Scent Classification by K Nearest Neighbors using Ion-Mobility Spectrometry Measurements

Scent Classification by K Nearest Neighbors using Ion-Mobility Spectrometry Measurements

Philipp Müller[a,*], Katri Salminen[b], Ville Nieminen[a], Anton Kontunen[a], Markus Karjalainen[a], Poika Isokoski[b], Jussi Rantala[b], Mariaana Savia[a], Jari Väliaho[a], Pasi Kallio[a], Jukka Lekkala[a], Veikko Surakka[b]

[a] *BioMediTech Institute and Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, P.O. Box 692, 33101 Tampere, Finland*
[b] *Research Group for Emotions, Sociality, and Computing, Faculty of Communication Sciences, 33014 University of Tampere, Tampere, Finland*

Abstract

Various classifiers for scent classification based on measurements using an electronic nose (eNose) have been studied recently. In general, classifiers rely on a static database containing reference eNose measurements for known scents. However, most of these approaches require retraining of the classifier every time a new scent needs to be added to the training database. In this paper, the potential of a K nearest neighbors (KNN) classifier is investigated to avoid the time-consuming retraining when updating the database. To speed up classification, a k-dimensional tree search in the KNN classifier and principal component analysis (PCA) are studied. The tests with scents presented to an eNose based on ion-mobility spectrometry (IMS) show that the KNN method classifies scents with high accuracy. Using a k-dimensional tree search instead of an exhaustive search has no significant influence on the misclassification rate but reduces the classification time considerably. The

*Corresponding author

Email addresses: philipp.muller@tut.fi (Philipp Müller), katri.salminen@uta.fi (Katri Salminen), ville.a.nieminen@tut.fi (Ville Nieminen), anton.kontunen@tut.fi (Anton Kontunen), markus.karjalainen@tut.fi (Markus Karjalainen), poika.isokoski@sis.uta.fi (Poika Isokoski), jussi.e.rantala@sis.uta.fi (Jussi Rantala), mariaana.savia@tut.fi (Mariaana Savia), jari.valiaho@tut.fi (Jari Väliaho), pasi.kallio@tut.fi (Pasi Kallio), jukka.lekkala@tut.fi (Jukka Lekkala), veikko.surakka@uta.fi (Veikko Surakka)

use of PCA-transformed data results in a higher misclassification rate than the use of IMS data when only the first principal components explaining 95% of the total variance are used but in a similar misclassification rate when the first principal components explaining 99% of the total variance are used. In conclusion, the proposed method can be recommended for classifying scents measured with IMS-based eNoses.

Keywords:
Machine learning, K nearest neighbours, Ion-mobility spectrometry, Scent classification

---

1. Introduction

Classifying scents measured with an electronic nose (eNose) is useful in several areas of interest. Traditionally, classification of eNose-measured scents has been used to maintain quality control in the food industry and to analyze curative and aromatic plants used in medicines, perfumes and cosmetics (e.g., Wilson and Baietto (2009) and references therein). One interesting area of application that could benefit considerably from the use of eNose measurements and classification of the data is scent production using olfactory displays. Olfactory displays refer to devices built to stimulate human olfaction while, for instance, watching a movie, interacting in a virtual reality, or using an educational interface (Nakamoto et al. (2008); García-Ruíz and Santana-Mancilla (2013)). Olfactory displays were developed because the sense of olfaction has a considerable role in human performance and behavior. For example, research suggests that different scents can evoke emotions and memories, modulate the functioning of the autonomic and central nervous system, and modulate the way we process information from other senses (García-Ruíz and Santana-Mancilla (2013); Vernet-Maury et al. (1999); Robin et al. (1999)). Olfactory displays are typically built to spray only a predetermined set of scents in the air. However, this approach limits the number of scents used, does not provide the system any kind of information about the quality of the scent (e.g., intensity or the identification of the scent), and most specifically, does not allow the olfactory display to adapt to communicate an unlabeled scent distantly from one user to another. The last approach in

particular requires the system to be able to measure scents and classify them rapidly.

In this paper, we focus on classifying scents, which is one of the key issues for our research on digitizing scents.[1] Analogous to cameras digitizing photons, the research is focused on developing a system that digitizes scents, thus, enabling us to save and transfer scents over time and space. Scent transfer has three steps. First, measurements from a scent source, such as a lemon or strawberry, are taken at location A. Second, the scent is classified based on the measurements, and the resulting label is transferred with the measurements to location B. Third, a synthetic scent is generated by a so-called scent synthesizer at location B. The synthetic scent is built from chemical key odor components and has to be as similar as possible to the original scent at location A. By using a scent classifier, the search space for the scent synthesizer can be narrowed down to a small number of key odor components that are likely to be part of the original scent. Thus, a good classifier speeds up the production of a synthetic scent and saves resources, because fewer key odor components need to be tested.

For the purpose suggested above, two different solutions are needed: analyzing scents and classifying them. Scents can be analyzed in different ways. One approach is to use high-demand laboratory instruments, such as mass spectrometers and gas chromatographs. The major disadvantages of these instruments are that they are difficult to miniaturize and they cannot be implemented cheaply as consumer-grade instruments. A low-cost and portable alternative is an eNose, which mimics the biological sense of smell and its communication with a biological brain (Kiani et al. (2016)). Traditional eNoses are based on a gas sensor array and the appropriate algorithms. However, ion-mobility spectrometry (IMS) technology can function in a similar way. In addition, IMS-based eNoses have been shown to be suitable for fast analysis (Loutfi et al. (2015) and references therein). For an overview of the major eNose techniques, we refer the reader to Wilson and Baietto (2009) and Loutfi et al. (2015) and references therein. In this paper, we use the Environics ChemPro 100i eNose, which is based on IMS and analyzes scents by separating and identifying ionized molecules in the gas phase based on their mobility in a carrier buffer gas (Utriainen et al. (2003)).

[1] Of course it is also of interest in the food industry, the cosmetics industry, biomedical applications, etc.

Unlike mass spectrometry, IMS does not require a vacuum. Furthermore, the sensors and electrodes of the IMS-based eNose do not age, unlike metal-oxide sensor-based eNoses. This means, that the IMS sensors experience signal drift mainly due to environmental changes.

Scents have been classified based on different eNose measurements using various approaches. In general, supervised learning methods are used. This means that measurements from known scents are stored in a training database. An unlabeled scent is then classified by comparing its measurements with the measurements in the database. Examples of classification algorithms used for quality assessment of medicinal and aromatic plant products can be found in Kiani et al. (2016) and references therein. In general, supervised learning trains the various classifiers. Classification methods include, but are not limited to, principal component analysis (PCA; e.g. in Mamat et al. (2011)), linear discriminant analysis (Martín et al. (2001)), canonical discriminant analysis (Seregély and Novák (2015)), discriminant functions analysis (Zheng et al. (2015)), hierarchical cluster analysis (Lin et al. (2013)), cluster analysis (Yang et al. (2009)), support vector machine (SVM; Längkvist et al. (2013)), fuzzy artificial neural networks (ANNs; Singh et al. (1996)), and multilayer perceptron (MLP)-type classifiers (e.g. in Zhang and Tian (2014)). For additional approaches and details we refer the reader to, for example, Kiani et al. (2016) and Loutfi et al. (2015) and references therein.

In the literature, to our knowledge, the ability to update scent training databases without using often time-consuming and cumbersome classifier retraining has not been considered. However, for real-world applications, such as digital scent transfer and quality control of cosmetic products, this property is desirable. Therefore, the use of the simple but effective K nearest neighbors (KNN) approach seemed to be a promising starting point because this method does not require generating the classifier in advance (Khan et al. (2002)). The fact that the KNN algorithm has no training phase is an advantage over its competitors, making it a better option than the methods mentioned above. For these methods adding new samples to the training database is possible but requires always retraining the classifiers. It is noteworthy that modern proposals for efficient KNN, such as the KNN-IS in Maillo et al. (2017), the clustering-based KNN in Gallego et al. (2018), and the efficient KNN in Zhang et al. (2018) also do not allow dynamic insertion of new samples meaning that they also require retraining when the training database is modified.

The standard KNN has been used by Martín et al. (2001) to characterize

vegetable oils and by Tang et al. (2010) to classify fruity scents. KNN compares the IMS sample for the scent, measured by the eNose, with the samples in the training database, picks the K training samples closest to the unlabeled scent's sample, and then classifies the scent based on the labels of these K training samples. This method has strong consistency results, and its misclassification rate can be controlled by adjusting K (Duda et al., 2001, p. 174 ff.).

The two major drawbacks of using the KNN classifier are that it is slow for large training databases, because it compares a new sample with all samples in the database,[2] and it can be fooled by irrelevant features. The latter drawback affects the majority of classifiers to some extent. Therefore, feature transformation and/or feature selection methods should be applied to the IMS samples before the classifier is learned and applied. To speed up the classification performed by KNN, three general techniques could be used: computing partial distances, prestructuring, and editing the training samples (Duda et al., 2001, p. 185 f.). (Duda et al. (2001), p. 185 ff.).

To understand how large the training database can get, let us consider the following example. For each scent from our scent library, we should store at least 5 minutes of IMS measurements, with a measurement frequency of 1 Hz, in the training database. In order to remove the influence of environmental conditions, measurement noise, etc., it would be ideal to repeat the measurements at least 5 times. This means that for each scent in the library we should store at least 1 500 IMS samples in the training database, although a much larger number of samples per scent would be desirable. Thus, it becomes obvious that a comprehensive training database would grow very large.

The contribution of this paper is threefold. First, the tests show that a KNN classifier is a suitable choice for classifying scents using ion-mobility spectrometry measurements, even for cases in which the IMS measurement data of different scents vary only slightly. At the same time, the KNN classifier works with a nonstatic training database of IMS measurements without the need to constantly retrain the classifier. The classifier is tested successfully with three chemicals and a wide selection of food scents. Second, quick classification of the original scent is desirable in many applications,

[2]This means that the classification time for KNN is linear to the number of samples in the training database (Abidin and Perrizo (2006)).

such as digital scent transfer and quality control. Therefore, this paper demonstrates that the classification time can be reduced considerably by using a prestructuring technique called k-dimensional trees (Bentley (1975). In the tests the classifier that uses k-dimensional trees needs approximately 1/8 of the time that the standard KNN needs). The test results show that this acceleration of the classification task does not cause any (significant) loss in classification accuracy. Third, the results show that principal component analysis (Duda et al., 2001, p. 580) can further accelerate the classification process without significant performance degradation. This furthermore implies that the channels of an IMS-based electronic nose are dependent. However, the tests also reveal that care has to be taken when choosing the number of principal components for classification to prevent loss of classification accuracy. Thus, the classifier presented in this paper is a crucial step in the development of a scent transfer system and will simplify and accelerate the production of synthetic scents that are copies of the scent to be transferred.

This paper is organized as follows. In Section 2, we describe the scent production method and the eNose we used, explain the KNN classifier and discuss ways to reduce the computation time for classifying scents. We test the classification performance of our KNN using data from scents presented to the eNose in different ways in Section 3. Finally, we discuss the test results, draw conclusions and give an outlook on future research in Section 4.

Notation: In this paper a denotes a scalar, $\mathbf{b}$ denotes a vector, and $\mathbf{C}$ denotes a matrix.

2. Methods

2.1. Description of scent production, eNose and data collection

In this paper, three different methods for presenting scents to an eNose are used. A carefully controllable olfactory display prototype is used to present three central chemical components of jasmine. In addition, seven food scents are presented on a plate and in a sealed jar to understand whether the presentation method affects eNose measurements and classification results.

The goal of olfactory display design is to produce synthetic scents that are close to the original scents by mixing significantly fewer odorous chemicals than comprise the originals. Jasmine, for example, consists of more than ten odorous molecules (Edris et al. (2008)). However, it is likely that jasmine can be synthesized using only three of its components: benzyl acetate, cis-

Jasmone, and indole. These components are used as the starting point for controlled scent production for the tests in Section 3.

For the test, we developed a scent production prototype that transfers liquid chemical compounds into the gas phase. In this olfactory display, which is shown in Fig. 1, the carrier gas, flow intensity, and scent intensity can be carefully controlled to mix key odor components, and the prototype can produce several scents in the gas phase reliably and continuously for a long time.

Scent samples are produced in three separate channels using programmable Newera NE-500 syringe pumps. The syringes are connected to evaporation units, which contain heat-resistant polyether ether ketone (PEEK) tubes that lead to the surface of the ceramic heater elements (see Fig. 2).

The scent concentration in the carrier gas is adjusted by changing the pumping speed of the pumps. The power of the heater elements is controlled with an Arduino power control unit. A linear relationship between the heater's power and the pumping speed has been discovered. Thus, in order to achieve stable evaporation the power of the heater elements needs to be adjusted according to the pumping speed.

The olfactory display has also an output for the ChemPro 100i eNose (Environics (2017)) shown in Fig. 1. The eNose is based on ion-mobility spectrometry. It analyzes scents by grouping ionized molecules in the gas phase based on their mobility in a carrier buffer gas. The ChemPro 100i produces channel data from 16 ion electrodes. Data sequences from 14 channels (electrodes 1 to 7 and 9 to 15) are used to classify scents. Electrodes 8 and 16 are used to control the carrier gas flow speed and therefore, provide no useful information for scent classification.

The Environics ChemPro 100i has a sampling frequency of 2 Hz. However, previous analysis of the channel readings revealed that the measurements do not change markedly within half a second. Therefore, for better readability of the data we used a sampling frequency of 1 Hz instead.

Due to the design of the olfactory display, the measurements for each channel require a few minutes to stabilize. Fig. 3 shows the channel 4 response for benzyl acetate with a concentration of 15%, with propylene glycol as diluent. In this example, it takes approximately 80 sec for the measurement signal to reach the steady state level (of 95). In the remainder of the paper, we distinguish the phase in which the signal stabilizes (the transient phase aka the stabilization phase) and the phase in which the signal is stable (the stable phase).

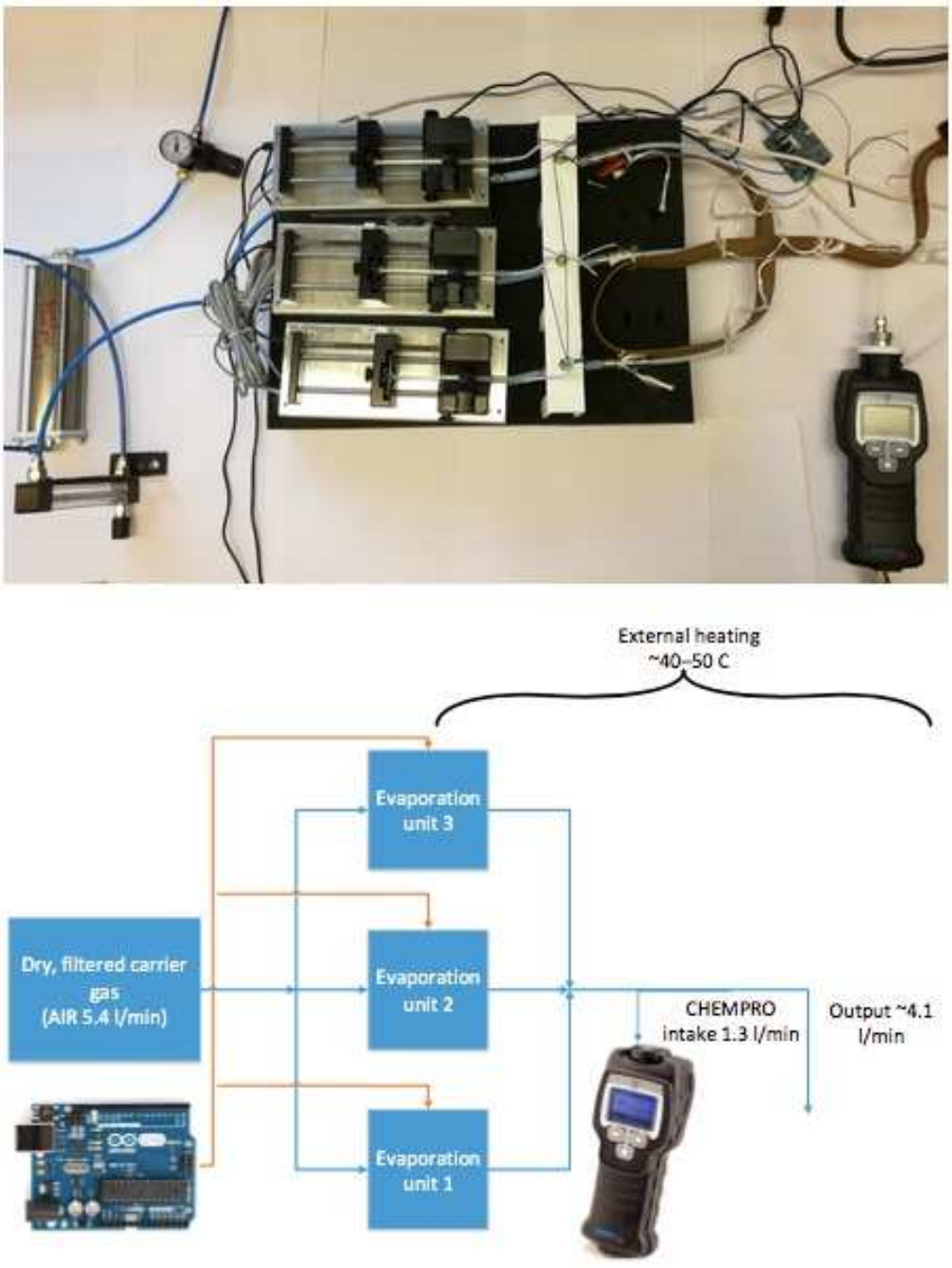

Figure 1: Compact olfactometer used for the tests in Section 3. The programmable Newera NE-500 syringe pumps can be seen in the center of the upper figure, and the ChemPro 100i eNose can be seen in the lower right corner. The lower figure summarizes the olfactometer's working principle.

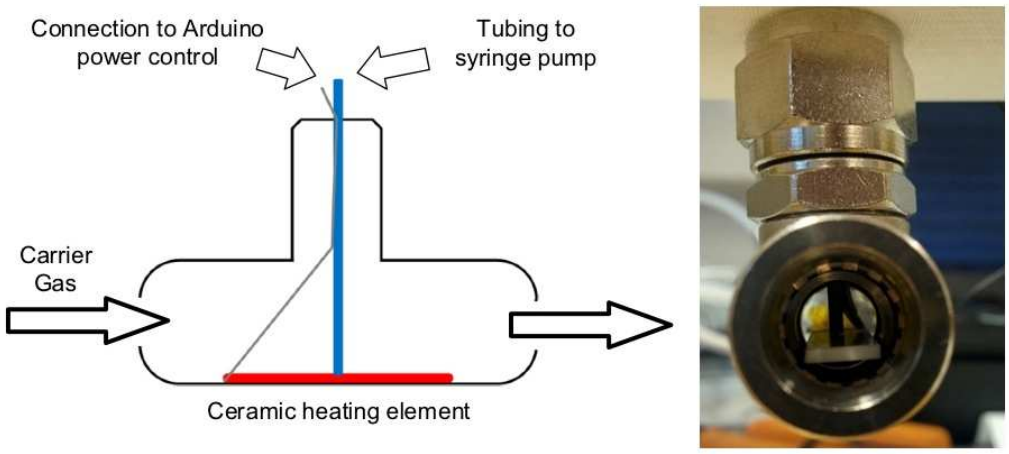

Figure 2: Ceramic heater elements used by the olfactory display.

In order to avoid skewed class distributions in the training database, each scent has the same (or at least very similar) number of training samples in the database.

2.2. K nearest neighbors classifier in a nutshell

For classifying an unlabeled scent the K nearest neighbors classifier is used. The idea behind KNN is to compare the scent's 14-dimensional IMS sample $\mathbf{x}^{(\mathrm{us})} = [x_1^{(\mathrm{us})} .. x_{14}^{(\mathrm{us})}]$ with IMS training samples in a training database, find the K training samples closest to $\mathbf{x}^{(\mathrm{us})}$ and label the scent based on the labels of the K closest training samples.

The training database contains N IMS samples $\mathbf{x}_i = [x_{i,1} .. x_{i,14}]$, $i = 1, .., N$ and their corresponding labels (i.e. the name of the scents they belong to). The closeness between the new sample $\mathbf{x}^{(\mathrm{us})}$ and the ith training sample is computed as the Euclidean distance between the two, which is defined as

$$d_{\mathrm{E}}(\mathbf{x}^{(\mathrm{us})}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{14} \left(x_{ij} - x_j^{(\mathrm{us})} \right)^2}. \quad (1)$$

The analyzed scent is then classified as belonging to the same scent as the majority of its K closest neighbors (i.e., the samples for which $d_{\mathrm{E}}(\mathbf{x}^{(\mathrm{us})}, \mathbf{x}_i)$ is minimal). For example, let two neighbors belong to lemon and one to orange. Then the scent will be labeled as lemon.

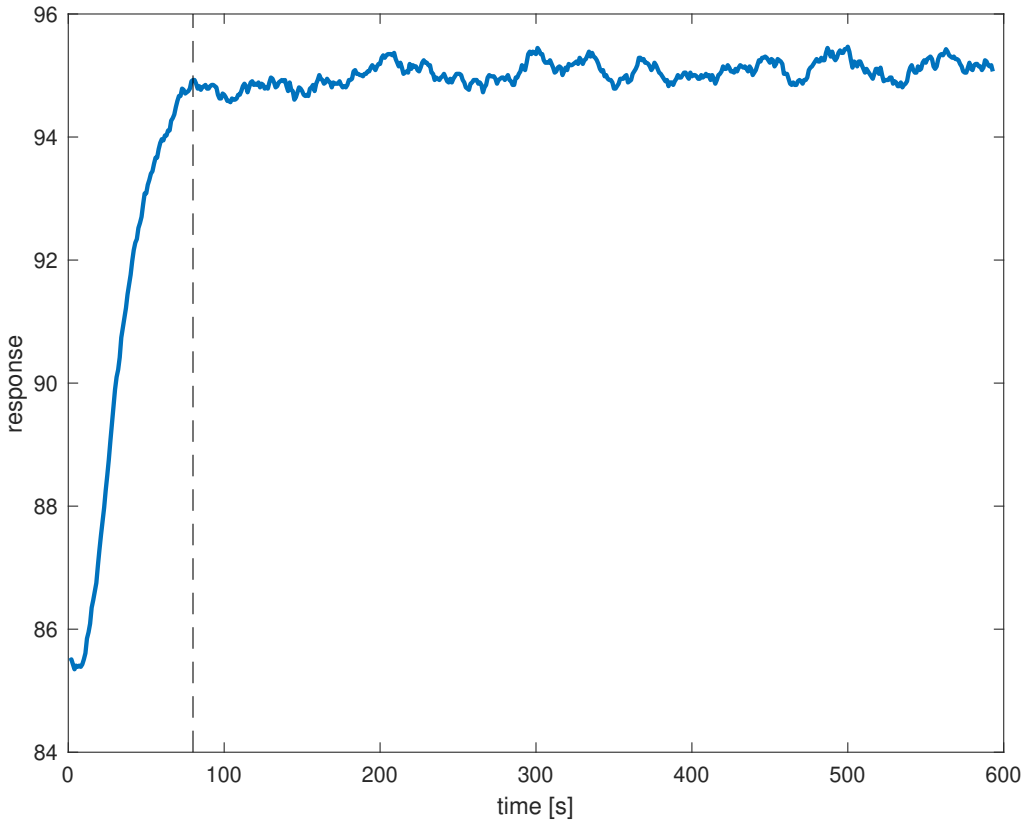

Figure 3: Temporal response on IMS channel 4 for 15% benzyl acetate with propylene glycol as diluent. The IMS reading stabilizes after approximately 80 seconds (dashed line).

In order to avoid ties in the majority vote, K should be odd (Duda et al., 2001, p. 183). However, even for an odd K, we might get a tie, for example, if the three nearest neighbors are from three different scents. In this classifier, in such a case we choose the label of the closest training samples as the label for the scent. This means that we use the nearest neighbor (NN), which is a special case of the KNN with $K = 1$, to break the tie.[3] The pseudo-code for classification with KNN is given in Algorithm 1. We discuss the algorithm details in the remainder of this section. For a more thorough discussion of KNN estimation, we refer the reader to (Duda et al., 2001, p. 174 ff.).

2.3. Choice of K and data preprocessing

For classification, the choice of K is crucial. In general, a large K is preferred as it reduces the effect of measurement noise on the classifier. However, a large K also weakens the boundaries between the scent classes, which could result in poorer classification performance. Using a small K has advantages. For example, if $N \to \infty$, then all K nearest neighbors will converge to the new sample $\mathbf{x}^{(\mathrm{us})}$ (Duda et al., 2001, p. 183). Thus, one could simply use $K = 1$ and could assume that the nearest neighbor is sufficiently close to the new sample. However, in real-world applications N is finite: in this case, a compromise between a reliable estimate for the scent's label, which favors large K, and all K neighbors close to $\mathbf{x}^{(\mathrm{us})}$, which favors a small K (Duda et al., 2001, p. 184). Typically, K is 3, 5 or 7 (Khan et al. (2002)).

Instead of using a fixed K for a test sample $\mathbf{x}^{(\mathrm{us})}$, optimal K values for each $\mathbf{x}^{(\mathrm{us})}$ can be determined. For example, Wang et al. (2006) propose the confident nearest neighbor rule, which is based on statistical confidence. Rather than setting K before classification, the user defines a confidence level p for the classifier's label. The classifier then adjusts K such that the label for $\mathbf{x}^{(\mathrm{us})}$ is believed to be correct with probability p or higher. Cheng et al. (2014) propose a KNN algorithm based on sparse learning. The method takes into account the correlation between samples and learns a specific test sample's K value by reconstructing it as a linear combination of training samples before determining the sample's label using KKNN. The drawback of these two and other methods, which also search for optimal K values, is that the optimization process is often time-consuming. Therefore, Zhang et

[3]Using weights based on the distances to the new sample for the K nearest neighbors is another option to avoid as it almost always results in a tie. Note that this weighted KNN differs from the weighted KNN presented by Voulgaris and Magoulas (2008).

Algorithm 1 Classification by KNN

Input: training data $\mathbf{X} = (x_{ij})$ and corresponding scent labels $\mathbf{s} = (s_i)$ with $i = 1, .., N$ and $j = 1, .., 14$, number of nearest neighbours K, window length w for sliding moving average (set $w = 1$ to use raw data), sample $\mathbf{x}^{(\text{us})}$ from unlabeled scent, percentage of total variance p explained by PCA features

Offline phase:

(i) Smooth training data: compute $\tilde{x}_{ij}(t)$ for w using (2) for all i and j

(ii) Standardise training data: compute $\{\mu_j\}_{j=1}^{14}$, $\{\sigma_j\}_{j=1}^{14}$, and $\bar{x}_{ij}$ for all i and j using (3)

(iii) Transform data by PCA (optional): compute $\mathbf{Y} = (y_{ij})$, $\boldsymbol{\mu}_{\text{PCA}}$, $\mathbf{C}$

(iv) Generate k-d trees for IMS and PCA-transformed data (optional)

Online phase:

(v) Smooth data from scent: compute $\tilde{x}_j^{(\text{us})}$ for w using (2) for all j if more than one sample is available

(vi) Standardise data from scent: compute $\bar{x}_j^{(\text{us})}$ for all i using (3), $\{\mu_j\}_{j=1}^{14}$, and $\{\sigma_j\}_{j=1}^{14}$

(vii) Transform data by PCA (optional): compute $\mathbf{y}^{(\text{us})} = (y_j^{(\text{us})})$ using (4), $\boldsymbol{\mu}_{\text{PCA}}$, and $\mathbf{C}$

(viii) Find K training samples closest to $\bar{\mathbf{x}}^{(\text{us})}$ based on IMS measurements, choose label $s^{(\text{us})}$ based on labels of the labels of the K nearest neighbours in $\mathbf{s}$ by majority vote

(viii) Find K training samples closest to $\mathbf{y}^{(\text{us})}$ based on PCA-transformed features that explain $p\%$ of the variance, choose label $s_{\text{PCA}}^{(\text{us})}$ based on labels of the labels of the K nearest neighbors in $\mathbf{s}$ by majority vote (optional)

Output: $s^{(\text{us})}$ and (optional) $s_{\text{PCA}}^{(\text{us})}$

al. (2018) propose an approach that builds a decision tree for predicting the optimal values of K for different test samples. The drawback of their method is that it requires a training stage, unlike KNN with a fixed K. Thus, we refrain from using these approaches and work with a fixed K for any test sample.

In order to reduce the effect of measurement noise on the classifier, instead of using a large K, we apply a sliding moving average (MA) to the training data and the scent's IMS samples. The reasons for using an MA are twofold. First, we are interested only in the time domain. Ideally, the reading on any channel should grow linearly to a stable level and then stay at this level until the eNose is switched off or no more scent is presented to it. Second, the ultimate goal is online classification, meaning we must constantly switch between taking measurements and classifying the scent based on these measurements. A sliding MA allows us to smooth the data online and classify the scent a few seconds after a new measurement is taken. The IMS measurement $x_{ij}(t)$ at time t is replaced with the MA smoothed:

$$\tilde{x}_{ij}(t) = \frac{1}{w} \sum_{\tau=-\frac{w-1}{2}}^{\frac{w-1}{2}} x_{ij}(t+\tau), \tag{2}$$

where w is the window length of the sliding MA, which is an odd natural number. If there are not enough samples before and/or after the ith sample with the same label (i.e., from the same scent) available, then we average over fewer samples and adjust w accordingly.

The KNN classifier uses the Euclidean distance for measuring the closeness between the new sample and the training samples. This is problematic, because the absolute values and the fluctuations in the IMS readings differ for the 14 channels. Therefore, we standardize the smoothed data by centering it and dividing it by the standard deviations of all measurements for any IMS channel. This means that for IMS channel j mean μ_j and standard deviation σ_j are computed over all training measurements on channel j. Then the normalized IMS measurement

$$\bar{x}_{ij} = \frac{\tilde{x}_{ij} - \mu_j}{\sigma_j}. \tag{3}$$

is computed. Samples from the scent are standardized accordingly using the means $\{\mu_j\}_{j=1}^{14}$ and standard deviations $\{\sigma_j\}_{j=1}^{14}$. The standardized sample is denoted as $\bar{\mathbf{x}}^{(\mathrm{us})}$.

Another crucial point that has to be taken care of in data preprocessing is the composition of the training database. Due to the use of majority voting for classifying a scent, for each (known) scent in the database, approximately the same number of training samples must be stored. Otherwise, one scent might dominate another scent in the database, causing the first scent's samples to be among the K closest samples to the new sample due only to the large number (Voulgaris and Magoulas (2008)). If the scent distribution is skewed, that is, that the number of samples of various scents varies markedly, weighted KNN (WKNN) could be used. In WKNN, the K closest training samples are weighted by their distance to the new sample.

2.4. KNN's search method

The computational complexity of KNN described above depends on the number of samples N in the training base and the dimensionality k of the samples. In this paper, $k = 14$ because measurements from the 14 IMS channels are used. For better classification accuracy, N should be large. If the aim is to classify a large variety of scents using the KNN classifier, then we need IMS samples from any of these scents, which would result in a large N and high computational complexity (Moreno-Seco et al. (2003)).

However, quick classification of a scent is desirable. Therefore, the computational complexity of KNN has to be reduced. Three different techniques can be used: computing partial distances, prestructuring, and editing the training samples (for details, see (Duda et al., 2001, p. 185 f.)).

For the presented KNN, we choose a prestructuring technique called k-dimensional trees (k-d trees aka multidimensional binary search trees), which was introduced in Bentley (1975). Note that k has no connection to K, and that generally $k \neq K$ will hold. The idea of the k-d tree search is to split the training data into subsets, use the binary k-d tree to address the new sample to a certain subset, and choose the nearest neighbors only from the training samples in this subset. A k-d tree with depth k splits the original data set into 2^k subsets. This means that the training database is split into $2^{14} = 16,384$ subsets, and the new sample is compared only with training samples from the subset to which the sample is addressed by the k-d tree.

One property of k-dimensional trees is that they are especially suited for low-dimensional, real-valued data, such as the IMS data. The major advantage, however, is that new nodes can be added to an existing k-d tree, which means that there is no need to retrain the whole tree when updating the training database with measurements from new scents. The algorithm

for adding new nodes to an existing k-d tree is described in detail in Bentley (1975). The major drawback is that k-d tree can miss the true nearest neighbors, because the k-d tree search is an approximate method. However, for a large N this method generally works well, because as $N \to \infty$ all nearest neighbors will converge to the new sample (Duda et al., 2001, p. 183).

2.5. Feature transformation and reduction

Although the data are already low-dimensional, techniques for reducing the dimensionality could be considered. Furthermore, we should address the problem that KNN can be fooled by irrelevant features, that is, IMS channels whose data are irrelevant for classification.[4]

A technique that addresses both issues at once is PCA. The object of PCA is to find a lower-dimensional representation that accounts for the variance of the features (Duda et al., 2001, p. 580). This new, improved representation of the full data space is provided by the eigenvectors of the covariance matrix that have the largest eigenvalues (Duda et al., 2001, p. 580 ff.). In general, the covariance matrix has only a few large eigenvalues, and the dimensions associated with the remaining, small eigenvalues contain only noise (Duda et al., 2001, p. 568). The dimensionality of the new data generated by PCA is chosen based on the value of the total variance in the original data that should be explained by the transformed features. For example, assume that at least 95% of the total variance should be explained. If the three transformed features with the highest eigenvalues explain 95% or more of the variance, then only these three features will be used for classification, and the rest can be discarded. This means that PCA can transform and reduce the features.

We use PCA to transform the observed IMS channel measurements, which most likely are correlated, into a set of new, artificial measurements that are linearly uncorrelated by orthogonal transformation. The drawback of PCA is that we end up with a set of new features that are difficult to interpret. This means that we cannot directly see on which IMS channels a certain scent is observed and on which only air or the diluent is detected.

Before classifying scents, PCA should be applied to samples in the training database that are in the offline phase. This requires us to be able to transform the IMS samples measured for the new scent into the same format as the

[4] An irrelevant channel is, for example, a channel whose reading is independent of the analyzed scent.

PCA-transformed training data. When transforming the training data with PCA, we obtain as side products the empirical means for all 14 channels $\boldsymbol{\mu}_{\mathrm{PCA}} = [\mu_1^{(\mathrm{PCA})} .. \mu_{14}^{(\mathrm{PCA})}]$ and the 14-by-14 matrix $\mathbf{C}$ that contains the principal component coefficients. A new standardized IMS sample $\bar{\mathbf{x}}^{(\mathrm{us})}$ from an unlabeled scent can then be transformed into a 14-dimensional PCA-transformed sample with

$$\mathbf{y}^{(\mathrm{us})} = (\bar{\mathbf{x}}^{(\mathrm{us})} - \boldsymbol{\mu})\mathbf{C}. \tag{4}$$

In the same way new IMS training data can be transformed into new PCA-transformed training data. This means, adding new scents to the training database does not require modifying the existing PCA-transformed data.

For classifying a scent using PCA-transformed data, we can use KNN. Depending on the percentage of total variance p that should be explained by the PCA features we use only data from the first n_{PCA} transformed features. This means the closeness between the new sample $\mathbf{y}^{(\mathrm{us})}$ and the ith PCA-transformed training sample $\mathbf{y}_i$ is computed with

$$d_{\mathrm{E}}(\mathbf{y}^{(\mathrm{us})}, \mathbf{y}_i) = \sqrt{\sum_{j=1}^{n_{\mathrm{PCA}}} \left(y_{ij} - y_j^{(\mathrm{us})}\right)^2}. \tag{5}$$

2.6. Comparison of KNN and alternative methods

KNN has three strengths. First, despite its simplicity it is very effective and can give very good results if the training data contain IMS readings similar to the reading of the scent. Second, KNN has strong consistent results. By increasing the number of training samples, the misclassification rate can be reduced, and K can be used to control the misclassification rate. Third, KNN enables us to constantly add new training samples without the need to retrain the classifier.

The two main drawbacks of KNN are its slowness for large datasets, because its classification time is linear to the size of the dataset, and its vulnerability to irrelevant features. In order to mitigate the influence of these two drawbacks on the presented KNN classifier, we use a k-d tree search to accelerate the search process. This way, we need to compare the new sample only to a subset of training samples. PCA removes the irrelevant features. It generates a new set of features that contribute to the variance in the data. By choosing, for example, features that explain 99% of the data's total variance irrelevant features can be completely removed.

An alternative is a multiclass SVM (M-SVM) classifier. These classifiers are accurate and in general do not overfit the data, which means that they can handle new data well. M-SVMs are fast classifiers, because they need to be trained only once, and afterward, the training data can be discarded. However, for this application we need to update the classifier once new training data are available. Another strength of M-SVMs is that they handle complex, nonlinear classification generally well, which could be beneficial for IMS data.

The weaknesses of M-SVMs are that the parameters are difficult to interpret, and its training and tuning are time-consuming. Thus, constantly updating the classifier for new training data is not recommended. In addition, M-SVMs can be slow when the training data contain large numbers of different scents, because the scents are separated pairwise by SVMs. This means for a classifier that has to distinguish m different scents the M-SVM consists of m or $\frac{m(m-1)}{2}$ SVMs, depending on the strategy.

ANNs could also be used for classification and perform well for nonlinear data with a large number of features. Furthermore, their classification time is independent on size of the training database. However, also for low-dimensional data, such as the 14-dimensional IMS measurements studied in this paper ANNs could be used. Neural network techniques have two challenges that need to be considered. First, the number of free parameters need to be chosen such that the network generalizes well. With too few parameters the training data is learned inadequately and with too many parameters the classifier will generalize poorly (Duda et al., 2001, p. 283). Second, they require large amount of data compared to the network size. The data sets studied in the next section would not suffice. For the application studied in this paper it is furthermore problematic that the network would need to be retrained every time the training database would be modified.

Other classification methods can be and have been used for scent classification. The interested reader is referred to, for example, the references in Section 1. These methods are not discussed in detail in this paper.

3. Classification results

This section assesses the classification performance of the KNN algorithm using IMS data and PCA-transformed data. In the first experiment three organic compounds are presented to the eNose using the olfactory display in Subsection 2.1, which ensures that only the organic compound and the

diluent are measured. In the second experiment seven food scent sources are presented to the eNose on a plate and in a sealed jar. In latter case also other scents present in the experimental facilities during the experiment are measured simultaneously with the food scents. by the eNose. This experiment was done in order to study how sensitive the developed classification system is to random distractions, which are common in real-world use. The measurement data for both experiments is available for download (Müller et al. (2018)).

In order to test the KNN classifier with a database that is as large as possible, all the data is used for the different runs of both experiments. This means that the training database is static during each run, and any classifier could be applied without the need to retrain it. However, the training databases for any two runs differ from each other, which would require retraining the classifier for any run. Furthermore, for later application in real world the training database has to be flexible. Therefore, it would be beneficial if the classifier does not need to be retrained any time the database is modified (e.g. a new scent is added).

3.1. Tests with key odor components of jasmine

In this subsection, the classifier is tested with the three key odor components of jasmine: benzyl acetate (BEA) using concentrations of 40 and 70 parts per million (ppm), cis-Jasmone (CIS) using concentrations of 39 and 69 ppm, and indole (IND) using concentrations of 4 and 7 ppm. For all three scents, propylene glycol (PG) is used as the diluent. BEA, CIS, and IND were purchased from Sigma Aldrich®. Their Chemical Abstract Service (CAS) Registry Numbers are 140-11-4 (BEA), 488-10-8 (CIS), and 120-72-9 (IND), respectively.

The data consist of 10 measurement sets for 10 min for each component-concentration (CC) combination, measured with a frequency of 1 Hz. The measurement sets for any specific CC combination were collected in the following way. After the first measurement set was collected, the eNose was cleaned by pumping only air for approximately 20 sec. Then the second measurement set was collected. The eNose was cleaned again, and the third set was collected, etc. Therefore, only the first measurement set of any SC combination contains data from the full transient phase. Sets 2 to 10 contain only the last fraction of the data from the transient phase, and we cannot say for certain how large this fraction is.

Before the classification test, the data are smoothed using a sliding moving average with window length $w = 11$ and normalized. As a compromise, $K = 3$ is used. For searching the three nearest neighbors, exhaustive and k-d tree searches are applied. For the test a setup inspired by cross-validation (CV) is used. In each run different training and test sets are used. For example, in run 1 all measurement sets with identifier (ID) 1 are used as test data and all sets with IDs 2 to 10 are used as training data. Then, in run 2 all measurement sets with ID 2 are used as test data and all sets with IDs 1 and 3 to 10 are used as training data, etc.

Fig. 4(a) shows the misclassification rates in percent, that is, the percentage of test samples that are classified incorrectly, in each of the ten runs for both search methods. The two search methods provide the same misclassification rate,[5] which was expected because for each CC combination 5 400 training samples are stored in the database. As mentioned in Subsection 2.4, large number of training samples ensure that k-d tree search works well, because as $N \to \infty$ all nearest neighbors will converge to the new sample (Duda et al., 2001, p. 183). For classifying a test sample k-d tree search requires, on average, only 13.6% of the computation time of exhaustive search.

The misclassification rate in run 1 for both search methods is 14.6%, but in runs 2 to 10, less than 3% of the test samples are misclassified. The slightly higher misclassification rates in run 10 can be explained by the fact that the pumps of the olfactory display in this run are closed a few seconds before the end. After the pumps are closed, the IMS readings change instantly. The reason for the poor performance of the KNN in the first run is that the test sets (i.e. sets with ID 1) contain measurements from the transient phase, while the training sets lack data from this phase. For example, if we consider only test data samples that are taken in the first 3 min of the first cycles, which corresponds to the transient phase, then the misclassification rate is 42.0% (for data from the first 2 min and from the first minute, the misclassification rates are 54.4% and 78.1%, respectively). For the test samples that are taken after the first 3 min (in the stable phase), the misclassification rate is only 2.7%, which is close to the results for runs 2 to 10.

It can be concluded that it is important to have training data from the

[5] The k-d tree might still miss some or all of the true nearest neighbors, but that does not change the label the KNN classifier yields.

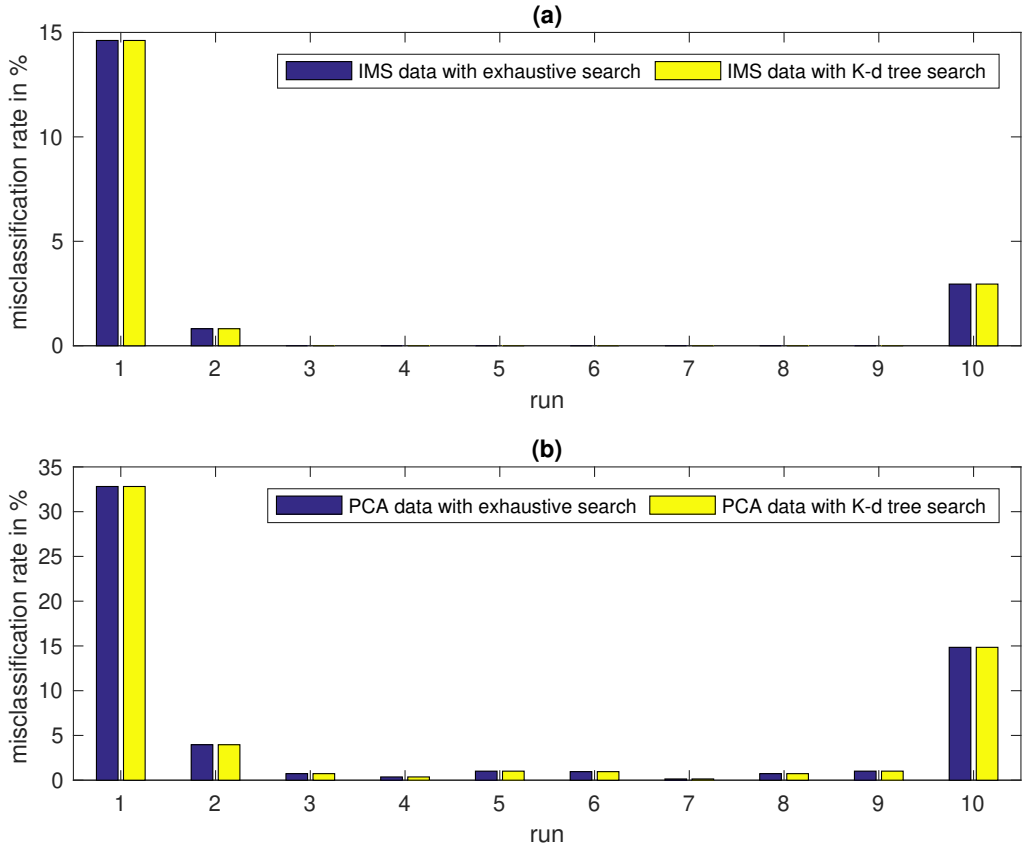

Figure 4: Misclassification rates in percent for the test in Subsection 3.1. 3.1. (a) shows the rates for the IMS measurements, and (b) shows the rates for the first two PCA-transformed components for classifying scents.

transient and stable phases to classify scents based on IMS readings from these two phases. Furthermore, this test shows that the scent concentration has a considerable influence on the IMS readings, and that the classifier distinguishes two scents even if their IMS values differ only slightly.

Classification with PCA-transformed data

Fig. 4(b) shows the percentage of misclassified test samples in each of the ten runs for the exhaustive and k-d tree searches, but this time, PCA-transformed data are used to classify the scents. In this test, the requirement is that the PCA features explain 95% of the total variance in the data. Using this value, each run ends up with the first two principal components, meaning that only 1/7 of the number of features from the test with IMS data is used. This reduces the computation time for classification considerably. Using PCA-transformed data with an exhaustive search requires 34.5% of the computation time required for the exhaustive search used on IMS data; using PCA-transformed data with the k-d tree search requires only 3.1% of the computation time using IMS data with an exhaustive search and 22.6% of the computation time using IMS data with a k-d tree search.

The reduction comes at the cost of higher misclassification rates in all ten runs. However, for runs 2 to 10 the classifier still performs well. In this run, the test data contain measurements from the transient and stable phases while the training data contain data only from the stable phase, which leads to the poor performance.

Thus, using PCA for transforming and reducing features is advisable as it helps to reduce the computation time considerably. Especially for real-world applications with considerably larger training datasets, PCA should be used to ensure reasonably fast classification of scents.

3.2. Tests with food scents

For these tests, seven sources of scents are presented to the eNose on a plate and in a sealed jar. The sources of the scents are cinnamon (crushed cinnamon sticks from Indonesia), coffee (crushed, non-brewed Roasted Arabica coffee beans), grape (red grape, Sharad from India), lemon peel (grated peel of a ripe lemon), pineapple (fresh and ripe pineapple), strawberry (sliced, fresh Marilyn strawberry from Spain), and vanilla (sliced dried vanilla fruit from Indonesia).

Each source is first put on a plate on a table approximately 2–3 cm from the eNose. For the second set, each source is placed in a sealed jar

and presented to the eNose by manually opening a valve that controls warm airflow. In both cases, five measurement sets for 5 min measurement frequency is again 1 Hz) are taken for each source of scent. In both sets, the scent source is approximately 5 ml.[6] The gap between taking two consecutive measurement sets for a source is set at 3 min to ensure that each set contains data from the transient phase. Because the scent reaches the IMS channels faster than when using the olfactory display in Subsection 3.1, the transient phase is considerably shorter in this test. In the sets tested within this subsection, it is approximately 20 to 30 sec long.

Visual inspection of the IMS readings for scents measured on a plate shows no significant variation for different scents. Thus, it could be expected that the classifier has more difficulty labeling correctly scents that are measured on a plate than correctly labeling scents that are measured from a closed jar.

A similar setup as described in Subsection 3.1 is used for this test. For each of the 14 scents (seven sources of scent in two presentation methods), one set is chosen as the test set and the remaining four sets as the training set. This is repeated five times, so that each set is the test set one time and part of the training set four times. As before, the data are smoothed using a sliding moving average with window length $w = 11.$ and normalized, and K is set to 3. Here a test scheme similar to the one in 3.1 and based on 5-fold CV is used.

Fig. 5(a) shows the percentage of misclassified test samples in each of the five runs for the exhaustive and k-d tree searches. Again, the two search methods provide the same results, meaning that using the k-d tree search is preferred due to its lower computational demand (on average, over the five runs 12.1% of the exhaustive search's computation time). The reason for the similar performance are as in the previous test the large number of training samples per scent (1 200 for any presentation method) and the low dimensionality of the data.

As in the test described in Subsection 3.1, the misclassification rate in the first run is considerably higher than in the remaining runs, although here all five sets contain data from the transient phase. However, when measuring data for sets 2 to 5, which are used as training sets in run 1, the

[6]The amount was measured using a teaspoon (1 US teaspoon equals 5 ml). We did not use water, alcohol, or anything to dilute the odorants.

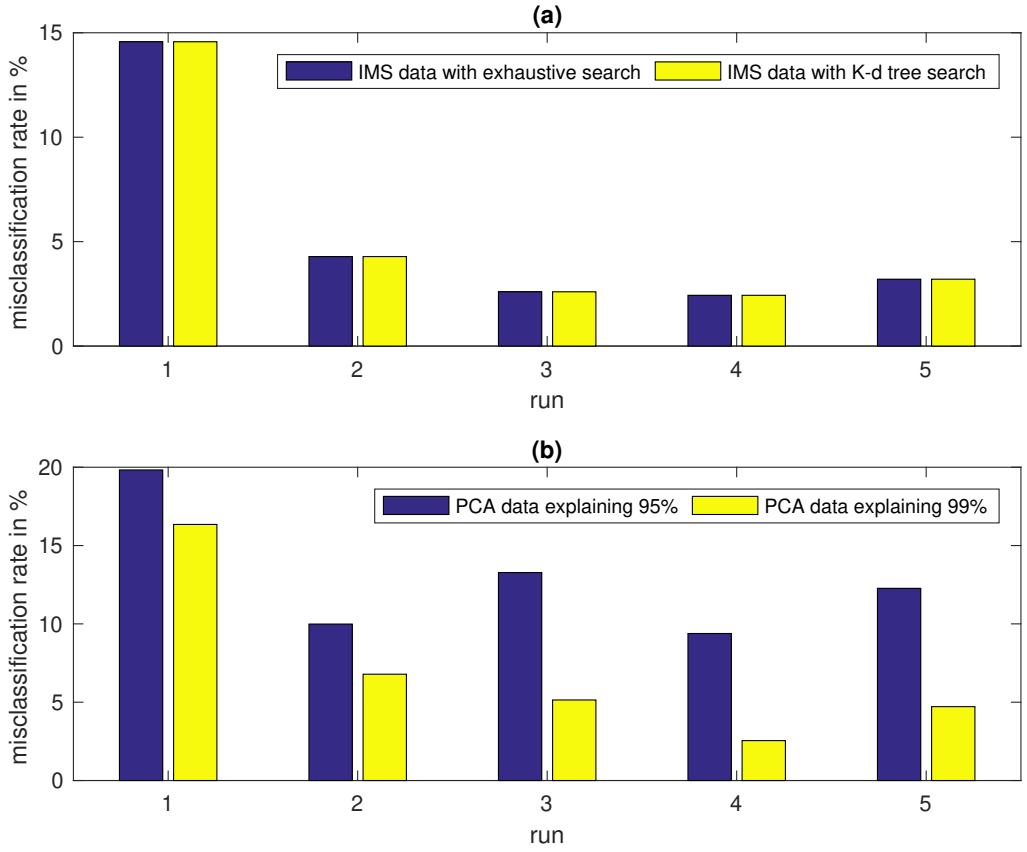

Figure 5: Misclassification rates in percent for the test in Subsection 3.2. (a) shows the rates for the IMS measurements and (b) shows the rates for the first three PCA-transformed components, which explain at least 95% of the total variance, and the first four PCA-transformed components, which explain at least 99% of the total variance, for classifying scents.

IMS channels were not completely cleansed.[7] If we consider, for example, only test data samples that were taken in the first 3 min of the first cycles, which corresponds to the transient phase, then the misclassification rate is 20.1%, compared to 14.6% when we use all the data from the first cycles (for data from the first 2 min and from the first minute the misclassification rates are 22.8% and 21.7%, respectively). For test samples that were taken after the first 3 min (mainly in the stable phase), the misclassification rate is only 6.0%, which is close to the results for runs 2 to 5.

Looking at the confusion matrices of the five runs, some typical misclassification patterns can be noticed. For example, in each run a considerable number of samples from strawberry on a plate are misclassified as coffee on a plate and vice versa. In run 1, in addition, 44.1% of all samples from grape in a jar are misclassified as lemon peel in a jar, and 35.0% of all samples from grape on a plate are misclassified as vanilla on a plate, which explains the considerably higher misclassification rate in run 1.

Table 1 contains a summary of all major misclassifications (i.e., at least 5% of all samples from one scent are misclassified as another scent) when we use the exhaustive search. Switching to the k-d tree search has no significant effect on those numbers. The majority of the misclassification happens for scents that are presented to the eNose on a plate. These results confirm what we expected when we looked at the IMS readings from the 14 scents. If a scent is presented in a sealed jar to the eNose, then the classifier has a higher chance of classifying the scent correctly compared to when the scent is presented on a plate. Nevertheless, the results for both presentation methods are encouraging.

From Table 1, it can be seen that a large portion of the strawberry and coffee samples presented on a plate are misclassified while no significant number of cinnamon samples is misclassified. When we map the 14-dimension IMS measurements for the five measurement sets of cinnamon, coffee, and strawberry on a plate to two dimensions we can get an idea why that is. For cinnamon, the samples from the five sets are grouped into one big cluster (the top of Fig. 6). In contrast, for coffee (the middle of Fig. 6) and strawberry (the bottom of Fig. 6) on a plate the plots show five distinctive clusters. This means that for coffee and strawberry on a plate the IMS readings vary

[7] Cleansing the channels completely after measuring one set would require impractical long breaks between taking measurements for the different sets.

Table 1: Major misclassifications in test in Subsection 3.2 using exhaustive search method and IMS data. Column *misclassifications* shows the percentage of samples from *true scent* that are misclassified as *predicted scent*.

run	misclassifications	true scent	predicted scent
1	44.1%	grape (jar)	lemon peel (jar)
	35.0%	grape (plate)	vanilla (plate)
	78.1%	strawberry (plate)	coffee (plate)
	15.5%	strawberry (plate)	pineapple (plate)
	24.2%	pineapple (plate)	lemon peel (plate)
2	40.7%	strawberry (plate)	coffee (plate)
	11.1%	coffee (plate)	vanilla (plate)
3	22.9%	coffee (plate)	strawberry (plate)
4	31.7%	coffee (plate)	strawberry (plate)
5	38.7%	coffee (plate)	strawberry (plate)
	5.1%	pineapple (jar)	grape (jar)

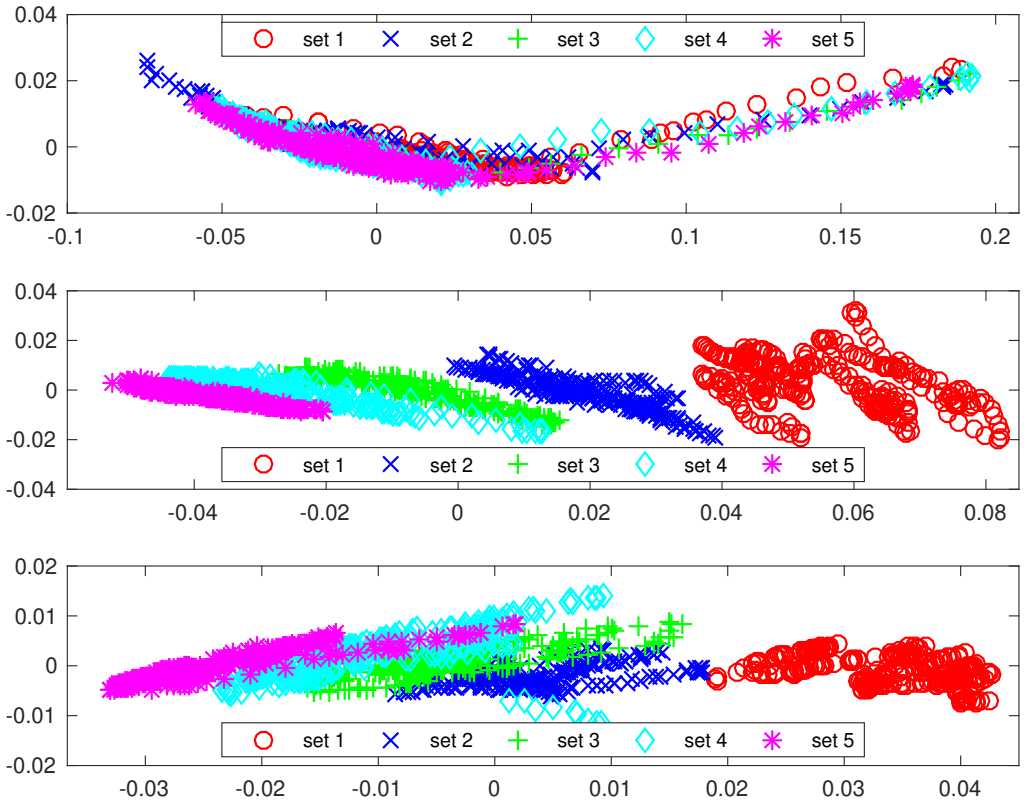

Figure 6: Metric multidimensional scaling of IMS data from five measurement sets of cinnamon (top), coffee (middle), and strawberry (bottom) presented to the eNose on a plate.

more than those for cinnamon on a plate. Visual inspection of the channel readings revealed that for coffee, strawberry, and pineapple the initial IMS readings (i.e., the first few measurements of each set) vary markedly for the five measurement sets, while for cinnamon the readings are similar for each set. Furthermore, we notice that for coffee, strawberry, and pineapple the initial IMS readings of one measurement set are usually close to the last IMS readings of the previous measurement set, which suggests that the 3-min gap was not long enough to cleanse the IMS channels. One reason could be that furaneol (aka strawberry furanone aka pineapple ketone) is a shared odor component for strawberry and pineapple. When we analyzed furaneol with the olfactory display, we noticed that it stayed in the tubes for a long time. Thus, it is likely that furaneol might have prevented the full cleansing of the IMS channels for strawberry and pineapple.

Classification with PCA-transformed data

Fig. 5(b) shows the percentages of misclassified test samples in each of the five runs for the exhaustive search applied to the PCA-transformed data. The results for the k-d tree search are similar and therefore, are omitted. When we use data only from the first three principal components, which explains at least 95% of the total variance of the data (blue bars in the figure), the misclassification rates increase markedly to unsatisfying levels of 10% to 20%. However, by also using data from the fourth principal component, which together with the first three components explains at least 99% of the variance, the misclassification rates can be reduced considerably (yellow bars in the figure). The misclassification rates are similar to those when we use IMS data, and at the same time, the classification process is speeded up because only 2/7 of the number of features from the test with IMS data are used.

A summary of all major misclassifications when we use data from the first four principal components and exhaustive search is presented in Table 2. The most stable misclassification patterns are classifying coffee on a plate as strawberry on a plate and vice versa, and strawberry on a plate as vanilla on a plate and vice versa. One possible reason for these misclassifications is that the three food scents have odor components with chemical similarities. However, this question remains open for further research.

Influence of K on the misclassification rate

As discussed in Subsection 2.3, the choice of K is a compromise between reliable estimates for the labels of unlabeled scents and the closeness of all

Table 2: Major misclassifications in test in Subsection 3.2 using exhaustive search method and data from first 3 principal components. Column *misclassifications* shows the percentage of samples from *true scent* that are misclassified as *predicted scent*.

run	misclassifications	true scent	predicted scent
1	12.5%	cinnamon (jar)	strawberry (jar)
	47.5%	grape (jar)	lemon peel (jar)
	35.0%	grape (plate)	vanilla (plate)
	37.0%	pineapple (plate)	lemon peel (plate)
	40.4%	strawberry (plate)	coffee (plate)
	6.7%	strawberry (plate)	pineapple (plate)
	48.2%	strawberry (plate)	vanilla (plate)
2	33.7%	coffee (plate)	vanilla (plate)
	59.9%	strawberry (plate)	coffee (plate)
3	20.5%	coffee (plate)	strawberry (plate)
	10.1%	coffee (plate)	vanilla (plate)
	26.6%	grape (jar)	pineapple (jar)
	7.4%	strawberry (plate)	coffee (plate)
4	25.9%	coffee (plate)	strawberry (plate)
	9.8%	grape (jar)	pineapple (jar)
5	43.8%	coffee (plate)	strawberry (plate)
	5.7%	pineapple (jar)	grape (jar)
	13.5%	vanilla (plate)	coffee (plate)

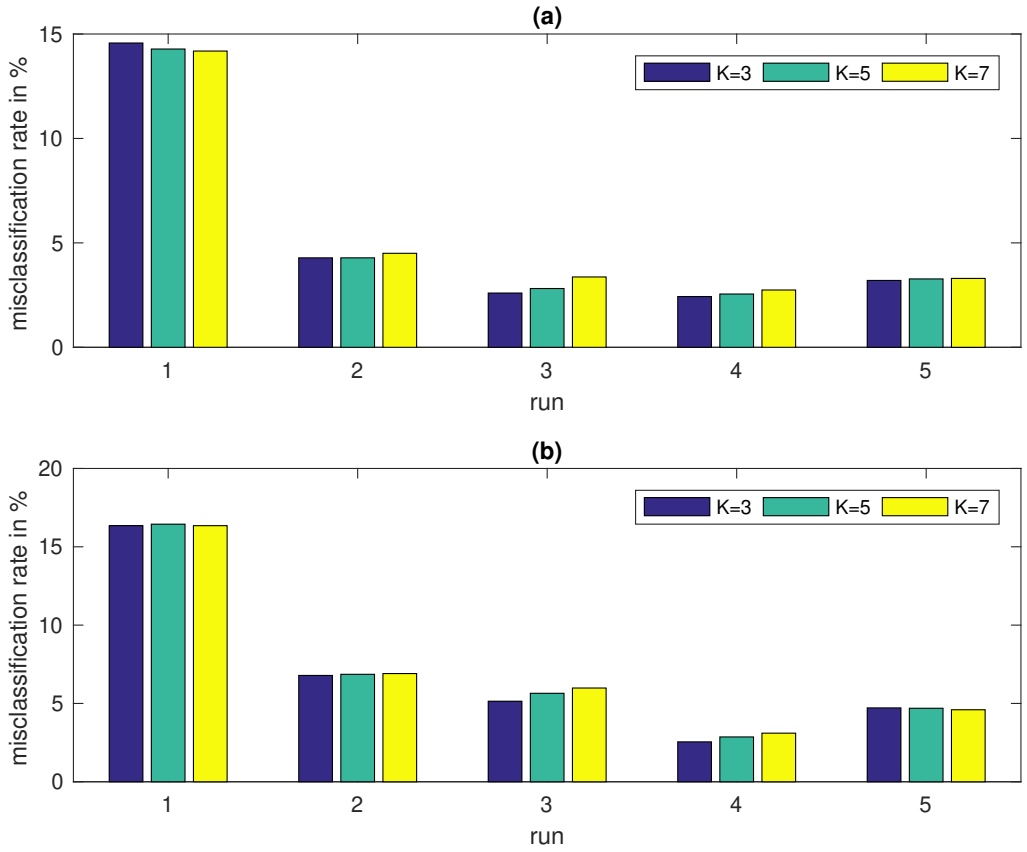

Figure 7: Misclassification rates in percent for test in Subsection 3.2 with varying K. (a) shows the rates when IMS measurements are used with the exhaustive search, and (b) shows the rates when first four PCA-transformed components are used, which explain at least 99% of the total variance with the exhaustive search.

K neighbors to $\mathbf{x}^{(\mathrm{us})}$ due to the finite number of training samples. To check how different values of K influence the misclassification rates, we repeat the tests of this subsection and set K to 5 and 7.

Fig. 7 shows the misclassification rates for K set to 3, 5, and 7 with the exhaustive search on the IMS data (upper plot) and on the first four principal components (lower plot), which explain 99% of the total variance. Again, the results using the k-d tree search are similar to those for the exhaustive search and therefore, are omitted due to space restrictions. From the figure, we can see that the influence of K, for the three options we tested, on the misclassification rate is marginal. Thus, we can set K at 3, because it is

slightly faster than the larger Ks. For exhaustive search on IMS data, K set to 5, on average, uses 102.6% and K set to 7, on average, uses 107.9% of the running time of K set to 3, because during the search more potential nearest neighbors have to be stored and compared.

4. Discussion and Conclusions

Our goal is to develop a scent classification algorithm for scents that will enable, for example, product quality control or digital transfer of scents in space and time. This system needs to consist of a scent analyzer, a scent classifier, and a scent synthesizer. Such a classifier needs to be reasonably fast and accurate, to update the training database without requiring classifier retraining, and be easily understandable for user friendliness.

This paper provided a detailed analysis of the scent classifier: a K nearest neighbors algorithm using ion-mobility spectrometry data from an electronic nose. In the classifier's basic form, i.e. when using the exhaustive search on IMS data, the KNN classifier performed well in the tests reported in Section 3. In most cases, KNN classified the scents correctly, regardless of how the scents were presented to the eNose.

In order to solve the problem of slow classification for large training databases, the k-dimensional tree search was applied. The results showed that this approach required only 12-13% of the time needed for the exhaustive search. In all tests, the k-d tree search yielded exactly the same misclassification rates, meaning that the reduction in search time did not attenuate the classification performance. Another advantage of the k-d tree search is that the tree can be generated in the offline phase, that is, before classifying a scent, and the tree can be updated when new samples are added to the training database instead of constructing a new tree.

The effect of principal component analysis on the misclassification rate was also tested. PCA revealed that the channel readings of the ion-mobility spectrometry-based eNose are dependent. Using only data from the first few principal components, which explained 99% of the total variance in the IMS data, only a minor increase in the misclassification rate was observed. At the same time, the search time can be reduced considerably by using PCA-transformed data instead of IMS data. Details were presented in Subsection 3.1. By using (4) IMS measurements from new scents can be transformed into PCA measurements and simply be added to the existing PCA training database, meaning that no retraining is required.

Finally, the influence of three suitable alternatives for K on the misclassification rate and the running time were tested. No significant influence of $K = \{3, 5, 7\}$ on either the misclassification rate or the running time was found.

Conclusion

Thus, we can conclude that the proposed KNN algorithm can be recommended for classifying scents based on IMS readings from an eNose, which is one of the main contributions of this paper. However, the data should first be smoothed, for example, with a sliding moving average to filter out noise in the data, and then normalized. In addition, the results show that by using k-dimensional tree search and principal component analysis the classification process can be tremendously accelerated. Furthermore, it is advisable to consider methods for reducing the size of the training database (see, e.g., Mainar-Ruiz and Perez-Cortes (2006) and references therein for an overview). Which of these methods works best for a database that is regularly updated by adding measurements from new scents needs further research. For example, it has to be ensured that from each scent the same or at least a similar amount of training data exists in the database to avoid the problem of skewed class distributions.

Outlook

Future research should consider how to reduce the misclassification rate further. Especially when a scent on a plate was presented to the eNose, the IMS readings often differed only slightly for various scents, which made it difficult for the KNN to classify the scent based on one IMS measurement or its PCA-transformed equivalent. One method we will in-vestigate is the use of sequences of IMS measurements over several seconds or minutes, instead of (smoothed) single measurements to classify scents. The aim is to check whether the temporal behavior of IMS readings for different scents follows different patterns, which can be used for classification. Furthermore, we will study the use of fuzzy KNN. Fuzzy KNN classifiers have been successfully applied, for example, for cardiac arrhythmia classification (Castillo et al. (2012)). Instead of a simple label, fuzzy KNN provides probabilities for all potential labels, and thus, information on how trustworthy the labels are. In addition, a conditional KNN classifier could be developed. For instance, if additional information on the scent source is available (e.g., "The scent source is a fruit"), then the nearest neighbors need to be searched only in a subset of the training set. In the example, only samples from fruits should

be considered in the search for K nearest neighbors.

In order to make the classification more generalizable IMS-based eNoses will be studied for device heterogeneity (aka signal shift). Device heterogeneity describes the phenomenon that two devices (even of the same brand and model) yield different results in identical sensing conditions (see e.g. Zhang et al. (2017)). If IMS-based eNoses indeed suffer from device heterogeneity, then device (aka shift) calibration methods have to be used to make the results from multiple eNoses comparable. For a short overview on device calibration methods the reader is referred to Vaupel et al. (2010); Haeberlen et al. (2004); Laoudias et al. (2012); Koski et al. (2010); Zhang et al. (2017) and references therein.

Furthermore, signal drift should be investigated. Although IMS sensors do not age (i.e. there responses do not get weaker), their readings depend on environmental factors, such as temperature and humidity. Therefore, the use of drift compensation methods will be studied. These methods can be divided into 1) component correction methods, 2) adaptive methods, and machine learning methods (see e.g. Zhang and Zhang (2015) and references therein).

For developing the scent transfer system mentioned in the introduction, it has to be studied how many scents can be distinguished by the KNN classifier using only IMS measurements. This number will be limited. However, it is important to note that only a limited number of scents need to be stored in the training database, because humans have difficulties to distinguish between, for example, the scent of jasmine oil and the scent of a mixture of its three key odor components that were tested in this paper (Surakka et al. (2016)). Furthermore, the final scent detection system could be enhanced by using additional chemical sensors.

Classification of scents has various real-world applications in areas such as the food industry, where classification includes detecting aroma compounds in dairy products and checking the quality of grains, eggs, meat, fish, and seafood (Wilson and Baietto (2009) and Baldwin et al. (2011) and references therein). There is even evidence that our approach can be used to distinguish different rooms in indoor localization (Müller et al. (2017)). Considering a wider and more ambitious perspective of digitizing human senses, the classification of odors is a necessary step required to develop systems that imitate human perception of scents. The present results showed promise to develop the science of odor further. As the famous scientist Alexander Graham Bell said in 1914: "But until you can measure their likeness and

differences, you can have no science of odor. If you are ambitious to find a new science, measure a smell." This is where we are heading.

Acknowledgment

This research was jointly carried out at Tampere University of Technology and University of Tampere, and was financially supported by both universities, the Academy of Finland (grant numbers 295432, 295433 and 295434) and the Finnish Cultural Foundation.

References

Abidin, T., Perrizo, W., April 2006. SMART-TV: a fast and scalable nearest neighbor based classifier for data mining. In: Proceedings of the 2006 ACM symposium on Applied computing (SAC '06). pp. 536–540.

Baldwin, E. A., Bai, J., Plotto, A., Dea, S., May 2011. Electronic noses and tongues: Applications for the food and pharmaceutical industries. Sensors 11 (5), 4744–4766.

Bentley, J. L., September 1975. Multidimensional binary search trees used for associative searching. Communications of the ACM 18 (9), 509–517.

Castillo, O., Melin, P., Ramírez, E., Soria, J., February 2012. Hybrid intelligent system for cardiac arrhythmia classification with fuzzy K-Nearest Neighbors and neural networks combined with a fuzzy system. Expert Systems with Applications 39 (3), 2947–2955.

Cheng, D., Zhang, S., Deng, Z., Zhu, Y., Zong, M., 2014. kNN algorithm with data-driven k value. In: Proceedings of the 10th International Conference on Advanced Data Mining and Applications (ADMA 2014). pp. 499–512.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification, 2nd Edition. Wiley-Interscience.

Edris, A. E., Chizzola, R., Franz, C., April 2008. Isolation and characterization of the volatile aroma compounds from the concrete headspace and the absolute of Jasminum sambac (L.) Ait.(Oleaceae) flowers grown in Egypt. European Food Research and Technology 226 (3), 621–626.

Environics, June 2017. Chempro 100i.
URL http://www.environics.fi/product/chempro100i/

Gallego, A. J., Calvo-Zaragoza, J., Valero-Mas, J. J., Rico-Juan, J. R., 2018. Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation. Pattern Recognition 74, 531–543.

García-Ruíz, M. Á., Santana-Mancilla, P. C., August 2013. Design, evaluation and impact of educational olfactory interfaces. In: AMCIS.

Haeberlen, A., Flannery, E., Ladd, A. M., Rudys, A., Wallach, D. S., Kavraki, L. E., September-October 2004. Practical robust localization over large-scale 802.11 wireless networks. In: MobileCom'04. Philadelphia, PA, USA.

Khan, M., Ding, Q., Perrizo, W., May 2002. K-nearest neighbor classification on spatial data streams using P-trees. In: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD '02). pp. 517–528.

Kiani, S., Minaei, S., Ghasemi-Varnamkhasti, M., March 2016. Application of electronic nose systems for assessing quality of medicinal and aromatic plant products: A review. Journal of Applied Research on Medicinal and Aromatic Plants 3 (1), 1–9.

Koski, L., Perälä, T., Piché, R., September 2010. Indoor positioning using wlan coverage area estimates. In: 2010 International Conference on Indoor Positioning and Indoor Navigation (IPIN).

Längkvist, M., Coradeschi, S., Loutfi, A., Rayappan, J. B. B., January 2013. Fast classification of meat spoilage markers using nanostructured ZnO thin films and unsupervised feature learning. Sensors 13 (2), 1578–1592.

Laoudias, C., Piché, R., Panayiotou, C. G., November 2012. Device signal strength self-calibration using histograms. In: 2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN). Sydney, NSW, Australia.

Lin, H., Yan, Y., Zhao, T., Peng, L., Zou, H., Li, J., Yang, X., Xiong, Y., Wang, M., Wu, H., October 2013. Rapid discrimination of Apiaceae plants

by electronic nose coupled with multivariate statistical analyses. Journal of Pharmaceutical and Biomedical Analysis 84, 1–4.

Loutfi, A., Coradeschi, S., Mani, G. K., Shankar, P., Rayappan, J. B. B., 2015. Electronic noses for food quality: A review. Journal of Food Engineering 144, 103–111.

Mailloa, J., Ramírez, S., Triguero, I., Herrera, F., February 2017. kNN-IS: An iterative Spark-based design of the k-nearest neighbors classifier for big data. Knowledge-Based Systems 117, 3–15.

Mainar-Ruiz, G., Perez-Cortes, J.-C., August 2006. Approximate nearest neighbor search using a single space-filling curve and multiple representations of the data points. In: The 18th International Conference on Pattern Recognition (ICPR'06). Vol. 2. pp. 502–505.

Mamat, M., Samad, S. A., Hannan, M. A., June 2011. An electronic nose for reliable measurement and correct classification of beverages. Sensors 11 (6), 6435–6453.

Martín, Y. G., Oliveros, M. C. C., Pavón, J. L. P., Pinto, C. G., Cordero, B. M., December 2001. Electronic nose based on metal oxide semiconductor sensors and pattern recognition techniques: characterisation of vegetable oils. Analytica Chimica Acta 449 (1–2), 69–80.

Moreno-Seco, F., Micó, L., Oncina, J., January 2003. A modification of the LAESA algorithm for approximated k-nn classification. Pattern Recognition Letters 24 (1–3), 47–53.

Müller, P., Ali-Löytty, S., Lekkala, J., Piché, R., October 2017. Indoor localisation using aroma fingerprints: A first sniff. In: 14th Workshop in Positioning, Navigation and Communication (WPNC'17). pp. 1–5.

Müller, P., Salminen, K., Nieminen, V., Kontunen, A., Karjalainen, M., Isokoski, P., Rantala, J., Savia, M., Väliaho, J., Kallio, P., Lekkala, J., Surakka, V., April 2018. Dataset for Müller et al. - "scent classification by K nearest neighbors using ion-mobility spectrometry".
URL http://urn.fi/urn:nbn:fi:csc-kata20180418151056882791

Nakamoto, T., Nimsuk, N., Wyszynski, B., Takushima, H., Kinoshita, M., Cho, N., October 2008. Reproduction of scent and video at remote site

using odor sensing system and olfactory display together with camera. In: IEEE Sensors. pp. 709–802.

Robin, O., Alaoui-Ismaïli, O., Dittmar, A., Vernet-Maury, E., January 1999. Basic emotions evoked by eugenol odor differ according to the dental experience. A neurovegetative analysis. Chemical Senses 24 (3), 327–335.

Seregély, Z., Novák, I., 2015. Evaluation of the signal response of the electronic nose measured on oregano and lovage samples using different methods of multivariate analysis. Acta Alimentaria 34 (2), 131–139.

Singh, S., L.Hines, E., Gardner, J. W., January 1996. Fuzzy neural computing of coffee and tainted-water data from an electronic nose. Sensors and Actuators B: Chemical 30 (3), 185–190.

Surakka, V., Lekkala, J., Levon, K., Kallio, P., Salminen, K., Nieminen, V., Karjalainen, M., Väliaho, J., Rantala, J., Kontunen, A., Savia, M., December 2016. From electrical scent analysis to digital scent production. In: 3rd World Congress of Digital Olfaction Society. http://urn.fi/urn:nbn:fi:csc-kata20180417150206263284, Tokyo, Japan.

Tang, K.-T., Chiu, S.-W., Pan, C.-H., Hsieh, H.-Y., Liang, Y.-S., Liu, S.-C., October 2010. Development of a portable electronic nose system for the detection and classification of fruity odors. Sensors 10 (10), 9179–9193.

Utriainen, M., Kärpänoja, E., Paakkanen, H., August 2003. Combining miniaturized ion mobility spectrometer and metal oxide gas sensor for the fast detection of toxic chemical vapors. Sensors and Actuators B: Chemical 93 (1–3), 17–24.

Vaupel, T., Seitz, J., Kiefer, F., Haimerl, S., Thielecke, J., September 2010. Wi-Fi positioning: System considerations and device calibration. In: 2010 International Conference on Indoor Positioning and Indoor Navigation (IPIN). Zurich, Switzerland.

Vernet-Maury, E., Alaoui-Ismaïli, O., Dittmar, A., Delhomme, G., Chanel, J., 1999. Basic emotions induced by odorants: a new approach based on autonomic pattern results. Journal of the Autonomic Nervous System 75 (2), 176 – 183.

Voulgaris, Z., Magoulas, G. D., February 2008. Extensions of the k nearest neighbour methods for classification problems. In: Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications (AIA '08). pp. 23–28.

Wang, J., Neskovic, P., Cooper, L. N., March 2006. Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. Pattern Recognition 39 (3), 417–423.

Wilson, A. D., Baietto, M., 2009. Applications and advances in electronic-nose technologies. Sensors 9 (7), 5099–5148.

Yang, Z., Dong, F., Shimizu, K., Kinoshita, T., Kanamori, M., Morita, A., Watanabe, N., June 2009. Identification of coumarin-enriched Japanese green teas and their particular flavor using electronic nose. Journal of Food Engineering 92 (3), 312–316.

Zhang, L., Liu, Y., Deng, P., July 2017. Odor recognition in multiple e-nose systems with cross-domain discriminative subspace learning. IEEE Transactions on Instrumentation and Measurement 66 (7), 1679–1692.

Zhang, L., Tian, F., July 2014. Performance study of multilayer perceptrons in a low-cost electronic nose. IEEE Transactions on Instrumentation and Measurement 63 (7), 1670–1679.

Zhang, L., Zhang, D., July 2015. Domain adaptation extreme learning machines for drift compensation in e-nose systems. IEEE Transactions on Instrumentation and Measurement 64 (7), 1790–1801.

Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R., 2018. Efficient kNN classification with different numbers of nearest neighbors. IEEE Transactions on Neural Networks and Learning Systems 29 (5), 1774–1785.

Zheng, S., Ren, W., Huang, L., February 2015. Geoherbalism evaluation of radix angelica sinensis based on electronic nose. Journal of Pharmaceutical and Biomedical Analysis 105, 101–106.