TECHNISCHE
UNIVERSITÄT
DRESDEN

Development of an automated data analysis process for measurement data from multiple sources for a new e-nose sensor

Timo Land

Diploma thesis

Submitted on: April 4, 2022

Statement of authorship

I hereby certify that I have authored this Diploma thesis entitled *Development of an automated data analysis process for measurement data from multiple sources for a new e-nose sensor* independently and without undue assistance from third parties. No other than the resources and references indicated in this thesis have been used. I have marked both literal and accordingly adopted quotations as such. During the preparation of this thesis I was only supported by the following persons:

Dr Robert Biele
Dr Leif Riemenschneider

Additional persons were not involved in the intellectual preparation of the present thesis. I am aware that violations of this declaration may lead to subsequent withdrawal of the degree.

Dresden, April 4, 2022

Timo Land

Contents

1 Introduction

In the 20th century, the digital capture of most human senses was achieved. Vision with cameras, hearing with microphones, touch with pressure gauges. Machine Learning techniques were successfully used to extract information from these sensors. In the field of electronic noses (e-nose) researchers try to digitize the sense of smell, or *olfaction*.

More recently, progress in the field of ML improved the information extraction from these sensors immensely. For example, Deep Neural Networks (DNNs) replaced and extended previous methods in the fields of computer vision[1], [2] and speech recognition[3], [4].

These advances have been mostly achieved with progressively more complex ML models being trained on datasets of increasing sizes. Less focus has been given to the quality of the data used for training[5]. However, gathering large datasets with electronic noses can be difficult and complex models do not necessarily offer the best performance. Furthermore, insight into the data of a e-nose system is important for development of the system and give feedback about the feasibility of different applications of said system. Therefore, a focus on *data quality* can help solving challenges of e-nose systems.

In this thesis an automated process for analysis of data quality for a new e-nose sensor is developed. Chapter 2 introduces into the field of electronics noses and presents the newly developed e-nose sensor used in this thesis. In chapter 3 the measurement setup used to acquire data, the feature extraction architecture and the resulting dataset are presented. Finally, multiple data quality dimensions are defined in chapter 4. The dimensions of ”Inner consistency” and ”Outer consistency” are further examined with the dataset at hand and a dashboard tool for interactive exploration of data quality is presented.

2 Electronic noses

In the human nose (see figure 2.1), odor sensations are induced by the interaction of odors with specialized receptors in the olfactory epithelium in the top of the nasal cavity. In the context of the human nose, odors are defined as volatile, hydrophobic compounds that have molecular weights of less than 300 daltons. Volatiles are true gases, liquids or solids in their vapor phase. The signals induced by the interaction of odors with the approximately 400 different types of olfactory receptors in the olfactory epithelium are transmitted to the olfactory bulb and ultimately to the brain. The brain processes the signals received[7], [8].

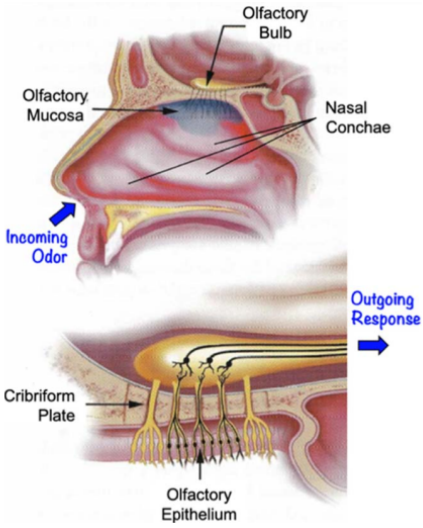


Figure 2.1: The human olfactory anatomy [6].

Researchers and sensor professionals have taken on the challenge to recreate human olfaction with sensors for decades, yet so far it is still not completely solved. The first step in the direction of the digitization of odors was the introduction of gas chromatography (GC). After its development in 1952, GC was quickly adopted to separate the individual compounds in complex odor mixtures with each detected compound appearing as a single peak in the analysis. However, GC analysis has been found to have limitations for characterizing odor quality since the amplitude of the peaks is not consistent with sensory relevance[6].

In the 1980s, one of the first systems described as an *electronic nose* (e-nose) was developed. It was a multi-array gas sensor. Instead of detecting individual gases with specific sensors, multiple partially specific sensors were used to detect and classify odors from the combined reaction. [9]. Figure 2.2 shows the analogy of such a system and the human olfaction.

A widely used definition of e-noses was published in 1994[10]: "An electronic nose is an instrument, which comprises an array of electronic chemical sensors with partial specificity and an appropriate pattern-recognition system, capable of recognizing simple or complex odours."

Electronic noses are used in a large range of applications. Examples are the food and beverage industry, agriculture and forestry, medicine and health-care, indoor and outdoor monitoring, military and civilian security, packaging, cosmetics, environmental monitoring and many more[11].

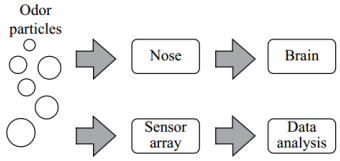


Figure 2.2: Analogy of the human olfaction and electronic noses[11]

Various technologies have been developed for the sensory part of such e-nose system. Here we try to give a brief overview over the most common technologies. More complete information can be found in [6], [11], [12]. See figure 2.3 for a direct comparison of the most common e-nose sensor technologies.

The largest group of technologies are chemoresistive sensors. Odors adsorb on the surfaces of the sensory elements and influence their resistances[12]. Most commonly used in industrial applications are chemoresistive metal-oxide (MOX) material gas sensors. MOX sensors fulfill many of the requirements for e-noses, but often have issues surrounding the range of odors detected, the environment and drift. A major drawback is their high power consumption due to high operating temperatures required[13].

Conducting polymer (CP) sensors operate in a similar way, but at room temperature, and are therefore more energy-efficient. However, they lack in selectivity and sensitivity[14].

Quartz crystal microbalances (QCM) operate through a change in natural resonance when odors adhere on a chemical sensitive layer on the device, therefore increasing its mass and altering its frequency[15]. QCMs are very sensitive, but integration into sensor electronics remains still a challenge.

More recently, sensors based on carbon nanomaterials and graphene have been devel-

Technology	CP	MOX	QCM	Nano
Sensitivity	Low	Average	High	Average
Selectivity	Low	Average	Average	Low
Portability	Good	Good	Good	Good
Cost	Low	Low	Low	Low
Trained Personnel	No	No	No	No
Sample Throughput	High	High	High	High
Speed	Real-Time	Real-Time	Real-Time	Real-Time
Pattern Recognition	Yes	Yes	Yes	Yes
Chemical Insight	No	No	No	No
Sensor Drift	Yes	Yes	Yes	Yes

(a)

Technology	IMS	Optical	GC-Sensor	GCMS
Sensitivity	High	Average	Average	High
Selectivity	Low	Good	Good	High
Portability	Good	Average	Average	Poor
Cost	Medium	High	Medium	High
Trained Personnel	No	Yes	Yes	Yes
Sample Throughput	Medium	Medium	Medium	Low
Speed	Real-Time/ Offline	Real-time Offline	Offline	Offline
Pattern Recognition	Yes	Yes	Yes	Yes
Chemical Insight	Yes	Yes	Yes	Yes
Sensor Drift	Minor	Minor	Minor	Minor

(b)

Figure 2.3: Comparison of common e-nose sensor technologies[6].

oped. These types of sensors offer high diversity and sensitivity at a small size and low power consumption. Challenges remain repeatability and reliability[16].

In some systems optical dyes are used as chemical sensors for odors. These devices produce a color change when exposed to an odor, which can be measured.

For higher-end systems technologies that measure physical properties of odorants are often used. For example, Ion Mobility Spectrometers (IMS) operate by ionising odors and then measuring the resultant ions in a high electric field[17].

Furthermore, existing technologies have been repurposed with new approaches. Instead of using a sensor array, a single sensor can be used as a ”virtual” array by sweeping its characteristic. For example, the operating temperature of a MOX sensor can be modulated producing higher dimensional information[18]. Such a sensor is called a tuneable gas sensor. Another approach is the combination of technologies. For example GC can be used as an odor filter in front of another sensor technology (GC-Sensor)[17].

However, the fundamentals of electronic noses are also still controversial. Depending on the technology used, the e-nose is blind to parts of the odor spectrum perceived by humans while certain odorless substances can be detected. This can lead to masking of relevant substances by irrelevant substances[19]. For this reason the term odor is not as clearly defined in the context of electronic noses as in the context of human olfaction. In this thesis it will be used to refer to the volatiles to be detected by the sensor. In addition, the use of e-nose systems results in information losses that must be taken into account. This poses increased challenges for the choice of sensors and the pattern recognition system[20].

2.1 The Smell Inspector

The e-nose sensor used in this thesis is the *Smell Inspector* - a multi-array gas sensor based on carbon nanotubes (CNTs) developed by the company Smartnanotubes Technologies. Figure 2.5a shows the handheld device.

Carbon nanotubes are cylinders of one layer (single-wall CNT) or multiple layers (multi-wall CNT) of graphene. Due to their unique physical properties they are used in composite materials, microelectronics, batteries, biotechnology, coatings and more[22].

CNTs are being researched in the field of gas sensory since the beginning of the 2000s. As shown in figure 2.4, they can be used as chemiresistive sensors. When odor molecules interact with the surface of the carbon nanotubes, the charge carrier density within the carbon nanotubes changes dependent on the molecules's characteristics, mainly its electric dipole moment. This changes the measurable resistance of the nanotubes[23]. Furthermore, the nanotubes can be *functionalized*. This means that the nanotubes are either modified in their structure or coated with for example polymers or metallic nanoparticles in order to create partial specificity to different molecules[24]. CNTs have a high surface-area-to-volume ratio. At the same time high currents only produce negligible heating losses. This enables high sensitivity to adsorbed molecules with very low power consumption[25].

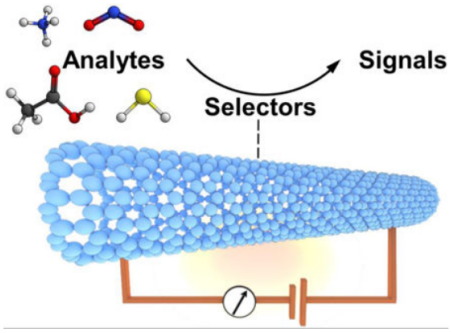
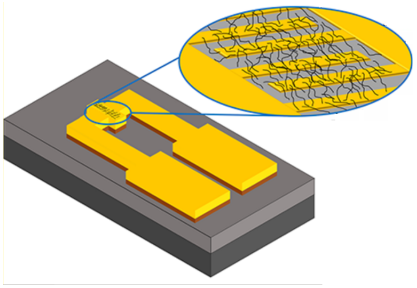


Figure 2.4: Illustration of a CNT used as a chemiresistor. The functionalization can act as an odor-selector with partial specificity[21].

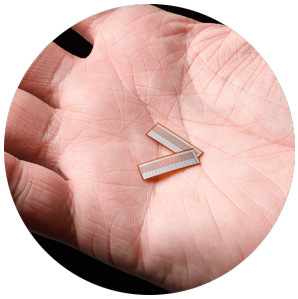
The founders of the company Smartnanotubes Technologies developed and patented a scaleable production process of single-walled semi-conducting carbon nanotubes with a high degree of purity[26]. The nanotubes produced where already used for detection of specific gases[27].

The Smell Inspector mainly consists of a measurement circuit and a microcontroller. The main purpose of the microcontroller is to control the measurement circuit and handle com-

(a)



(b)



(c)

Figure 2.5: (a) The handheld Smell Inspector device. (b) Illustration of an electrode pair connected by a network of carbon nanotubes[27]. (c) Two detector foils in a hand for size comparison.

munication with other devices via USB and Bluetooth.

The sensory elements of the Smell Inspector are 4 *detector foils* (see figure 2.5c). Each foil has 16 printed electrode pairs which are connected by a network of carbon nanotubes. Each electrode pair can be measured separately and therefore forms one measurement channel (see figure 2.5b). With 4 detector foils with 16 channels each the whole sensor has 64 measurement channels. The detector foils are exposed to the smell substances via an in-built air channel. A fan sucks the environmental air through this air channel exposing it to the detector foils.

A major advantage of the Smell Inspector is its modularity. The detector foils used in the sensors are exchangeable and the channels on each detector foil can be functionalized based on the set of substances to be distinguished. Each functionalization is developed to target a specific range of substances. As there are a lot of different types of functionalizations, they are internally numbered. In this thesis this internal numbering will be used to refer to the functionalizations. For example, the functionalization with the internal number 33 will be called functionalization 33 or in short *F*33.

Previous work has shown, that the sensor can be used for classification of volatile organic compounds (VOCs). For this a multi-class multi-label classification with multi-layer perceptrons (MLPs) and support vector machines (SVMs) was conducted. Three substances could be distinguished with an accuracy of over 91%. However, it was also established that the functionalizations used at the time were not able to distinguish the substances Isopropanol and Ethanol, which are chemically very similar. From this it was concluded that the sensor functionalizations should be systematically developed in order to improve the specificity of the e-nose system.

3 Data Acquisition

3.1 Measurement setup

In order to generate a dataset, an measurement setup was developed. The goal was to enable automatic measurements in a controlled environment, which can still be applied to open environments. Figure 3.1 shows an illustration of the measurement setup.

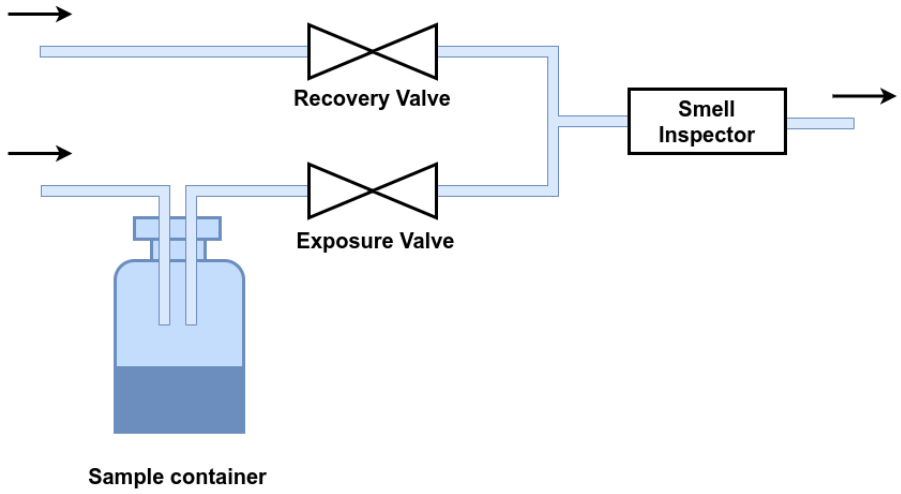


Figure 3.1: Illustration of the measurement setup. Depending on the state of the valves the air is sucked in by the fan in the Smell Inspector directly into the Smell Inspector or first through the sample container. When going through the sample container, the air acts as the carrier gas for the odor substance.

It consists of the Smell Inspector sensor connected to a T-split pipe connector. On one side of the T-split the pipe connects to the recovery valve and then opens to the environmental air. On the other side the pipe connects to the exposure valve and then opens into the headspace

of the sample container. The sample container has an inlet to the environmental air. The two valves are controlled via relays by a Raspberry Pi mini computer. When the recovery valve is closed and the exposure valve open, the fan in the Smell Inspector sucks environmental air through the sample container and subsequently the sensor device. With an odor source placed in the sample container the environmental air acts as the carrier gas for the odor. The sensor is exposed to the odor substance. This setup can be categorized as dynamic headspace sampling[28]. This means with a continuous air flow a constant odor concentration in the air flow is achieved. The concentration level is determined by the evaporation rate of the odor substance.

The evaporation rate of liquid odor substances is dependent on its surface area. The higher the surface area, the higher the evaporation rate. By placing the odor substances into small containers with different surface areas within the sample container the odor concentration can be changed. This allows for qualitative control of odor concentration for liquid odor substances.

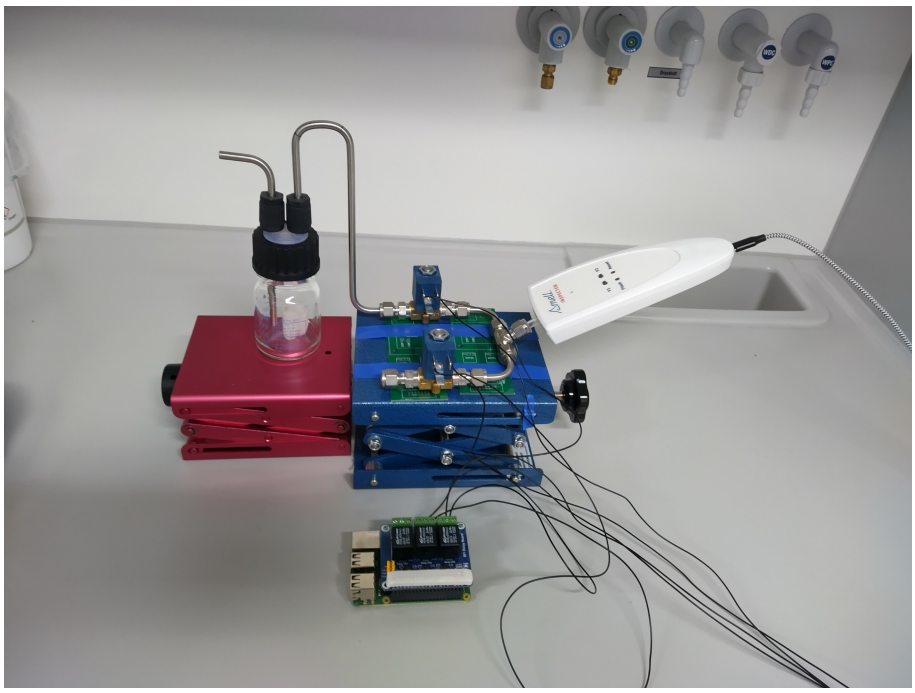


Figure 3.2: The measurement setup in the laboratory. The Smell Inspector device (right) sucks in air through the T-split pipe (middle) - either from the sample container (left) or directly from the environment. A Raspberry Pi (bottom) controls the valve states.

When the recovery valve is open and the exposure valve closed, clean environmental air is sucked into through the sensor. The odor is flushed from the sensor and the measurement

channels recover.

The measurements were conducted in a fume cupboard in order to avoid any interfering substances in the environmental air. The pipes are made of stainless steel, which is inert and therefore does retain any odor molecules. The sample container is made of heat resistant glass. Therefore it can be baked out in order to remove any odor rests when changing the odor substance. The sample container can be replaced while a recovery is running.

The setup is able to automatically perform measurements in a continuous cycle using a script running on the Raspberry Pi.

3.2 The dataset

The measurement setup was used to generate a dataset of odor measurements. One measurement is defined as follows:

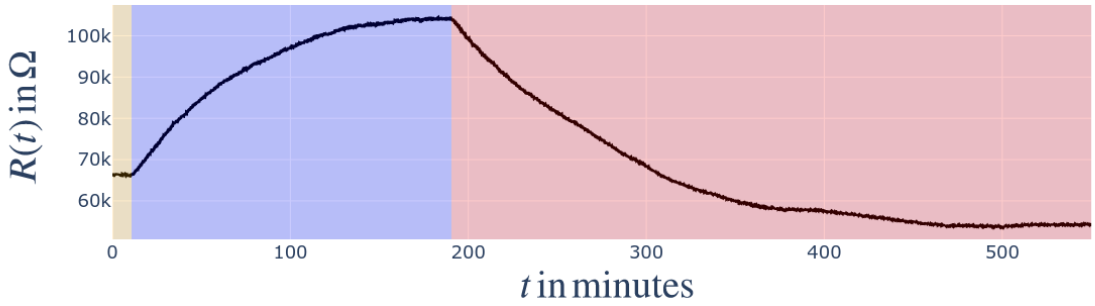


Figure 3.3: Resistance of one channel over time in a measurement. The three measurement phases are marked by the background color: Stabilization (orange), Exposure (blue) and Recovery (red).

First, no smell is present. All changes in the channel resistances are due to measurement noise and sensor drift. It is waited until the channel resistances stabilize. This is called the Stabilization phase. Then the exposure of the sensor with the odor begins. The surfaces of the sensor channels adsorb the odor molecules from the air flow and the channel resistances change accordingly. This is called the Exposure phase. When the odor molecules are flushed from the sensor, the odor molecules desorb from the sensor channels surfaces. The channel resistances return to their original values subject to drift variations. This is the Recovery phase. An example of the resistance of one channel and the three phases during a measurement can

be seen in figure 3.3.

An open question in the sensor development is the repeatability of measurements across detector foils - i.e. does a channel on one foil react similar to a channel on another foil if both have the same functionalization? As described in 2.1 the Smell Inspector has four slots for detector foils. In order to enable exploration of this question, detector foils with redundant functionalization were used. Two of these slots were filled with two detector foils with a combined eight functionalizations. The other two slots were filled with a different pair of detectors with the same functionalizations as the first pair. These two pairs of detectors will be called detector set A and B. Measurements taken with this setup were interpreted as taking two measurements at the same time - each with one detector set. This allows for direct comparison of the measurement results.

Odor substance	Number of measurements	Sample state
Ethanol	11	liquid
Eugenol	21	liquid
Guajacol	12	liquid
Isophoron	8	liquid
Toluol	16	liquid
Vanillin	15	solid
Coffee	13	solid

Table 3.1: Overview over the measurements taken with the detector sets A and B. The number of measurements were taken with each detector set.

For each odor substance and concentration an exposure duration was chosen that facilitates the availability of the static feature S in order to enable comparison of static and transient features (see section 3.4). This resulted in exposure times between 30 minutes up to 4 hours for some odors. For these substances with slow reactions fewer measurements were taken due to time constraints. Table 3.1 gives an overview over the dataset. In total 96 measurements of 7 odors were taken with each detector set.

For the liquid odor substances measurements at different concentration levels were taken

using the qualitative control of the setup. All substances measured except coffee are simple odors, which means the odor only consists of one molecule. The brand used for the coffee measurements was "NESCAFÉ Dolce Gusto Caffe Crema Grande".

3.3 Feature Extraction Architecture

The measurement data described in the previous section was stored in the *Application database*. This database is a MongoDB[29] database storing raw measurement values of the sensor channels and various meta information like the detectors used, timestamps of the valve switches or values of the environmental sensors (humidity and temperature). In order to analyze and classify measurements, features have to be extracted from them. In ML, features are measurable characteristics of a phenomenon. For analysis and comparison of measurements, it makes sense to generate one set of features for each measurement. A simple feature for a measurement could be for example the resistance of a channel at the end of the Exposure phase.

The Smell Inspector is to be tested for a wide range of possible applications. In this process different organizations will take measurements with the sensor. Therefore the goal was to implement a feature extraction architecture that is able to process the measurements automatically and generate features that can be used for feedback on these possible applications as well as classification. Figure 3.4 shows the overall architecture of the feature extraction.

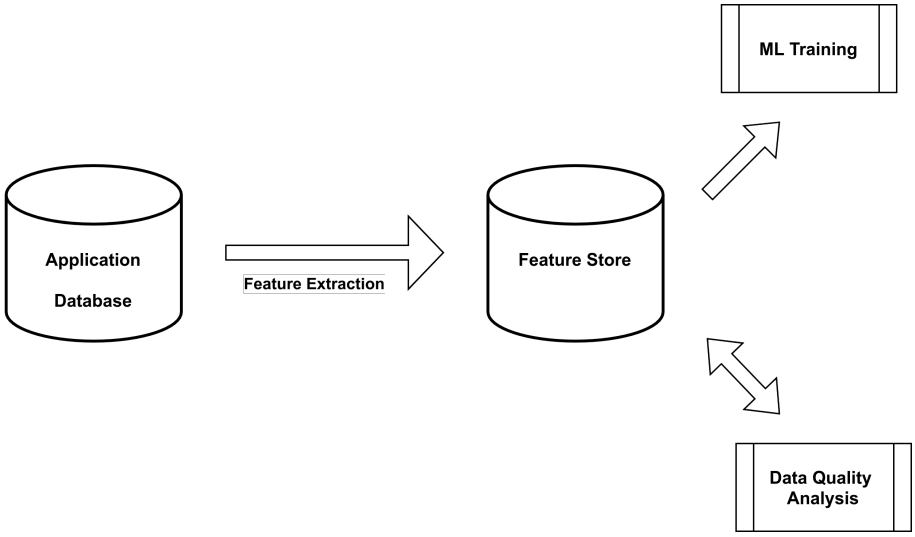


Figure 3.4: Overview of the feature extraction

The feature extraction is decoupled from feature usage with the *Feature Store* pattern[30]. Features extracted are stored separately in the Feature Store. A Feature Store can house multiple *feature sets* derived from multiple sources and with different formats. A typical structure of a Feature Store can be seen in 3.5. In this thesis, the only data source of the Feature Store is the Application database and the only feature set contains one set of features for each measurement. Therefore this is a simplified usage of the Feature Store pattern. It is still useful, because features stored in the Feature store don't have to be regenerated every time ML Training or Data Analysis is performed. Furthermore, it is likely, that additional data sources will be added in the future and additional feature sets will be created. For example, if time series models were to be trained, this would require a feature set containing some form of time series for each measurement.

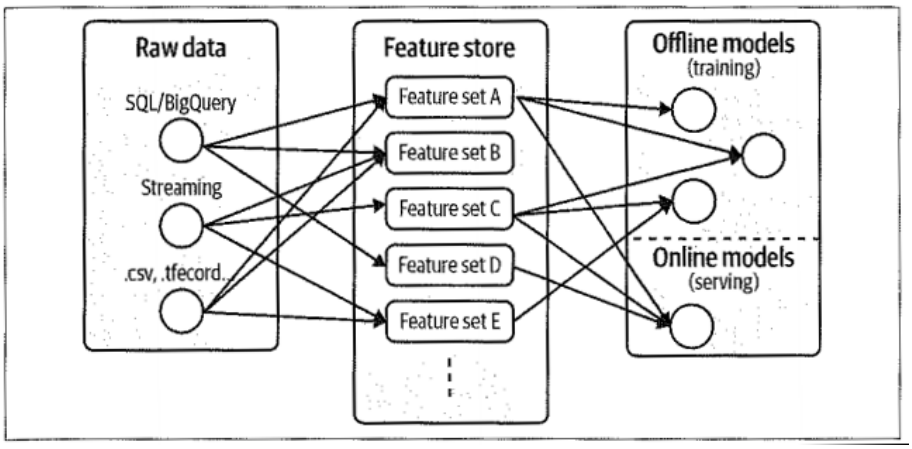


Figure 3.5: A Feature Store provides a bridge between raw data sources and model training and serving[30].

A feature extraction pipeline was implemented in Python. The pipeline is based on the scikit-learn package[31]. Each step of the data transformation and feature extraction is implemented derived class of the scikit-learn *BaseEstimator* and *TransformerMixin* classes. This enables using the scikit-learn *Pipline* class for implementing the pipeline.

Figure 3.6 shows a flow chart of the feature extraction pipeline. The *FieldImputer* imputes missing values in the meta data that are necessary in the rest of the pipeline. With the measurement setup the measurements are taken in a continuous cycle. Therefore, the raw data loaded from the Application Database is split into the separate measurements based on the meta data using the *MeasSplitter* class. The *PhaseLabeler* class labels the time series data with the three phases.

At this point in the pipeline, there is one time series of absolute resistance vectors $\vec{R}(t_i)$ for

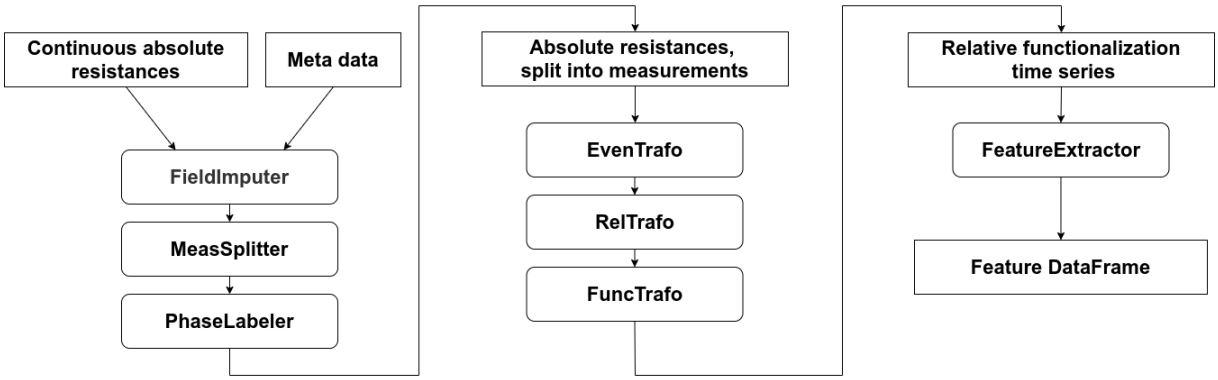


Figure 3.6: Flow Chart of the feature extraction pipeline

each measurement. Each component of this vector represents one channel. As the data from the two detector sets are interpreted as separate measurements, the resistance vectors of one measurement have 32 dimensions. Due to possible communication delays and other factors, $\vec{R}(t_i)$ is an unevenly spaced discrete time series.

$$\vec{R}(t_i) \in \mathbb{R}^{32} \text{ with } t_i \in \mathbb{T}_n \tag{3.1}$$

$$\mathbb{T}_n = \{(t_1 < t_2 < ... < t_n) : t_i \in \mathbb{R}, 1 \leq i \leq n\} \tag{3.2}$$

The class *EvenTrafo* is used to generate the evenly spaced time series of absolute resistances

$$\vec{R}(k) \in \mathbb{R}^{32} \text{ with } k \in \{0, 1, 2, ..., N\} \text{ and } t_k = t_0 + k * \Delta t. \tag{3.3}$$

This is achieved with linear interpolation. A time interval Δt of 1.5 seconds was used as this is the measurement interval of the Smell Inspector device.

The evenly spaced time series of absolute resistances is then transformed into the time series of relative resistances

$$\vec{r}(k) = \frac{\vec{R}(k) - \vec{R}_0}{\vec{R}_0} \tag{3.4}$$

using the *RelTrafo* class. The resistance vector at the start of the Exposure phase is used as the base resistance vector $\vec{R}_0$.

Each measurement channel can malfunction at some point in time. Therefore the redundancy of the functionalizations in the same detector set is used by calculating the time series of relative functionalization vectors $\vec{f}(k)$. There are eight distinct functionalizations in each detector set. Therefore the functionalization vectors have eight dimensions. $\vec{r}(k)$ is transformed into $\vec{f}(k)$ with the class *FuncTrafo* by picking the median value of the relative resistances for each functionalization.

$$\vec{f}(k) \in \mathbb{R}^8 \tag{3.5}$$

The last step of the pipeline is the feature extraction. The Feature DataFrame is extracted by the *FeatureExtractor* class. For each feature type to be extracted a BaseTransformer class can be implemented and passed as an argument to the FeatureExtractor class. Eight features are extracted from one feature type - one feature for each functionalization. The Feature DataFrame has one row for each measurement and one column for each feature. Its content is stored in the Feature Store.

The advantages of this pipeline design is its flexibility and extensibility. The pipeline can be reused for generation of a different feature set with only minor adjustments. For example, the features generated for this thesis are based on the relative functionalization vector time series. Another feature set could be based on the time series of the absolute resistance vectors. This feature set could be generated by reusing the existing pipeline without the RelTrafo and FuncTrafo classes. Furthermore, an existing feature set could be extended with new features by implementing the classes for the generation of the new features and rerunning the existing pipeline with these new feature generation classes.

3.4 Feature Engineering

In this section, the features extracted by the Feature Extraction pipeline are discussed. In order to simplify the explanations, the features will be presented as feature types independent of the functionalization. For each feature type presented, eight features are extracted - one for each functionalization. The feature types are calculated based on the relative functionalization value time series $f(k)$, which represents one component of the relative functionalization vector time series $\vec{f}(k)$.

3.4.1 Static features

The most common feature used in the recognition systems of e-nose systems based on chemiresistors is the steady state response ΔR. The fully recovered sensor is exposed to a substance at t_0. Due to the exposure the sensor resistance change in a transient. It is waited until the resistances have stabilized. The steady-state response can now be calculated with $\Delta R_{MAX} = \lim_{t \to \infty} R(t) - R(t_0)$[9]. Because of the drift common in chemiresistors, in practice it is waited until the slope of $R(k)$ is reasonably small. In the literature it can also be found as the plateau height S. Often S is calculated as the change of the relative resistance. Here, S is calculated based on $f(k)$ with k_{start} being the start of the Exposure phase and k_{end} the end.

$$S = f(k_{end}) - f(k_{start}) \tag{3.6}$$

Figure 3.7a shows an example of the derivation of the feature S from the time series $f(t)$. According to different rankings this steady-state feature is one of the most informative features [32]. However, in practice there are two problems with capturing this feature. On one hand, it can take a long time until this feature is available as the transient can be very slow. With the sensors used, transients of multiple hours have been measured. On the other hand, the long-term drifts of chemoresistive sensors can distort the feature.

3.4.2 Transient features

Features based on the transient can be used in two ways. First, they can speed up the required measurement duration as they are available much earlier than steady-state features. Alternatively, the transient has been shown to contain additional information to steady-state features[33][34][35]. A combination of steady-state and transient features can be used to improve recognition accuracy.

A transient feature that has been shown to be able to act as a substitute as well as an additional information source for MOX e-nose systems is the feature E_α - the peak value of the first derivative filtered with a exponential moving average $ema_\alpha(k)$ [36] Formula 3.7 shows the iterative calculation of E_α given the discrete time series of relative functionalization values $f(k)$. The resulting signal $ema_\alpha(k)$ for one measurement example is shown in figure 3.7b. With the assumption that the exposure process of the sensor channels with odor substances can be modeled by a sum of exponential functions, $ema_\alpha(k)$ has one single peak. This peak of this signal is the feature E_α.

$$\begin{gathered} E_\alpha = max(\, ema_\alpha(k)\,) \\ ema_\alpha(k) = (1-\alpha) * ema_\alpha(k-1) + \alpha * (ema_\alpha(k) - ema_\alpha(k-1)) \\ \text{with } 0 < \alpha \leq 1 \text{ and } ema_\alpha(0) = 0 \end{gathered} \tag{3.7}$$

The exponential moving average is a low-pass filter with an exponential decay as the impulse response. The parameter α determines the exponent of this decay and is therefore also called the exponent of the filter. With an exponent $\alpha = 1$ the signal $ema_\alpha(t)$ is equal the unfiltered first derivative. This creates a very noisy, but fast signal. With $0 < \alpha < 1$ the noise of the signal can be filtered out, however the peak of the signal is available after a longer period. Figure 3.8 shows the signal ema_α with three value of the filter exponent.

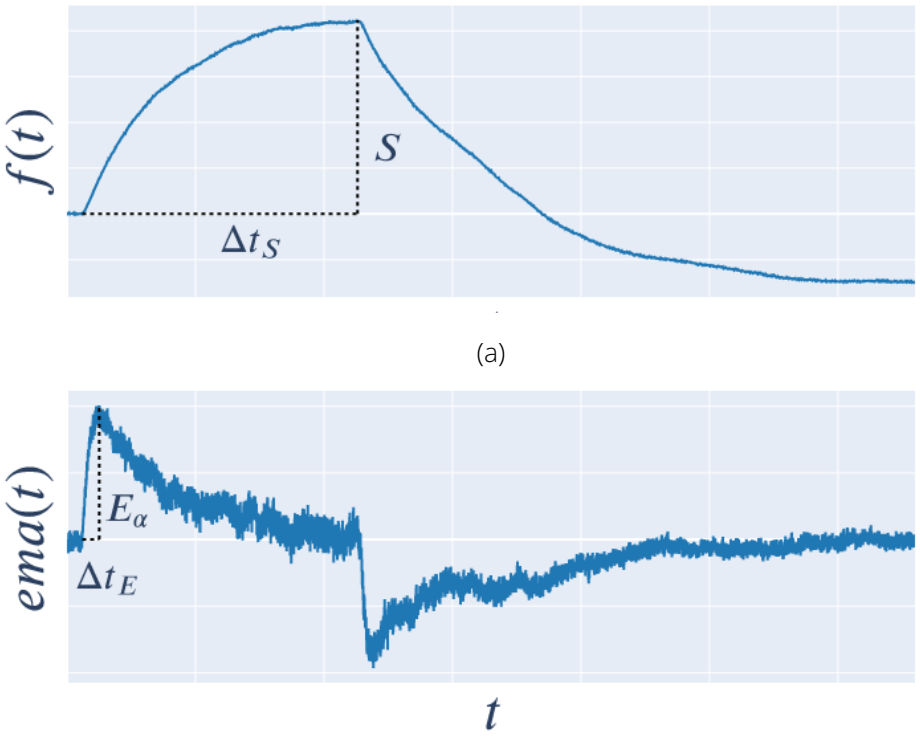


Figure 3.7: Derivation of (a) feature S from relative functionalization time series $f(t)$ and of (b) feature E_α from time series $ema_\alpha(t)$.

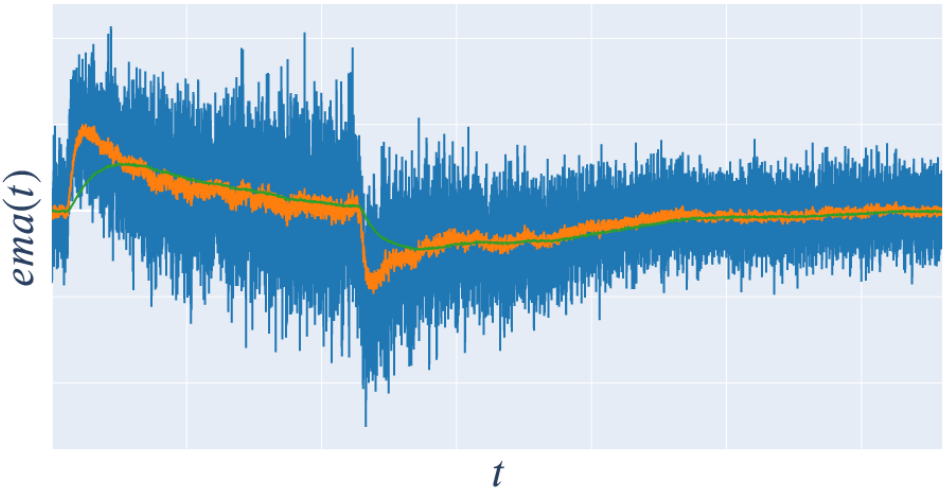
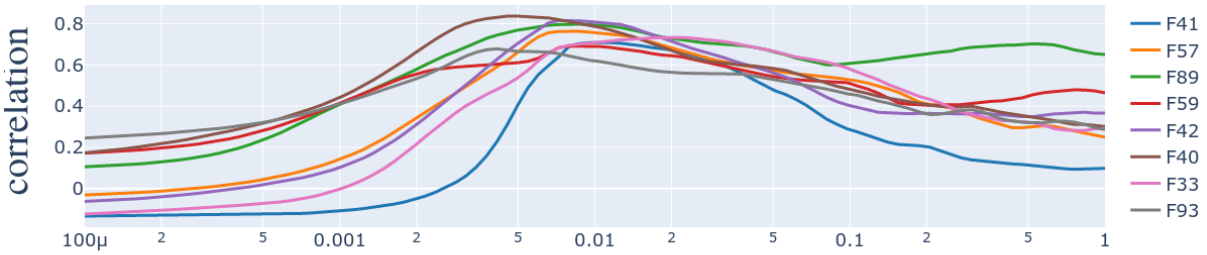
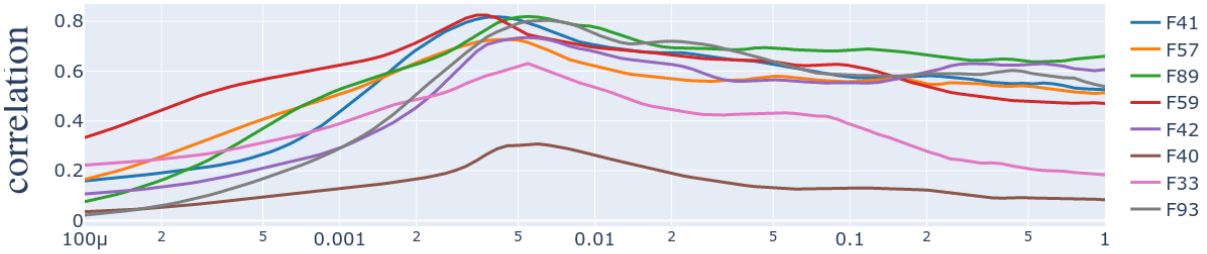


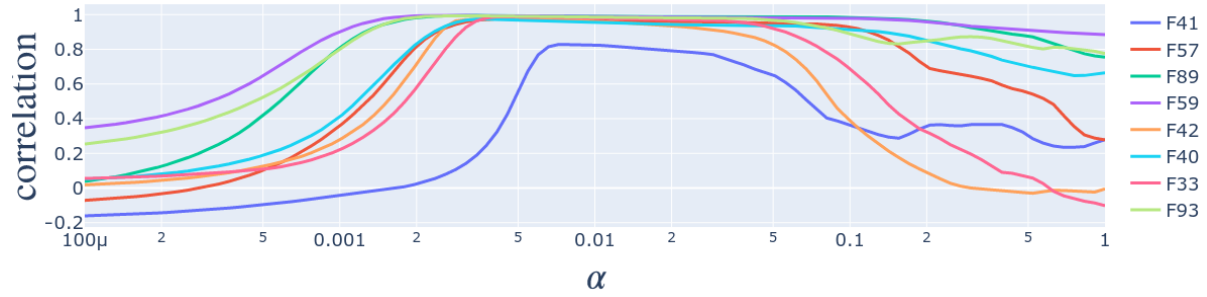
Figure 3.8: $ema_\alpha(t)$ derived from $f(t)$ in figure 3.7a with filter exponent values $\alpha = 0.1$ (blue), $\alpha = 0.01$ (orange) and $\alpha = 0.001$ (green). With a lower filter exponent value the signal gets less noisy. However, the peak E_α is available later. As exemplified by the blue function excessive noise can delay the feature availability.



(a)



(b)



(c)

Figure 3.9: (a) and (b): Weighted average correlation between the feature types S and E_α for all odor substances over a wide range of filter exponents α for detector set A and B. (c): Correlation for the odor substance Eugenol with detector set A.

In the context of analysis of data quality both applications of the transient signal can useful. On one hand the feature E_α allows the examination of measurements which did not reach the steady-state peak S. On the other a combination the transient and steady-state features can improve the assessment of different dimensions of data quality.

Figures 3.9a and 3.9b show the correlation between the feature types E_α and S over a range of filter exponent values α. The correlation between the features is calculated for each odor separately and then the mean weighted by the relative number of measurements for each odor is calculated for each functionalization. The curves for the functionalization have a similar form: For $\alpha < 0.005$ a low correlation is observed. With α in the range of (0.005; 0.1) a higher correlation is reached. For $\alpha > 0.1$ the correlation drops down again. This can be explained

by the noise, which grows in relation to the signal strength.

It should be noted that the average correlation is lowered by odors which only cause small reaction for some functionalizations. These functionalizations will have a high noise to signal ratio in the signal $ema_\alpha(t)$. The correlation for these functionalizations will be close to zero and lower the average correlation of all odor substances. Figure 3.9c shows the correlation scores for the odor substance Eugenol. For this substance all functionalization show a big reaction. Therefore a high correlation for all functionalizations is observed.

Figure 3.10: Average classification score of the 5-fold cross-validation (blue) and test set score (orange) using the feature type E_α over a range of filter exponents α for detector set A (top) and B (bottom). The black dotted line marks the test set score using the feature type S for the corresponding dataset.

In order to assess how suited the feature E_α is in a classification task a machine learning experiment based on the dataset of detector sets A and B was carried out. For each α in the range previously used 8 features E_α are available (one feature for each functionalization). A stratified test set of 25% of the dataset of these features was created. The remaining 75% were used to train the classification pipeline. The classification pipeline is based on the scikit-learn library[31] and was established in a previous work. First, the features are normalized. Then a linear support vector machine is used for automated feature selection. Finally a support vector machine with a radial basis function kernel is trained as the classifier. A grid search with 5-fold cross-validation is conducted to find the optimal hyperparameters for each α and the resulting model evaluated on the test set. The following code shows the code of the pipeline

implementation and the parameter grid used for the grid search.

```
from sklearn.feature_selection import SelectFromModel
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import GridSearchCV
from sklearn.pipeline import Pipeline
from sklearn.svm import LinearSVC, SVC

feature_selector =  SelectFromModel(estimator=LinearSVC(C=1.))
classifier_svm = SVC(random_state=random_state, probability=True)
pipeline = Pipeline([('scaler', StandardScaler()),
                     ('feature_selector', feature_selector),
                     ('cl', classifier_svm)])

parameter_grid = [
    {'cl__C': [0.1, 0.5, 1., 10., 100.],
     'cl__kernel': ['rbf'],
     'cl__gamma': ['scale', 'auto'],
     'cl__decision_function_shape': ['ovr'],
     'feature_selector__max_features': [2, 3, 4, 5, 6, None]
    }
]
```

For the evaluation of the results the macro-averaged f1-score was used. The f1-score is a well known metric in the field of machine learning. It is the harmonic mean of the precision and the recall. When the f1-score is macro-averaged the score for each class is calculated separately and then averaged. This improves the evaluation of unbalanced datasets as it penalizes models that perform bad on underrepresented classes[37].

Figure 3.10 shows the validation and test f1-score for both detector sets. For both detector sets a f1-score over 85% percent was reached on the test set. The development of the performance over a changing α is similar to the development of the correlation of S and E_α.

Based on these results, the two filter exponents $\alpha_0 = 0.005$ and $\alpha_1 = 0.02$ were selected for further investigations. Table 3.2 compares the results of the feature types S and E_α with the

exponents selected as well as combinations of these feature types. The feature type E_α is able to perform better than the feature S. A possible explanation for this is its the drift-resilience. Combining S and E_α as well as combining E_α with multiple values of α is able to improve the performance further.

Feature type set	f1-score	
	detector set A	detector set B
(S)	0.919	0.802
(E_{α_0})	0.922	0.749
(E_{α_1})	0.911	0.894
(S, E_{α_1})	0.927	0.964
($E_{\alpha_0}, E_{\alpha_1}$)	0.959	0.838
($S, E_{\alpha_0}, E_{\alpha_1}$)	0.968	0.964

Table 3.2: Classification results on the test set using different sets of feature types.

Figure 3.11 provides information about the feature availability. It displays the average ratio $\frac{\Delta t_S}{\Delta t_{E_\alpha}}$. Δt_S is the time from the exposure start until the feature S is available and Δt_{E_α} the time until the feature E_α is available. In the range of high correlation Δt_{E_α} is between 20% and 30% of Δt_S for most functionalizations.

In conclusion, it can be confirmed that the feature E_α can be used to substitute as well as a complement to the feature S not only for sensors based on MOX, but also for sensors based on carbon nanotubes. It can help the acceleration of feature availability as well as supplement information to the static feature S.

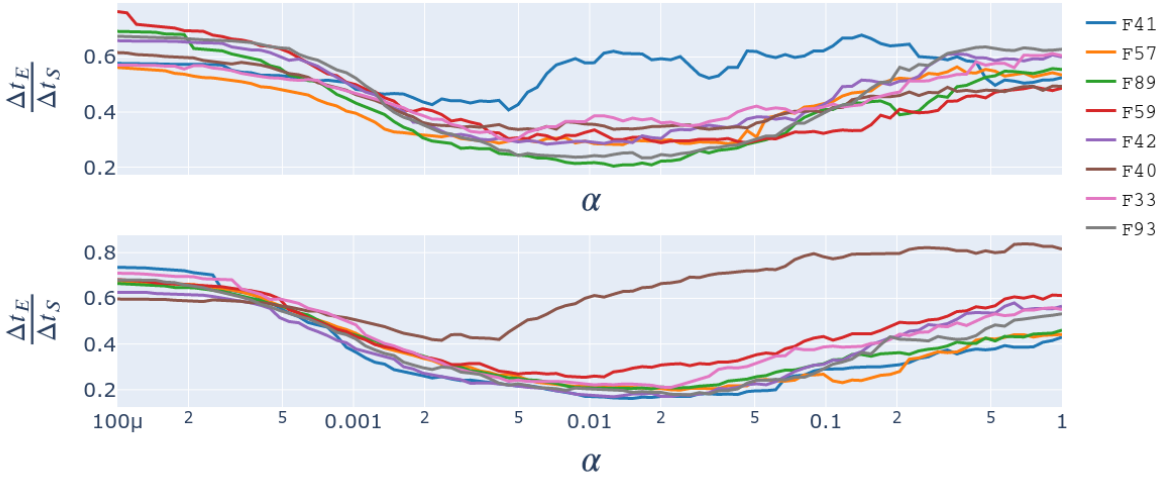


Figure 3.11: Average ratio between Δt_{E_α} and Δt_S over a wide range of filter exponents α for detector set A (top) and B (bottom).

4 Data analysis

4.1 Data quality

In order to test possible applications for the newly developed e-nose sensor, measurements will be taken by various sources and with various sets of odor substances. The term source in this context is used to describe the combination of the setup and environment a measurement is taken in and the sensor a measurement is taken with. Analysis of the quality of measurements originating from various sources and ultimately the feasibility of the applications requires a mostly automated process.

The term *data quality* is used in reference to a set of characteristics that data should own, such as accuracy - the extent to which the data are correct, reliable and certified for the task at hand - or currency - the extent to which data are up-to-date for a task[38]. Data quality has been used in computer science as a set of dimensions of "fitness for use" since the 90's of the 20th century. There have been different proposed sets of data quality dimensions. However, there is no agreed upon general set of dimensions. This is mostly because the use of data quality dimensions is highly depend on the usage of the data[39] [40].

The quality of data in the field of machine learning has been an oversight for a long time[41]. The solution to data with bad quality was gathering more data and using complex models that could be trained on this data. However, data gathering can be an expensive task. Recently Andrew Ng - a well-known researcher and educator in the field of machine learning - urged the machine leaning community to focus more on the quality of the data used to train models. He argues that it is in many cases more efficient to focus on the quality than the quantity of the

data. Ng describes the improvement of data quality in a dataset as an iterative process that requires appropriate tooling [5]. A major focus here is on the accuracy of labels. As the labels of many datasets are crowd-sourced the quality of the labels can have an important impact on the performance of models trained with such datasets[42], [43]. A recent field of research investigates how data quality propagates through the real-world ML process[44].

In the case of e-nose systems data gathering is a time intensive task. Due to the required duration of exposure and recovery, taking new measurements can take a lot of time. Furthermore, the measurement setup and environmental conditions influence the measurement results and therefore the data quality. Therefore, the assessment of the quality of the measurement data is an important step in the exploration of applications of the sensor system.

In order to assess the data quality for the given e-nose system, four dimensions were defined: Inner consistency, Outer consistency, label quality and measurement definiteness. The following enumeration defines the four data quality dimensions and lists questions related to these dimensions.

1. Inner consistency: The Inner consistency describes the consistency of measurements from one source.
    - Do measurements from the same source result in similar reactions?
    - Is more data necessary?
    - Which substances can be distinguished by which functionalizations?
    - Did the sensor degrade?
2. Outer consistency: The Outer consistency describes the consistency and transferability of measurements from multiple sources.
    - Do measurements from multiple sources return similar reactions?
    - Can models trained on data from one source be trained on data from other sources?
    - Is more data necessary?
3. Label quality: The label quality describes the accuracy and consistency of labels.
    - Are the labels accurate and consistent? Are there label aliases?
    - How are hierarchical label relationships dealt with?
    - How are complex odors labeled?
4. Measurement definiteness

- Are environmental influences known?
- Is the measurement protocol known?
- Which assumptions can be made about a measurement?

In the following sections methods to assess the Inner and Outer consistency are explored.

4.2 Inner consistency

As defined previously, the Inner consistency describes the consistency of measurements from one source. For this purpose, data from the detector set A is used to demonstrate the methods used.

A consistent e-nose sensor reacts consistently when exposed to the same odor. Therefore, an obvious choice is to examine this consistency is to examine it for each odor separately. Furthermore, the sensor's reaction can possibly differ in consistency between the different functionalizations.

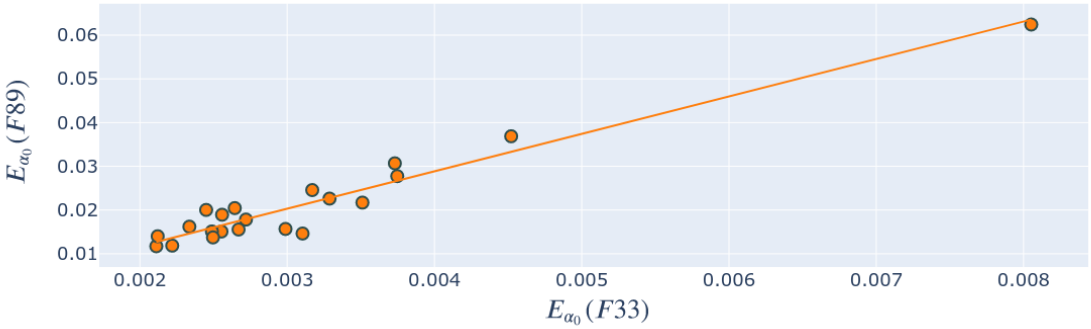


Figure 4.1: Scatter plot of the feature E_{α_0} with the functionalizations 89 and 33 for the odor substance Eugenol. The two features have a correlation of $r = 0.967$ for this feature.

Figure 4.1 shows the feature E_{α_0} of the two functionalizations f89 and f33 extracted from measurements of Eugenol. The trendline shows the linear relationship with Pearson's Correlation Coefficient $r = 0.967$.

In contrast, figure 4.2 displays an example of a functionalization pair with low correlation ($r = 0.146$).This low correlation could possibly occur due to environmental influences or erratic sensor channels. Which case applies cannot be determined from this analysis. However

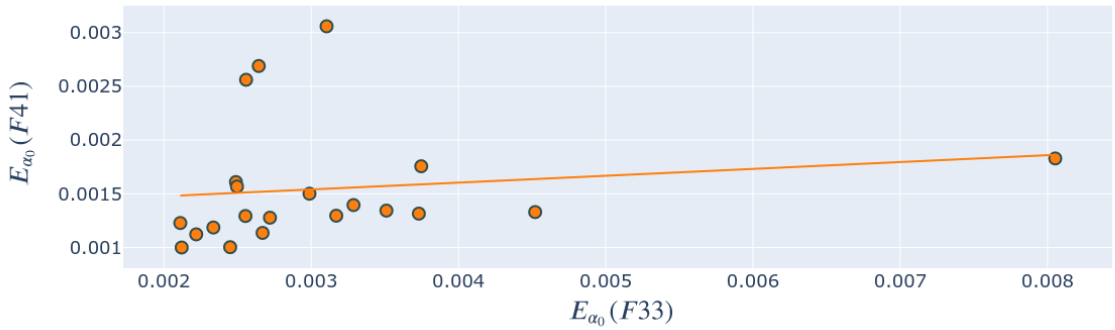


Figure 4.2: Scatter plot of the feature E_{α_0} for the functionalizations 41 and 33 for measurements of the odor substance Eugenol. Due to outliers the two features have a low correlation score of r = 0.146. Functionalization 41 of detector set A is unreliable.

it is a good starting point for further investigation.

Figure 4.3 visualizes this correlation analysis for all functionalization pairs. It is a heatmap of the correlation coefficient for each functionalization pair for the odor Eugenol. All pairs including functionalization 41 have a low correlation. All other pairs have high correlation scores. Analysis with other odor substances and the features E_{α_1} and S reveal similar results.

But is there enough data to draw conclusions from these correlations? The dataset used to verify potential application should ideally reflect the theoretical true population of measurements in this application. Most importantly, it should be made sure, that the dataset reflects the range of environmental conditions to be expected. If this is the case, statistical tests can be used to judge whether the amount of measurements is sufficient.

Assuming bivariate normal distribution of the two features, the t-test can be used to show significance of the relation[45]. However, this test only rejects the Nullhypothesis of a true population correlation $\rho = 0$. Calculating the confidence interval of ρ based on the correlation coefficient r and the number of measurements N gives more information about the "completeness" of the dataset. This can be done using Fisher's z-transform[45]. For the relation in figure 4.1 with a correlation coefficient $r = 0.967$ based on $N = 22$ measurements ρ is with a confidence of 95% in the range of 0.921 to 0.986. This small range of the confidence interval suggests enough measurements were taken.

It can be deducted that the measurements were executed in a repeatable manner. The functionalization 41 of detector set A can be flagged as unreliable. A linear relationship between the working functionalizations can be established. This relationship seems to be connected

to the substance concentration. However, only given qualitative control over the substance concentration in the measurement setup, this cannot be further investigated.

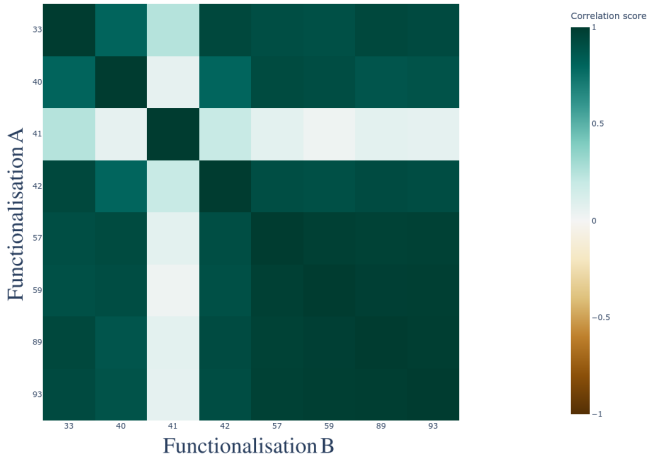


Figure 4.3: Heatmap of the correlation scores of measurements taken of Eugenol for all functionalization pairs (Feature type: E_{α_0}).

4.2.1 Linear separability score

With the correlation analysis it can be examined how consistent the measurements are for each odor separately. In the end, the e-nose system is supposed to be used for classification tasks of multiple substances. Therefore, the part of the Inner consistency should be analysis that takes the classification task into account.

Here a method is presented which will be called linear separability score. It is loosely based on the *linear separability* - a well known characteristic of two set of points in the euclidean space. These two set of points are called linearly separable if there exists a hyperplane which separates all points of the first set from the second set.

It is important to note, that the following approach does not attempt to provide optimal classification results, but an intuitively understandable metric, which can be used in the analysis of a set of measurements of the e-nose system.

Given a set of features X and a set of odor classes Ψ a score is to be calculated, that expresses how well linearly separable the classes are in the euclidean space of X. This metric should not neglect how well the hyperplane decision surface generalizes to unknown measurements.
First, let us consider the case of $|X| = 2$ and $|\Psi| = 2$. In words: We try to separate two odor

classes only given two features. We can fit a Linear Discrimant Analysis (LDA) classifier to find a linear decision surface in the euclidean space spanned by X that separates the two classes optimally. LDA is a robust method for finding a linear decision surface in $\mathbb{R}^n$ for two classes[46]. By using stratified 5-fold cross-validation the generalization of separability can be tested. The macro f1-score is used as the validation metric. The linear separability score is calculated as the average validation result of cross-validation. Figure 4.4 shows the decision surface created by the LDA trained in one iteration of the cross-validation for the two substances coffee and eugenol. This pair of functionalizations shows a linear separability score of 1 for the two substances. Similar to the correlation score, a heatmap can be used to visualize the linear separability scores of all functionalization pairs of one feature type at once. Figure 4.5a shows this heatmap for the two odor substances Coffee and Eugenol.

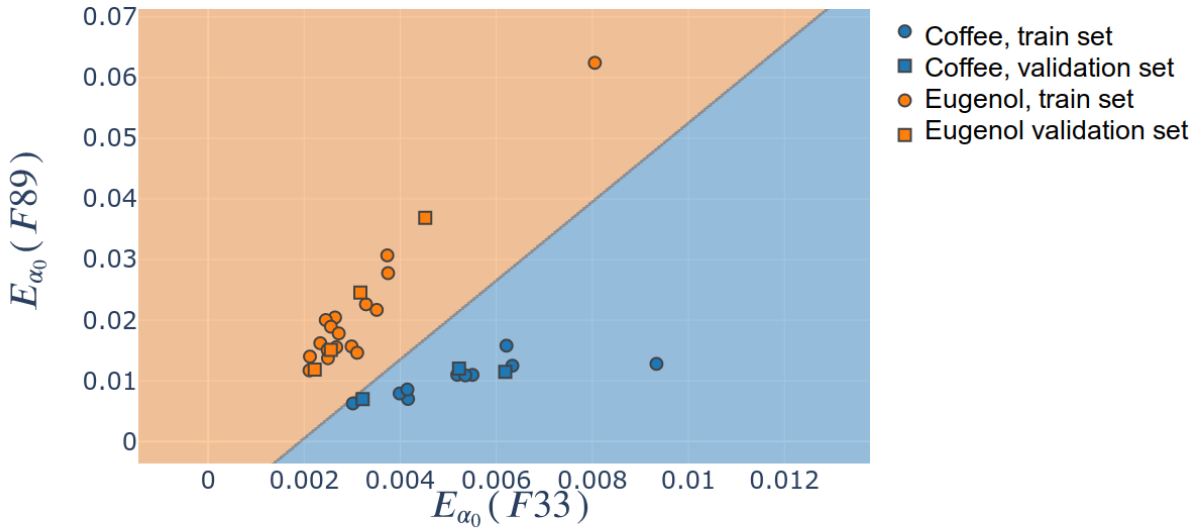
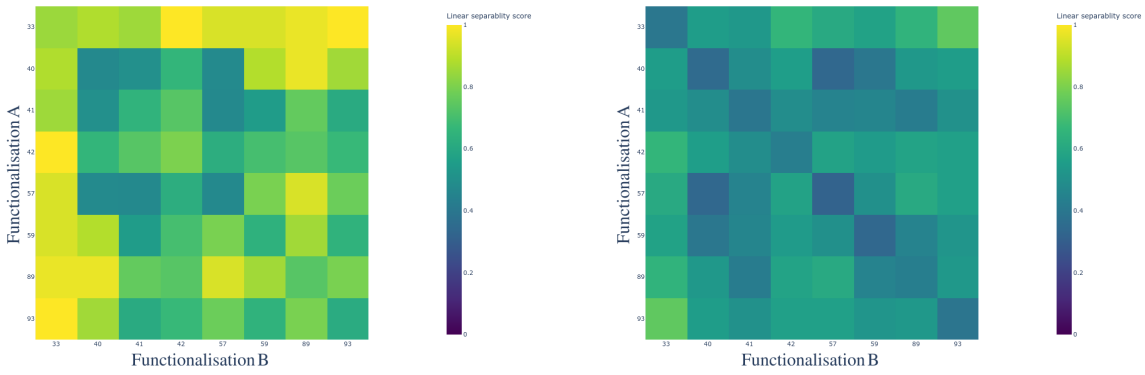


Figure 4.4: Training set, validation set and resulting decision surface of one iteration of the 5-fold cross-validation in the linear separability score calculation. Classes: Coffee and Eugenol. Features $E_{\alpha_0}(F93)$ and $E_{\alpha_0}(F33)$.

The approach for a odor class set size $|\Psi| = 2$ could be adjusted to the case of $|\Psi| > 2$ by fitting one LDA classifier all classes. However, this approach generalizes bad to unknown measurements. Instead, a *one-vs-one* multiclass strategy is used. For each class pair (ψ_1, ψ_2) with $\psi_1 \in \Psi$, $\psi_2 \in \Psi$ and $\psi_1 \neq \psi_2$ a LDA classifier is fitted to distinguish ψ_1 and ψ_2. For prediction the majority vote of this model ensemble is used. This approach generalizes good



(a) (b)

Figure 4.5: Heatmap of the linear separability scores for the classes (a) Coffee and Eugenol and (b) all 7 odor classes using feature sets of size 2 (Feature type: E_{α_0})).

to unknown measurements, because the pairwise distinction of classes tends to create linear hyperplanes that capture the linear relationship of the functionalizations (see figure 4.4).

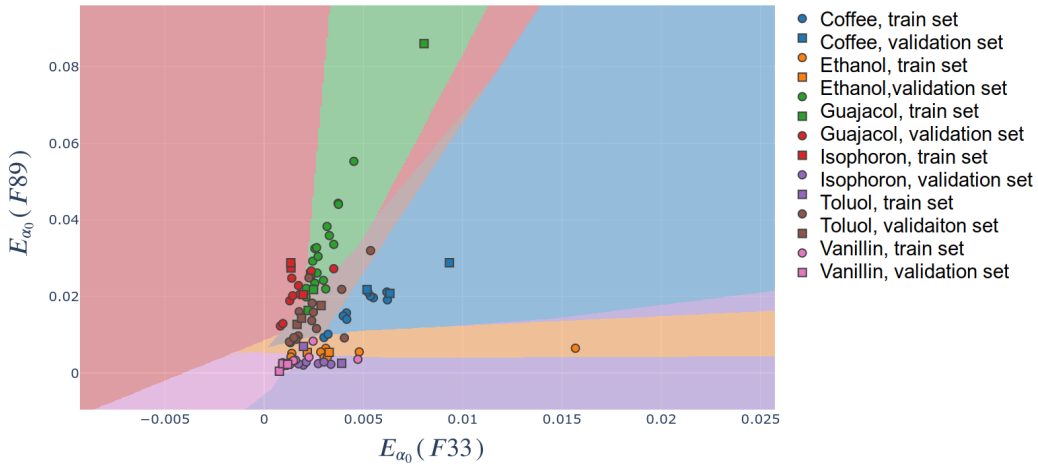


Figure 4.6: Feature points and decision surfaces for all odor substances. (Feature type: E_{α_0})

Figure 4.6 shows the decision surface for the feature $E_{\alpha_0}(F33)$ and $E_{\alpha_0}(F93)$ using all classes. This feature pair has an overall linear separability score of 0.709. Overall, the scores for all functionalization pairs of E_{α_0} (see figure 4.5b) are worse for the separation of the 7 odor classes than the case of two classes, which is to be expected.

For further analysis of these multiclass cases, the approach presented can be used for the case of a feature set size $|X| > 2$ without any changes. This allows for the selection of a functionalization set, which is optimal for a specific set of odor substances. The tables 4.1 shows

the top 5 functionalization sets for the feature type E_{α_0} for the detector sets A. Interestingly, two of the top 5 still contain F41 even though it was deemed unreliable. However, the best set consists of F33, F89 and F93. The ML pipeline used in section 3.4.2 was trained with only features for these three functionalizations (feature types: S, E_{α_0} and E_{α_1}). This resulted a test score of 0.964 for detector set A and 0.968 for detector set B, which is almost the same as the results based on the full set of functionalizations. This shows, that the linear separability score was able to determine a very informative set of functionalizations.

functionalization set	linear separability score
(33, 41, 57, 59, 93)	0.718
(33, 41, 42, 59, 93)	0.718
(33, 57, 89, 93)	0.731
(33, 57, 59, 89, 93)	0.731
(33, 89, 93)	0.751

Table 4.1: Top 5 linear separability scores for functionalization sets of size 5 or smaller for the feature type E_{α_0} using detector set A

One drawback of this approach is the complexity: For each iteration of the cross-validation $\frac{|\Psi|*(|\Psi|-1)}{2}$ classifiers have to be trained. This means a training complexity of $O(|\Psi|^2)$. However, in practice this is no major concern. The linear separability score is meant to be a metric for functionalization selection for a specific set of odor classes to be distinguished - not a classification model for a wide range of odors.

Another problem of the linear separability score is the dependency on the substance concentration. A odor substance producing a high reaction in the sensor channel resistances can be easily separated from a substance producing a low reaction. One cannot infer from this that these two substances can be distinguished at any concentration. Figure 4.7 shows the feature scatter plot of E_{α_0} for the substances Eugenol and Guajacol. These substances are chemically very similar. Therefore a similar sensor response is to be expected. However, due to its lower boiling point Eugenol creates a higher substance concentration in the air flow of the measurement setup. The overall sensor channel response in the Eugenol measurements is higher and therefore there is only little intersection between the point cloud of ”Eugenol” and ”Guajacol”. This results in a high linear separability score, giving false feedback for the functionalization

selection.

This problem can be addressed by using features that do not contain information about the substance concentration. Assuming the linear relationship between concentration and sensor channel reaction discussed in the previous chapter, the fraction of two features $\frac{F_a(F_A)}{F_a(F_B)}$ retain the information if the functionalization pair F_A and F_B while removing any concentration information. Figure 4.8 shows the resulting feature points for the same functionalization combination as used in figure 4.7. As intented, the intersection of the two classes is high. As the heatmap in figure 4.9 shows, there are still good candidates for separation of the the two substances even though they were measured at different concentrations.

This result should be regarded with caution. Like in the previous section, further investigation requires data with more precise control of the substance concentration.

In this section it was attempted to answer the four questions posed for the Inner consistency. In conclusion, the similarity of measurements with detector set A was shown using Pearson's Correlation Coefficient. A strong linear relation was found between all functionalization pairs except F41. The confidence interval of the true correlation coefficient ρ was used to judge the sufficiency of the amount of data. Furthermore, the linear separability score was proposed as a metric for analysis of multiple odors at once. The feature set F33, F89 and F93 was shown to contain sufficient information to distinguish the 7 odor classes with a 0.964 for detector set A and 0.968 for detector set B. Whether the detector did degrade over time could be subject for future research.

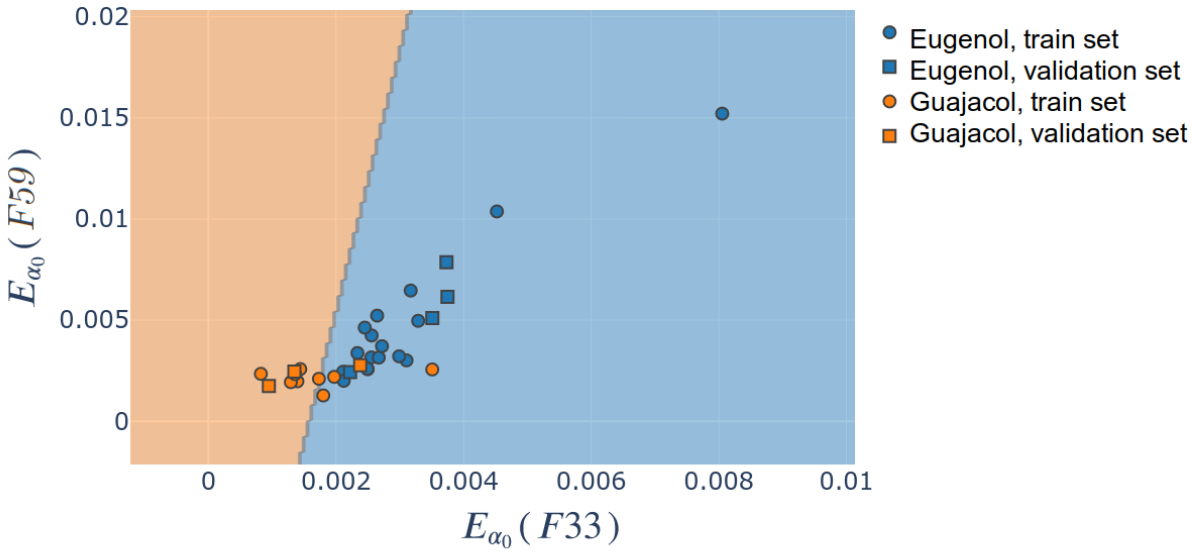


Figure 4.7: Feature points and decision surface for the feature E_{α_0} and the functionalizations 33 and 59. Even though the two odor substances can be separated with high accuracy the decision surface can be expected to generalize bad to other concentrations.

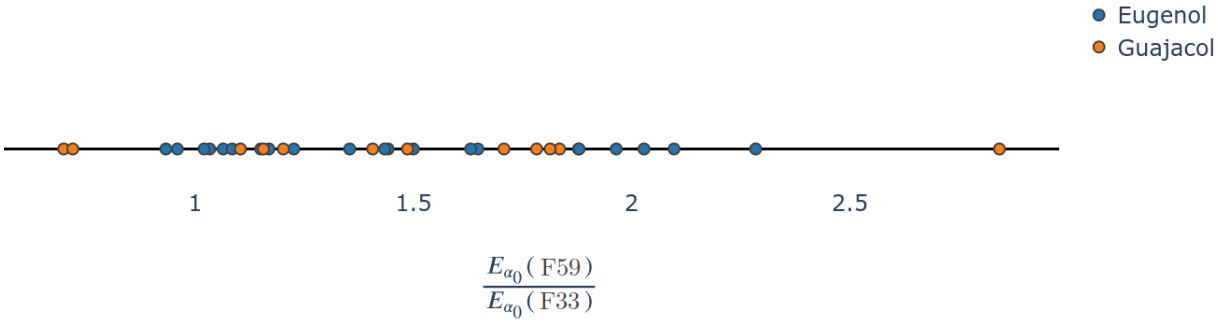


Figure 4.8: Ratio of $E_{\alpha_0}(F59)$ to $E_{\alpha_0}(F33)$ for the classes Eugenol and Guajacol. Due to the intersection of the two classes, they can be expected to not be separable at a similar reaction level.

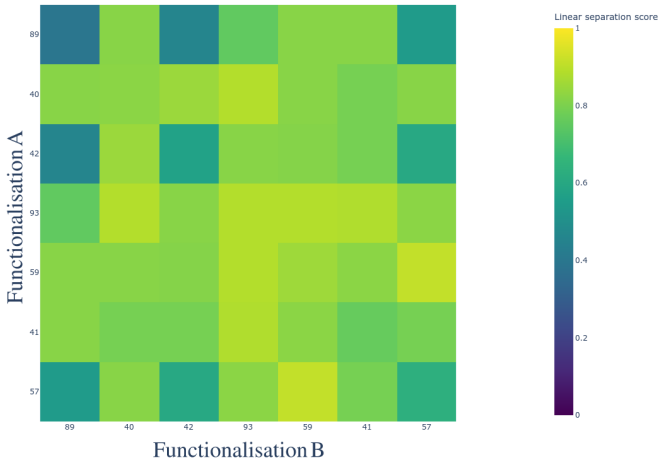


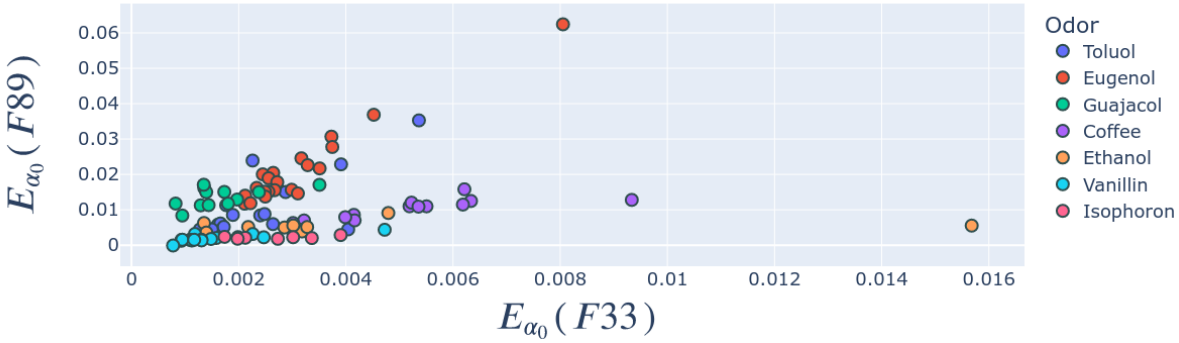
Figure 4.9: Heatmap of the linear separability score for measurements taken of Eugenol and Guajacol using features relative to F33 (Feature type: E_{a_0}).

4.3 Outer Consistency

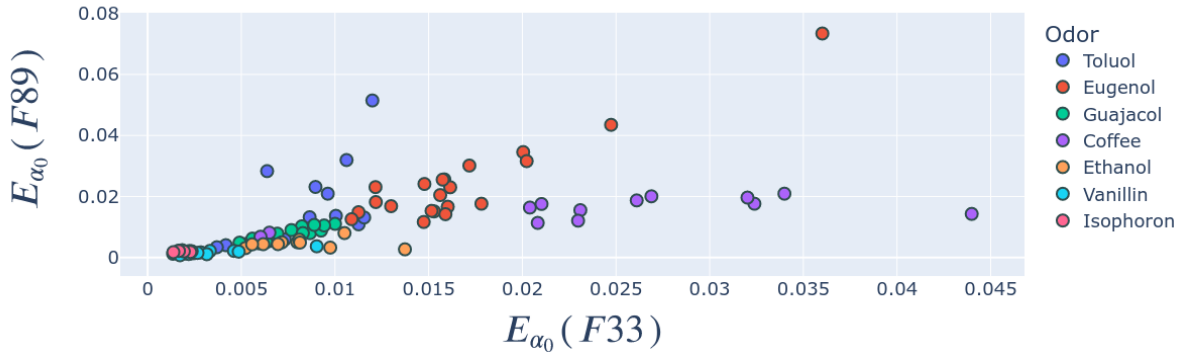
The Outer consistency describes the consistency of measurements from multiple sources. As defined previously, a source is the combination of the sensor and the environment.

As a simple test for the transferability of models from one sensor to another, the models trained on the reduced functionalization set of one detector set (see section 4.2) were tested on the full dataset set of the other detector set. This transfer does not work. The model trained on the measurements of detector set A has a f1-score of 0.172 on the data of detector set B. The transfer in the other direction produces a f1-score of 0.233.

Comparing the data of the two detector sets, one problem is clear. While there are similarities in the relations of the odors (see figure 4.10 for the comparison of one feature pair), the scale of the reactions are different. This could be due to differences in the CNT networks. A network with a higher surface area allows more odor molecules to be adsorbed, therefore causing a bigger change in the resistance. Furthermore the concentrations of the functionalization molecules on the CNT surfaces could be different. A possible solution for this problem could be calibration of the detector sets. Due to time constraints this was not explored further.



(a)



(b)

Figure 4.10: Comparison of the feature pair $E_{\alpha_0}(F33)$ and $E_{\alpha_0}(F89)$ of (a) detector set A and (b) B. Most odors are found in similar locations in relation to the other odors, although some differences can be seen. Differences in the scales make models trained on data from one detector set not directly transferable to the other set.

4.4 Dashboard

The methods presented in the previous sections can be used to examine the Inner consistency of measurement data. However, in order to iteratively improve the data quality, the appropriate Tooling is necessary. For this reason an interactive dashboard was implemented using Python with Plotly and Dash. Figure 4.11 shows the interface of the dashboard.

The data to be displayed can be adjusted by multiple loading and filtering settings. Like in the previous sections, heatmaps are used to display information about the correlation and linear separability scores for multiple functionalizations and feature types at once (top left and right). By clicking on one of the tiles in the heatmaps, a scatter plot of the feature pair selected can be displayed (bottom left). By selecting a data point in this scatter plot the specific measurement behind this point is presented on the bottom right.

This layout allows for interactive data exploration from the high abstraction of feature and functionalization wide effects to examination of specific measurements.

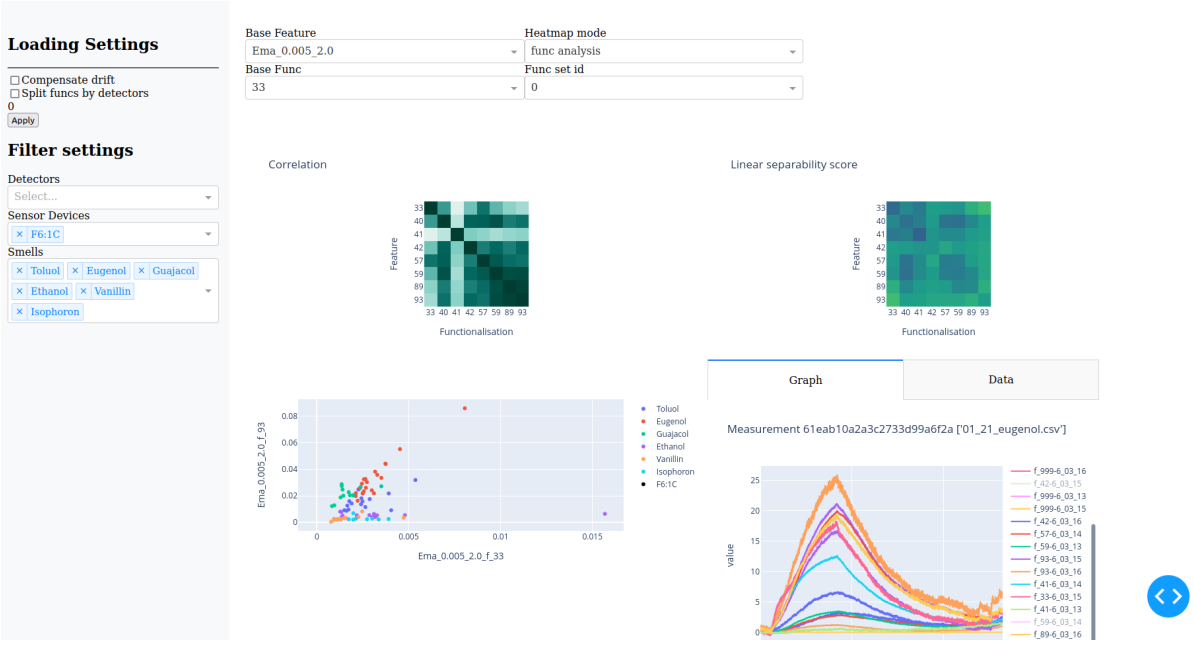


Figure 4.11: Screenshot of the dashboard

5 Discussion

In this thesis the implementation of an automated process for analysis of data quality presented. It was successfully demonstrated, that the transient feature type E_α can be adapted for use with the newly developed e-nose device Smell Inspector as a replacement and supplement of the static feature type S. The combination of the feature types S, E_{α_0} and E_{α_1} instead of only using the feature type S improved the f1-score on the test set from 0.919 to 0.968 for detector set A and from 0.802 to 0.964 for detector set B.

Four data quality dimensions for measurement data of an e-nose system were defined: Inner consistency, Outer consistency, Label quality and Measurement definiteness. For the Inner consistency - the consistency of measurements originating from one source - the repeatability of measurements of one odor was shown. A linear relation between most functionalizations was established. The linear separability score was introduced for analysis of the distinguishability of measurements of multiple odors. Based on this metric, the size of the functionalization set was reduced from eight to three while retaining a high classification score (0.964 for detector set A and 0.968 for detector set B).

These methods provide feedback for further development of the functionalizations. Additional research of the Inner consistency under changing environmental conditions and with quantitatively controlled odor concentrations could provide additional insights.

Further research is necessary in the Outer consistency. Models trained on one detector set can not be transferred directly for usage on another detector set. A possible solution could be calibration of the detector sets.

A dashboard was created for interactive data exploration. It allows for analysis of measurement data using the methods examined in this thesis - thus enabling a feedback loop for improvement of data quality.

Bibliography

[1] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A Simple Deep Learning Baseline for Image Classification?" *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, Dec. 2015, Conference Name: IEEE Transactions on Image Processing, ISSN: 1941-0042. DOI: 10.1109/TIP.2015.2475625.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, Apr. 2018, Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2017.2699184.

[3] D. Wang, X. Wang, and S. Lv, "An Overview of End-to-End Automatic Speech Recognition," en, *Symmetry*, vol. 11, no. 8, p. 1018, Aug. 2019, Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2073-8994. DOI: 10.3390/sym11081018. [Online]. Available: https://www.mdpi.com/2073-8994/11/8/1018.

[4] J. Li, "Recent Advances in End-to-End Automatic Speech Recognition," *arXiv:2111.01690 [cs, eess]*, Feb. 2022, arXiv: 2111.01690. [Online]. Available: http://arxiv.org/abs/2111.01690.

[5] R. Sagar, *Big Data To Good Data: Andrew Ng Urges ML Community To Be More Data-Centric And Less Model-Centric*, en-US, Apr. 2021. [Online]. Available: https://analyticsindiamag.com/big-data-to-good-data-andrew-ng-urges-ml-community-to-be-more-data-centric-and-less-model-centric/.

[6] J. A. Covington, S. Marco, K. C. Persaud, S. S. Schiffman, and H. T. Nagle, "Artificial Olfaction in the 21st Century," *IEEE Sensors Journal*, vol. 21, no. 11, pp. 12 969–12 990, Jun. 2021,

Conference Name: IEEE Sensors Journal, ISSN: 1558-1748. DOI: 10.1109/JSEN.2021.3076412.

[7] S. S. Schiffman and T. C. Pearce, "Introduction to Olfaction: Perception, Anatomy, Physiology, and Molecular Biology," en, in *Handbook of Machine Olfaction*, Section: 1 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/3527601597.ch1, John Wiley & Sons, Ltd, 2002, pp. 1–31, ISBN: 978-3-527-60159-2. DOI: 10.1002/3527601597.ch1. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/3527601597.ch1.

[8] A. Sharma, R. Kumar, I. Aier, R. Semwal, P. Tyagi, and P. Varadwaj, "Sense of Smell: Structural, Functional, Mechanistic Advancements and Challenges in Human Olfactory Research," *Current Neuropharmacology*, vol. 17, no. 9, pp. 891–911, Sep. 2019. DOI: 10.2174/1570159X17666181206095626.

[9] K. Persaud and G. Dodd, "Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose," en, *Nature*, vol. 299, no. 5881, pp. 352–355, Sep. 1982, Number: 5881 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/299352a0. [Online]. Available: https://www.nature.com/articles/299352a0.

[10] J. W. Gardner and P. N. Bartlett, "A brief history of electronic noses," en, *Sensors and Actuators B: Chemical*, vol. 18, no. 1, pp. 210–211, Mar. 1994, ISSN: 0925-4005. DOI: 10.1016/0925-4005(94)87085-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0925400594870853.

[11] D. Karakaya, O. Ulucan, and M. Turkan, "Electronic Nose and Its Applications: A Survey," en, *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 179–209, Apr. 2020, ISSN: 1751-8520. DOI: 10.1007/s11633-019-1212-9. [Online]. Available: https://doi.org/10.1007/s11633-019-1212-9.

[12] S. Y. Park, Y. Kim, T. Kim, T. H. Eom, S. Y. Kim, and H. W. Jang, "Chemoresistive materials for electronic nose: Progress, perspectives, and challenges," en, *InfoMat*, vol. 1, no. 3, pp. 289–316, 2019, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/inf2.12029, ISSN: 2567-3165. DOI: 10.1002/inf2.12029. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/inf2.12029.

[13] A. Oprea, D. Degler, N. Barsan, A. Hemeryck, and J. Rebholz, "3 - Basics of semiconducting metal oxide–based gas sensors," en, in *Gas Sensors Based on Conducting Metal Oxides*, ser. Metal Oxides, N. Barsan and K. Schierbaum, Eds., Elsevier, Jan. 2019, pp. 61–165, ISBN: 978-0-12-811224-3. DOI: 10.1016/B978-0-12-811224-3.00003-2.

[Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128112243000032.

[14] M. A. Ryan, A. V. Shevade, H. Zhou, and M. L. Homer, "Polymer–Carbon Black Composite Sensors in an Electronic Nose for Air-Quality Monitoring," en, *MRS Bulletin*, vol. 29, no. 10, pp. 714–719, Oct. 2004, Publisher: Cambridge University Press, ISSN: 1938-1425, 0883-7694. DOI: 10.1557/mrs2004.208. [Online]. Available: https://www.cambridge.org/core/journals/mrs-bulletin/article/abs/polymercarbon-black-composite-sensors-in-an-electronic-nose-for-airquality-monitoring/5DE83A8B4AE290081F8AB0298

[15] X. Huang, Q. Bai, J. Hu, and D. Hou, "A Practical Model of Quartz Crystal Microbalance in Actual Applications," en, *Sensors*, vol. 17, no. 8, p. 1785, Aug. 2017, Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1424-8220. DOI: 10.3390/s17081785. [Online]. Available: https://www.mdpi.com/1424-8220/17/8/1785.

[16] S. Orzechowska, A. Mazurek, R. Świsłocka, and W. Lewandowski, "Electronic Nose: Recent Developments in Gas Sensing and Molecular Mechanisms of Graphene Detection and Other Materials," en, *Materials*, vol. 13, no. 1, p. 80, Jan. 2020, Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1996-1944. DOI: 10.3390/ma13010080. [Online]. Available: https://www.mdpi.com/1996-1944/13/1/80.

[17] S. Wang, H. Chen, and B. Sun, "Recent progress in food flavor analysis using gas chromatography–ion mobility spectrometry (GC–IMS)," en, *Food Chemistry*, vol. 315, p. 126 158, Jun. 2020, ISSN: 0308-8146. DOI: 10.1016/j.foodchem.2019.126158. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308814619323106.

[18] A. P. Lee and B. J. Reedy, "Temperature modulation in semiconductor gas sensing," en, *Sensors and Actuators B: Chemical*, vol. 60, no. 1, pp. 35–42, Nov. 1999, ISSN: 0925-4005. DOI: 10.1016/S0925-4005(99)00241-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925400599002415.

[19] F. Röck, N. Barsan, and U. Weimar, "Electronic Nose: Current Status and Future Trends," *Chemical Reviews*, vol. 108, no. 2, pp. 705–725, Feb. 2008, Publisher: American Chemical Society, ISSN: 0009-2665. DOI: 10.1021/cr068121q. [Online]. Available: https://doi.org/10.1021/cr068121q.

[20] P. Boeker, "On 'Electronic Nose' methodology," en, *Sensors and Actuators B: Chemical*, vol. 204, pp. 2–17, Dec. 2014, ISSN: 0925-4005. DOI: 10.1016/j.snb.2014.07.087. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925400514009277.

[21] V. Schroeder, S. Savagatrup, M. He, S. Lin, and T. M. Swager, "Carbon Nanotube Chemical Sensors," *Chemical reviews*, vol. 119, no. 1, pp. 599–663, Jan. 2019, ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.8b00340. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6399066/.

[22] M. F. L. De Volder, S. H. Tawfick, R. H. Baughman, and A. J. Hart, "Carbon Nanotubes: Present and Future Commercial Applications," *Science*, vol. 339, no. 6119, pp. 535–539, Feb. 2013, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.1222453. [Online]. Available: https://www.science.org/doi/10.1126/science.1222453.

[23] S. Kumar, V. Pavelyev, P. Mishra, and N. Tripathi, "A Review on hemiresistive gas sensors based on Carbon Nanotubes: Device and Technology transformation," *Sensors and Actuators A Physical*, vol. 283, Sep. 2018. DOI: 10.1016/j.sna.2018.09.061.

[24] G. W. Hunter, S. Akbar, S. Bhansali, M. Daniele, P. D. Erb, K. Johnson, C.-C. Liu, D. Miller, O. Oralkan, P. J. Hesketh, P. Manickam, and R. L. V. Wal, "Editors' Choice—Critical Review—A Critical Review of Solid State Gas Sensors," en, *Journal of The Electrochemical Society*, vol. 167, no. 3, p. 037 570, Feb. 2020, Publisher: The Electrochemical Society, ISSN: 1945-7111. DOI: 10.1149/1945-7111/ab729c. [Online]. Available: https://doi.org/10.1149/1945-7111/ab729c.

[25] H. Chang, J. D. Lee, S. M. Lee, and Y. H. Lee, "Adsorption of NH3 and NO2 molecules on carbon nanotubes," *Applied Physics Letters*, vol. 79, no. 23, pp. 3863–3865, Dec. 2001, Publisher: American Institute of Physics, ISSN: 0003-6951. DOI: 10.1063/1.1424069. [Online]. Available: https://aip.scitation.org/doi/10.1063/1.1424069.

[26] V. Bezugly, E. Bezugly, V. Khavrus, D. Krylov, and G. Cuniberti, "Verfahren zum Wachstum von vertikal ausgerichteten einwandigen Kohlenstoffnanoröhren mit gleichen elektronischen Eigenschaften sowie zum Vervielfältigen von einwandigen Kohlenstoffnanoröhren mit gleichen elektronischen Eigenschaften," de, DE102014212077A1, Dec. 2015. [Online]. Available: https://patents.google.com/patent/DE102014212077A1/ko.

[27] L. A. Panes-Ruiz, M. Shaygan, Y. Fu, Y. Liu, V. Khavrus, S. Oswald, T. Gemming, L. Baraban, V. Bezugly, and G. Cuniberti, "Toward Highly Sensitive and Energy Efficient Ammonia Gas Detection with Modified Single-Walled Carbon Nanotubes at Room Temperature," *ACS Sensors*, vol. 3, no. 1, pp. 79–86, Jan. 2018, Publisher: American Chemical Society. DOI: 10.1021/acssensors.7b00358. [Online]. Available: https://doi.org/10.1021/acssensors.7b00358.

[28] J. Burlachenko, I. Kruglenko, B. Snopok, and K. Persaud, "Sample handling for electronic nose technology: State of the art and future trends," en, *TrAC Trends in Analytical Chemistry*, vol. 82, pp. 222–236, Sep. 2016, ISSN: 0165-9936. DOI: 10.1016/j.trac.2016.06.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165993616300279.

[29] *MongoDB: The Application Data Platform*, en-us. [Online]. Available: https://www.mongodb.com (visited on 02/18/2022).

[30] V. Lakshmanan, S. Robinson, and M. Munn, *Machine Learning Design Patterns: Solutions to Common Challenges in Data Preparation, Model Building, and MLOps*, en. O'Reilly Media, Incorporated, Oct. 2020, Google-Books-ID: Y52uzQEACAAJ, ISBN: 978-1-09-811578-4.

[31] 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," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[32] M. Pardo and G. Sberveglieri, "Comparing the performance of different features in sensor arrays," en, vol. 123, no. 1, pp. 437–443, Apr. 2007, ISSN: 0925-4005. DOI: 10.1016/j.snb.2006.09.041. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925400506006411.

[33] L. Carmel, S. Levy, D. Lancet, and D. Harel, "A feature extraction method for chemical sensors in electronic noses," en, *Sensors and Actuators B: Chemical*, Proceedings of the Ninth International Meeting on Chemical Sensors, vol. 93, no. 1, pp. 67–76, Aug. 2003, ISSN: 0925-4005. DOI: 10.1016/S0925-4005(03)00247-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925400503002478.

[34] S. Huang, A. Croy, L. A. Panes-Ruiz, V. Khavrus, V. Bezugly, B. Ibarlucea, and G. Cuniberti, "Machine Learning-Enabled Smart Gas Sensing Platform for Identification of Industrious Gases," en, *Advanced Intelligent Systems*, vol. n/a, no. n/a, p. 2 200 016, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aisy.202200016, ISSN: 2640-4567. DOI: 10.1002/aisy.202200016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202200016.

[35] M. Schmuker, V. Bahr, and R. Huerta, "Exploiting plume structure to decode gas source distance using metal-oxide gas sensors," en, *Sensors and Actuators B: Chemical*, vol. 235, pp. 636–646, Nov. 2016, ISSN: 0925-4005. DOI: 10.1016/j.snb.2016.05.098. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925400516307833.

[36] M. K. Muezzinoglu, A. Vergara, R. Huerta, N. Rulkov, M. I. Rabinovich, A. Selverston, and H. D. I. Abarbanel, "Acceleration of chemo-sensory information processing using transient features," en, *Sensors and Actuators B: Chemical*, vol. 137, no. 2, pp. 507–512, Apr. 2009, ISSN: 0925-4005. DOI: 10.1016/j.snb.2008.10.065. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925400508007119.

[37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[38] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, 16:1–16:52, Jul. 2009, ISSN: 0360-0300. DOI: 10.1145/1541880.1541883. [Online]. Available: https://doi.org/10.1145/1541880.1541883.

[39] M. Scannapieco and T. Catarci, "Data quality under a computer science perspective," *Journal of The ACM - JACM*, vol. 2, Jan. 2002.

[40] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, vol. 39, no. 11, pp. 86–95, Nov. 1996, ISSN: 0001-0782. DOI: 10.1145/240455.240479. [Online]. Available: https://doi.org/10.1145/240455.240479.

[41] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, "“Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21, New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–15, ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445518. [Online]. Available: https://doi.org/10.1145/3411764.3445518.

[42] M. Lease, "On Quality Control and Machine Learning in Crowdsourcing.," Jan. 2011.

[43] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident Learning: Estimating Uncertainty in Dataset Labels," *arXiv:1911.00068 [cs, stat]*, Apr. 2021, arXiv: 1911.00068. [Online]. Available: http://arxiv.org/abs/1911.00068.

[44] C. Renggli, L. Rimanic, N. M. Gürel, B. Karlaš, W. Wu, and C. Zhang, "A Data Quality-Driven View of MLOps," *arXiv:2102.07750 [cs]*, Feb. 2021, arXiv: 2102.07750. [Online]. Available: http://arxiv.org/abs/2102.07750.

[45] M. Gollwitzer, M. Eid, and M. Schmitt, *Statistik und Forschungsmethoden*. Mar. 2013, ISBN: 978-3-621-27834-8. [Online]. Available: https://content-select.com/de/portal/media/view/519cc0c7-24e0-4d29-8d98-23ef5dbbeaba?forceauth=1.

[46] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, en. Springer Science & Business Media, Aug. 2009, Google-Books-ID: tVIjmNS3Ob8C, ISBN: 978-0-387-84858-7.