Improved detection of chemical substances from colorimetric sensor data using probabilistic machine learning

Lasse L. Mølgaard[a], Ole T. Buus[a], Jan Larsen[a], Hamid Babamoradi[b], Ida L. Thygesen[b], Milan Laustsen[b], Jens Kristian Munk[b], Eleftheria Dossi[c], Caroline O'Keeffe[c], Lina Lässig[d], Sol Tatlow[e], Lars Sandström[f], and Mogens H. Jakobsen[b]

[a]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[b]Department of Micro- and Nanotechnology, Technical University of Denmark, Denmark
[c]Centre for Defence Chemistry, Cranfield University, United Kingdom
[d]Securetec Detektions-Systeme AG, Germany
[e]Pro Design Electronic GmbH, Germany
[f]Gammadata Instrument AB, Sweden

ABSTRACT

We present a data-driven machine learning approach to detect drug- and explosives-precursors using colorimetric sensor technology for air-sampling.

The sensing technology has been developed in the context of the CRIM-TRACK project. At present a fully-integrated portable prototype for air sampling with disposable sensing chips and automated data acquisition has been developed. The prototype allows for fast, user-friendly sampling, which has made it possible to produce large datasets of colorimetric data for different target analytes in laboratory and simulated real-world application scenarios.

To make use of the highly multi-variate data produced from the colorimetric chip a number of machine learning techniques are employed to provide reliable classification of target analytes from confounders found in the air streams. We demonstrate that a data-driven machine learning method using dimensionality reduction in combination with a probabilistic classifier makes it possible to produce informative features and a high detection rate of analytes. Furthermore, the probabilistic machine learning approach provides a means of automatically identifying unreliable measurements that could produce false predictions.

The robustness of the colorimetric sensor has been evaluated in a series of experiments focusing on the amphetamine pre-cursor phenylacetone as well as the improvised explosives pre-cursor hydrogen peroxide. The analysis demonstrates that the system is able to detect analytes in clean air and mixed with substances that occur naturally in real-world sampling scenarios.

The technology under development in CRIM-TRACK has the potential as an effective tool to control trafficking of illegal drugs, explosive detection, or in other law enforcement applications.

Keywords: artificial nose, colorimetric sensor, machine learning

1. INTRODUCTION

The aim of the CRIM-TRACK project[1] is to demonstrate a working sensing device that can be developed into a portable, miniaturized, automated, rapid, low cost, highly sensitive, and simple sniffer and detection unit, based on a disposable micro-colorimetric chip. The unit can be used for identification of drug precursors and home-made explosives. The project combines highly advanced disciplines, like organic chemistry, micro fabrication and hardware technology, machine learning and signal processing techniques. It will provide customs officers, police and other authorities with an effective tool to control trafficking of illegal and dangerous chemicals.

Further author information: Lasse L. Mølgaard: E-mail: llmo@dtu.dk, Telephone: +45 45253431

The main objective is to develop a miniaturized, "sniffer" system based on colorimetric sensor technology. The method utilizes chemo-selective compounds (dyes) to discover and record color changes of dyes in the presence of target molecules. It should give nearly real-time information about potential illegal substances through detection of their vapor with high specificity and sensitivity, low cost, without extensive user training. Target analytes include illicit drugs and their precursors, commercial and military explosives, improvised explosives and their precursors, and energetic materials in general.

We present results using the developed functional prototype version of the sniffer system applied in controlled experiments in laboratory environments.

This paper focuses on the detection of two illegal substances; phenylacetone also known as benzyl methyl ketone (BMK), which is an important precursor chemical and H_2O_2 selected as a precursor chemical in the preparation of improvised explosive such as TATP. We evaluate the detection performance for the two target analytes compared to confounder substances such as water, acetone, and gasoline that will occur naturally in real life application of the sniffer system. These disturbance substances are investigated to develop a robust detection method for our target substances in the diverse environments that the final sniffer system may be used in.

Colorimetric sensing technology has been used successfully for visualizing and detecting the presence of target molecules in complex backgrounds.[2,3] In our approach building an effective detection tool we apply a statistical classification framework to the sensor response to provide a simple yes/no answer concerning whether a target analyte is present in the air sample.

Statistical classifiers have been applied to colorimetric sensors with good results[4] e.g. using K-nearest neighbor classifiers. We show that we can provide a high detection rate by use of feature selection methods in combination statistical classification methods. We compare the use of probabilistic logistic regression, K-nearest neighbor classifier, and a random forest classifier. We demonstrate a high detection ability of the sensor to detect BMK and H_2O_2 molecules even in mixtures with confounding substances.

2. METHODS

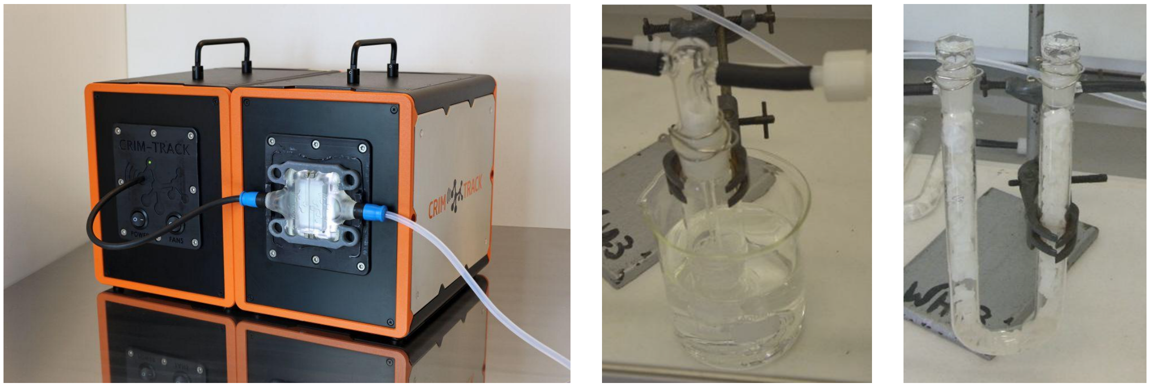

Figure 1. The apparatus used to lead air sample over colorimetric sensor chip and acquire image data.

2.1 Experimental procedure

The experiments were performed using the CRIM-TRACK prototype[1] shown in figure 1. The apparatus is equipped with an air sampling system that pumps an airflow through an exchangeable flow chamber, where the colorimetric chip is exposed to the air sample. The colorimetric chip uses a selection of 26 chemo-selective chemicals (dyes) that change color in presence of target molecules. These dyes are spotted on the chip in 15x15 spot pattern as shown in figure 2. The chips are single-use and are easily exchangeable.

Two separate experimental campaigns were carried out to investigate the detectability of the target substances: BMK and H_2O_2:

1) The first experiment concerned with the detection of BMK versus confounding substances acetone, diesel, gasoline, seawater, and water. The air samples sucked into the machine were generated using two methods depending on the volatility of the substances. Both methods use synthetic bottled air that is led through the target substance to produce saturated air samples. The first sample generation method leads the synthetic air through a glass gas-washing bottle containing the target analyte shown in figure 1. This method was used for the confounder substances. The second method used to generate BMK sample airflow utilized a glass U-tube containing polyester fleece soaked with BMK. The exact concentration of the samples in the airflow was unknown. We exposed our sensor to the resulting analyte air stream for 5 minutes. To evaluate the colorimetric response robustness and repeatability, 14 measurements of BMK, clean air, and water were performed. For each of the confounding substances 7 samples were recorded. Finally, air mixtures with BMK and each of the confounding substances were measured. Each of these mixtures were repeated 7 times. The dataset thereby includes 118 samples.

2) The second experiment was performed to test the detectability of H_2O_2 compared to water vapor. The experiment uses a 30% dilution of H_2O_2 and samples of water vapor to test the specificity of the sensor when distinguishing between water and H_2O_2. The experimental setup is analogous to the former where samples are generated using the bubbler setup. Each of the target analytes were measured 10 times. As a further refinement, the analyte airflow was mixed with a synthetic airflow to ”dilute” the concentration of the target analyte. In total 87 samples were included for analysis.

2.2 Data acquisition and processing

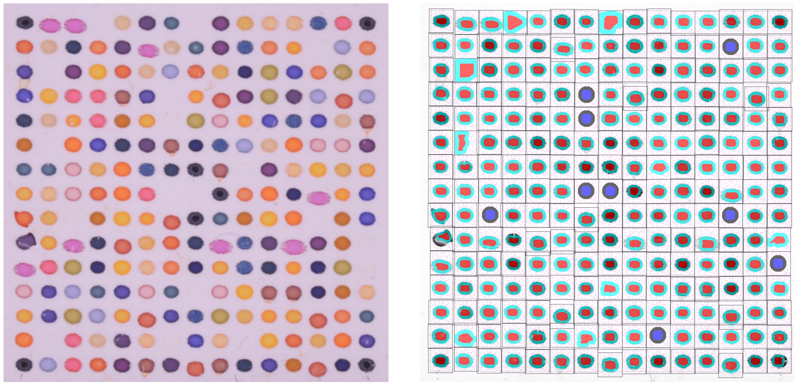

Figure 2. Left panel shows a sample of image of dyes acquired by the apparatus, and right panel shows the results of spot localization on the left image by the image processing pipeline. The software localizes each spot and the stable center of the spot is marked in red. The empty spaces (that were left because of a malfunctioning dye) on the chip are marked in blue and black by the software. Furthermore, the software also extracts the local background of spots for normalization purposes.

The CRIM-TRACK sniffer system[1] provides an integrated system for obtaining color images of the sensor chip. The color images are recorded as a red, green, and blue tuple (RGB). For each measurement the apparatus records one image before the chip is exposed to the air sample to be analyzed. After five minutes of exposure to the air stream, another image is recorded. Thus, each measurement constitutes an image pair. In order to measure color changes for each of the dye spots on the chip, an image processing pipeline was implemented that localizes each spot on the chip and aligns the two recorded images. It further, localizes each of the dye spots that consists of up to a few hundred pixels, which is reduced to a single color change value for each color channel by calculating the median value.

2.3 Color correction

The color values extracted in the procedure above will be influenced by the changing lighting intensity produced by the lighting in prototype and unevenness of the light source. The color change values are therefore normalized with the white background color of the chip in a region around the spot.

The design of the chip includes eight replicate spots of each dye. The color change for each dye is therefore calculated by calculating the median over the eight repetitions. This has proven to be an efficient way to make the measurements robust towards outlier values for single spots.

2.4 Data visualization

The color changes are visualized as absolute bar charts, which gives a compact representation of the color changes for all three color channels for each dye. Another representation of the color changes uses principal component analysis (PCA). PCA will find latent variables that are linear combinations of the measured color changes with maximum variance. The linear combinations are often denoted principal components.

2.5 Classification algorithms

Earlier work has applied different classification algorithms for colorimetric sensors.[5] The investigation showed that K-nearest neighbor (KNN) classifiers was be effective for the purpose, as well as logistic regression. In this work we have also tested a random forest classifier.

2.5.1 K-nearest neighbor

Despite its simplicity the Knearest neighbor is an effective classification technique,[6] which works as follows: when testing an unknown data point, the Euclidean distances for all known points are calculated. The classes of the closest K points are then identified and the unknown point is classified using majority voting of these known points. In the event of a tie, the algorithm uses the nearest neighbor among the tied classes to break the tie selecting the closest point as the class. All possible values of K are probed during model selection.

2.5.2 Logistic regression

The logistic regression model[6] estimates the posterior probability of a positive class given a measurement. The model is written as

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{\exp(\mathbf{w}_k^T\mathbf{x})}{1+\exp(\mathbf{w}_k^T\mathbf{x})} \tag{1}$$

where $p(\mathcal{C}_k|x)$ is the probability of the positive class $\mathcal{C}$ given a data point $\mathbf{x}$.

The two methods can be somewhat prone to overfit given the high dimensionality of our dataset. The input feature dimensionality to these algorithms is therefore reduced using the PCA method mentioned above. Tuning of parameters for the methods was performed by 10-fold cross validation procedure.

2.5.3 Random forest

The random forest classifier[7] is an example of an ensemble method that uses a collection of weak learning algorithms (decision trees) to achieve high classification performance. We employ a bootstrap aggregation method. The method draws a number of random bootstrap sample of the data points to train individual decision tree classifiers. To predict the class of a new sample, the predictions from all trees are averaged to get the overall prediction of the random forest classifier.

3. RESULTS AND DISCUSSION

3.1 H_2O_2 visualization

Figure 3 shows examples of the red channel color changes for the H_2O_2samples. Each panel in the figure shows the color changes for the red color channel for each dye. Each dot depicts the mean over all repetitions for each dilution level of the analyte. Each sample is shown with errorbars representing the 95% confidence interval of the mean. As an example the color change for the MISC17 dye are larger for H_2O_2 than for the water and air samples and this effect is significant for all dilution levels.

Visual inspection of these color changes indicates that MISC17, TAM4, TAM5, and TAM9 could be used to detect H_2O_2, however, remaining dyes might also have discriminative power and all will be included principled statistical produce described above in section 2.

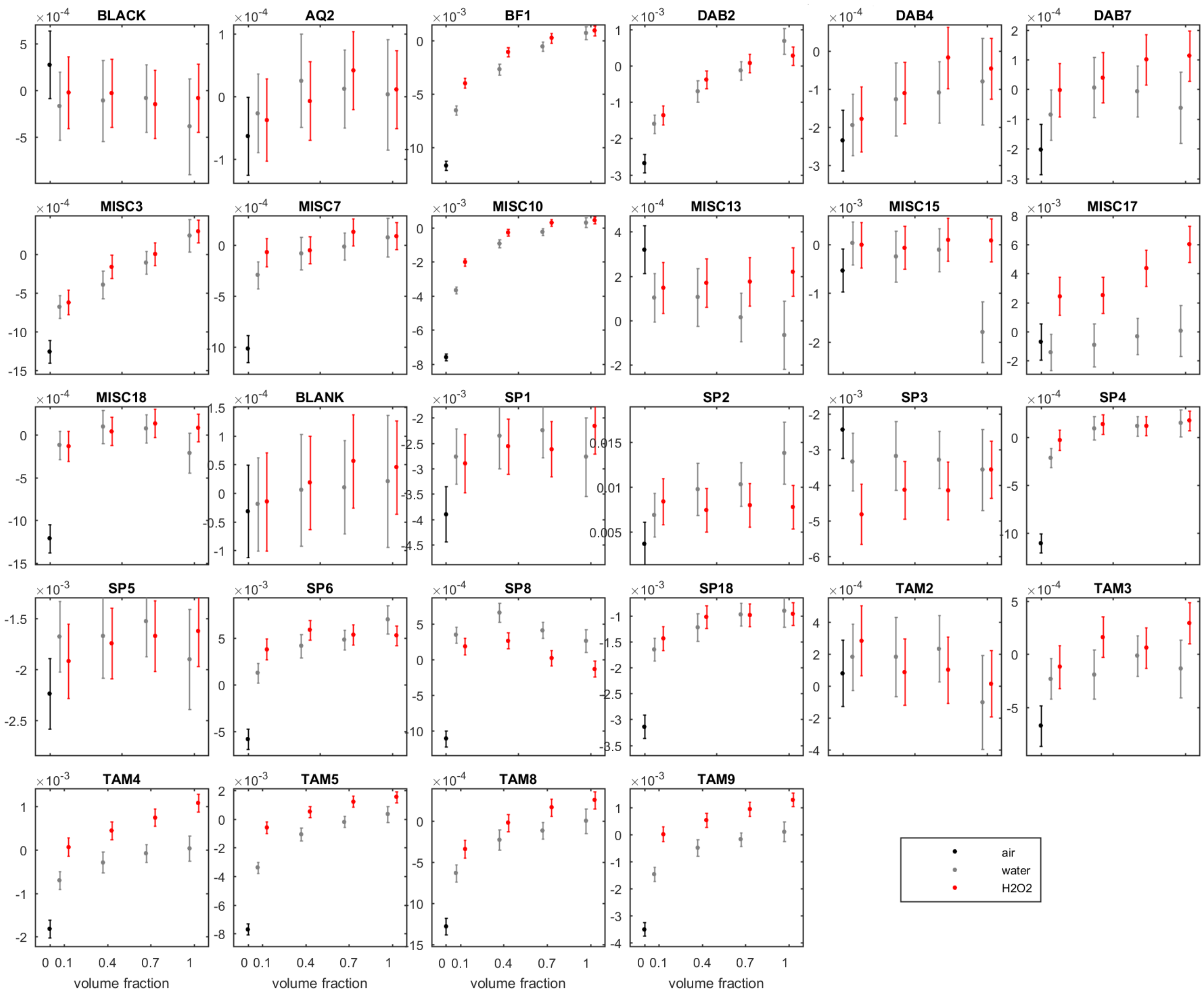

Figure 3. Color changes for the red channel for H_2O_2 experiments. Each dot depicts the mean over all repetitions for each experimental setup with errorbars showing the 95% confidence interval of the mean. The visual inspection of these color changes reveals that the dyes MISC17 and TAM4 can be used to separate H_2O_2 from the water samples for all dilution levels. Dye names are internally used shorthands.

Using the PCA method for dimensionality reduction on all color channels and dyes provides a mean of assessing the separability of the classes using the full information present in the color change values. Figure 4 shows that the combination dyes on the chip provides a good separation between Air and the H_2O_2 and water samples. Furthermore, there is also a good separation between H_2O_2 and water samples. As the samples are diluted with air to lower concentrations, it is evident that the color changes get closer to the clean air samples.

3.2 BMK visualization

Similarly, the PCA analysis of the BMK experiments is shown in figure 5. The left panel shows that the BMK samples are separable from the confounder substances. The first principal component clearly aligns with the separation of BMK from the rest of the analytes, while the second principal component discriminates between the confounder analytes. The right panel of Figure 5 presents the results of experiments which includes samples of mixture of BMK and confounders. These mixed samples confuse the general picture but the samples containing BMK are still distinguishable from the clean confounder samples.

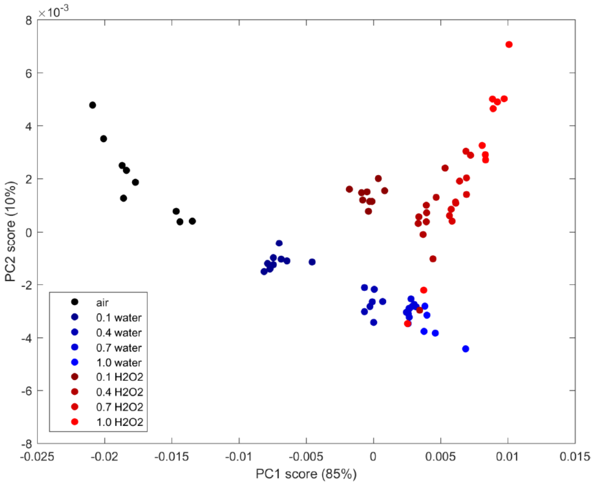

Figure 4. Principal component plots of H2O2 experiment results. The plot shows that the H_2O_2 samples are detectable. The diluted samples of both water and H_2O_2 are seen to become more similar to the air samples. However, even the lowest dilution level of the analytes can be separated from air and the overlap between water and H_2O_2 samples is small.

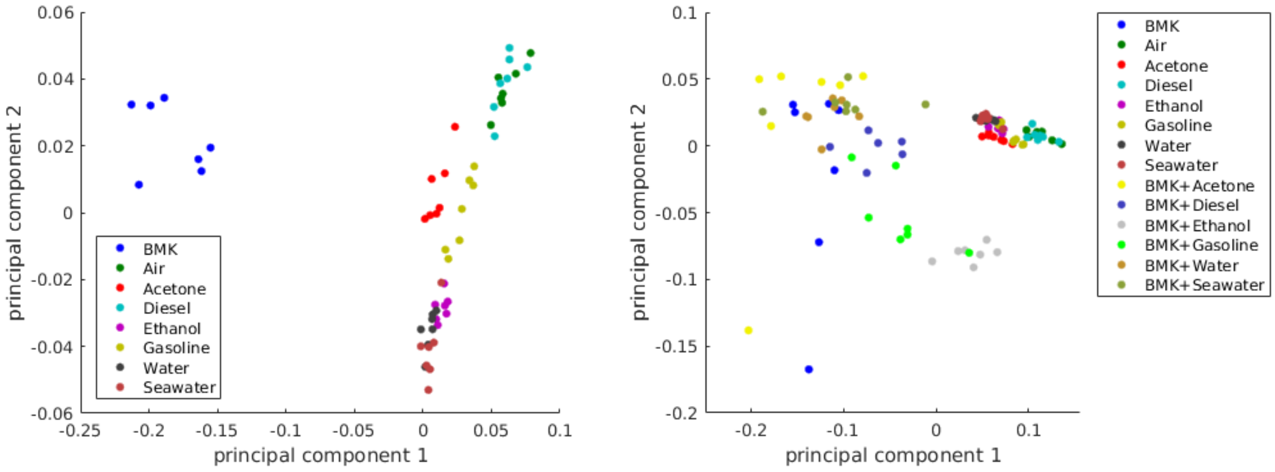

Figure 5. Principal component plots of BMK experiment results. Both plots show that the BMK samples are detectable. The addition of confounding substances to the airflow shown in the right plot changes the sensor response but the BMK samples are still separable.

3.3 Classification results

The three presented classifiers were trained on data from two experimental campaigns. The performance of the classifiers was evaluated using 10-fold cross validation, to get an estimate of the detection rate on unseen data. The detection rates are shown in table 1. The H_2O_2 dataset uses data for all dilution levels of the analytes. The BMK data are split into two datasets: one containing only samples where one analyte is present in the airflow, and one where all samples are included.

The KNN and random forest classifiers perform very well on all datasets, achieving high detection rates close to perfect classification. The logistic regression on the other hand performs much poorer, mainly because of a very high number of false positives.

The main goal of using the statistical classifier is to provide robust detection of the target analytes regardless of the possible confounding substances that could be present in the airflow. To test the ability to robustly detect

Table 1. Overall cross validation detection rate for the three evaluated classifiers.

	KNN	logist. reg.	RF
H_2O_2	94.3%	75%	95%
clean BMK	97.3%	53.2%	100%
mixed BMK	100%	71%	100%

Table 2. Confusion matrices for the three classifiers on the H_2O_2 dataset.

		KNN		**log.reg.**		**RF**	
		H_2O_2	confounder	H_2O_2	confounder	H_2O_2	confounder
True class	H_2O_2	36	4	39	1	38	2
	confounder	1	46	20	27	2	45

the target analyte, we train the classifier on the single analyte samples from the BMK experimental campaign, and then evaluate detection performance on the mixed samples. The random forest classifier detection rate for BMK in this setup is 86% while the KNN classifier detection rate is 53%. The significant drop in performance for the KNN indicates that this classifier has overfitted to the training data. The overfitting may stem from the fact that the KNN classifier uses PCA-based feature extraction that sometimes can be influenced by variance inflation.[8]

The performance drop for the random forest classifier is not as severe but it is clear that the colorimetric sensor response is not completely selective towards the BMK molecules. On the other hand the signal for BMK is still detectable in mixtures with the confounders as shown with the high detection rates achievable shown in table 1.

4. CONCLUSION

We present a data-driven machine learning approach to detect drug- and explosives-precursors using colorimetric sensor technology for air-sampling.

The use of the CRIM-TRACK fully-integrated prototype for air sampling has enabled the generation of standardized data for the target analytes H_2O_2 and BMK. The experiments have demonstrated the possibility of detecting the target analytes in complex mixtures of confounding substances that will occur in real-world applications, e.g. when sampling luggage or containers in a harbor or airport setting.

We have evaluated three statistical classification algorithms: K-nearest neighbors, logistic regression, and random forest classifiers. The evaluation showed that the random forest classifier generally performs best and is also able to detect the target analytes when the air sample is ”polluted” with confounder substances.

The technology under development in CRIM-TRACK has the potential as an effective tool to control trafficking of illegal drugs, explosive detection, or in other law enforcement applications.

ACKNOWLEDGMENTS

The CRIM-TRACK project is funded by EU FP7 Grant Agreement Number 313202.

REFERENCES

[1] Munk, J. K., Buus, O. T., Larsen, J., Dossi, E., Tatlow, S., and Jakobsen, M. H., "CRIM-TRACK: Sensor system for detection of criminal chemical substances," in [*Proc. SPIE*], **9652**, 965208–965208–5 (2015).

[2] Janzen, M. C., Ponder, J. B., Bailey, D. P., Ingison, C. K., and Suslick, K. S., "Colorimetric sensor arrays for volatile organic compounds," *Analytical Chemistry* **78**(11), 3591–3600 (2006). PMID: 16737212.

[3] Kostesha, N. V., Alstrøm, T. S., Johnsen, C., Nilesen, K. A., Jeppesen, J. O., Larsen, J., Jakobsen, M. H., and Boisen, A., "Development of the colorimetric sensor array for detection of explosives and volatile organic compounds in air," (2010).

[4] Alstrøm, T. and Larsen, J., *Assessing Miniaturized Sensor Performance using Supervised Learning, with Application to Drug and Explosive Detection*, PhD thesis (2013).

[5] Alstrøm, T. S., Raich, R., Kostesha, N. V., and Larsen, J., "Feature extraction using distribution representation for colorimetric sensor arrays used as explosives detectors," in [*2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*], 2125–2128 (March 2012).

[6] Bishop, C. M., [*Pattern Recognition and Machine Learning*], Springer, Secaucus, NJ, USA (2006).

[7] Breiman, L., "Random Forests," *Machine Learning* **45**(1), 5–32 (2001).

[8] Kjems, U., Hansen, L. K., and Strother, S. C., "Generalizable singular value decomposition for ill-posed datasets," in [*Advances in Neural Information Processing Systems*], 549–555 (2001).