SCENT OF HEALTH (S-O-H): OLFACTORY MULTIVARIATE TIME-SERIES DATASET FOR NON-INVASIVE DISEASE SCREENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Exhaled breath analysis has become an advantageous alternative to traditional medical diagnostic methods. Electronic nose (eNose) sensors can enable low-cost, non-invasive disease screening from exhaled breath. Still, progress is limited by small, site-specific datasets and sensor-specific temporal artifacts (e.g., baseline drift). In this paper, we introduce Scent of Health, the largest printed-metal-oxide eNose clinical dataset with curated temporal splits. We also introduce breath diagnosis as a realistic multivariate time-series task with temporally stratified splits that mimic deployment. We provide a reproducible benchmark, including classical algorithms with handcrafted features, convolutional neural networks with data augmentation, and specialized time series classification methods, and show that, while these methods offer useful inductive biases, substantial gaps remain in robustness and generalization under drift and limited labels. Our findings demonstrate that machine learning for data from eNose can achieve clinically relevant performance in detecting malignant lung neoplasms and differentiating respiratory diseases. The substantial sample size of this dataset addresses a critical gap in research and provides a valuable resource for developing and validating disease classification models and olfactory data representation.

1 INTRODUCTION

Recent advances in artificial intelligence have been mainly driven by progress on richly annotated, high-volume datasets and architectures that can exploit temporal and high-dimensional structure (e.g., transformers and modern convolutional models). Yet, despite impressive successes in vision and language, specific sensory modalities remain underserved by openly available benchmarks and accompanying algorithmic studies. One of the most notable is the chemical sensing modalities underlying olfaction and breathomics. Exhaled breath contains complex mixtures of volatile organic compounds that encode clinically relevant metabolic information, and electronic-nose (eNose) technology offers a practical, portable route to digitize these signals for non-invasive diagnostics (Lee et al., 2024). However, the literature relies mainly on small cohorts or site-specific collections, which limits the development of robust representation learning and reliable clinical evaluation for breath-based screening (Li et al., 2023). The scarcity of olfactory data stems from two primary challenges: the intrinsic difficulty of capturing odor information and the lack of standardized, affordable smell-digitization technology. While highly accurate, gold-standard methods like gas chromatography-mass spectrometry are often prohibitively expensive and lack the usability for large-scale data collection. Consequently, researchers are searching for more cost-effective, accurate, and scalable sensor technologies, such as electronic noses, to obtain reliable odor fingerprints. One promising direction is the on-chip printing of tailored metal oxide nanomaterials, a technique that enables the fabrication of dense arrays of analyte-specific microsensors. The resulting devices produce complex, data-rich response patterns to volatile compounds, creating a robust input for AI models in olfactory analytics (Goikhman et al., 2022; Gohel et al., 2024b).

However, two technical gaps impede progress in machine learning for breathomics. First, the field lacks large, well-curated clinical collections that span multiple pathologies and support rigorous out-of-distribution and cross-site evaluation; this scarcity makes it difficult to study model generalization, sensor drift adaptation, and clinically meaningful performance thresholds. Recent efforts

to assemble clinical breath molecule catalogs and to perform cross-site validation show promise but remain limited in scale or scope for broad benchmarking (Kuo et al., 2024). Second, eNose outputs are naturally multivariate time series with sensor-specific dynamics, cross-sensor correlations, and measurement artifacts (e.g., baseline drift), so standard image/text architectures are not directly optimal without careful representation design and domain-aware augmentation.

In this work, we address both gaps. We introduce a large clinical eNose breathomics collection and cast odor diagnosis as a multivariate time-series learning problem with realistic temporal and deployment challenges. Our dataset enables the development and evaluation of both classical and modern time-series techniques, from strong feature-based tabular learners to dedicated time-series architectures, and supports validation against temporal sensor drift and week-wise splits that mimic realistic deployment shifts. To establish informative baselines and highlight the algorithmic challenges that olfactory signals present, we evaluate a spectrum of approaches that have shown strong performance on time-series tasks: random-kernel convolutional methods, deep convolutional ensembles, and recent self-supervised/contrastive representation learning approaches for time series. These methods illustrate complementary trade-offs between speed, sample efficiency, and robustness to temporal perturbations; they also point to promising directions (data augmentation, pre-training, domain adaptation) for future work on sensor-based diagnostics.

The key contributions of this work are the following:

- **S-O-H (Scent of Health)**: A novel and extensive medical olfactory dataset comprising 1,027 patients across eight distinct groups (control and seven clinically significant disease groups), together with recommended train/test splits that control for temporal drift and realistic validation, it is the largest and most diverse dataset for this type of eNose device, capturing breath samples via a unique 17-sensor array of printed metal oxide (ZnO) microsensors on a temperature-controlled chip.
- **Benchmarking and reproducible baselines**: We provide a comprehensive benchmark for odor classification as a multivariate time-series analysis problem. It includes classical feature-based learners, and self-supervised representation baselines, evaluated under splits that expose drift and sample-size limitations.
- **Practical analysis of deployment challenges**: We quantify the effects of sensor drift and temporally concentrated sampling, and we report cross-validation strategies that minimize leakage while reflecting clinical deployment scenarios. These findings align with recent cross-site studies and underscore the need for domain adaptation in eNose applications.

2 Related Work

Exhaled breath analysis as a non-invasive diagnostic method and a way to monitor disease progression has advantages over other traditional methods, such as blood and urine analysis. Exhaled air contains volatile organic compounds (VOCs), which are the end products of organic matter transformations in the body, and changes in the composition of VOCs can be used to diagnose diseases. In the last decade, this technology has been actively introduced into clinical practice as an alternative to traditional research methods, including gas chromatography/mass spectrometry, since gas chromatography and mass spectrometry are quite labor-intensive, expensive, and have low portability (Chen et al., 2021).

eNose technology for breath analysis represents a rapidly advancing field with significant potential to transform medical diagnostics. Substantial progress has been made in demonstrating the clinical validity of this approach for various conditions, particularly lung cancer. The technology offers numerous advantages, including non-invasiveness, rapid results, cost-effectiveness, and potential for point-of-care testing. The eNose system is not inferior to this method and can detect mixtures of volatile metabolites even in low concentrations, without identifying individual chemicals. Machine learning methods allow the electronic nose to accurately identify odors using qualitative and quantitative analysis (Li et al., 2023).

Recently developed sensors (Goikhman et al., 2022) for the eNose applications allow for a large-scale study of the applicability of technology in diagnosing diseases by exhaled air. Previously presented works on the diagnosis of diseases by exhaled air focused on diseases of certain organ groups, for example, the digestive tract (Tiele et al., 2019), lungs and respiratory tract (Baldini et al.,

Table 1: Latest studies on eNose for medical diagnostics

Target disease	Sample size	Metrics (%)	Reference
COPD	56	Acc. 82	(Rodríguez-Aguilar et al., 2019)
Malignant neoplasm of lungs	145	Spec. 84	(Van de Goor et al., 2018)
Chronic renal failure	98	Acc. 86	(Kalidoss et al., 2021)
Tuberculosis	224	Spec. 87	(Bruins et al., 2013)
Diabetes mellitus	240	Acc. 93	(Weng et al., 2023)
Asthma	38	ROC AUC 80	(Tenero et al., 2020)
Malignant neoplasm of rectum	210	ROC AUC 84	(van Keulen et al., 2020)

2020), excretory system (Capuano et al., 2025), or used the data from a limited number of patients or did not use analog technologies of the eNose. General trends in the field include the prevalence of limited datasets, typically comprising tens to hundreds of patients. Table 1 features recent studies that employed eNose technology, primarily with metal oxide sensor arrays, for disease diagnostics via exhaled breath analysis. Regrettably, in most similar studies, datasets are not disclosed, which could otherwise have helped advance the research in olfactory modality for medical diagnostics and other applications. Most studies focus on the applicability of the electronic nose to respiratory diseases. These investigations generally examine either individual diseases or groups of conditions united by their anatomical location (Mortazavi et al., 2025).

3 eNose System Description

3.1 Multielectrode chip design

In this study, a multielectrode chip with 18 Pt (150 nm)/Ti (5 nm) strip co-planar electrodes is utilized to analyze the exhaled breath of patients. The chip, 10×10 mm^2, represents a silicon crystal with a silica layer of ca. 500 nm. Each pair of electrodes, distanced by 50 μm and with a functional material in between, forms an individual sensor segment, 17 in total. On-chip made two meander-shaped thermoresistors and two meander-shaped heaters enabled to control precisely the temperature of the chip surface during gas sensing measurements (Gohel et al., 2024b; Abayarathne et al., 2025; Gohel et al., 2024a). Subsequently, the prepared chip was wired to the ceramic card by ultrasonic bonding and installed in a gas-tight chamber with a chamber volume of 0.76 cm^3. The ceramic card with the chip was connected to a custom-made printed circuit board (PCB) to operate the sensor array and acquire the output signal at a sampling rate of ca. 0.4 Hz. An IR pyrometer Kelvin Compact 1200D. was used to tune the temperature of the chip before and after the tests. Before testing the exhaled breath samples, the chip was kept at 300 ± 5 °C for 24 h in an air atmosphere for stabilization. The operational temperature of the multielectrode chip was maintained at 300 ± 5 °C.

3.2 Synthesis and on-chip printing of functional materials

The synthesis of functional materials for this study included the following route. A solution of lithium hydroxide ($LiOH$, 0.315 g, 0.075 mol) was added to 25 mL of absolute ethanol, using a dropping funnel. Afterwards, the obtained solution was added to solution of zinc nitrate ($Zn(NO_3)_2$), 1.49 g, 0.005 mol) and either indium/silver/cerium nitrate ($In(NO_3)_3$/$AgNO_3$/$Ce(NO_3)_4$) or nickel acetate ($Ni(CH_3COO)_2$, 0.00025 mol) in 25 mL of absolute ethanol. The addition was performed with vigorous stirring while the solution was cooled to 2 °C in an ice-water bath (Ge et al., 2017). The mixture was then stirred for 2 hours under the same conditions. Afterward, the precipitate was purified by centrifugation and rinsing alternately with ethanol and cyclohexane in 6 consecutive cycles. The precipitate was dried in dry air at 60 °C, then annealed in a furnace at 200 °C for 2 hours to finally get the corresponding powders. All chemicals were of at least analytical purity. The synthesized functional materials, i.e., zinc oxide or metal-doped zinc oxides (ZnO, $In-ZnO$, $Ag-ZnO$, $Ce-ZnO$, and $Ni-ZnO$) were placed on the top of the chip using a printing approach. The materials are deposited onto the chip surface using a REGEMAT 3D BIO V1 liquid bioprinter. As inks, particle suspensions are prepared with a particle mass ratio of 5 wt. % in an aqueous ethylene glycol solution (chemically pure, 60 wt. %).

As a result, printed lines with an average width of ca. 300 μm were obtained, each covering three sensors. The prepared chip was annealed at 90 °C to remove the residual solvent.

4 Data Collection Protocol

4.1 Participant Enrollment and Screening

The study cohort comprised patients with specific target nosologies and healthy volunteers. All potential participants were informed about the study's objectives and procedures. Those who agreed to participate provided written informed consent (see A.1) prior to any study-related activities. The study protocol was approved by the Ethics Committee (see Section 7). A physician-researcher screened each potential participant against predefined inclusion and exclusion criteria (detailed in A.2). Only individuals who met all inclusion criteria and none of the exclusion criteria were enrolled in the study.

4.2 Pre-Sampling Preparation and Data Logging

Upon enrollment, the physician completed a standardized participant questionnaire to record demographic and clinical data, including a unique study identification number, full name, year of birth, gender, clinical diagnosis (coded with ICD-10 World Health Organization (2019)), and corresponding Electronic Health Record (EHR) number.

Prior to breath sample collection, participants adhered to a standardized pre-sampling protocol designed to minimize confounding variables. This included a minimum 4-hour fasting period, abstinence from smoking for at least 4 hours, abstinence from alcohol for 48 hours, and the avoidance of perfumes and other strong odors on the day of sampling.

4.3 Breath Sample Collection and Analysis

Exhaled breath samples were collected using individual, sterile, disposable 2-liter bags, chosen for their biocompatibility and standard medical-grade quality. A key methodological constraint was that each participant could provide only a single sample for the entire study; the provision of multiple samples for the investigation of different diseases was not permitted.

Following this preparation, participants rested in a seated position for five minutes before providing a single deep exhalation into the bag. The electronic nose (eNose) device was initialized, with system status confirmed via indicator lights.

For measurement, the sampling bag was connected to the eNose's intake port via a cross-shaped valve. An airtight seal was verified by a pressure sensor triggered upon gentle compression of the bag. The measurement cycle was initiated from the software interface, with the multivariate time-series sensor data automatically saved to a database upon completion. Each sampling bag was discarded after a single use, and the eNose's sensing chamber was purged with clean, dry air between samples to prevent cross-contamination. Samples were analyzed immediately or stored at room temperature for no more than 4 hours to ensure sample integrity.

4.4 Collected data

Patient data is stored in a JSON structure. The primary sensor for olfactory analysis is the "eNose", which records a multivariate time series of approximately 15 minutes in duration across 17 distinct channels. Auxiliary sensors within the device simultaneously monitor environmental parameters, including temperature, pressure, humidity, and CO_2 levels, which are stored in separate JSON fields.

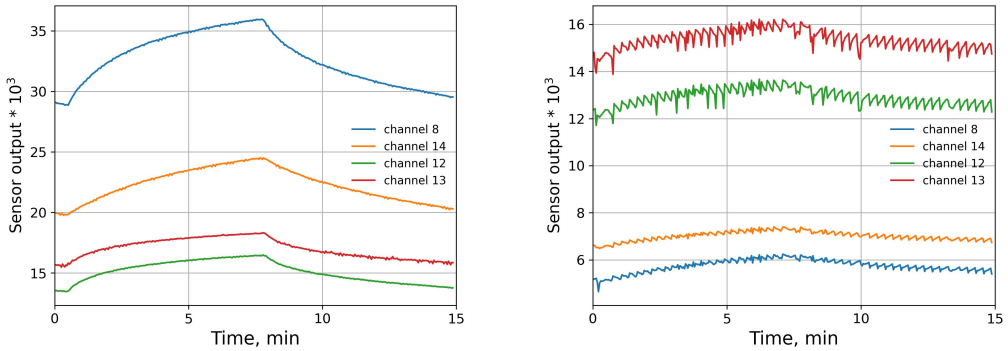
A total of 1027 samples were collected. The distribution of the patient cohort is presented in Table 2. The minimum age of patients in the study is 18, and the maximum is 89. In general the distribution by gender was as follows: 567 (55.4%) women and 457 (44.6%) men.

The sensor output exhibits substantial variation in magnitude across its 17 channels, with each channel degrading at a unique rate over time (Fig. 1). The peak response for all channels occurs near the

Table 2: Distribution and demographic characteristics of the S-O-H dataset samples by diagnostic group

Diagnostic Group (ICD-10 Code)	N	%	Mean Age	Age SD	Male, %
Healthy Individuals (Z00)	164	16.0	38.77	12.76	0.25
Hepatitis B/C (B18)	138	13.5	50.88	13.44	0.55
Gastritis and Duodenitis (K29)	138	13.5	52.32	16.27	0.36
Non-alcoholic Fatty Liver Disease (K76)	128	12.5	47.96	16.21	0.39
Diabetes Mellitus Type II (E11)	128	12.5	60.40	11.92	0.30
Chronic Renal Failure (N18)	128	12.5	59.82	14.61	0.50
COPD (J44)	100	9.8	64.16	10.43	0.64
Lung Cancer (C34)	100	9.8	66.71	8.63	0.73
Total	**1027**	**100.0**	**52.9***	**15.1***	**0.41***

Note: SD = Standard Deviation.
* Weighted average or overall proportion for the entire cohort.



(a) 2 days of sensor exploitation

(b) 53 days of sensor exploitation

Figure 1: An example of sensor output and signal deterioration over time.

timestamp labeled ”endTimeGases” in the associated JSON file, while the ”durationSec” parameter records the total length of the measurement.

4.5 Validation scheme

The data was collected over 11 weeks (Fig. 2). As exhaled air samples were gathered from outpatient patients, ensuring consistent collection for each disease became challenging. The experiment was designed so that the analysis of patients with specific conditions was heavily concentrated in time. Due to sensor degradation and drift over time, data leakage in the disease classification was possible. To minimize this effect, we suggested an individual train/test split for each of the eight conditions. There were two criteria for selecting patients for validation. Firstly, the positive validation samples should be distant from the positive training samples. Secondly, the negative validation samples should include samples that are close in time to the negative train samples. We divided the experiments into weeks and used these chunks to create the train/test splits. The data collection period spanned over eleven weeks, three of which were used for validation, and the rest for training. The suggesed train/test split is provided in Table 3.

5 Experimental Evaluation

5.1 CNN-based Odor Signal Map Classification

The key idea behind this approach is to treat time series data as images Semenoglou et al. (2023), Hatami et al. (2018). To address sensor degradation over time, the time series data is smoothed using the weightlet transformation and then normalized using min-max scaling. Next, polynomial features

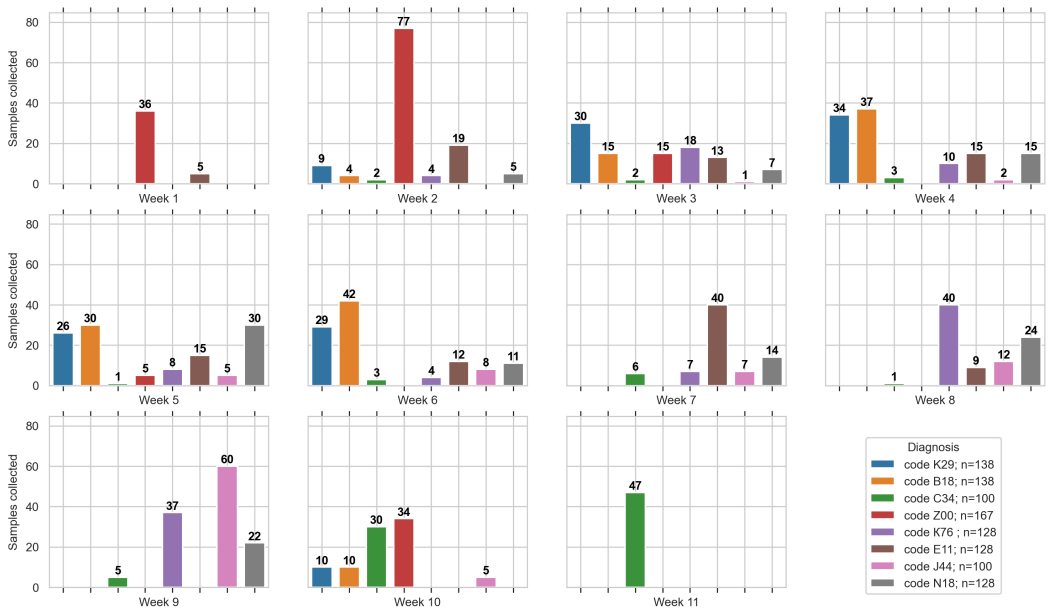


Figure 2: Distribution of the samples collected from patients throughout the study.

Table 3: Suggested train/test split. Every sample from each week is put into either the train or the test subsets. Regarding a particular ICD-10 code as a positive class, "+" and empty indicate weeks selected for test and train, respectively

ICD-10 code	Z00	E11	K29	K76	B18	C34	N18	J44
Week								
1		+	+					
2		+	+	+	+		+	
3		+	+	+	+		+	
4				+			+	
5								+
6								+
7								+
8						+		
9	+					+		
10	+				+	+	+	
11	+							+

are extracted from the smoothed and normalized time series using the PolynomialFeatures function from the sklearn.preprocessing library. These features are then combined with the original data to create a floating-point matrix that can be interpreted as an image. This result of the aggregation is illustrated in Fig. 3. A neural network with several convolutional and fully connected layers processes the data and classifies it. The inference time for the whole process is less than one second on a laptop using only the CPU, so the algorithm could be implemented on embedded systems.

5.2 Tabular Odor Classification with CatBoost

Our hypothesis for this approach was grounded in the assumption that readings from the sensors exhibit a curve pattern with a saturation plateau, which can be modeled as a kinetic curve. Additionally, we used other basic statistical features to describe our time series data Faouzi (2022). Prior to any preprocessing, the time series data is clipped according to the 'startTimeGases' and 'endTimeGases' parameters from the breath analysis experiment. The time series data is smoothed with

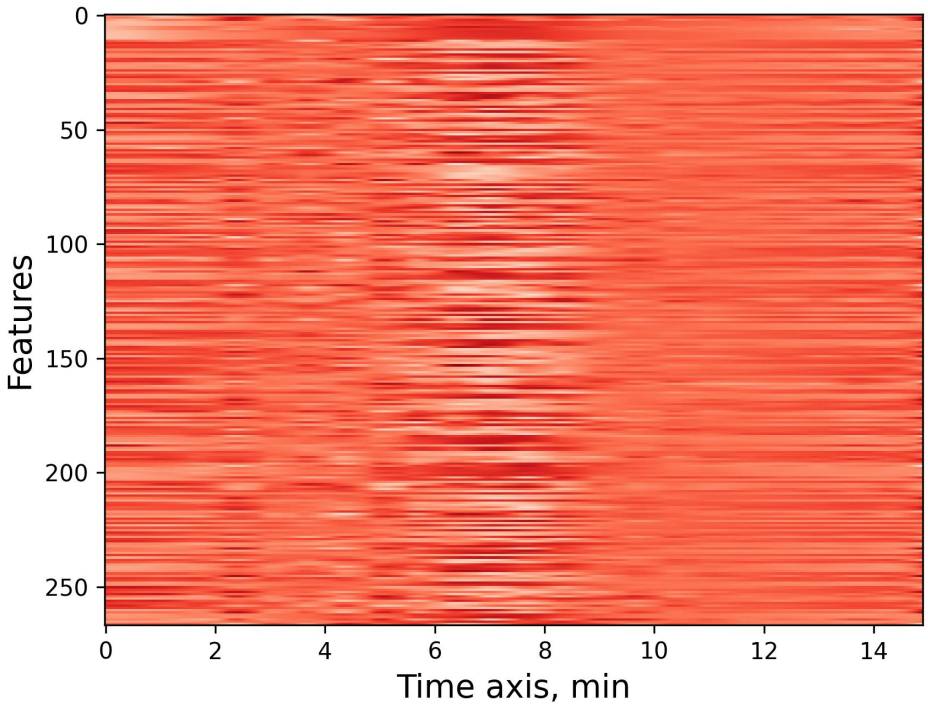


Figure 3: The interpretation of time series features as an image.

a median filter from scipy.ndimage library to address sensor degradation over time and subsequently normalized using min-max scaling independently for each 1D-time series.

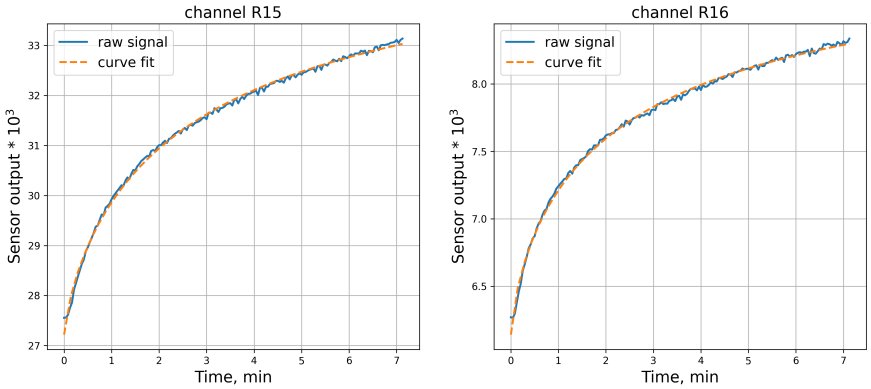


Figure 4: An example of fitted function for the signal time series of interest.

$$f(t) = R_0 + \frac{R_{\max} - R_0}{1 + \left(\frac{t_{50}}{t}\right)^k} \tag{1}$$

Initially we fitted each time series with a function (1) using curve_fit method from scipy.optimize library (Fig. 4). From the four resulting function parameters we selected two sets of features:

Table 4: Binary classification results for the proposed train/test split, ROC AUC (**best metrics**)

Model	Features	Z00	E11	K29	K76	B18	C34	N18	J44
CatBoost	logfit_4	0.454	0.535	0.407	0.663	0.689	0.515	0.598	0.496
	logfit_2	0.482	0.536	0.507	**0.669**	0.657	0.492	0.531	0.467
	stats_5	0.513	0.557	0.464	0.642	**0.703**	0.52	0.553	0.443
	stats_3	0.513	0.557	0.464	0.642	**0.703**	0.52	0.553	0.443
CNN	polynomial	**0.635**	0.548	0.564	0.600	0.615	**0.713**	**0.697**	0.508

- logfit_4: R_{max}, R_0, t_{50}, k;
- logfit_2: t_{50}, k.

We also compared this approach to two sets of basic statistical parameters:

- stats_3: minimum, maximum, mean, median and standart deviation values of each time series;
- stats_5: mean, median and standart deviation values of each time series.

The feature engineering methods were applied exclusively to eight specific channels from the eNose array, as these channels exhibited consistent signaling over time, a crucial requirement for effective feature extraction. The resulting feature vectors for each patient comprised 32, 16, 24, and 40 elements, respectively, corresponding to each feature set. We adopted the following hyperparameters for CatBoost classification: 1500 iterations, learning rate 0.05, loss function 'Logloss', max tree depth 10, L2 leaf regularization 8. The resulting binary classification results (Table 4) indicate two key insights for this straightforward approach. First, the inclusion of the minimum and maximum time series values does not significantly affect model performance, as the preprocessing steps render these parameters ineffective. Second, the fitted function parameters appear to be suboptimal and fail to fully capture the information contained in the time series data.

5.3 Results discussion

The results shown in Table 4 highlight the significant potential for disease screening using the metal oxide electronic nose (eNose) sensor. The CNN-based method demonstrates promise in distinguishing between healthy individuals and those with specific conditions, successfully identifying four out of seven targeted conditions. While the feature-based approach performs better in classifying fatty liver disease (K76) and hepatitis (B18), convolutional contrastive learning shows promising results for malignant neoplasms of the lungs (C34) and chronic renal failure (N18). The best classification results were obtained for hepatitis and malignant lung formations, with ROC AUC values surpassing 0.7.

In the future, we intend to gather more data from a variety of sensors and conduct experiments to ensure that disease screenings are evenly distributed over time. To effectively demonstrate the statistical significance of our results, a multi-fold experimental setup will be essential. In the current experimental conditions, this was not possible due to the uneven distribution of disease screenings over time.

Our baseline methods did not demonstrate significant effectiveness in classifying certain conditions, such as gastritis, duodenitis, diabetes mellitus, and chronic obstructive pulmonary disease. It may be necessary to explore different sensor configurations to improve the classification of these diseases.

6 Conclusion

This paper presents Scent of Health, a large clinical eNose breathomics collection with a reproducible benchmark and a suite of baselines that expose the practical challenges of olfactory time-series, particularly sensor drift, limited labeled data, and cross-site variability. Our experiments show that while modern time-series classifiers provide strong starting points, significant gaps remain in

robustness and generalization under realistic temporal splits, highlighting the need for targeted augmentation, domain-adaptive pretraining, and sensor-aware modeling. By releasing the data, splits, and code, we aim to catalyze machine-learning research on olfactory data, from improved time-series architectures and pre-training strategies to robust adaptation techniques for sensor networks, and to accelerate the translation of eNose technology toward reliable, non-invasive disease screening. The future work will focus on scalable pretraining across sites, principled drift-correction methods, and prospective clinical validation to move eNose systems from promising prototypes toward reliable real-world tools.

7 Ethics statement

Ethics approval from a local Ethics Committee at the ***Medical Research Organization*** was received before the start of the study. Prior to their inclusion in the study, all participants provided written informed consent.

8 Reproducibility statement

Once the paper is accepted, we will make our dataset publicly available and provide a DOI. For review purposes, the csv and json files are temporarily available at: `https://figshare.com/s/98fe218ce5b256fd3ed3`, and `https://figshare.com/s/cab6ea5f4b034b86218f`. The code is available at `https://anonymous.4open.science/r/enos-8FD9/README.md`.

References

HMI Abayarathne, VR Gohel, NP Simonenko, K Zamansky, G Meng, TL Simonenko, V Zaytsev, V Kondrashov, S Yu Kottsov, A Moklokov, et al. "equivalent" multisensor array based on zirconium doped zinc oxide for indoor environment monitoring. *Journal of Environmental Chemical Engineering*, 13(3):116807, 2025.

Chiara Baldini, Lucia Billeci, Francesco Sansone, Raffaele Conte, Claudio Domenici, and Alessandro Tonacci. Electronic nose as a novel method for diagnosing cancer: a systematic review. *Biosensors*, 10(8):84, 2020.

Marcel Bruins, Zeaur Rahim, Albert Bos, Wendy WJ van de Sande, Hubert Ph Endtz, and Alex van Belkum. Diagnosis of active tuberculosis by e-nose analysis of exhaled air. *Tuberculosis*, 93(2): 232–238, 2013.

Rosamaria Capuano, Valerio Allegra, Alexandro Catini, Gabriele Magna, Manuela Di Lauro, Giulia Marrone, Antonio Agresti, Sara Pescetelli, Massimo Pieri, Roberto Paolesse, et al. Disposable sensor array embedded in facemasks for the identification of chronic kidney disease. *ACS sensors*, 2025.

Ting Chen, Tiannan Liu, Ting Li, Hang Zhao, and Qianming Chen. Exhaled breath analysis in disease detection. *Clinica Chimica Acta*, 515:61–72, 2021.

Johann Faouzi. Time series classification: A review of algorithms and implementations. *Machine Learning (Emerging Trends and Applications)*, 2022.

Yuru Ge, Zhong Wei, Yushu Li, Jiang Qu, Baiyi Zu, and Xincun Dou. Highly sensitive and rapid chemiresistive sensor towards trace nitro-explosive vapors based on oxygen vacancy-rich and defective crystallized in-doped zno. *Sensors and Actuators B: Chemical*, 244:983–991, 2017.

Vishalkumar Rajeshbhai Gohel, Margarita Chetyrkina, Andrey Gaev, Nikolay P Simonenko, Tatiana L Simonenko, Philipp Yu Gorobtsov, Nikita A Fisenko, Darya A Dudorova, Valeriy Zaytsev, Anna Lantsberg, et al. Multioxide combinatorial libraries: fusing synthetic approaches and additive technologies for highly orthogonal electronic noses. *Lab on a Chip*, 24(16):3810–3825, 2024a.

Vishalkumar Rajeshbhai Gohel, Andrey Gaev, Nikolay P Simonenko, Tatiana L Simonenko, Elizaveta P Simonenko, Anna Lantsberg, Valeriy Zaytsev, Albert G Nasibulin, and Fedor S Fedorov. Gas sensing beyond classification: Analysis of gas mixtures using multisensor array based on al-doped zinc oxide. *Microchemical Journal*, 206:111547, 2024b.

Boris V Goikhman, Fedor S Fedorov, Nikolay P Simonenko, Tatiana L Simonenko, Nikita A Fisenko, Tatiana S Dubinina, George Ovchinnikov, Anna V Lantsberg, Alexey Lipatov, Elizaveta P Simonenko, et al. Quantum of selectivity testing: detection of isomers and close homologs using an azo based e-nose without a prior training. *Journal of Materials Chemistry A*, 10(15): 8413–8423, 2022.

Nima Hatami, Yann Gavet, and Johan Debayle. Classification of time-series images using deep convolutional neural networks. In *Tenth international conference on machine vision (ICMV 2017)*, volume 10696, pp. 242–249. SPIE, 2018.

Ramji Kalidoss, Snekhalatha Umapathy, and Usha Rani Thirunavukkarasu. A breathalyzer for the assessment of chronic kidney disease patients' breathprint: Breath flow dynamic simulation on the measurement chamber and experimental investigation. *Biomedical Signal Processing and Control*, 70:103060, 2021.

Ping-Hung Kuo, Yue-Chen Jhong, Tien-Chueh Kuo, Yu-Ting Hsu, Ching-Hua Kuo, and Yufeng Jane Tseng. A clinical breathomics dataset. *Scientific Data*, 11(1):203, 2024.

Meng-Rui Lee, Mu-Hsiang Kao, Ya-Chu Hsieh, Min Sun, Kea-Tiong Tang, Jann-Yuan Wang, Chao-Chi Ho, Jin-Yuan Shih, and Chong-Jen Yu. Cross-site validation of lung cancer diagnosis by electronic nose with deep learning: a multicenter prospective study. *Respiratory Research*, 25(1): 203, 2024.

Ying Li, Xiangyang Wei, Yumeng Zhou, Jing Wang, and Rui You. Research progress of electronic nose technology in exhaled breath disease analysis. *Microsystems & Nanoengineering*, 9(1):129, 2023.

Sajjad Mortazavi, Somayeh Makouei, Karim Abbasian, and Sebelan Danishvar. Exhaled breath analysis (eba): A comprehensive review of non-invasive diagnostic techniques for disease detection. In *Photonics*, volume 12, pp. 848. MDPI, 2025.

Maribel Rodríguez-Aguilar, Sofía Ramírez-García, Cesar Ilizaliturri-Hernández, Alejandro Gómez-Gómez, Evelyn Van-Brussel, Fernando Díaz-Barriga, Susanna Medellín-Garibay, and Rogelio Flores-Ramírez. 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.

Artemios-Anargyros Semenoglou, Evangelos Spiliotis, and Vassilios Assimakopoulos. Image-based time series forecasting: A deep convolutional neural network approach. *Neural Networks*, 157: 39–53, 2023.

Laura Tenero, Marco Sandri, Michele Piazza, Giulia Paiola, Marco Zaffanello, and Giorgio Piacentini. Electronic nose in discrimination of children with uncontrolled asthma. *Journal of Breath Research*, 14(4):046003, 2020.

Akira Tiele, Alfian Wicaksono, Jiten Kansara, Ramesh P Arasaradnam, and James A Covington. Breath analysis using enose and ion mobility technology to diagnose inflammatory bowel disease—a pilot study. *Biosensors*, 9(2):55, 2019.

Rens Van de Goor, Michel van Hooren, Anne-Marie Dingemans, Bernd Kremer, and Kenneth Kross. Training and validating a portable electronic nose for lung cancer screening. *Journal of Thoracic Oncology*, 13(5):676–681, 2018.

Kelly E van Keulen, Maud E Jansen, Ruud WM Schrauwen, Jeroen J Kolkman, and Peter D Siersema. Volatile organic compounds in breath can serve as a non-invasive diagnostic biomarker for the detection of advanced adenomas and colorectal cancer. *Alimentary pharmacology & therapeutics*, 51(3):334–346, 2020.

Xiaohui Weng, Gehong Li, Ziwei Liu, Rui Liu, Zhaoyang Liu, Songyang Wang, Shishun Zhao, Xiaotong Ma, and Zhiyong Chang. A preliminary screening system for diabetes based on in-car electronic nose. *Endocrine Connections*, 12(3), 2023.

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.

A Appendix

A.1 Patient Informed Consent Form

Information for a research participant

Dear participant in the study!

You are invited to participate in the "Medical digital nose" research work. Participation in this study is voluntary; if you refuse, this will not affect the quality of the provision of medical care to you. You can stop participating in the study at any time. Your participation in the study can also be stopped at any time by your attending physician, research doctor, or study participant. You will not have any direct benefits from participating in this study, in addition to the fact that the data obtained can subsequently serve for the development of medical science. The purpose of this scientific research is to check the hypothesis of the existence of a dependence between the molecules of the air exhaled by a person and diseases, to create a prototype of a medical AI approval for quick mass screening/early diagnosis of diseases. The data obtained during the study will allow the "digital nose" to accurately identify a set of substances released with exhaled air in patients with certain diseases. This will further simplify the diagnosis, treatment, and conduct of patients with these diseases.

You have been proposed to participate in this study because you have previously established one of the following diagnoses: ——-. You may also be offered participation in a group of healthy people.

This scientific study is conducted by the research team ————————-. In total, the study is planned to include 1024 participants.

If you agree to take part in this study:

1. You will be asked questions regarding your demographic data (age and gender) and the presence of bad habits.
2. If the research doctor, based on the data of your medical card, decides that you can take part in the study, you will be offered to exhale in a special bag equipped with an individual trunk, for 1 minute under the researcher's observation.

Afterward, the air from this bag will be analyzed in a special device, and the data obtained will be used to train ML models. No other medical procedures or subsequent visits are provided during the study. Therefore, this study does not bear any additional risks for you compared to the risks in everyday life or during an outpatient medical examination and testing. No payments for your participation in this study are provided. All information obtained during the above scientific research will be strictly confidential and processed with strict compliance with the norms of the current legislation on protecting medical secrets and individuals' personal data. The data obtained during the study, including medical information, will be impersonated by a research doctor. The information received during this scientific research can also be analyzed and designed in a scientific publication. The information identifying you will not be used anywhere, and it will be impossible to connect this data with you by establishing your personality.

An independent ethics committee approved the conduct of this scientific research at ————————————.

If you have questions regarding your rights as a participant in the study, you can contact the representative of the Independent Ethics Committee observing the study in this research center:

Full name of the contact person: ———— Phone number: ————- Contact information: The main researcher is the head of the department ————————-.E-mail: ———— phone: ————-

Informed consent of the research participant

I, ————————— (surname, name, patronymic of the patient) have read information about the scientific research "Medical digital nose", and I agree to participate in it. I confirm that the essence, purpose, and risks of scientific research were explained to me clearly and in detail by a doctor or another research team member. I had enough time to decide on participation in the study. I had the opportunity to ask all the questions I was interested in, and I received comprehensive answers to each one. I understand that I can at any time, at my desire, abandon further participation in the study, and if I do this, this will not affect my subsequent treatment and the attention of doctors. I permit researchers and the scientific organization conducting the study to process all personal data I reported and information about the current state of my health received by the medical institution, both during the implementation of the above scientific research and previously available. I voluntarily agree that my depersonalized data obtained during scientific research will be used for scientific purposes and published under the condition of compliance with the rules of confidentiality. I allow the medical personnel participating in the study to contact me with potentially necessary additional information on the further state of my health and proposals for participation in new research. I received a copy of "information for the patient and the patient's informed consent."

Full name of the participant in the study

Signature of the research participant; Date

Full name of a research doctor

Signature of a research doctor; Date

A.2 Inclusion and exclusion criteria for the study

Inclusion criteria for the study:

1. Men and women aged 18 to 88 years;
2. The presence or the absence of established diseases for a group of patients or conditionally healthy volunteers, respectively:
   - Non-alcoholic fatty liver disease: K76*
   - Gastritis and duodenitis: K29*
   - Chronic obstructive pulmonary disease: J44*
   - Diabetes mellitus type 2: E11*
   - Chronic renal failure: N18*
   - Malignant neoplasm of the lungs: C34*
   - Hepatitis B and C: B18*
   - Conditionally healthy - patients without specified ICD codes in the EHR
3. Availability of signed and dated informed consent from the patient/conditionally healthy volunteer to participate in the study;
4. No pregnancy or lactation at the time of the study (according to the patient).

Non-inclusion criteria:

1. Inability to fully inhale and/or exhale the required volume;
2. Other diseases and conditions that, in the opinion of the researcher (healthcare professional), may affect the study results or negatively affect the patient's condition.

Exclusion criteria for the study:

1. Withdrawal of informed consent by a research participant;
2. Inability of a research participant to understand and follow the instructions of a health care professional.