |
| Accuracy | 97% | 94% | 79% | 98% | 98% | 76% |
| Precision | 100% | 24% | 55% | 100% | 100% | 75% |
| Recall | 74% | 100% | 100% | 98% | 86% | 10% |
| F1 Score | 85% | 39% | 71% | 99% | 92% | 86% |

## V. CONCLUSION

This paper applied a PCA-K-means method to exploit PT-IRIS based trace explosives with overlapping spectral absorption bands. We intend to explore the underlying patterns that affect the clustering performance using top principal components. We also strive to investigate the effectiveness of the clustering algorithm on different analytes and substrates. The principal component analysis (PCA) was used to reduce the dimension of data space to the top principal components feature (PC1-PC2) space, and thus the most prominent features or patterns were revealed. Then we used the K-mean clustering algorithm to classify them into four analytes and two substrates. We used the performance evaluation matrices to measure the accuracy of classification. The experimental results demonstrated that the combination of the principal component analysis and K-means clustering algorithm are efficient for achieving dimensional reduction and clustering on highly overlapped photo-thermal infrared imaging data. The F1 score of the classification of RDX, PC, Copper, TNT, DNT, and PE is 85%, 39%, 71%, 99%, 92%, and 86%, respectively.

## ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF) grants: HRD #1505509, HRD #1533479, and DUE #1654474.

## REFERENCES

[1] C. Kendziora, R. Furstenberg, M. Papantonakis, V. Nguyen, J.Stepnowski, and R. McGill, "Advances in standoff detection of trace explosives by infrared photo-thermal imaging," *Proc. SPIE 7664, Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XV*, 76641J, 2010.

[2] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, "Simultaneous spectral-spatial feature selection and extraction for hyperspectral images," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 16-28, 2018.

[3] M. Liu, C. Xu, Y. Luo, C. Xu, Y. Wen, and D. Tao, "Cost-sensitive feature selection by optimizing f-measures," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1323-1335, 2018.

[4] X. Wen, L. Shao, W. Fang, and Y. Xue, "Efficient feature selection and classification for vehicle detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 508-517, 2015.

[5] M. Dash and H. Liu, "Feature selection for classification," Intell. Data Anal., vol. 1, no. 1-4, pp. 131-156, 1997.

[6] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," Eur. J. Oper. Res., vol. 206, no. 3, pp. 528–539, Nov. 2010.

[7] N. Zhang and K. Leatham, "Feature selection based on SVM in photo-thermal infrared (IR) imaging spectroscopy classification with limited training samples", *WSEAS Transactions on Signal Processing*, ISSN / E-ISSN: 1790-5052 / 2224-3488, vol. 13, Art. #33, pp. 285-292, 2017.

[8] N. Zhang, J. Xiong, J. Zhong, and K. Leatham, "Gaussian process regression method for classification for high-dimensional data with limited samples", *The 8th International Conference on Information Science and Technology (ICIST 2018)*, Cordoba, Granada, and Seville, Spain, June 30-July 6, 2018.

[9] N. Zhang, J. Xiong, J. Zhong, and L. A. Thompson, "Feature selection method using BPSO-EA with ENN classifier", *The 8th International Conference on Information Science and Technology (ICIST 2018)*, Cordoba, Granada, and Seville, Spain, June 30-July 6, 2018.

[10] N. Zhang and L. A. Thompson, "An intelligent clustering algorithm for high dimensional and highly overlapped photo-thermal infrared imaging data," *Fall 2016 ASEE Mid-Atlantic Regional Conference*, Hofstra University, Hempstead, NY, October 21-22, 2016.

[11] N. Zhang, "Cost-sensitive spectral clustering for photo-thermal infrared imaging data," *2016 Sixth International Conference on Information Science and Technology (ICIST)*, Dalian, pp. 358 – 361, May 6-8, China, 2016.

[12] J. F. Ramirez Rochac and N. Zhang, "Reference clusters based feature extraction approach for mixed spectral signatures with dimensionality disparity," *10th Annual IEEE International Systems Conference (IEEE SysCon 2016)*, Orlando, Florida, pp. 1 – 5, April 18-21, 2016.

[13] J. F. Ramirez Rochac and N. Zhang, "Feature extraction in hyperspectral imaging using adaptive feature selection approach," *The Eighth International Conference on Advanced Computational Intelligence (ICACI2016)*, Chiang Mai, Thailand, pp. 36-40, 2016.

[14] L. I. Smith, A Tutorial on Principal Components Analysis, 2002.