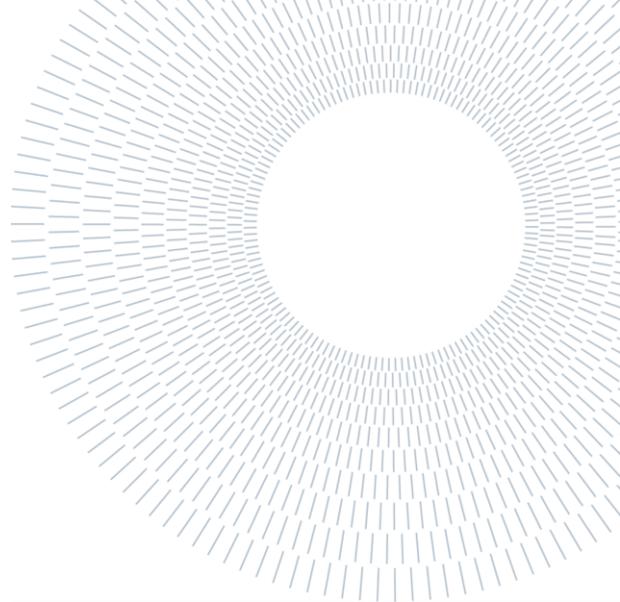




**POLITECNICO
MILANO 1863**

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE



EXECUTIVE SUMMARY OF THE THESIS

Addressing the issue of gas sensor reproducibility: development of a calibration transfer methodology for urine headspace analysis

TESI MAGISTRALE IN CHEMICAL ENGINEERING – INGEGNERIA CHIMICA

AUTHOR: Michela Cassinerio

ADVISOR: Laura Capelli

CO-ADVISOR: Beatrice Julia Lotesoriere

ACADEMIC YEAR: 2023-2024

1. Introduction

Electronic Noses (E-Noses) face a significant challenge due to the intrinsic non-reproducibility of gas sensors, and especially Metal Oxide Semiconductor (MOS) sensors. While they appear identical at a macroscopic level, microscopic variations in their surface result in non-reproducible responses, limiting their adoption in clinical or industrial contexts.

E-Noses characterize the odor fingerprint of complex gaseous mixtures by reacting with volatile organic compounds (VOCs). To do this, they require a specific training phase for each device and for each application. During this phase, sensors are exposed to known samples, and their resistance changes are recorded to develop classification or regression models. However, building these models is resource- and time-intensive, involving sample collection, data preprocessing and computationally expensive training. The non-reproducibility of MOS sensors

means that a model developed for one E-Nose (“master”) performs poorly when used to predict samples analyzed by another one (“slave”), even if the devices are nominally identical. Each new instrument, for this reason, must be trained individually. Additionally, due to sensor drift (gradual changes in sensor responses over time) even a single device requires frequent recalibration over time [1]. In medical diagnostics, despite significant advancements in recent years focused on identifying VOC biomarkers in biological fluids, the identification of specific biomarkers remains challenging. Literature often reveals conflicting findings regarding the biomarkers associated with particular diseases. What has emerged is that diseases are likely characterized by pools of biomarkers present in varying concentrations, rather than individual ones. In this context, E-Noses offer a cost-effective alternative to traditional analytical techniques (such as GC-MS) by analyzing VOC mixtures holistically, acting as a black box that captures global odor profiles without the need of preliminarily identifying the

specific compounds present in gaseous mixtures. Several studies have demonstrated their potential in diagnosing diseases like lung cancer, prostate cancer and diabetes through breath and urine analysis. However, most studies remain at the proof-of-concept stage, lacking scalability due to sensor non-reproducibility and high costs.

Calibration transfer (CT) techniques have been developed to eliminate the need for recalibrating each new instrument individually, allowing the calibration model established on a "master" E-Nose to be directly applied to a "slave" E-Nose. One of the available CT techniques is Direct Standardization (DS). DS aligns the master and the slave using a small number of transfer samples (TS). However, the variability of real samples complicates CT if used as TS, as differences in their composition add to the variability of the instruments. This highlights the need for calibrants that mimic real sample stimulation while being standardized and reproducible. This work focused on the following aspects.

- **Developing PLS-DA classification models** to classify real urine mixtures (pure, with acetone, and with 4-heptanone to simulate different physio-pathological conditions). Models were created for individual E-Noses and for pairs of E-Noses (to study the impact of duplicated sensors).

- **Formulating synthetic urine mixtures as reproducible calibrants** for CT, testing their ability to stimulate sensors similarly to real urine and also evaluating their sensory profiles compared to real urine.

- **Building a CT setup with nominally identical E-Noses**, developing algorithms to select optimal TS, and implementing DS across four master-slave combinations.

2. Materials and Methods

2.1. Real urine batches and additions

In this study, CT techniques using synthetic urine were investigated to transfer between nominally identical E-Noses the classification models developed to distinguish between different classes of real urine samples. The different classes were obtained in the following way. Multiple samples of real urine were collected anonymously from different individuals to prepare six distinct batches, each exhibiting natural variability. Each of

these six batches was then divided into three classes to be classified. A portion was kept unaltered (pure urine); in the two other portions 4-heptanone ($0.625 \mu\text{L}/L_{\text{urine}}$) and acetone ($3.25 \mu\text{L}/L_{\text{urine}}$) were added, respectively, in quantities that are in the range of those naturally found in human urine. The idea was to mimic in this way the varying levels of VOC biomarkers found in urine as indicators of various pathologies.

2.2. Experimental setups: sampling system and E-Noses setup for Calibration Transfer

The headspace of urine samples (real or synthetic) was generated using a dual-line sampling system: one line collected urine headspace in a 6L Nalophan™ bag, while the other created a 10L reference air bag for baseline establishment and restoration. Both lines included bubblers (20 mL of urine or distilled water) in a 60°C oven, connected to compressed air for VOC stripping and PermaPure™ membranes to control humidity, minimizing humidity effects on the E-Nose sensors. The CT setup used three parallel E-Noses in PEEK chambers: one for preliminary discrimination of urine classes (pure, with acetone, and with 4-heptanone) and testing synthetic mixtures, while all three were used for CT studies. Electrovalves controlled flow through ambient air, reference air (for baseline), and urine headspace lines, with three vacuum pumps generating flow (300 mL/min). Sampling times were 15 min for urine and 25 min for reference air, with analysis phases set at 4 min (baseline), 3 min (sample), and 4 min (baseline restoration).

The six types of MOS sensors used in the three PEEK chambers E-Noses (A, B and C) are TGS2602, TGS2603, TGS2610, TGS2611, TGS2620, and TGS2600 (from Figaro Engineering Inc). Each of the three PEEK chambers contains eight sensors, with two of the six sensor types duplicated.

Apart from the physical chambers A, B, and C, "virtual" chambers were generated by eliminating sensor duplicates from the three physical chambers: the first three virtual chambers (V1, V2, and V3) were formed by removing duplicates (from A, B, and C, respectively), while the fourth virtual chamber (V4) grouped all duplicate sensors together. This allowed obtaining 4 nominally identical arrays of sensors made of the same 6 types of sensors, to study CT between them.

2.3. E-Nose classification models: Data Processing

On raw sensor resistances, these preprocessing techniques were applied: normalization by subtracting the baseline established during the reference air “before” step (R-R0) and a moving average filter. After that, 26 different features for each sensor were extracted from the resistance curves over time, representative of both steady-state and transient conditions [3, 4].

The classification models were developed for each of the seven previously described chambers: chambers A, B, and C, the “physical” chambers, and V1, V2, V3, and V4, the “virtual” chambers with no sensor duplicates. Additional classification models were constructed by pairing the sensors of the virtual chambers (e.g., V1 with V2, etc.) to investigate how classification and calibration transfer performance varies under different configurations, particularly by duplicating each of the installed sensors. The process is as follows for each configuration. The dataset is divided into 50 random partitions, each with a 70% training set (~50 samples) and a 30% test set (~20 samples). For each partition, PLS-DA (Partial Least Squares Discriminant Analysis) is applied to the training set, after scaling and centering, using leave-one-out (LOO) cross-validation. The VIP (Variable Importance in Projection) metric is calculated for each feature and used as a feature selection tool by selecting features with $VIP \geq 0.5$. The selected features, based on VIP scores, are used as inputs for another PLS-DA model. The next PLS-DA model is trained again, after scaling and centering, on the 70% training subset using Leave-One-Out Cross-Validation and tested on the 30% test set. The number of latent variables (LVs) tested in the analysis ranged from 1 to 10. This process is repeated for all 50 partitions, then the average accuracies (CV and test) and their 95% confidence intervals are calculated for each configuration.

2.4. Synthetic Urine Mixtures as E-Nose calibrant: Laboratory Protocol

Synthetic urine mixtures were developed as TS. They were analyzed using E-Nose, PID, and sensory tests. The primary goal was to ensure that the synthetic mixtures stimulated the E-Nose similarly to real urine. Additionally, it was

evaluated whether the synthetic mixtures closely resemble real urine from a sensory perspective.

The study of a synthetic urine recipe based on theoretical research and thermodynamic calculations [2], forms the basis of this work. The initial formulation, focused on urinary biomarkers, was refined through an optimization process. Several variants, with compound concentrations within literature-reported ranges, were tested to identify those most closely matching real urine in E-Nose responses. Sensor response curves were visually analyzed, and features extracted from these curves were evaluated using PCA. A Photoionization Detector (PID) was used to measure VOC concentrations, ensuring alignment with real urine. Three variants, recipes #4 1:5, #6, and #7, showing the highest overlap with real urine, were selected for further study. Their compositions are detailed in Table 1.

Table 1.: Composition of the three selected synthetic mixtures (#4 1:5, #6 and #7).

	#4 1:5	#6	#7
Liquid Species:	$\mu\text{L}/\text{L}_{\text{water}}$	$\mu\text{L}/\text{L}_{\text{water}}$	$\mu\text{L}/\text{L}_{\text{water}}$
4-heptanone	0.104	0.104	0.0234
Acetone	0.26833	0.6613	0.6613
Trimethylamine	2.05	2.05	0.2945
Acetaldeide	0.27	0.4968	0.4968
2-butanone	0.915	0.915	0
Methanol	1.837	4.5151	4.5151
Acetic acid	1.38	1.5915	1.5915
Isobutyric acid	1.68	1.68	1.2358
Propionic acid	0.41166	0.4117	0.1671
Ammonia	24.5	24.5	24.5
Solid Species	$\text{mg}/\text{L}_{\text{water}}$	$\text{mg}/\text{L}_{\text{water}}$	$\text{mg}/\text{L}_{\text{water}}$
P-cresol	9.2333	9.2333	1.4454

The synthetic urine mixtures are prepared using an ice bath to prevent the evaporation of volatile compounds. Distilled water is placed in a glass flask, and liquid compounds are carefully added with a micro-pipette, with special attention to problematic substances like trimethylamine, which is pre-cooled to favor adhesion to the plastic tips. Solid p-cresol is added after weighing, and highly volatile acetaldehyde is weighed in a small amount of water for easier transfer. The final mixtures are stored at -18°C in small aliquots.

2.5. Sensorial analysis

Two sensory tests were conducted to evaluate the similarity of synthetic urine mixtures to pure real urine: a Descriptive Test and a Difference from Control Test. In the Descriptive Test, 58 panelists rated descriptors (“offensive”, “pungent”, “intense”, “fishy”, “sweet”, “rotten eggs”, “chemical/acetone-like” and “ammoniacal”) on a 0-6 scale for three synthetic urine mixtures samples and a pure real urine sample, with results normalized by intensity and analyzed using radar charts and PCA. The Difference from Control Test assessed global odor similarity, with 58 panelists rating the difference between pure real urine (reference) and the three synthetic mixtures on a 0-6 scale. ANOVA, post-hoc LSD and PCA were used for statistical analysis. Both tests ensured unbiased evaluations by randomizing sample order and labeling.

2.6. Calibration transfer and selection of transfer samples

Direct Standardization (DS) is the technique chosen for the study of CT. It's based on the idea that the relationship between the master and slave sensor responses (S_M and S_S , respectively) can be described by a transformation matrix, F . The general relationship is expressed by (1).

$$S_M = S_S F; \quad (1)$$

The transformation matrix is estimated with the so-called transfer samples (TS) analyzed on both the instruments, by relating the matrix of responses of the slave instrument ($S_{M_{TS}}$) to the matrix of responses of the master instrument ($S_{S_{TS}}$), through the use of the pseudo-inverse (S_S^+) [5] (2).

$$F = S_{S_{TS}}^+ S_{M_{TS}}; \quad (2)$$

The data from the new analyses performed with the slave are corrected by multiplying them by that transformation matrix, before being predicted using the master's model. A critical aspect of this technique is the selection of the TS, measured on both master and slave, to use to calculate the transformation matrix. The methods used in this study to select them are summarized in Table 2.

Table 2.: Transfer samples selection methods

Method	Possible Inputs	Description
Kennard-Stone (KS)	-Selected features	Samples ordered based on

Algorithm (Mahalanobis distance)	-2 PCs scores -3 PCs scores	Mahalanobis distance (descending order), to homogeneously cover the analysis space.
Extremes + Densest Cluster	-2 PCs scores -3 PCs scores	Selection of 2 spatial extremes + increasing number of central points from the densest cluster identified using DBSCAN.
Random Selection	/	Randomly ordered samples.

Each of the three methods is applied using as TS or all 3 synthetic mixtures together or each of the three synthetic mixtures separately. Each method constructs ordered data frames of potential TS based on its specific logic. In each iteration, an incremental number of TS is selected from these data frames. These TS are used to calculate the correction matrix, F , used to adjust the slave Nose's dataset of real urine samples (pure, with 4-heptanone, or with acetone). The corrected dataset is then predicted using the master's calibration model. The outputs of this process are: prediction accuracies of the slave datasets corrected with DS at each iteration and the number of TS used in each iteration.

3. Results

The study demonstrated the feasibility of differentiating three classes of real urine (pure urine, urine with acetone, and urine with 4-heptanone) using an E-Nose, despite batch-to-batch variability. PCA score plot confirmed separation between these classes.

Classification models developed for individual (physical and virtual) and paired PEEK chambers (V1+V2, V1+V3, etc., to analyze the effect of having all sensors duplicated) achieved good accuracy, with paired chambers showing slightly higher accuracy and improved robustness due to sensor duplication. In particular, the average prediction accuracies on the test dataset, averaged over 50 train-test partitions for each configuration (with their 95% confidence interval in parentheses), are as follows: for A 76.11% (73.35%-78.86%), for B 78.42% (75.89%-80.95%), for C 75.47% (72.58%-78.37%), for V1 74.84% (72.17%-77.52%), for V2 76.95% (74.57%-79.32%), for V3 73.47% (70.51%-

76.43%), for V4 73.47% (70.81%-76.14%), for V1+V2 81.26% (78.70%-83.83%), for V1+V3 78.42% (75.63%-81.21%), for V1+V4 77.26% (74.52%-80.01%), for V2+V3 78.42% (75.91%-80.94%), for V2+V4 81.37% (78.60%-84.14%), and for V3+V4 73.16% (70.18%-76.14%). The confidence intervals for the average accuracies are narrow, demonstrating consistency in the dataset, in particular over time (the samples were analyzed in a 4-month time period).

Among the multiple synthetic mixtures developed, three formulations (Synthetic #4 1:5, Synthetic #6, and Synthetic #7) showed the closest match in both sensor response and PCA score space (they overlap with real urine mixtures samples). PID measurements further validated these findings, with Synthetic #7 and Synthetic #4 1:5 closely resembling pure urine, while Synthetic #6 aligned more with urine containing 4-heptanone and acetone.

The sensory tests did not provide univocal results, indicating that different recipes could represent the variability of urine. A radar chart of the descriptor test results showed that synthetic mixtures #6 and #7 had similar sensory profiles, while #4 1:5 differed more. Real urine significantly differed in some descriptors but overlapped with synthetic mixtures in others. PCA indicated that all synthetic mixtures differed from real urine along the first principal component (PC), but #6 was closest to real urine along the second. The Difference from Control Test, supported by ANOVA and post-hoc LSD, revealed that #4 1:5 was the least different from real urine, followed by #7 and #6, which were statistically similar. PCA further confirmed #4 1:5 as the closest to real urine along the second PC, while all synthetics differed along the first.

Then, CT was studied with 40 synthetic urine samples across the three selected recipes as potential TS. Over the 50 partitions, one model developed with the virtual chamber V1 (master) was used to predict analyses performed with V2, V3, and V4 (slaves). Similarly, one model developed with the pair V1+V2 (master) was used to predict analyses performed with the pair V3+V4 (slave). These two PLS-DA models are developed with 7 and 8 latent variables, respectively. As expected, using a model developed with one chamber/pair of chambers to predict samples analyzed with others resulted in significantly lower accuracy (37.14%-54.93% among the different master-slave configurations) compared to

the target one (the external prediction accuracy of samples analyzed with the same Nose used to develop the model, 78.95%). This shows the impact of sensor non-reproducibility. Applying DS with TS improved accuracy to 75.38%-80% comparable to the target prediction accuracy (78.95%). The main results are in Table 3.

Table 3.: CT best results for each of the 4 master-slave combinations.

Master-slave	Prediction accuracy without DS correction	Prediction accuracy with DS correction and n° of transfer samples	Method and synthetic mixture(s) used
V1→V2	37.14%	80% 15 TS	Extremes + Dense Cluster, 3 PCs. 3 mixtures together.
V1→V3	40.85%	77.46% 19 TS	KS, 3 PCs. 3 mixtures together.
V1→V4	54.93%	75.38% 12 TS	KS, 2PCs. 3 mixtures together.
V1+V2 → V3+V4	54.93%	76.92% 8 TS	KS, 2PCs. 3 mixtures together.

Some considerations are now presented.

- **Preferred methods of TS selection**

No single method outperformed others, but KS with 2 PCs and Extremes + Dense Cluster with 3 PCs showed strong performance. Random selection performed poorly, highlighting the effectiveness of the proposed approaches.

- **Inputs of transfer sample selection**

Using PC scores (2 or 3 PCs) as input, rather than features, improved performance. PCA maximizes variance in the dataset, by transforming the features into a set of PCs that capture the most significant patterns in the data, focusing on the most informative aspects and reducing noise.

- **Presence of sensor duplicates**

The highest accuracy among the various selection methods tested for V1+V2 → V3+V4 (76.92%) was achieved with only 8 TS, significantly fewer than the ones required to obtain the best performances in terms of accuracy for the transfers involving single chambers (15, 19 and 12 TS, respectively). This shows the benefits of using paired

chambers (with all sensors duplicated), which enhance model robustness and reduce the number of TS needed.

- **Different synthetic urine mixtures**

When tested individually, synthetic recipes #6 and #7 achieved a comparable number of successes, while synthetic #4 1:5 did not perform well. However, combining all three mixtures yielded the best results in all the four cases compared to using a single mixture. This makes sense because the three synthetic mixtures, which differ in the concentrations of the compounds, could reflect both the natural variability of urine in healthy conditions (due to factors like diet and hydration) and the different biomarker levels associated with various pathophysiological conditions in patients. Since all three mixtures are easy to prepare and store, using them as three standard mixtures for CT is a valid approach.

4. Conclusions

This work addresses the critical challenge of sensor reproducibility in Electronic Noses, which limits their widespread adoption, particularly in medical diagnostics. Real urine mixtures with varying biomarker levels (acetone, 4-heptanone) were prepared to simulate different physiological conditions and successfully classified using E-Nose analysis, despite batch variability. Synthetic urine formulations were developed and evaluated through E-Nose, PID measurements, and sensory tests, with three mixtures (#4 1:5, #6, and #7) identified as most similar to real urine.

CT using DS significantly improved prediction across all tested scenarios, with accuracies rising from 37.14%-54.93% to 75.38%-80%, matching the master E-Nose's performance (78.95%). The sensory analysis and CT results did not yield a univocal outcome to identify a single synthetic recipe as more promising as standard calibrant. In CT, using all three synthetic mixtures together gave better results. The varying compositions of the three recipes could represent the different diagnostic classes (biomarkers level) or the intrinsic variability of healthy urine: developing three standard synthetic mixtures and using TS from all three is a valid and advantageous approach.

Future research could explore additional recipes starting from the results of this work: #6 and #7 performed better compared to #4 1:5. TS selection

methods could also be refined to find a single optimal one.

In conclusion, the value of this study consists in the general methodology developed to address the problem of sensor non-reproducibility. By proposing a valid approach that uses synthetic mixtures as calibration standards and DS, this work provides a basis for overcoming one of the main barriers to the widespread adoption of E-Noses, in particular in diagnostics but more in general in all their fields of applications. By enabling the transfer of a fixed calibration model to other nominally identical devices, the methodology developed could play a role in the certification process for biomedical devices. Certification, to meet regulatory standards and ensure reliability, requires reproducibility across devices, meaning that different instruments must perform with the same diagnostic performance under the same conditions.

References

- [1] L. Zhang, F. Tian, and D. Zhang, "Electronic Nose: Algorithmic Challenges."
- [2] M. Cucciniello *et al.*, "Development and testing of novel sampling systems for biological fluids characterization by electronic noses."
- [3] A. M. Tischer, M. Elici, L. Capelli, B. J. Lotesoriere, and C. Bax, "A novel sampling method for eNose analysis of urine headspace aimed at prostate cancer diagnosis."
- [4] S. Zhang, C. Xie, M. Hu, H. Li, Z. Bai, and D. Zeng, "An entire feature extraction method of metal oxide gas sensors," *Sens Actuators B Chem*, vol. 132, no. 1, pp. 81–89, May 2008, doi: 10.1016/j.snb.2008.01.015.
- [5] B. Igne and C. R. Hurburgh, "Standardisation of near infrared spectrometers: Evaluation of some common techniques for intra-and inter-brand calibration transfer," *J Near Infrared Spectrosc*, vol. 16, no. 6, pp. 539–550, 2008, doi: 10.1255/jnirs.819.