MELBA Journal Machine Learning for Biomedical Imaging

Scent of Health (S-OH): Olfactory Multivariate Time Series Dataset for Non-Invasive Disease Screening

Ivan Poddiakov [1] Ivan Kruzhilov [2], Galina Zubkova [1], Svetlana Erofeeva [3], Andrey Savchenko [1,4], Pavel Blinov [5]

1 Sber AI Lab, Moscow, Russia
2 Moscow Power Engineering Institute, Moscow, Russia
3 Moscow Regional Research and Clinical Institute (MONIKI), Moscow, Russia
4 National Research University Higher School of Economics, Nizhny Novgorod, Russia
5 SQB AI, Tashkent, Uzbekistan

Abstract

The Scent of Health (S-OH) dataset is the largest publicly available collection of clinical electronic nose (eNose) data for non-invasive disease screening via exhaled breath analysis. It comprises 1,234 patients across nine diagnostic groups (healthy controls and eight diseases: hepatitis, gastritis, fatty liver disease, diabetes, chronic renal failure, COPD, lung cancer, and tuberculosis), each providing a 17-channel multivariate time series of approximately 895 seconds recorded at 0.4 Hz. The dataset includes explicit temporal annotations over 13 consecutive weeks and two clinical sites, enabling reproducible research on sensor drift and cross-site generalization. All data are anonymized and provided in an AI-ready format with predefined temporal train/test splits. Baseline validation includes a 3-layer LSTM classifier for lung cancer screening, achieving AUC 0.691 under temporal evaluation, alongside additional CNN-based methods. The dataset and code are released under the MIT License and is publicly available at `https://doi.org/10.57967/hf/9752`. S-OH is designed to support a range of machine learning tasks, including binary and multi-class time series classification, drift-robust learning, and demographic bias analysis for breath-based diagnostics.

Keywords

eNose, breath analysis, multivariate time series, disease screening, medical dataset, sensor drift, open data

Article informations

`https://doi.org/10.59275/j.melba.2026-7d42`

Volume 2026, Received: 2026-06-19, Published 2026-09-21

Corresponding author: ivanpodd@gmail.com

Special issue: MICCAI Open Data 2026 x MELBA

Guest editors: AT, MS et al.

1. Background

Electronic nose (eNose) technology has emerged as a promising non-invasive tool for disease screening through exhaled breath analysis. Exhaled breath contains complex mixtures of volatile organic compounds (VOCs) that reflect metabolic processes, and eNose sensors can capture these patterns without identifying individual compounds. This offers a portable, cost-effective, and rapid alternative to traditional methods such as gas chromatography-mass spectrometry (GC-MS), which remain prohibitively expensive and impractical for large-scale clinical deployment. The ability to detect human diseases through breath analysis could enable low-cost, point-of-care screening, particularly in resource-limited settings.

Despite growing interest in breath-based diagnostics, the field lacks large-scale, publicly available clinical datasets that can support robust machine learning (ML) research. Existing studies typically collect small cohorts, often fewer than 100 patients (Rodríguez-Aguilar et al., 2019; Tenero et al., 2020), targeting a single pathology (Van de Goor et al., 2018; Bruins et al., 2013), use inconsistent evaluation protocols, and rarely share raw data (Acevedo et al., 2021). This fragmentation makes it impossible to compare methods across studies, reproduce reported results, or build upon prior work.

Public olfactory datasets focus predominantly on molecular odor perception for computational chemistry (Sharma et al., 2022; Lalis et al., 2024; Feng et al., 2025; Lee et al., 2023) rather than clinical disease screening. Publicly avail-

able eNose datasets for disease detection are still scarce, with existing examples limited to small cohorts (e.g., only 78 measurements from 34 participants for COPD) (Acevedo et al., 2021)), which is insufficient for training robust ML models or evaluating generalization across multiple conditions. No publicly available dataset combines (i) large-scale multi-pathology clinical breath data, (ii) explicit temporal annotations to study sensor drift, and (iii) multi-site collection to assess cross-site generalization.

The **Scent of Health (S-OH)** dataset addresses this gap by providing the largest publicly available clinical eNose dataset to date. It comprises 1,234 patients (over 7.5 million sensory signal points) across nine diagnostic groups, with explicit temporal splits (13 consecutive weeks) and two clinical sites, enabling reproducible research on sensor drift, demographic confounders, and multi-disease discrimination. To the best of our knowledge, S-OH is the first open-access clinical breathomics resource designed specifically for time series classification with drift-aware evaluation protocols.

While eNose data are not imaging data in the classical sense, exhaled breath analysis occupies an established niche within the medical image computing community. Sensor array outputs can be treated as spatiotemporal maps of chemical response, and the algorithmic challenges (multivariate time-series classification, domain shift due to sensor drift, multi-site generalization, and demographic confounding) are structurally identical to those encountered in dynamic medical imaging (e.g., functional MRI, 4D CT perfusion). S-OH is thus directly relevant to the MICCAI community's core interests in robust, generalizable, and fair medical signal processing.

2. Summary

The Scent of Health (S-OH) dataset aims to accelerate the development of robust, generalizable machine learning models for non-invasive disease screening using electronic nose technology. Our vision is to establish a community benchmark that enables fair comparisons of time-series classification methods, promotes reproducibility in clinical breathomics research, and ultimately facilitates the translation of eNose technology into clinical practice. The dataset is designed to support research on sensor drift, multi-disease discrimination, and demographic confounding—key challenges that currently hinder real-world deployment.

S-OH serves researchers and practitioners in machine learning, time series analysis, and medical diagnostics. It is particularly valuable for those working on sensor-based disease screening, drift-robust learning, and multi-class classification in healthcare settings. Clinicians interested in breath-based diagnostics may also use the dataset to validate eNose technology for specific disease targets.

The dataset is AI-ready: each sample is provided as a structured JSON object containing a 17-channel multivariate time series with a consistent sampling rate (0.4 Hz) and fixed duration (895 seconds). The accompanying metadata CSV file includes patient demographics, diagnosis codes, collection week, and clinical site, enabling stratified analysis and controlled experiments. Temporal train/test splits are pre-defined to support drift-aware evaluation. No additional preprocessing is required to ingest the data into standard time series classifiers (e.g., LSTM, InceptionTime) or feature-based models (e.g., CatBoost, XGBoost). Example code for loading the data, applying the recommended splits, and reproducing baseline results is provided in the accompanying repository.

The dataset requires approximately 230 MB of storage. Users should have basic familiarity with Python and common data science libraries (e.g., json, pandas, numpy, scikit-learn) to load and process the data. For deep learning experiments, a standard GPU is recommended but not required; baseline models can be trained on a modern CPU within hours.

3. Discussion

3.1 Strengths

The S-OH dataset offers several distinctive strengths. First, it is the largest public clinical eNose collection to date, with 1,234 patients spanning nine diagnostic groups, substantially larger than existing public datasets. Second, the dataset includes explicit temporal annotations over 13 consecutive weeks, enabling drift-aware evaluation through predefined temporal train/test splits. Third, data collection across two clinical sites supports initial investigations into cross-site generalization. Fourth, the inclusion of multiple diseases (hepatitis, gastritis, fatty liver disease, diabetes, chronic renal failure, COPD, lung cancer, and respiratory tuberculosis) allows researchers to study multi-disease discrimination and disease-specific VOC signatures within a single coherent dataset.

3.2 Limitations and Biases

Several limitations should be acknowledged:

1. **Demographic representation.** The dataset is derived from two clinical sites in a single geographic region. Age and gender distributions vary across disease groups (e.g., lung cancer and COPD cohorts are older and predominantly male), reflecting real-world epidemiology but introducing potential confounding between diagnosis and demographic factors.

2. **Sensor hardware.** Both clinical sites used the same eNose device model; cross-device generalization remains

to be validated.

3. **Class imbalance.** Sample sizes vary across diseases (100-256 patients per condition), which may limit statistical power for conditions with weaker VOC signals.

4. **Single sample per patient.** Each participant provided only one breath sample, preventing analysis of within-patient temporal variability or longitudinal monitoring.

3.3 Recommendations for Responsible Use

Users of the S-OH dataset should adhere to the following recommendations. First, always use the provided temporal splits for evaluation, as random cross-validation substantially overestimates performance (a 0.05-0.12 AUC drop has been observed). Second, report demographic-stratified performance (by age, gender, and site) to assess model fairness and generalizability. Third, compare sensor-based models against simple demographic-only baselines (age, gender) to distinguish genuine VOC signals from demographic confounders. Fourth, when developing new methods, evaluate on all disease groups rather than cherry-picking conditions where performance is high. Finally, respect the MIT license and cite the dataset appropriately.

3.4 Planned Extensions

We plan several extensions to the S-OH resource. First, we will add more clinical sites to increase geographic and population diversity. Second, we will collect longitudinal samples from the same patients to study within-subject variability and disease progression. Third, we will integrate additional sensor modalities (e.g., GC-MS reference measurements on a subset of samples) to enable VOC identification and cross-validation of eNose signals. Fourth, we will maintain a public leaderboard for time series classification benchmarks to track progress and encourage community participation.

3.5 Comparison with Parallel Initiatives

To our knowledge, no other public clinical eNose dataset matches S-OH in size, diagnostic diversity, and temporal annotation. Existing public eNose datasets for disease detection are limited to small, single-pathology cohorts without temporal validation splits. While molecular olfactory datasets (e.g., OlfactionBase (Sharma et al., 2022), M2OR (Lalis et al., 2024)) provide valuable resources for computational chemistry, they do not contain clinical breath time series and cannot be used for disease screening research. S-OH thus complements existing resources by addressing an underserved area at the intersection of medical diagnostics and time series machine learning.

4. Resource Availability

4.1 Data Location

The S-OH dataset (with accompanying code) is permanently archived and publicly available at `https://doi.org/10.57967/hf/9752`. The dataset is openly accessible without registration or authentication. Users may download the JSONs and CSV files manually from the repository, and a `README.md` file provides detailed instructions for data loading and usage.

4.2 Versioning

Version 1.0.0 is the initial release; future versions will be documented with release notes in the repository, and previous versions will remain accessible.

4.3 Licensing

The S-OH dataset is released under the **MIT License** (SPDX identifier: MIT; full text: `https://opensource.org/licenses/MIT`), which permits unrestricted use, reproduction, modification, and redistribution for any purpose, including commercial use, provided the original copyright notice is retained. The full license text is included in the data repository as `LICENSE.txt`. The accompanying code repository is independently licensed under the MIT License.

4.4 Potential Use Cases

The S-OH dataset is designed to support a range of machine learning tasks and clinical applications. Target ML tasks include, but are not limited to:

- **Binary disease classification:** One-vs-rest classification for each of the eight diagnostic groups (e.g., lung cancer vs. others, COPD vs. others).
- **Multi-class time series classification:** Simultaneous discrimination among multiple disease classes using multivariate sensor data.
- **Drift-robust learning:** Development and evaluation of domain adaptation, style transfer, or self-supervised methods to mitigate sensor drift over time.
- **Demographic bias analysis:** Studying model performance across age, gender, and clinical site subgroups to assess fairness and generalizability.
- **Temporal validation:** Benchmarking models under realistic deployment conditions using the provided week-wise train/test splits.

In the clinical domain, the dataset enables research on non-invasive disease screening via exhaled breath analysis.

Target conditions include lung cancer, COPD, hepatitis, fatty liver disease, diabetes, chronic renal failure, gastritis, and tuberculosis. For example, a researcher developing a drift-robust classifier for lung cancer screening can use the provided temporal splits to evaluate generalization across time, compare against demographic-only baselines, and benchmark against the provided methods.

4.5 Ethical Considerations

The study protocol was approved by the Independent Ethics Committee at Moscow Regional Research and Clinical Institute (MONIKI), approval No. 09 dated June 5, 2025. The same protocol was also used for data collection at the Central Research Institute of Tuberculosis (CRIT). All participants provided written informed consent prior to enrollment. The informed consent form explicitly permitted the use of anonymized data for research purposes and open release of de-identified breath samples and associated metadata.

The dataset is classified as low sensitivity under institutional data protection guidelines. No direct identifiers (names, addresses, dates of birth, medical record numbers) are included. Age is provided as integer years. The dataset does not include children (minimum age 15). Participants aged 15 years and older were included in the study. This is consistent with the national legal framework, in which individuals aged 14 and older hold a national identity passport and are entitled to provide informed consent to medical interventions.

No other vulnerable populations (e.g., prisoners, cognitively impaired individuals) were enrolled. Collection dates are anonymized; only week numbers (1-13) and timestamps relative to the start of collection are retained. Clinical site information is provided, but no further location data is included. No free-text fields or comments are present in the dataset. For questions regarding dataset access, ethics, or data protection, please contact the corresponding author.

5. Methods

5.1 Data Details

Participants were enrolled prospectively over 13 consecutive weeks at two clinical sites in Moscow, Russia: Moscow Regional Research and Clinical Institute (SiteA) and Central Research Institute of Tuberculosis (SiteB). Inclusion criteria were: (i) age 15 years or older, (ii) confirmed diagnosis of one of the target conditions according to ICD-10 codes, or healthy status for the control group, and (iii) ability to provide written informed consent. Exclusion criteria were: (i) acute respiratory infection within the past two weeks, (ii) pregnancy, (iii) known malignancy other than lung cancer (for cancer patients), and (iv) inability to comply with the pre-sampling protocol.

Data were collected prospectively between June and September 2025. The collection period spanned 13 consecutive weeks: 11 weeks at SiteA (MONIKI) covering all disease groups except tuberculosis, followed by 2 weeks at SiteB (CRIT) enrolling only tuberculosis patients (A15) and healthy controls.

The final dataset comprises 1,234 patients across nine diagnostic groups (healthy controls plus eight diseases). No ethnicity or race data were collected. Table 1 provides the complete demographic breakdown by diagnosis. All eligible participants who provided informed consent were included; no additional filtering was applied after data collection.

5.2 Methods Used for Data Creation

Breath samples were collected and analyzed using an electronic nose device with a 17-channel metal-oxide-semiconductor sensor array printed on a single chip (see Figure 1). The chip comprises 17 co-planar Pt/Ti electrode pairs with functional materials (ZnO and metal-doped ZnO: In-ZnO, Ag-ZnO, Ce-ZnO, Ni-ZnO) deposited via microplotter printing (Fedorov et al., 2020). The sensor array is enclosed in a gas-tight chamber (volume 0.76 cm^3) with on-chip heaters maintaining a stable operating temperature of 300±5°C. The sampling rate is 0.4 Hz, producing approximately 895 seconds (≈15 minutes) of recording per sample. Auxiliary sensors (ze03 for O_3, mhz14 for CO_2, ze08 for CO, bme280 for pressure, temperature, and humidity) were also recorded at the same sampling rate.

Before sampling, participants adhered to a standardized pre-sampling protocol: 4-hour fasting, no smoking for at least 4 hours, no alcohol for 48 hours, and avoidance of perfumes or strong odors on the day of sampling. Each participant provided a single exhaled breath sample collected in a sterile, disposable 2-liter bag with a check valve. The collection procedure was performed by trained laboratory staff in a ventilated room. For analysis, each sample bag was connected to the eNose intake port via a sealed valve, and an internal pumping system provided constant, controlled airflow over the sensor array. Between samples, the sensing chamber was purged with clean, dry air to prevent cross-contamination. All samples were analyzed within 4 hours of collection.

To preserve the original sensor characteristics, including drift, no additional filtering or normalization was applied. The raw signals are provided as-is to allow users to implement their own preprocessing pipelines. Disease labels were assigned based on clinical diagnoses using ICD-10 codes (World Health Organization, 2019), and verified by attending physicians at each clinical site. For the healthy control group (Z00), participants underwent a standard

Table 1: Distribution and demographic characteristics of the S-OH dataset by diagnostic group.

Diagnostic Group (ICD-10)	N	%	Age (mean)	Age (std)	Male (%)
Healthy individuals (Z00)	256	20.7	40.6	12.9	21.5
Diabetes mellitus type II (E11)	128	10.4	60.4	11.9	30.5
Gastritis and duodenitis (K29)	138	11.2	52.3	16.3	35.5
Non-alcoholic fatty liver disease (K76)	128	10.4	48.0	16.2	39
Hepatitis B/C (B18)	138	11.2	50.9	13.4	55
Malignant neoplasm of lungs (C34)	100	8.1	66.7	8.6	73
Chronic renal failure (N18)	128	10.4	59.8	14.6	50
COPD (J44)	100	8.1	64.2	10.4	64
Respiratory tuberculosis (A15)	118	9.5	41.9	16.1	62.7
Total or Average	**1,234**	**100**	**53.9**	**13.4**	**44.1**

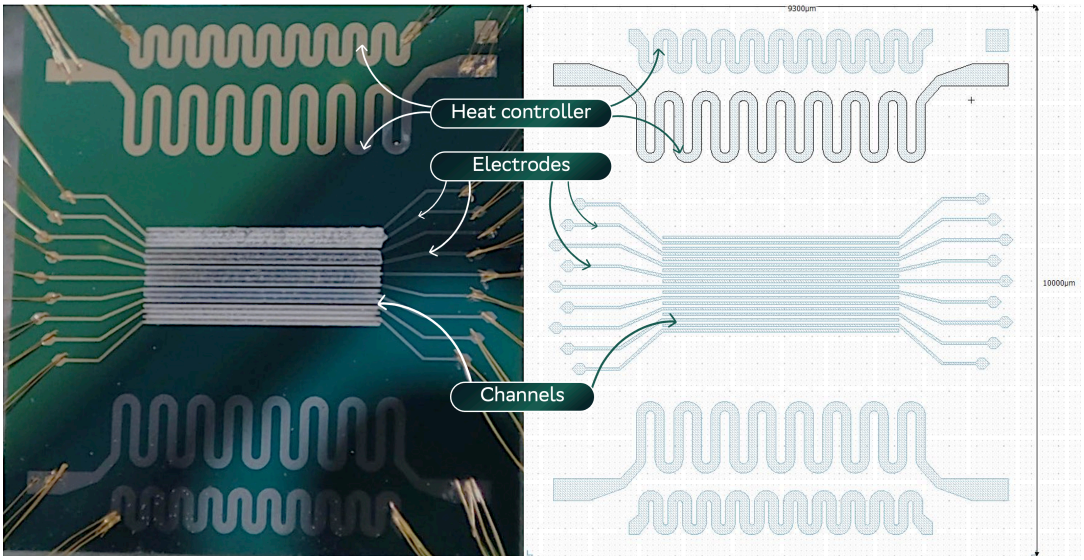


Figure 1: eNose chip: physical implementation and schematic layout of the 17-element ZnO microsensor array.

health screening including medical history, physical examination, and routine laboratory tests to confirm the absence of the target diseases. Diagnoses were not validated by central adjudication; however, both clinical sites follow national diagnostic guidelines.

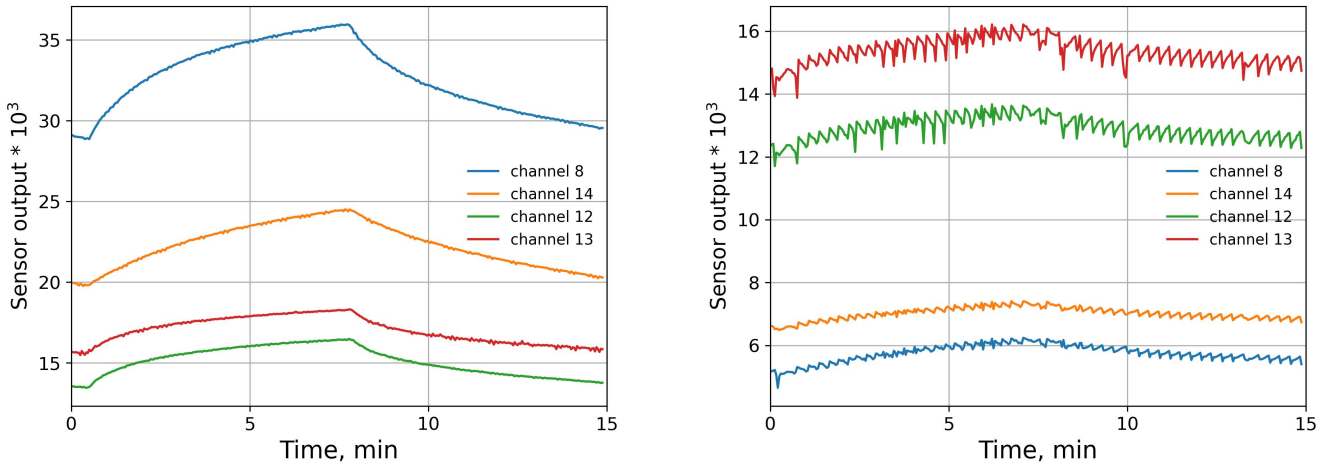
Data format. The dataset is provided in two primary files: (i) a CSV metadata file (`metadata.csv`) containing patient demographics, diagnosis codes, collection week, and clinical site; and (ii) a JSON data file (`data/manifest.json`) that indexes per-patient time series data. The full multivariate time series for each patient is stored as an individual JSON file in a subdirectory named by ICD-10 diagnosis code (e.g., `data/C34/patient_131.json`). Each patient JSON file contains the 17-channel eNose signals, auxiliary sensor readings, and temporal metadata (start/end timestamps, duration). This structure facilitates efficient loading of specific patient groups or individual samples without parsing a single large JSON file.

Known artifacts and limitations. The eNose sensor array exhibits progressive signal deterioration over time (sensor drift), characterized by baseline shifts and increased signal fluctuations (see Figure 2). This drift is a known phenomenon in metal oxide gas sensors and is intentionally preserved in the released dataset to enable research on drift-robust methods. No drift correction has been applied. Additionally, the dataset includes only one sample per patient, limiting longitudinal analyses.

6. Validation

6.1 Sensor Calibration and Temperature Control

Prior to the study, the multielectrode chip was preconditioned by heating at 300±5 °C for 24 hours in an air atmosphere to ensure stable sensor performance. During routine data collection with patients, the chip was maintained at a constant operational temperature (300±5 °C) using on-chip meander-shaped thermoresistors and heaters, with the temperature verified by an infrared pyrometer (Kelvin



(a) After 2 days of operation.

(b) After 53 days of operation.

Figure 2: Progressive sensor signal deterioration.

Compact 1200D). No recalibration was performed on the production line; all temperature settings were controlled automatically by the device.

6.2 Metadata Validation

Metadata consistency was verified using a custom Python script that checks:

- Age range (15-89 years) and plausible gender encoding.
- ICD-10 code consistency with `D_class` and `D_bin_class` mappings.
- Temporal consistency: `start` < `end` < `duration`.
- Unique patient IDs and no duplicate entries.

All validation scripts are included in the repository.

6.3 Temporal Split Verification

The temporal splits (Table 2) were verified to ensure:

- No patient appears in both training and test sets.
- Training and test weeks are disjoint per disease.
- The test set for each disease contains at least 10 samples.

These checks passed for all nine conditions.

6.4 Example ML/AI Use Case

To demonstrate the dataset's usability, we provide two complete use-case examples in the repository.

Table 2: Train (empty) / test (+) split by ICD-10 codes across weeks. For each diagnostic group, we use different temporal segments.

Week	Z00	E11	K29	K76	B18	C34	N18	J44	A15
W1		+	+						
W2		+	+	+	+		+		
W3		+	+	+	+		+		
W4				+			+		
W5								+	
W6								+	
W7								+	
W8						+			
W9	+					+			
W10	+				+	+	+		
W11	+							+	
W12									
W13									+

6.4.1 *Long Short-Term Memory (LSTM) Classifier*

LSTMs are well-suited for sequential data and can capture long-range temporal dependencies in breath signals without manual feature engineering (Hochreiter and Schmidhuber, 1997). The example included in the repository demonstrates training an LSTM classifier for lung cancer (C34) detection and includes the following steps:

- Data loading and preprocessing (min-max normalization).
- Temporal split application (weeks 8-10 for testing, all the rest for training).
- Model definition (3-layer LSTM with hidden size 128).

- Training loop with early stopping.
- Evaluation (ROC AUC, confusion matrix).

The script achieves an AUC of 0.691 on the test set. This example can serve as a template for users adapting the dataset to their own ML tasks.

6.4.2 Convolution-Based Time Series Classification

The key idea behind this approach is to treat multivariate time series as images, enabling the application of powerful convolutional architectures. Following the paradigm of time series classification via image representations (Homenda et al., 2024; Costa et al., 2024), we transform our sensor signals into 2D matrices using wavelet smoothing and polynomial feature extraction. While 1D convolutions are computationally efficient, they require learning temporal features from scratch. Our 2D approach allows transfer learning from ImageNet-pretrained models, which has been shown to improve performance on small medical datasets. Recent work has also demonstrated the effectiveness of pre-trained vision models for time series data encoded as images (Zhai et al., 2023).

Each 17-channel time series undergoes the following preprocessing pipeline to create an image-like representation (Fig. 3):

1. **Wavelet smoothing:** wavelet transform is applied to each channel to reduce high-frequency noise while preserving the overall signal shape.
2. **Normalization:** each channel is independently normalized via min-max scaling to reduce absolute baseline variability and sensor drift effects.
3. **Polynomial feature extraction:** polynomial features (degree 3, including interaction terms) are extracted from the normalized signals to capture non-linear dynamics between channels.
4. **Matrix construction:** original signals and polynomial features are combined into a floating-point matrix of size (374×337), which is interpreted as an image (height = number of features, width = time steps).

This representation preserves both the temporal structure of the original signal and the non-linear cross-channel relationships, while enabling direct application of computer vision architectures.

To evaluate the suitability of image-based time series representations for eNose data, we trained both a custom HeightWise CNN (400k parameters) and a fine-tuned ResNet18 (11M parameters) on the polynomial feature matrices. We evaluate two convolutional approaches:

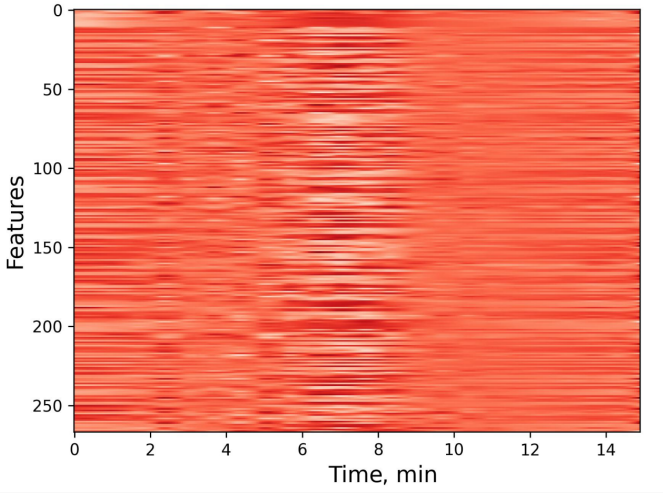


Figure 3: Time series as an image representation.

- **HeightWise CNN:** a custom convolutional network specifically designed for time series-as-image representations. The architecture consists of three 2D convolutional layers with temporal kernel sizes of 15, 10, and 20, progressively increasing the number of channels from 32 to 64 to 128, followed by adaptive pooling and two fully connected layers. The network is trained from scratch on the polynomial feature matrices.
- **ResNet18:** a pre-trained ResNet18 model (He et al., 2016), fine-tuned for binary classification using the same floating-point matrix input. The input is resized to 224×224 to match the expected input dimensions of the pre-trained model.

The HeightWise CNN achieved a ROC AUC of 0.606 for classifying healthy controls (Z00) versus all other diseases, while ResNet18 achieved 0.604. The comparable performance of the two architectures suggests that (i) the task of distinguishing healthy individuals from mixed disease cohorts is inherently challenging due to the heterogeneity of the healthy population, and (ii) model capacity alone does not guarantee improved performance on this data modality. Both models achieve inference times of under 3 second on CPU, making them suitable for deployment on embedded systems or point-of-care devices.

7. Conflicts of Interest

The authors declare no financial or non-financial conflicts of interest relevant to this work. No funding was received from manufacturers of the electronic nose device used in this study.

8. Acknowledgements

The authors thank the patients who participated in this study, the medical staff at both clinical sites for their assistance with data collection, and the technical team for developing and maintaining the electronic nose device.

References

Cristhian Manuel Durán Acevedo, Carlos A Cuastumal Vasquez, and Jeniffer Katerine Carrillo Gómez. Electronic nose dataset for copd detection from smokers and healthy people through exhaled breath analysis. *Data in Brief*, 35:106767, 2021.

Marcel Bruins, Zeaur Rahim, et al. Diagnosis of active tuberculosis by e-nose analysis of exhaled air. *Tuberculosis*, 93(2):232–238, 2013.

Henrique V. Costa, André G. R. Ribeiro, and Vinicius M. A. Souza. Fusion of image representations for time series classification with deep learning. In *International Conference on Artificial Neural Networks*, pages 235–250. Springer, 2024.

Fedor S Fedorov, Nikolay P Simonenko, Vanessa Trouillet, Ivan A Volkov, Ilya A Plugin, Dmitry P Rupasov, Artem S Mokrushin, Ilya A Nagornov, Tatiana L Simonenko, Ivan S Vlasov, et al. Microplotter-printed on-chip combinatorial library of ink-derived multiple metal oxides as an "electronic olfaction" unit. *ACS Applied Materials & Interfaces*, 12(50):56135–56150, 2020.

Dewei Feng, Carol Li, et al. Smellnet: A large-scale dataset for real-world smell recognition. *arXiv preprint arXiv:2506.00239*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Wladyslaw Homenda, Agnieszka Jastrzebska, Witold Pedrycz, and Mariusz Wrzesien. Time series classification with their image representation. *Neurocomputing*, 573:127214, 2024.

Maxence Lalis, Matej Hladiš, et al. M2or: a database of olfactory receptor–odorant pairs for understanding the molecular mechanisms of olfaction. *Nucleic Acids Research*, 52(D1):D1370–D1379, 2024.

Brian K Lee, Emily J Mayhew, et al. A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381 (6661):999–1006, 2023.

Maribel Rodríguez-Aguilar, Sofía Ramírez-García, et al. Ultrafast gas chromatography coupled to electronic nose to identify volatile biomarkers in exhaled breath from chronic obstructive pulmonary disease patients: A pilot study. *Biomedical Chromatography*, 33(12):e4684, 2019.

Anju Sharma, Bishal Kumar Saha, et al. Olfactionbase: a repository to explore odors, odorants, olfactory receptors and odorant–receptor interactions. *Nucleic Acids Research*, 50(D1):D678–D686, 2022.

Laura Tenero, Marco Sandri, et al. Electronic nose in discrimination of children with uncontrolled asthma. *Journal of Breath Research*, 14(4):046003, 2020.

Rens Van de Goor, Michel van Hooren, et al. Training and validating a portable electronic nose for lung cancer screening. *Journal of Thoracic Oncology*, 13(5):676–681, 2018.

World Health Organization. International statistical classification of diseases and related health problems, 10th revision, 2019. URL `https://icd.who.int/browse10/2019/en`. Version: 2019.

Shuang Zhai et al. Time series as images: Vision transformer for irregularly sampled time series. *Advances in Neural Information Processing Systems*, 36, 2023.