

Received 16 July 2024, accepted 23 August 2024, date of publication 2 September 2024, date of current version 23 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3453496

RESEARCH ARTICLE

Classification of Volatile Organic Compounds by Differential Mobility Spectrometry Based on Continuity of Alpha Curves

ANTON RAUHAMERI¹, ANGELO ROBIÑOS², OSMO ANTALAINEN³, TIMO SALPAVAARA¹, JUSSI RANTALA⁴, VEIKKO SURAKKA⁴, PASI KALLIO¹, (Member, IEEE), ANTTI VEKHOJA¹, AND PHILIPP MÜLLER¹

¹Faculty of Medicine and Health Technology, Tampere University, 33720 Tampere, Finland

²Laboratory of Molecular Science and Engineering, Åbo Akademi University, 20500 Turku, Finland

³Olfactomics Oy, 33720 Tampere, Finland

⁴Faculty of Information Technology and Communication Sciences, Tampere University, 33720 Tampere, Finland

Corresponding author: Anton Rauhameri (anton.rauhameri@tuni.fi)

This work was supported in part by the Academy of Finland under Grant 323498, Grant 323529, and Grant 323530; and in part by the Suomen Kulttuurirahasto under Grant 50221583.

ABSTRACT Classification of volatile organic compounds (VOCs) is of interest in many fields. Examples include but are not limited to medicine, detection of explosives, and food quality control. Measurements collected with so-called electronic noses can be used for classification and analysis of VOCs. One type of electronic noses that has seen considerable development in recent years is Differential Mobility Spectrometry (DMS). DMS yields measurements that are visualized as dispersion plots that contain traces, also known as alpha curves. Current methods used for analyzing DMS dispersion plots do not usually utilize the information stored in the continuity of these traces, which suggests that alternative approaches should be investigated. In this work, for the first time, dispersion plots were interpreted as a series of measurements evolving sequentially. Thus, it was hypothesized that time-series classification algorithms can be effective for classification and analysis of dispersion plots. An extensive dataset of 900 dispersion plots for five chemicals measured at five flow rates and two concentrations was collected. The data was used to analyze the classification performance of six algorithms. The highest classification accuracy of 88% was achieved by a Long-Short Term Memory neural network, which supports the hypothesis that interpreting DMS measurements as sequential data is beneficial and outperformed classification algorithms traditionally used for DMS-based VOC identification.

INDEX TERMS Classification, differential mobility spectrometry, long-short term memory, machine learning, neural networks.

I. INTRODUCTION

Classification of scents is of interest in many fields. Notable application areas include but are not limited to medicine, food quality control, detection of warfare agents, and digitalization of scents (see e.g. [1] and references therein). For example, in the medical field, scents emitted by evaporated tissue can be used for distinguishing pathological and healthy tissue [2], [3] and identifying the tumor type during surgical

operation [4]. Likewise, in food quality control scents can be used for checking the maturity level of fruits and vegetables or ensuring freshness of meat products [5]. Accurate classification and analysis of scents will open new possibilities for, for example, diagnosing cancer patients [6], and digitalizing and transmitting scents over the internet [7].

An emerging field for applying classification of scents relates to extended reality (XR) research and development. XR applications aim at the creation of digital twins of the real world with which humans can interact in the same way as in a real world [8]. For such developments measurement and

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li¹.

analysis of the outside world is important, for example, for remote work and operations. Measurement and classification of real-world scents is needed in order to be able to reproduce them with the help of an olfactory display (OD) in virtual reality (e.g., [9], [10], [11]). The present work is part of a more extensive project focusing on sensing and reproducing scents to virtual reality. We are developing scent classifications, olfactory displays, and virtual reality (VR) environments. This paper focuses on improving classification of five scents to enable us using the full potential of the five channels in our OD prototype.

Various approaches for sensing scents exist that often rely on detecting and analyzing volatile organic compounds (VOC). VOCs in air can be separated and measured due to their different physical properties. One of the separation methods for VOCs is differential mobility spectrometry (DMS). Several manufacturers produce commercial DMS devices. In this study the IonVision from Olfactomics [12] was used. The sample in the IonVision is ionized with 4.9 keV photons from an x-ray source in a sequence of atmospheric pressure reactions. The sequence is understood to start from electron extraction from N_2 and O_2 molecules and interaction between neutral molecules in air, leading quickly to the formation of protonated water clusters $(H_2O)_nH^+$ called positive reactant ions (RIP), and $O_2^-(H_2O)_n$ called negative reactant ions (RIN) [13]. When the reactant ions collide with neutral analyte molecules, product ions are generated. The formation of product ions depends on the concentration of reactant ions and neutral analyte, as well as the proton affinity of the reactants. The higher the proton affinity, the more probable is the formation of such product ions. The ionized molecules then enter a drift region, which acts as a filter. In this region, ions are carried by neutral gas flow and are separated in an asymmetrically oscillating electric field based on their field dependence. The passband of the DMS ion filter is controlled by a separation field which is defined by an amplitude of oscillating voltage, and compensation field, which is defined by the DC component added to the oscillating field. The ions that pass the filter are detected in Faraday plate detector following immediately after the drift region. The DMS filter operates by scanning both separation and compensation fields. DMS is often used as a sample pre-separation method in mass spectrometer analysis or as a detector in Gas Chromatography Spectroscopy (see e.g. [14] and [15]).

However, DMS can also be used independently for VOC analysis [16]. A measurement of DMS is often represented as a matrix of numbers, where each row represents measurements with a fixed separation voltage U_{sv} and varying compensation voltages U_{cv} . When visualized, the matrix is commonly referred to as a dispersion plot. It is important to note that ion mobility spectrometry, including DMS, is an experimental technique and without prior testing it is impossible to infer from the dispersion plot what chemical was measured. Instead, prior knowledge of the sample as well as a pre-trained classification algorithm are necessary

to identify the measured chemical with the dispersion plot as input. Various machine learning methods have been used recently for the classification of chemicals based on dispersion plots. Lepomäki et al. [16] used shrinkage linear discriminant analysis (sLDA), support vector machines (SVM), and a convolutional neural network (CNN) to differentiate between five porcine tissue types introduced to DMS through laser desorption. The classification accuracies were 79.8% (sLDA), 79.0% (SVM), and 86.4% (CNN). The authors also demonstrated that the mean and standard deviation of the dispersion plots for each type of tissue had noticeable visual differences in intensity. Hence, the authors already expected high classification accuracy for CNN. However, in general, dispersion plots of different VOCs cannot be distinguished by visual inspection alone, making the accurate classification more challenging than what was demonstrated in [16].

In a recent study conducted by Haapala et. al. [4], linear discriminant analysis (LDA) achieved an average accuracy of 85% in binary classification of isocitrate dehydrogenase (IDH) mutation in gliomas based on dispersion plots. However, their dataset consisted of only 352 samples, which made it challenging to apply more complex classification algorithms. With such limited data, it is difficult to develop a neural network that provides accurate classification.

There are many other works trying to classify or analyze DMS dispersion plots with the help of machine learning algorithms. For example, [17] explored a comparative analysis of neural network platforms by classifying VOCs with tandem DMS. In their work, the protonated monomers were isolated by applying a strong electric field and analyzed at the second DMS filter aiming to create fragments, and using the fragments as basis for neural network analysis. Although, the setup differs significantly from the one presented in this paper, the classification accuracy obtained with the neural network developed by the authors was over 90% for familiar compounds and 64% for unfamiliar. As opposed to work presented in [17], only one DMS filter was used in the current work. Thus, our setup is simpler and produces mostly native product ions and not field induced fragments. Additionally, using the tandem DMS for that specific application suggests that the authors of [17] had some prior information about the samples being analyzed.

Another work by [18] aimed at predicting dispersion curves yielded by DMS. The authors used a Random Forest regression algorithm, and achieved high accuracies¹ for different types of alpha curves. Their setup and sample preparation methods, however, differed significantly from the work presented in this paper. In [18] DMS was only used as a filter before measurements were taken by a mass spectrometer. Furthermore, nitrogen was used as carrier gas, which reduces contamination compared to ambient air as carrier gas due to lower trace chemicals content in nitrogen. The disadvantage is, however, that using nitrogen requires

¹ Authors measured accuracy by the Mean Absolute Error (MAE).

higher electric fields. Finally, the authors knew in advance mass-to-charge ratio (m/z) and collision cross-section (CCS), which simplifies the prediction task considerably.

Rajakpse et al. [19] demonstrated a promising method for classification of dispersion plots with partial least squares - discriminant analysis. They showed that the first two principal components represented over 65% of variability in the dispersion plots and are able to reveal clear clusters of chemicals. However, compared to the work presented in this paper, the samples in [19] were prepared using pure nitrogen as dilution gas.

The works mentioned above, and many other, demonstrate development of DMS methods and high interest of algorithmic analysis of the measurement results. Different research groups use various DMS techniques for obtaining reliable results of VOC analysis. However, all these methods differ significantly in their setups, sample preparations, etc. This work is aimed to present a classification algorithm applied to a simple setup with weakly controlled environmental conditions. Additionally, the dataset is published to enable interested readers to compare their methods with the methods presented in this work.

DMS data, in general, have high dimensionality. Measurements for tens or even hundreds of different separation and compensation voltages can be collected for only a single dispersion plot. Therefore, using regularization [20] and principal component analysis [21] to avoid the curse of dimensionality is suggested. In this paper, we explore an alternative approach to address the curse of dimensionality. Our hypothesis is that dispersion plots can be interpreted as a set of sequential measurements. In order to verify our hypothesis, a new dataset containing five chemical solutions measured at five carefully controlled flowrates was collected. This was done in order to produce variation and generate multilabel data. Based on the hypothesis, for example, a 40-by-200 matrix can be interpreted as sequence with 40 measurements,² where each measurement is of dimension 200. This means that at each step, values for 200 features are available. As a result, there is no need to apply principal component analysis (PCA) and lose the information preserved between measurement points. However, ion separation is independent of the order of separation and compensation field sequences, and the sequence could be run in random order or using only a limited amount of U_{sv} - U_{cv} pairs to save data acquisition time. For convenience of configuration, parameters are usually scanned in increasing order.

The contribution of this article is threefold. First, new results are presented for classifiers that at their best can distinguish between the different chemical dilutions measured at different flow rates with an accuracy of 89% on the collected data set. To our knowledge, there are no carefully controlled studies that reported accuracies of 89% or higher for multilabel machine learning problems using DMS

²Which means each row is one measurement of 200 features and each column a sequence of 40 measurements for one feature

measurements. Second, it is demonstrated how time-series deep learning networks (Long-Short Term Memory, LSTM) can be applied to DMS data, and that they outperform other well-known classifiers. Third, the collected data set is shared openly to enable anyone interested to repeat our tests and compare their own classifiers with the algorithms proposed in this paper. In section II the data gathering process is described. Section III describes the collected data set. Section IV briefly describes algorithms used for classification of DMS dispersion plots. Section V discusses results of the applied methods. In Section VI, an outlook and plans for further improvements are discussed.

II. DATA ACQUISITION

As discussed in Section I, our current olfactory display prototype can currently emit mixtures of up to five chemicals, thus, five chemicals were selected that each on their own have a distinct scent that could be used in a VR without mixing it with other chemicals. For example, ethyl 2-methylbutyrate and (-)-Carvone smell like strawberry and mint respectively. Thus, the analysis in this paper was limited to five chemicals.

A. MATERIALS

Carvone(+,-) ($C_{10}H_{14}O$, 98%), ethyl 2-methylbutyrate ($C_7H_{14}O_2$, 99%), methyl cyclopentenolone ($C_6H_8O_2$, $\geq 98.0\%$), 2-phenylethanol ($C_8H_{10}O$, $\geq 99.0\%$), n-butanol ($C_4H_{10}O$, 99.9%), and propylene glycol ($C_3H_8O_2$, $\geq 99.5\%$) from Merck were used as stimuli for DMS measurements. Except for methyl cyclopentenolone, which is solid, all chemicals were in liquid form. n-Butanol served as the reference substance while propylene glycol was used to dilute the concentrated scents. All chemicals used in this work are presented in Table 1.

B. SAMPLE PREPARATION AND MEASUREMENT

The experiment utilized containers and apparatus that were thoroughly cleaned to prevent contamination of the samples. Solutions of percent volume (%v/v) and percent weight (%w/w) concentrations were prepared from liquid and solid substances, respectively. In the case of undissolved solid chemicals, the flasks were sonicated until virtually no solids remained. Two concentration levels per chemical were then prepared as samples: 1/100 (1% v/v or w/w) for a stronger and 1/10 000 (0.01 % v/v or w/w) for a milder scent.

A custom-built odor display [22] was used to vaporize the sample solutions, to adjust the intensity of the odor gas by mixing it with clean air and to deliver the diluted odor flow to the DMS. In short, filtered air was pumped into two channels (odor and dilution air) and the air flows were adjusted using two mass flow controllers (MFCs). A scented airflow was created by passing a known volume flow of air through an odor source. This channel could be combined with the dilution air flow (five standard liters per minute (SLPM)) with a three-way valve. The DMS was connected in parallel to the output of the system to take in samples of scented airflow.

A glass vial filled with 2 mm diameter glass beads and odorant solution as in Figure 1 acted as odor source. The beads were added to ensure that the air bubbles travelling from the bottom of the vial spend an equal amount of time soaking up the analyte from the solution before measurement. Schematically this strategy aimed to improve the repeatability of the results.

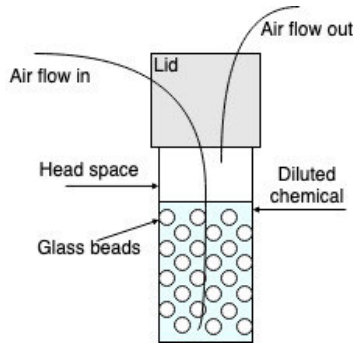


FIGURE 1. Schematic representation of chemical put into the vial.

The samples were measured in random sequence, with increasing flowrates, to account for instrument drift. During one day five sets of chemicals were measured. Each set contained one chemical measured at 8, 16, 32, 64 and 128 sccm flow rates mixed with room air. Thus, each chemical at a certain flow rate was measured once per day. The procedure was repeated on 35 days. In total 900 samples of five chemicals measured with five different flow rates at two different concentrations were collected. Carvone was measured for one additional session due to suspicion that measurements of one session were faulty. However, both measurement sets of Carvone were included in the final dataset since our suspicion was disproved. Propylene glycol was also measured in every five sets of samples to monitor the baseline and to check for any cross-contamination. Each set of chemicals was repeatedly measured for five days to further account for within-day fluctuations in the odor display, temperature and humidity, which could affect concentration. For a more detailed schedule of the measurements, the reader is referred to Section 3 in the supplementary material.

III. DATA DESCRIPTION

A dispersion plot (two examples are shown in Figure 2) represents a matrix of size $U_{sv} \times U_{cv}$, where rows represent compensation voltages for a fixed separation voltage U_{sv} and columns represent separation voltages for a fixed compensation voltage U_{cv} . For this study, ranges for compensation and separation voltages that would optimize classification accuracy were unknown in advance. Hence, the default values proposed by Olfactomics for their IonVision DMS device were used, meaning that separation voltage ranged from 200 V to 700 V, and compensation voltage ranged from -1 V to 9 V. The frequency of the separation voltage was 1 MHz with 20% duty cycle and rectangular waveform. The gap

width of the DMS sensor³ used was 0.25 mm, leading to a high electric field strength of $(D - 1)U_{sv} d^{-1}$ and a low field strength of $-D U_{sv} d^{-1}$, where D is the duty cycle.⁴ Duty cycle in IonVision is tunable, and here it was set to 22%.

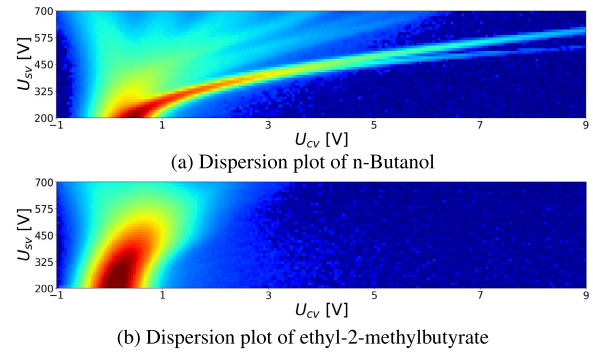


FIGURE 2. Examples of dispersion plots of two chemicals, n-Butanol and ethyl-2-methylbutyrate showing positive ions, measured by Olfactomics IonVision device. For better visualization cubic root was applied to the raw values. The typical pattern in dispersion plots of all chemicals, except those of ethyl-2-methylbutyrate, is shown in (a). Ethyl-2-methylbutyrate patterns in dispersion plots resembled subfigure (b).

In all measurements, of all chemicals, drift in the DMS device output ion current values (referred later as intensity) was observed. Figure 3 shows the drift of the measurements ordered by time. The spikes in Figure 3a are measurements of ethyl-2-methylbutyrate.

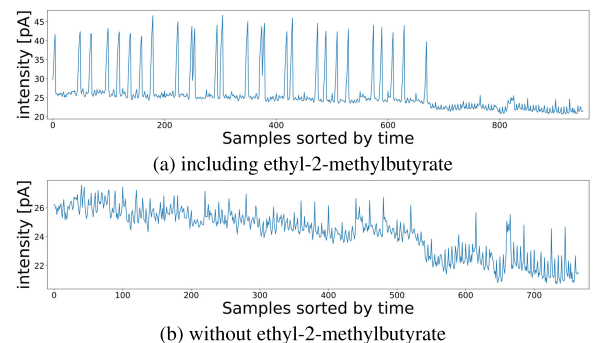


FIGURE 3. Drift of the average intensity over all compensation voltages at $U_{sv} = 200$ V.

Visual inspection revealed no noticeable differences amongst dispersion plots of the different chemicals, with exception of ethyl-2-methylbutyrate (E2MB). Dispersion plots of E2MB varied considerably from each other and Figure 2b shows one example. Varying environmental conditions on different days prevented the dispersion plots of one chemical measured multiple times at the same concentration and flow rate to be identical. Figure 4 displays the humidity level for nBuOH during the measurement campaign. As can be seen, the humidity level trended during the measurement campaign. Changes in humidity affect DMS [24] as they

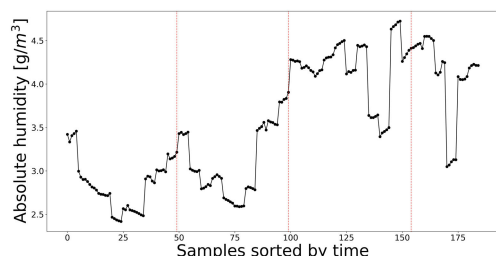
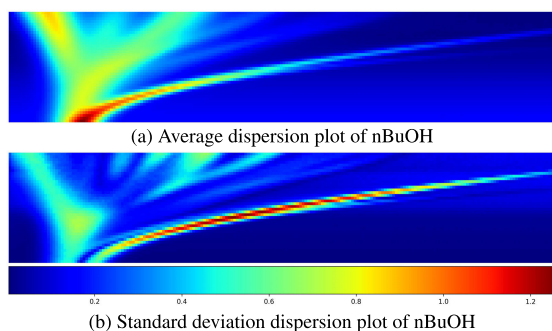
³Increasing the gap decreases the electric field. Here a gap of 0.25 mm was used to obtain a high electrical field with practical low electrical potential.

⁴The low and high field strengths are thus 1 600 V/cm and 22 400 V/cm respectively.

TABLE 1. List of used chemicals, their abbreviations and Chemical Abstract Service (CAS) Registry Numbers. The full names of chemicals used in this study are obtained from [23].

Chemical	Number of samples	Scent	Full name	Abbreviation	CAS	Purity %
$C_{10}H_{14}O$	200	strong minty	Carvone, (+,-)-	Carvone	6485-40-1	≥ 98.0
$C_7H_{14}O_2$	175	strong strawberry	ethyl 2-methylbutyrate	E2MB	7452-79-1	≥ 99.0
$C_6H_8O_2$	175	sugary/nutty	methyl cyclopentenolone	MCP	80-71-7	≥ 98.0
$C_8H_{10}O$	175	rosy	2-phenylethanol	2PEtOH	60-12-8	≥ 99.0
$C_4H_{10}O$	175	moderate alcoholic	n-butanol	nBuOH	71-36-3	≥ 99.9

change the mobility of ionized molecules. In this study, the humidity ranged from 2.6 g/m^3 (3350 ppm_v) to 4.5 g/m^3 (6050 ppm_v). Based on Figure 5 and Table 2 in [25] proton affinity varies between 1040 and 1090 kJ/mol. Although proton affinity of water clusters was high, which affects sensitivity of the DMS, change of proton affinity was low, and, thus the change in sensitivity can be neglected. Standard deviations calculated for a set of dispersion plots of nBuOH measured with the same flow rate revealed combinations of compensation and separation voltages for which DMS measurements differed considerably (see Figure 5). For generating Figure 5 the dispersion plots of nBuOH of the same flow rate were stacked and pixel-wise average response values and corresponding standard deviations were calculated. The dispersion plots were normalized for better visualization. For average and standard deviation dispersion plots of other chemicals and flow rates readers are referred to Section 5 in the supplementary material.

**FIGURE 4.** Moisture level of nBuOH during measurements. The red dotted lines depict moisture filter maintenance.**FIGURE 5.** Dispersion plots showing (a) average appearance and (b) dissimilarities of nBuOH with the flowrate 32.

IV. CLASSIFICATION METHODS

In this section the preprocessing techniques as well as the six algorithms for DMS-based classification are explained. Section 3 of the supplementary material lists for each classifier its strengths and weaknesses. All presented algorithms

contain several hyperparameters. For those hyperparameters that, in the authors' experience, have the largest impact on performance various values were tested using the GridSearch algorithm. The best values from this search are also presented in Section 3 of the supplementary material. For the remaining hyperparameters their default values, specified in the Scikit-learn documentation, were used.

A. PREPROCESSING

All algorithms described in this section were implemented in Python using libraries scikit-learn and Keras. In order to provide stable accuracy estimation, repeated stratified k -fold cross-validation (RSCV) was used. The "stratified" in this abbreviation indicates that a balanced ratio of classes in training and test sets was maintained, while "repeated" indicates that each fold was repeated N times, and the average result was returned. The parameter k defines the number of folds in the dataset and is different from the K used in the K -Nearest Neighbors (KNN). In this study $k = 10$ was used.

All dispersion plots in the collected dataset were scanned with a wide range of U_{sv} and U_{cv} because no prior information was available on the ranges in which the dispersion plot would show reactions for the various chemicals. The U_{sv} ranged from 200 V to 700 V with a step size of 12.82 V, and U_{cv} ranged from -1 V to 9 V with a step size of 0.05 V. Therefore 200 different compensation voltages and 40 different separation voltages were checked, resulting in dispersion plots of dimension 40-by-200. Consequently, the dispersion plots contain much redundant information. Anttalainen et al. [26] described the different parts of a dispersion plot and proposed a method for determining the part of the dispersion plot in which separation of ions can be observed. The authors also stated that row-wise Shannon entropy measure can be used to find the optimal section of a dispersion plot with a good signal-to-noise ratio. While low entropy values indicate that a row contains mainly noise, high entropy values indicate that a row contains a meaningful signal [26].

From Figure 2a it can be seen that the branching starts at high U_{sv} and low U_{cv} indicating that separation of ions has started. The analysis with Shannon entropy on a small selection of dispersion plots confirmed it and therefore dispersion plots were clipped in a first preprocessing step to ranges of $U_{sv} = [443.59 \text{ V}, 700.00 \text{ V}]$ and $U_{cv} = [-1.00 \text{ V}, 4.03 \text{ V}]$. The resulting shape of the clipped dispersion plots was 20-by-100. Figure 6a shows the clipped dispersion plot (indicated by a white rectangle) for a dispersion plot

of 2PEtOH. Figure 6b shows row-wise entropy values for the same dispersion plot, with the green area on the plot indicating the rows with highest entropy. Thus, the top-left part of the dispersion plots was selected for further analysis, except for the Convolutional Neural Network (CNN). For CNN the dispersion plots were not clipped due to using convolutional kernels coupled with maxpooling layers that reduce drastically the size as the image moves through the network. The convolutional layers are used for capturing features of the images and the maxpooling layers are used to impose shift invariance of the scene. Thus we ensured that possible shift in the dispersion plots will be captured by the neural network. Additionally, after the first pair of convolutional and maxpooling layers the next convolutional layer can capture surroundings of a pixel where the filter was put. This means that clipping the dispersion plots prior to processing in the CNN is not beneficial.

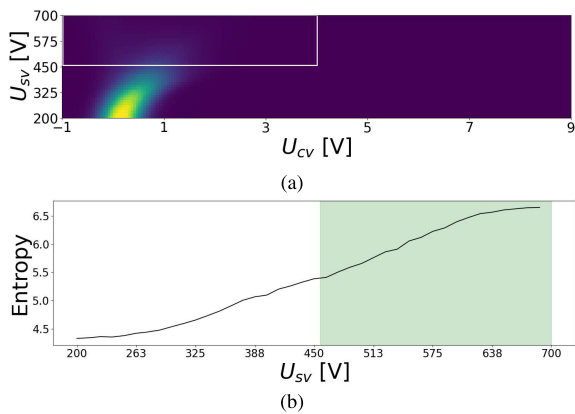


FIGURE 6. Reducing area of the dispersion plot used for analysis. Full measured dispersion plot of 2-phenylethanol is shown in (a). Row-wise entropy values of that dispersion plot are given in (b). The area filled with green shows the highest entropy. Based on (b) the part of the dispersion plot marked by the white rectangle in the upper left corner of (a) is selected for further processing.

The second preprocessing step was normalization of the data to ensure that all values were between -1 and 1 . Since each dispersion plot represents a series of measurements with different separation voltages, the dispersion plots were normalized row-wise by subtracting the mean and dividing by the standard deviation. Originally, this method was proposed in [27] but there rows were scaled to values between 0 and 1 .

The data was compressed with PCA before applying KNN, LDA, ExtraTrees Classifier (ETC), and multilayer perceptron (MLP). PCA is a feature transformation technique that can be defined as the orthogonal projection of the data onto a lower dimensional linear space, known as the principal subspace, such that the variance of the projected data is maximized [28, p.561]. In our tests applying PCA to the full data set resulted in the first 25 orthogonal principal components, out of the 2 000 features in the data,⁵ explaining

⁵Originally, a dispersion plot has the dimensionality of 8 000. After clipping the dispersion plot with the method described previously, its dimensionality reduces to 2 000.

over 99% of the variability in the data (see Figure 7). PCA was used because applying the aforementioned classification algorithms to the uncompressed data resulted in significantly worse results. After normalization and PCA the data was ready for classification by the algorithms described in the following subsections.

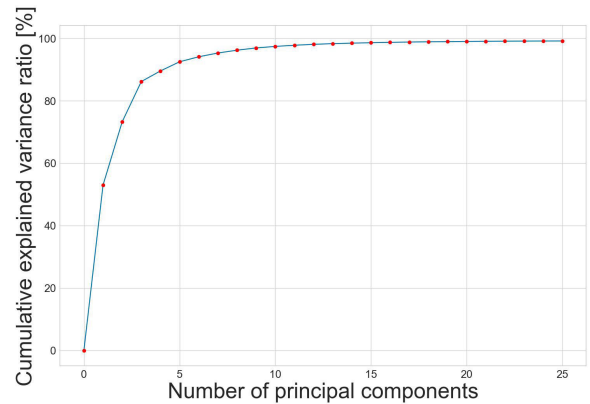


FIGURE 7. Ratio of cumulative variance in original data explained by different numbers of principal components.

B. EXTRATREES CLASSIFIER

The ExtraTrees Classifier is an extension of the popular Random Forests classifier, which in turn is an ensemble of the decision tree (DT) algorithm [29]. The main idea behind the DT algorithm is to construct a binary tree for the feature space of a dataset. However, the DT algorithm suffers from high training variance and overfitting. The ensemble of trees in the Random Forest Classifier is used to overcome these problems by generating a large, random set of such decision trees (hence named a random forest) and averages over their predictions to obtain the final prediction [30, p. 344]. Trees are chosen such that they fit the underlying dataset optimally. The ETC, in contrast, aims at high variation amongst the decision trees even if it requires sacrificing accuracy of fit. Random Forest-type algorithms are often treated as black-box methods [31].

In this work, the ETC algorithm from the scikit-learn library [32] for Python was used with two parameters: $n_estimators$ and $criterion$. The parameter $n_estimators$ specifies the number of trees in the ensemble. Choosing this parameter value is noncritical, and generally, a large number of trees is selected [30, p. 341] with the default value being 100. The $criterion$ parameter accepts three values, namely $gini$, $entropy$ and log_loss , and is required for defining decision nodes in the tree. For our application either $gini$ or $entropy$ would be suitable choices, with no considerable difference between them except their computational complexity [33]. Thus, in this work $gini$ was used for splitting the nodes. For the test, the optimal parameter set was selected using the grid search technique. The grid search technique is a method for finding the best combination of parameters for a classifier by simply iterating over all combinations

of the predefined parameter grid. By means of the Out-Of-Bag score, it was found that 100 estimators were enough to achieve a cross-validated accuracy of 80%. The remaining parameters were set to default values given in the scikit-learn implementation.

C. K NEAREST NEIGHBORS CLASSIFIER

The K Nearest Neighbors classifier is effective for many machine learning tasks. Despite its simplicity, this algorithm has performed well in many different problems. One example of applying KNN to ion mobility spectrometry data is described in [34]. The idea behind this algorithm is to find K training points x_1, \dots, x_K with known labels that are closest with respect to a distance measure to a query point x_0 without label. Based on the labels of the K neighbors a label for the query point is derived using majority vote [35, p. 463]. Generally, this algorithm has two main parameters: the number of nearest neighbors K used to vote the predicted class and the distance metric. For the tests in this paper the Euclidean distance, being the default distance measure, was chosen, despite being susceptible to the curse of dimensionality for high-dimensional data. This choice was motivated by lack of knowledge on the suitability of alternative distance measures. By using PCA-transformed data the number of features was reduced from 2000 to 25. Since our dataset contains 900 samples, 25-dimensional data is not considered high-dimensional data [36, p.XV]. The problem of selecting the right number of closest neighbors is known as the model selection problem [37, p.15]. Selecting only one voting neighbor can result in overfitting [37, p.15]. In this paper, $K=3$ was selected based on the grid search algorithm.

Drawbacks of the KNN classifier include large memory requirement and its slowness for databases with a large number of training points if distances between query point and all labeled points are calculated, and its tendency to focus on irrelevant features. The latter weakness can be mitigated by applying a feature transformation, such as PCA, or feature selection method to the dataset before running the KNN on it. For tackling the slowness prestructuring, editing the data, or computing only partial distances are valid options [35]. One prestructuring technique that has been successfully used in [34] for classification of scents based on ion mobility spectrometry are k-dimensional trees [38], where k is the depth of the tree and should not to be confused with the K from the KNN. The idea of this technique is to split the dataset into subsets by generating a binary tree. A new unlabeled sample is then moved along this tree until it reaches a leaf node, i.e. a subset. Its label is then determined based on the K nearest neighbors that are within this subset.

D. LINEAR DISCRIMINANT ANALYSIS

Linear discriminant analysis has demonstrated its effectiveness in many problems, including classification based on DMS dispersion plots. For example, it has shown its ability to differentiate between two types of porcine tissue

from surgical smoke measured by DMS with classification accuracy close to 93% [3]. The idea behind LDA is to find a decision boundary between classes of samples such that inter-class variability is maximized and intra-class variability is minimized. The advantages of LDA are its simplicity, that it provides a closed-form solution, and, its small number of hyperparameters. The two most important hyperparameter are *solver* and *shrinkage*. The solver parameter was set to *lsqr*, which stands for Least Squares. The *shrinkage* parameter is a form of regularization that improves estimation of covariance matrices. In this work the shrinkage-parameter was set to *auto* to determine an optimal shrinkage parameter analytically according to the approach proposed in [39].

E. MULTILAYER PERCEPTRON

Multilayer perceptron is a non-linear machine learning algorithm that has the ability to approximate decision boundaries in case of complicated and highly non-linear problems [28, p.225]. In this paper, MLP is used as a baseline classifier for comparison with more sophisticated neural networks that are described in the next subsections. MLP is an extension of the ordinary binary perceptron algorithm. The idea behind the perceptron algorithm is to weigh inputs and pass the weighted inputs into a nonlinear function. An MLP consists of several layers of perceptrons, and is often called Artificial Neural Network. Usually, an MLP has at least three layers: input, hidden, and output layer. Training is performed by backpropagating derivatives from the output layer to the input layer. For a more detailed explanation of MLP the reader is referred to [28].

The implemented architecture of MLP is shown in Figure 8. For finding the optimal parameter set grid search algorithms implemented specifically for neural networks were used. The network consisted of three hidden dense layers, each followed by a dropout layer and a batch normalization. The dense layers were initialized with the *glorot_uniform* initializer [40] and the layer regularization parameter was set to L2 [41]. The dropout rate for all dropout layers was set to 0.1, which means that during the training 10% of neurons were randomly disabled. The dropout approach enables the neural network to better generalize and avoid overfitting [42].

F. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural networks are a special kind of artificial neural network, which is intended for processing data that has a known grid-like topology. Such data can be images, which can be thought of as a 2-D grid of pixels [41]. CNNs have demonstrated their suitability for image classification, object recognition, etc. For example, Anttalainen et al. applied a CNN regression model for detecting the concentration of lecithin based on DMS measurements [43].

The dispersion plots studied in this paper can be interpreted as images. Thus, it was hypothesized that a CNN could classify chemicals based on dispersion plots with high

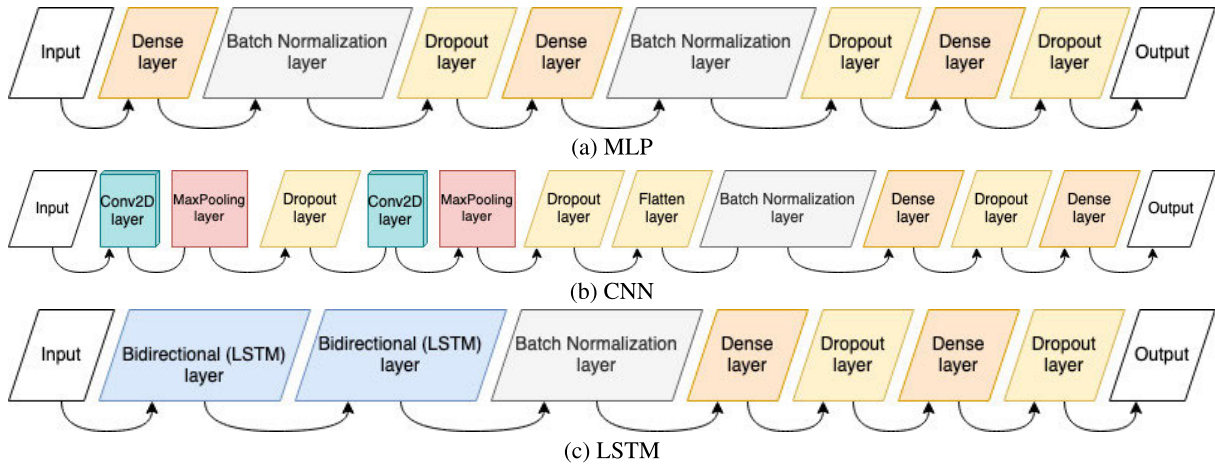


FIGURE 8. Architectures of the artificial neural networks described in Section IV.

accuracy. The employed CNN architecture is schematically shown in Figure 8b. It contained two convolution layers with 16 and 8 feature maps accordingly. Each convolutional layer was followed by MaxPooling and dropout layers. The end of the network contained two hidden dense layers.

Convolutional neural networks have many parameters to be tuned: number of convolution layers, kernels, activation functions, regularizers, architecture of the network, etc. Having enough computation capacity, it is very convenient to tune a neural network by iterating different combinations of the hyperparameters. For the tests in this paper, the optimal combination of hyperparameters was found by applying grid search.

G. LSTM NEURAL NETWORK

Long short-term memory is another extension of artificial neural networks [44]. Instead of neurons as in MLP or convolutions as in CNN, long-short term memory (LSTM) layers are used for capturing features, i.e. temporal patterns. The difference between a general neuron and LSTM neuron is that each LSTM neuron has a shortcut that retains information from previous time steps. Thus, the LSTM neural network can exploit historic information. The so-called bidirectional LSTM layers extend their functionality allowing reversing of the current time steps for capturing prior as well as posterior information.

A DMS dispersion plot represents a set of measurements performed with different settings. That is, the separation voltage increases from bottom to top and the compensation voltage increases from left to right. The different combinations of separation and compensation voltages are measured as sequences and can hence be interpreted as sequential information. Thus, it was hypothesized that LSTM neural networks are capable of capturing sequential features and yielding high classification accuracy. To our knowledge, there are no publications discussing applying sequential models to DMS dispersion plots, and we are providing the first experimental with LSTM.

Before entering the LSTM network, each dispersion plot was preprocessed by row-wise normalization and clipped to only retain the area containing the most entropy as described in Subsection IV-A. The clipping resulted in 20 samples with 100 features each. The optimal set of parameters was found by grid search since no rule of thumb exist on how to define the number of neurons in the LSTM layer, learning rate, momentum, activation function, etc. The implemented architecture of the LSTM network is visualized in Figure 8c. The LSTM neural network contained 8 bidirectional LSTM neurons on the first layer and 256 bidirectional neurons on the second LSTM layer. The two hidden dense layers contained 700 and 500 neurons, respectively. Additionally, the network had two dropout layers between the hidden layers with a dropout rate of 10%.

V. RESULTS AND DISCUSSION

All six classification algorithms were tested with two different cross-validation techniques. The first was Repeated Stratified 10-fold cross-validation. Here, stratified means that the algorithm maintains the same proportion of data in both training and testing sets, repeated means that every cross-validation iteration is repeated N times and the average result is presented, and 10-fold means that the dataset is divided into ten subsets. The algorithm trains on nine subsets (810 samples) and leaves the tenth subset for testing (90 samples). This process is repeated for each subset. The second used cross-validation technique was Leave-One-Group-Out cross-validation based on flowrates. This algorithm splits the dataset into subset based on groups. Here, grouping was done with respect to flow rate. On the first iteration the algorithms were trained on data for flow rates of 8, 16, 32, and 64 sccm, and tested with data for flow rate of 128 sccm. In the next iteration the algorithm was trained on data from 8, 16, 32, and 128 sccm, and tested on data from 64 sccm, etc. The dataset contained for every flow rate a total of 180 samples from the five chemicals. Therefore, in each iteration of the Leave-One-Group-Out

TABLE 2. The table contains classification results. Columns two, and three show cross-validation accuracies with the standard deviations for cross-validation of fold size 10 (CV10) and Leave-One-Group-Out (GCV) in percents. Columns 4 to 8 contain average classification accuracies for each chemical with the standard deviations.

	CV accuracy [%]		Accuracy for each chemical [%]				
	CV10	GCV	nBuOH	Carvone	E2MB	2PEtOH	MCP
ETC	81.1 ±3.6	82.5 ±8.4	87.7 ±7.6	79.0 ±8.1	95.6 ±4.8	74.9 ±7.9	71.8 ±7.5
KNN	74.7 ±3.9	75.8 ±14.5	81.1 ±9.0	74.3 ±8.3	93.9 ±6.0	68.8 ±8.3	60.8 ±9.5
LDA	44.7 ±4.9	46.3 ±6.7	50.2 ±12.4	33.1 ±7.2	97.6 ±4.9	41.6 ±9.5	21.5 ±10.0
MLP	78.0 ±4.0	79.6 ±13.3	84.9 ±7.9	75.6 ±10.0	96.8 ±4.6	75.1 ±9.4	65.5 ±8.5
CNN	69.3 ±6.1	75.4 ±13.9	74.5 ±9.7	67.8 ±12.1	89.9 ±9.7	64.3 ±12.7	57.1 ±10.4
LSTM	88.4 ±3.5	91.0 ±9.5	91.3 ±6.8	86.7 ±7.8	96.9 ±4.3	85.6 ±7.6	84.9 ±8.7

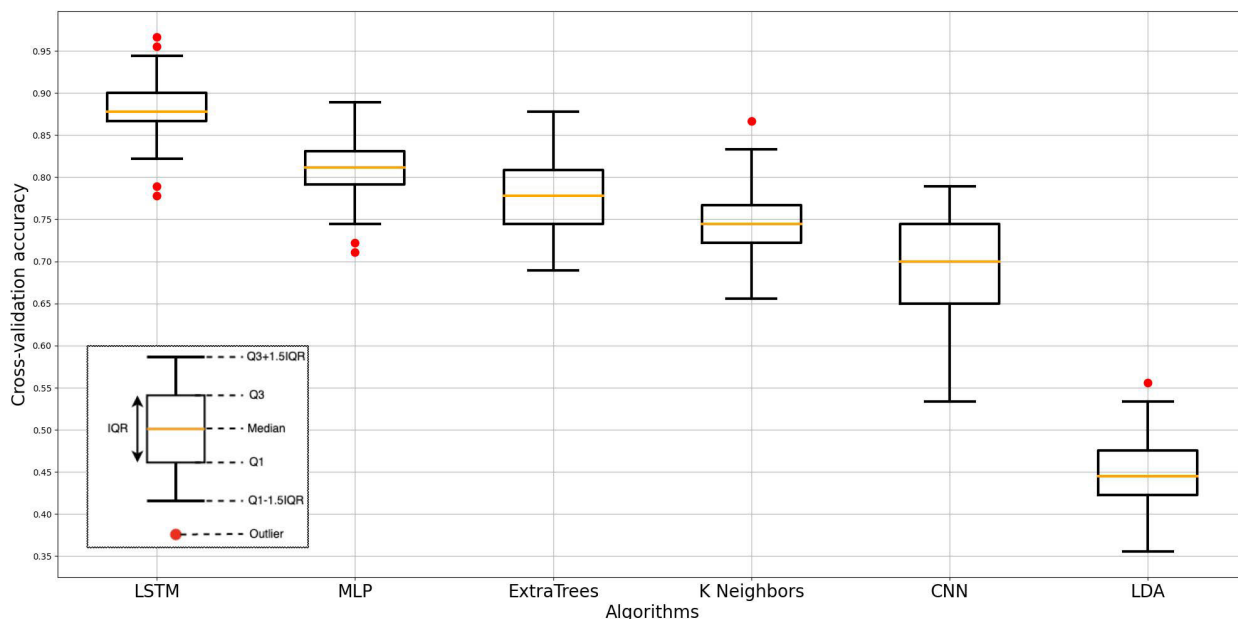


FIGURE 9. Boxplots plot showing cross-validation results for each algorithm. The box extends from the first (Q1) to the third (Q3) quartile. The orange line depicts the median (the second quartile, Q2). The distance between Q1 and Q3 forms the interquartile range (IQR). All data points that lie beyond 1.5 times the IQR are considered to be outliers.

cross-validation 720 samples were used for training and 180 samples for testing.

Table 2 shows cross-validation results for both Repeated Stratified 10-fold and Leave-One-Group-Out cross-validation (GCV). Classification accuracy was measured as the ratio of correctly classified chemicals. Additionally, cross-validation results can be compared from the box plot (Figure 9). The boxes in the box plot span interquartile ranges ($IQR = Q3 - Q1$, 75th percentile - 25th percentile). The orange line in the box shows median (Q2, 50th percentile). The whiskers extend by 1.5 IQR and the red dots beyond the whiskers are considered as outliers. The outlier data were not excluded from evaluation to not artificially increase the accuracy levels. Also, the accuracies from each cross-validation were collected and compiled into confusion matrices (see Section 4 in the supplementary material). Each matrix element in the confusion matrices contains the average accuracy for a particular chemical and the corresponding standard deviation over all cross-validation iterations. From the confusion matrices true positive, true negative, false positive and false negative rates can be derived. Furthermore,

Section 4 in the supplementary material contains for each classifier recall/sensitivity, specificity, and precision rates.

The tests indicated that distinguishing methyl cyclopentenolone from the remaining four chemicals is the most challenging task (see column *MCP* in Table 2) and distinguishing ethyl-2-methylbutyrate from the remaining chemicals is the easiest task (see column *E2MB*), independent on the used classifier. The high accuracies in column *E2MB* can be explained by the higher concentrations of *E2MB*, as observed in Figure 2b.

PCA was applied to the data prior to classification for all algorithms except CNN and LSTM. Since PCA transforms the original data into a subspace that changes the original meaning of the features, it is impossible to say what impact each of the original features had on the classification results for most of the algorithms. For more detailed information on performance of the classifiers the reader is referred Section 4 in the supplementary material.

LDA was selected as baseline classifier due to its comparatively high classification accuracy in [3] and [4]. The inferior performance of LDA compared to both [3] and [4]

can be explained by the fact that in this paper non-binary classification was required, while in [4] the classification problem was binary. Furthermore, resolutions and concentrations of the samples differed considerably. Samples in [4] were 1200-dimensional, while in this paper they were 8000-dimensional.

Applying deep neural networks to a dataset of 810 (Repeated Stratified 10-fold cross-validation) or 720 (Leave-One-Group-Out cross-validation) training samples is challenging. The most common problem observed for small datasets is overfitting, which means that the neural network memorizes the data. However, the difference between “small” and “big” dataset is not obvious, as it depends on many factors. Moreover, classification and detection problems may need less data than regression problems [45, p. 2]. Nevertheless, techniques that effectively reduce overfitting have been applied in this work. The neural networks described in this manuscript contained regularisation on each layer as described in the supplementary material. Additionally, dropout layers were used as can be seen in Figure 8. The training and validation history of the neural networks is shown in Section 4 of the supplementary material. Each plot displays both accuracy and loss curves for the training and validation sets. The accuracy and loss curves demonstrate convergence for all three networks and the behavior of the curves implies that no overfitting occurred during training. The level of regularization for avoiding overfitting and the behavior of the validation and loss curves show that the results are trustworthy. However, more tests with larger datasets and more complex chemicals are planned for the future.

As can be seen from Table 2), the LSTM neural network achieved the highest classification accuracy for both 10-fold cross-validation (88.4%) and for group cross-validation (91.0%), which supports our hypothesis that interpreting dispersion plots as series of measurements evolving sequentially is beneficial. This allows the LSTM network to capture relations between measurements within different separation voltages and consequently yield considerably higher accuracies for nBuOH, Carvone, 2PEtOH and MCP than any other method. It furthermore achieves accuracies similar to the best methods for E2MB.

LDA achieved the best accuracy in discriminating E2MB from other chemicals. This can be explained by the visual differences in E2MB dispersion plots (Figure 2) compared to dispersion plots of the other chemicals.

VI. CONCLUSION AND OUTLOOK

In this work, a new method for classification of DMS dispersion plots was presented. The novel idea is to interpret dispersion plots as a set of measurements evolving sequentially. Hence, a time-series approach for classification was suggested. Following this idea, for the first time, to our knowledge, an LSTM model was applied to dispersion plots and achieved multilabel classification accuracy of 89%. It was compared with other proven classification techniques,

but none of them yielded the same or higher accuracy in either Leave-One-Group-Out or 10-fold cross-validation tests. Even when considering classification for single chemicals, LSTM consistently yielded the highest or close to highest accuracy.

DMS measurements are very sensitive to temperature and humidity changes due to their impact on the mobility of ionized molecules [46, p. 250]. The state-of-the-art approach for dealing with the impact of environmental conditions is to normalize the measurement conditions, which is often insufficient. Therefore, additional data should be collected at both stable and varying environmental conditions at both stable and varying environmental conditions to enable more extensive testing of the LSTM neural network. Another possible topic of research is extracting alpha curves and use them as input for the LSTM as it potentially increase the classification accuracy [47]. To use the alpha-curves approach to its fullest potential, the impact of changing environmental conditions on dispersion plots needs to be studied thoroughly.

In this paper the LSTM model was tested with dispersion plots of different chemicals. However, based on our experience with dispersion plots of other VOCs, such as sweat samples from Alzheimer and Parkinson patients, cancer tissues, etc., it is reasonable to assume that the LSTM method will work well with all types of dispersion plots.

A. ABBREVIATIONS AND ACRONYMS

APPENDIX

DATA AND SUPPLEMENTARY MATERIAL

The dataset used for this research is available to download in [48].

ACKNOWLEDGMENT

The authors thank Tuan-Anh Tran for his help in collecting the data. Osmo Anttalainen is a shareholder of Olfactomics Oy Ltd., a company that developed IonVision-differential mobility spectrometer. The other authors declare no competing interests.

REFERENCES

- [1] A. D. Wilson and M. Baietto, “Applications and advances in electronic-nose technologies,” *Sensors*, vol. 9, no. 7, pp. 5099–5148, Jun. 2009. [Online]. Available: <https://www.mdpi.com/1424-8220/9/7/5099>
- [2] G. Lubes and M. Goodarzi, “GC-MS based metabolomics used for the identification of cancer volatile organic compounds as biomarkers,” *J. Pharmaceutical Biomed. Anal.*, vol. 147, pp. 313–322, Jan. 2018, doi: 10.1016/j.jpba.2017.07.013.
- [3] A. Kontunen, M. Karjalainen, A. Anttalainen, O. Anttalainen, M. Koskenranta, A. Vehkaoja, N. Oksala, and A. Roine, “Real time tissue identification from diathermy smoke by differential mobility spectrometry,” *IEEE Sensors J.*, vol. 21, no. 1, pp. 717–724, Jan. 2021. [Online]. Available: <https://trepo.tuni.fi/handle/10024/134787>
- [4] I. Haapala, A. Rauhameri, A. Roine, M. Mäkelä, A. Kontunen, M. Karjalainen, A. Laakso, P. Koroknay-Pál, K. Nordfors, H. Haapasalo, N. Oksala, A. Vehkaoja, and J. Haapasalo, “Method for the intraoperative detection of IDH mutation in gliomas with differential mobility spectrometry,” *Current Oncol.*, vol. 29, no. 5, pp. 3252–3258, May 2022. [Online]. Available: <https://www.mdpi.com/1718-7729/29/5/265>
- [5] M. Hernández-Mesa, A. Escourrou, F. Monteau, B. Le Bizec, and G. Dervilly-Pinel, “Current applications and perspectives of ion mobility spectrometry to answer chemical food safety issues,” *TrAC, Trends Anal. Chem.*, vol. 94, pp. 39–53, Sep. 2017.

- [6] I. Haapala, M. Karjalainen, A. Kontunen, A. Vehkaoja, K. Nordfors, H. Haapasalo, J. Haapasalo, N. Oksala, and A. Roine, "Identifying brain tumors by differential mobility spectrometry analysis of diathermy smoke," *J. Neurosurgery*, vol. 133, no. 1, pp. 100–106, Jul. 2020. [Online]. Available: <https://thejns.org/view/journals/j-neurosurg/133/1/article-p100.xml>
- [7] D. Panagiotakopoulos, G. Marentakis, R. Metzidakos, I. Deliyannis, and F. Dedes, "Digital scent technology: Toward the Internet of Senses and the metaverse," *IT Prof.*, vol. 24, no. 3, pp. 52–59, May 2022. [Online]. Available: <https://ieeexplore.ieee.org/document/9811514>
- [8] N. S. Archer, A. Bluff, A. Eddy, C. K. Nikhil, N. Hazell, D. Frank, and A. Johnston, "Odour enhances the sense of presence in a virtual reality environment," *PLoS ONE*, vol. 17, no. 3, Mar. 2022, Art. no. e0265039. [Online]. Available: <https://dx.plos.org/10.1371/journal.pone.0265039>
- [9] V. Nieminen, M. Karjalainen, K. Salminen, J. Rantala, A. Kontunen, P. Isokoski, P. Müller, P. Kallio, V. Surakka, and J. Leikkala, "A compact olfactometer for IMS measurements and testing human perception," *Int. J. Ion Mobility Spectrometry*, vol. 21, no. 3, pp. 71–80, Sep. 2018. [Online]. Available: <http://link.springer.com/10.1007/s12127-018-0235-1>
- [10] K. Salminen, J. Rantala, P. Isokoski, M. Lehtonen, P. Müller, M. Karjalainen, J. Väliäho, A. Kontunen, V. Nieminen, J. Leivo, A. A. Telembeci, J. Leikkala, P. Kallio, and V. Surakka, "Olfactory display prototype for presenting and sensing authentic and synthetic odors," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Boulder, CO, USA, Oct. 2018, pp. 73–77, doi: [10.1145/3242969.3242999](https://doi.org/10.1145/3242969.3242999).
- [11] P. Müller, K. Salminen, A. Kontunen, M. Karjalainen, P. Isokoski, J. Rantala, J. Leivo, J. Väliäho, P. Kallio, J. Leikkala, and V. Surakka, "Online scent classification by ion-mobility spectrometry sequences," *Frontiers Appl. Math. Statist.*, vol. 5, p. 39, Jul. 2019. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fams.2019.00039/full>
- [12] Olfactomics Web. (Apr. 18, 2023). *Olfactomics Oy*. [Online]. Available: <https://olfactomics.fi>
- [13] A. Good, D. A. Durden, and P. Kebarle, "Mechanism and rate constants of ion–molecule reactions leading to formation of $H^+(H_2O)_n$ in moist oxygen and air," *J. Chem. Phys.*, vol. 52, no. 1, pp. 222–229, Jan. 1970, doi: [10.1063/1.1672668](https://doi.org/10.1063/1.1672668).
- [14] F. Martinelli, R. Scalenghe, A. Giovino, P. Marino, A. A. Aksenov, A. Pasamontes, D. J. Peirano, C. E. Davis, and A. Dandekar, "Proposal of a *Citrus* translational genomic approach for early and infield detection of Flavescence dorée in *Vitis*," *Plant Biosyst. Int. J. Dealing Aspects Plant Biol.*, vol. 150, no. 1, pp. 43–53, Jan. 2016. [Online]. Available: <http://www.tandfonline.com/doi/full/10.1080/11263504.2014.908976>
- [15] S. L. Coy, E. V. Krylov, B. B. Schneider, T. R. Covey, D. J. Brenner, J. B. Tyburski, A. D. Patterson, K. W. Krausz, A. J. Fornace, and E. G. Nazarov, "Detection of radiation-exposure biomarkers by differential mobility prefiltered mass spectrometry (DMS-MS)," *Int. J. Mass Spectrometry*, vol. 291, no. 3, pp. 108–117, Apr. 2010. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1387380610000308>
- [16] M. Lepomäki, A. Anttalainen, A. Vuorinen, T. Tolonen, A. Kontunen, M. Karjalainen, A. Vehkaoja, A. Roine, and N. Oksala, "Laser desorption tissue imaging with differential mobility spectrometry," *Experim. Mol. Pathol.*, vol. 125, Apr. 2022, Art. no. 104759.
- [17] P. E. Fowler, T. Bernat, J. Z. Pilgrim, and G. A. Eiceman, "Neural network classification of mobility spectra for volatile organic compounds using tandem differential mobility spectrometry with field induced fragmentation," *Analytica Chim. Acta*, vol. 1252, Apr. 2023, Art. no. 341047.
- [18] C. Ieritano, J. L. Campbell, and W. S. Hopkins, "Predicting differential ion mobility behaviour in silico using machine learning," *Analyst*, vol. 146, no. 15, pp. 4737–4743, Jul. 2021.
- [19] M. Y. Rajapakse, E. Borrás, D. Yeap, D. J. Peirano, N. J. Kenyon, and C. E. Davis, "Automated chemical identification and library building using dispersion plots for differential mobility spectrometry," *Anal. Methods*, vol. 10, no. 35, pp. 4339–4349, 2018. [Online]. Available: <https://pubs.rsc.org/en/content/articlelanding/2018/ay/c8ay00846a>
- [20] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989. [Online]. Available: <https://www.jstor.org/stable/2289860>
- [21] N. Altman and M. Krzywinski, "The curse(s) of dimensionality," *Nature Methods*, vol. 15, no. 6, pp. 399–400, 2018.
- [22] J. Rantala, P. Müller, T. Salpavaara, J. Verho, T. Ryyänen, P. Isokoski, A. Rauhameri, N. Toimela, A. Vehkaoja, J. Leikkala, P. Kallio, and V. Surakka, "On characterizing olfactory displays," in *Proc. Smell, Taste Temp. Interfaces Workshop CHI*, 2023, p. 4. [Online]. Available: <https://researchportal.tuni.fi/en/publications/on-characterizing-olfactory-displays>
- [23] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, "PubChem 2023 update," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D1373–D1380, Jan. 2023, doi: [10.1093/nar/gkac956](https://doi.org/10.1093/nar/gkac956).
- [24] Z. Safaei, "Application of differential ion mobility spectrometry for detection of water pollutants," Lappeenranta-Lahti Univ. Technol. LUT, Lappeenranta, Finland, Tech. Rep. Acta Universitatis Lappeenrantaensis 915, 2020.
- [25] Z. Safaei, G. A. Eiceman, J. Puton, J. A. Stone, M. Nasirikheirabadi, O. Anttalainen, and M. Sillanpää, "Differential mobility spectrometry of ketones in air at extreme levels of moisture," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, Apr. 2019, Art. no. 5593, doi: [10.1038/s41598-019-41485-7](https://doi.org/10.1038/s41598-019-41485-7).
- [26] O. Anttalainen, J. Puton, A. Kontunen, M. Karjalainen, P. Kumpulainen, N. Oksala, Z. Safaei, and A. Roine, "Possible strategy to use differential mobility spectrometry in real time applications," *Int. J. Ion Mobility Spectrometry*, vol. 23, no. 1, pp. 1–8, Apr. 2020.
- [27] J. Virtanen, A. Anttalainen, J. Ormiskangas, M. Karjalainen, A. Kontunen, M. Rautiainen, N. Oksala, I. Kivekäs, and A. Roine, "Differentiation of aspirated nasal air from room air using analysis with a differential mobility spectrometry-based electronic nose: A proof-of-concept study," *J. Breath Res.*, vol. 16, no. 1, Dec. 2021, Art. no. 016004, doi: [10.1088/1752-7163/ac3b39](https://doi.org/10.1088/1752-7163/ac3b39).
- [28] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York, NY, USA: Springer, 2006.
- [29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).
- [30] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R* (Springer Texts in Statistics), 2nd ed., New York, NY, USA: Springer, 2021.
- [31] M. Mashayekhi and R. Gras, "Rule extraction from random forest: The RF+HC methods," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)). Springer, Switzerland: Springer, 2015, pp. 223–237.
- [32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.
- [33] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the Gini index and information gain criteria," *Ann. Math. Artif. Intell.*, vol. 41, pp. 77–93, May 2004.
- [34] P. Müller, K. Salminen, V. Nieminen, A. Kontunen, M. Karjalainen, P. Isokoski, J. Rantala, M. Savia, J. Väliäho, P. Kallio, J. Leikkala, and V. Surakka, "Scent classification by K nearest neighbors using ion-mobility spectrometry measurements," *Expert Syst. Appl.*, vol. 115, pp. 593–606, Jan. 2019, doi: [10.1016/j.eswa.2018.08.042](https://doi.org/10.1016/j.eswa.2018.08.042).
- [35] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York, NY, USA: Springer, 2009.
- [36] C. Giraud, *Introduction to High-Dimensional Statistics*, 2nd ed., Boca Raton, FL, USA: CRC Press, 2021.
- [37] O. Kramer, *Dimensionality Reduction With Unsupervised Nearest Neighbors* (Intelligent Systems Reference Library), vol. 51. Berlin, Germany: Springer, 2013, doi: [10.1007/978-3-642-38652-7](https://doi.org/10.1007/978-3-642-38652-7). [Online]. Available: <https://link.springer.com/book/10.1007/978-3-642-38652-7>
- [38] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, Sep. 1975.
- [39] O. Ledoit and M. Wolf, "Honey, I shrunk the sample covariance matrix: Problems in mean-variance optimization," *J. Portfolio Manage.*, vol. 30, no. 4, p. 110, 2004.
- [40] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, in Proceedings of Machine Learning Research, Sardinia, Italy, Y. W. Teh and M. Titterton, Eds., May 2010, pp. 249–256. [Online]. Available: <https://proceedings.mlr.press/v9/glorot10a.html>
- [41] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: <http://www.deeplearningbook.org>
- [42] P. Baldi and P. Sadowski, "The dropout learning algorithm," *Artif. Intell.*, vol. 210, pp. 78–122, May 2014. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0004370214000216>

- [43] A. Anttalainen, M. Mäkelä, P. Kumpulainen, A. Vehkaoja, O. Anttalainen, N. Oksala, and A. Roine, "Predicting lecithin concentration from differential mobility spectrometry measurements with linear regression models and neural networks," *Talanta*, vol. 225, Apr. 2021, Art. no. 121926. [Online]. Available: <https://trepo.tuni.fi/handle/10024/136805>
- [44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [45] A. Safonova, G. Ghazaryan, S. Stiller, M. Main-Knorn, C. Nendel, and M. Ryo, "Ten deep learning techniques to address small data problems with remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 125, Dec. 2023, Art. no. 103569. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S156984322300393X>
- [46] G. A. Eiceman, Z. Karpas, and H. H. Hill Jr., *Ion Mobility Spectrometry*. Boca Raton, FL, USA: CRC Press, 2013.
- [47] A. Rauhameri, O. Anttalainen, J. Rantala, V. Surakka, A. Vehkaoja, and P. Müller, "Clustering of alpha curves in differential mobility spectrometry data," in *Proc. IEEE Int. Symp. Olfaction Electron. Nose (ISOEN)*, May 2022, pp. 1–3.
- [48] A. Rauhameri, A. Robiños, P. Müller, V. Surakka, A. Vehkaoja, and P. Kallio, Aug. 2024, "DMS measurements dataset for manuscript 'classification of volatile organic compounds by differential mobility spectrometry based on continuity of alpha curves,'" *Zenodo*, doi: [10.5281/zenodo.13373440](https://doi.org/10.5281/zenodo.13373440).



ANTON RAUHAMERI received the M.Sc. degree in robotics focusing on algorithms and artificial intelligence from Tampere University, Finland, in 2021, where he is currently pursuing the D.Sc. degree with the Faculty of Medicine and Health Technology. His research interests include differential mobility spectrometry, machine/deep learning, artificial intelligence, and applied mathematics.



ANGELO ROBIÑOS received the dual M.Sc. degrees in analytical chemistry from the University of Tartu, Estonia, and Åbo Akademi University, Finland, in 2022. He is currently pursuing the D.Sc. degree with the Laboratory of Molecular Science and Engineering and Natural Materials Technology, Åbo Akademi University. His research interests include chemiresistive gas sensors, lignin, hard carbon, and sodium-ion batteries.



OSMO ANTATALAINEN received the M.Sc. degree in energy technology from Lappeenranta University of Technology, Finland, in 1992. From 1994 to 2018, he was with Environics Oy, Mikkeli, Finland. Currently, he is continuing his career with a university spin-off company, Olfactomics Oy, Tampere, Finland. His research interests include ion mobility spectrometry and mixed signal electronics.



TIMO SALPAVAARA received the M.Sc. degree in electrical engineering and the D.Sc. degree in biomedical sciences and engineering from Tampere University of Technology (TUT), Tampere, Finland, in 2005 and 2018, respectively. His research interests include sensors and microfabrication.



JUSSI RANTALA received the M.Sc. degree in computer science and the Ph.D. degree in interactive technology from the University of Tampere, in 2007 and 2014, respectively. He is currently a Staff Scientist with Tampere University and a member of Tampere University Computer–Human Interaction Research Center (TAUCHI). His research interests include human–computer interaction, multisensory experiences, olfaction, and haptics.



VEIKKO SURAKKA has been a Professor of interactive technology, since 2007; and the Head of the Research Group for Emotions, Sociality, and Computing (<https://research.tuni.fi/esc/>). He and his group's research focuses especially on research on emotion, cognition, human-human and human-technology interaction research, and development of new technologies.



PASI KALLIO (Member, IEEE) received the M.Sc. degree in electrical engineering and the D.Tech. degree in automation engineering from Tampere University of Technology (TUT), Tampere, Finland, in 1994 and 2002, respectively. From 2008 to 2018, he was a Professor of automation engineering with the Faculty of Biomedical Sciences and Engineering in TUT. Since 2019, he has been a Professor of biomedical micro- and nanodevices with Tampere University, where he was the Vice Dean for Research with the Faculty of Medicine and Health Technology, from 2019 to 2023. He has authored more than 170 peer-reviewed articles and has 16 patent applications. His current research interests include microfluidics, microsensors, microrobotics, and their application in organ-on-chip, olfactory display, and fibrous material testing applications. He was a recipient of the Finnish Automation Society Award in 2009. He has co-founded three start-up companies. He was the Chair of the IEEE Finland Section, from 2012 to 2013.



ANTTI VEHKAOJA received the D.Sc. (Tech.) degree in automation science and engineering from Tampere University of Technology, Tampere, Finland, in 2015. He is currently an Associate Professor (tenure track) of sensor technology and biomeasurements with the Faculty of Medicine and Health Technology, Tampere University, Finland. He has authored more than 120 scientific articles, mainly in the area of biomedical engineering. His research interests include embedded measurement technologies for physiological monitoring and related signal processing and data analysis methods and the analysis of volatile organic compounds in biomedical applications.



PHILIPP MÜLLER received the M.Sc. degree in mathematics from the Chemnitz University of Technology, Germany, in 2010, and the D.Sc. (Tech.) degree in automation science and engineering from Tampere University of Technology, Finland, in 2016. He is currently a Senior Research Fellow with the Computing Sciences Unit, Faculty of Information Technology and Computing Sciences, Tampere University, Finland. His research interest include data analytics, model-based estimation, machine learning, and ion mobility spectrometry.

...