UC Irvine
UC Irvine Previously Published Works

Title

Breast cancer detection by analyzing the volatile organic compound (VOC) signature in human urine.

Permalink

Journal

Authors

Giró Benet, Judit
Seo, Minjun
Khine, Michelle
et al.

Publication Date

DOI

Copyright Information

scientific reports

OPEN

Breast cancer detection by analyzing the volatile organic compound (VOC) signature in human urine

Judit Giró Benet[1]✉, Minjun Seo[1], Michelle Khine[2], Josep Gumà Padró[3], Antonio Pardo Martnez[4] & Fadi Kurdahi[1]

A rising number of authors are drawing evidence on the diagnostic capacity of specific volatile organic compounds (VOCs) resulting from some body fluids. While cancer incidence in society is on the rise, it becomes clear that the analysis of these VOCs can yield new strategies to mitigate advanced cancer incidence rates. This paper presents the methodology implemented to test whether a device consisting of an electronic nose inspired by a dog's olfactory system and olfactory neurons is significantly informative to detect breast cancer (BC). To test this device, 90 human urine samples were collected from control subjects and BC patients at a hospital. To test this system, an artificial intelligence-based classification algorithm was developed. The algorithm was firstly trained and tested with data resulting from gas chromatography-mass spectrometry (GC–MS) urine readings, leading to a classification rate of 92.31%, sensitivity of 100.00%, and specificity of 85.71% (N = 90). Secondly, the same algorithm was trained and tested with data obtained with our eNose prototype hardware, and class prediction was achieved with a classification rate of 75%, sensitivity of 100%, and specificity of 50%.

The US female population is 165.92 million women. The US Preventive Services Task Force (USPSTF) recommends biennial screening mammography for women aged 50–74 and recommends against screening younger and older women due to current evidence being insufficient to prove its benefits[1]. This reduces BC mortality by 15%[2]. According to the CDC, only 66% of women over 40 years of age do attend BC screenings annually or biannually[3]. The main reason why women skip their mammogram appointment is pain[4]. In the US, 451,936 BCs are detected yearly[5], 150,000 of which are detected late-stage because of mammogram absenteeism or inefficiency of current methods. In fact, the impact of implementing a mammogram-based screening on breast cancer mortality has been under discussion for a very long time. A review by Moss et al. suggests that although most statistical studies report a decrease in breast cancer mortality over the years, such a decrease can also be observed before the implementation of screening and in age ranges excluded from screening[6].

If BC is detected in an early stage, its metastasis risk is minimal as well as the likelihood of a subsequent metastasis, patient suffering and death. If detected in a later stage, however, patients will require treatment and potentially a mastectomy, which has a notorious impact on women's mental health. These interventions have an approximate cost of $25,000 per patient. Considering that 80% of the US population is covered by insurance, health insurances could save over $3 billion yearly if all BCs detected in their patients were detected in an early stage. Additionally, apart from the mastectomy, if cancer progresses and the patient lives with metastasis, the cost of her treatment (monoclonal antibodies, sometimes conjugated with chemotherapy, cyclines inhibitors, hormone therapy, palliative radiotherapy...) is notorious. Blumen et al. report yearly costs per patient of up to $82,121 for BCs detected in stage I/II as opposed to $129,387–$134,682 if detected in stage III/IV, mainly due to chemotherapy[7].

BC is the leading cause of death by cancer in women under 40 years of age, especially among African–American women. Young women present a denser mammary tissue than post-menopause women, thus decreasing

[1]Center for Embedded Cyber-Physical Systems (CEPS), University of California Irvine (UCI), Irvine 92697, USA. [2]Department of Biomedical Engineering, University of California Irvine (UCI), Irvine 92697, USA. [3]South Catalonia Oncology Institute (IOCS), Sant Joan de Reus University Hospital, IISPV, Rovira i Virgili University, 43204 Reus, Spain. [4]Department of Electronic and Biomedical Engineering, Universitat de Barcelona (UB), 08028 Barcelona, Spain. ✉email: jgirbene@uci.edu

mammogram's sensitivity and specificity in small tumors[8]. In addition to that, their tissue is more sensitive to mammography's radiation dose[9]. Although its dose is not substantial enough to be considered harmful, biennially exposure to mammography could increase BC risk[10,11]. This observation, shared by many authors, started a debate on the worthiness of mammogram-based BC screenings[12]. Furthermore, in 2017, the World Health Organization (WHO) published the "WHO Position paper on mammography screening". Reference[13,14] stating an urgent need for a new radiation-free and sensitive BC screening solution. In other words, although mammography is currently the best solution to reduce BC mortality, there is still room for improvement.

Current methods for breast cancer screening.

Most healthcare systems base their BC screening programs on image-based technologies, i.e. mammography, ultrasonography and magnetic resonance imaging (MRI). The most common reason why women seek medical advice related to BC is having detected a mass in their breasts. However, 90% of these will be found to not be cancer but other benign lesions such as fibroadenoma[15]. Actually, image-based techniques typically fail to correctly identify tumors in women with fibrocystic breasts (breast tissue with healthy lumps) because of highly dense breast tissue[16]. Around 53–60% of women worldwide have this condition[17]. This fact does not only lead to numerous false positives but also increases the recommended screening frequency (and thus the potentially cancerogenic small radiation dose).

When the tumor is small or the mammary tissue is dense, MRI is typically the preferred approach, since it is highly specific and sensitive in addition to being a radiation-free technique. For this reason, BRCA1 or BRCA2-positive women (who present a high risk of developing BC) are screened from a younger age using MRI. Women with fibrocystic mastopathy (whose breasts have difficult tumor visualization) and patients whose mammogram results are inconclusive are typically screened with MRI as well[18]. However, screening 1 M women costs $640 M and $216 M if done with MRI and mammography respectively[19]. There is therefore a tendency of enhancing mammography by processing the resulting image with an AI-based classification algorithm. However, even if this solution can potentially greatly increase the mammogram's sensitivity, it does not solve another problem: this technique is painful to some women and thus causes screening absenteeism.

State of the art: electronic noses.

Even though there do exist some publications on odor-based cancer screenings—i.e. electronic noses or eNoses, the tech transference rate is extremely low. The first eNose, proposed by Persaud and Dodd[20] in Nature in 1982, was designed to loosely mimic the human olfactory pathway. An eNose is based on a sensor array that responds differently according to specific VOCs. In 2014, Asimakopoulos et al.[21] set the race towards an eNose approach for cancer screening with the first study on an eNose-based prostate cancer identification, achieving a specificity of 93% (n = 41). Furthermore, Guerrero-Flores et al.[22] describes a cervix cancer screening in which a dog detects cancer cells in blood by smelling it. Buszewski et al.[23] also presents a VOC-based screening from breath through canine smell and Blatt et al.[24] describes the same screening carried out by an eNose. Phillips et al.[25], describe a BC screening using an electronic device that analyzes the patient's breath. Burton et al.[26] and Guo et al.[27] among others report on some biomolecules that could allow a BC screening from urine samples. In this case, not only does it detect the presence of cancer but also the stage: I and II (early) versus III and IV (advanced). However, in these studies, the dogs/eNoses may have responded to odors associated with cancer, such as inflammation or metabolic products, rather than specifically to cancer itself[28]. Hence, the future of volatilome-based screening should not seek a cancer odor but rather a specific VOC pattern for each cancer type, which has already been proven feasible with BC among others[29].

Technologies for artificial odoring.

Some electronic noses have already reached the industry: vReCIVAő Breath Sampler by Owlstone Medical analyzes breath to detect lung cancer. Aeonose is a certified medical device that can screen colorectal cancer with a sensitivity of 95% and specificity of 64% and advanced adenomas with sensitivity and specificity of 79% and 59% respectively (N = 511) from exhaled-breath[30]. NASA has also developed an eNose, a smartphone-based device to monitor air quality inside spacecraft. And they have recently modified it to screen people for COVID-19 in a low-cost and efficient manner. Heracles by Alpha MOS is an electronic nose based on ultra-fast gas chromatography that classifies wines[31], olive oil[32] and other substances. Although being the best-performing approach according to most authors[33], metal-oxide sensors are not the only low-cost and portable solution for artificial odoring. An eNose based on a nanosensor array with gold nanoparticles (GNP)[34,35] or a quartz microbalance[36,37] are different types of sensors that have proven high sensitivity and specificity applied to BC and lung diseases respectively. Peng et al.[34] present a tailor-made array of cross-reactive nanosensors based on organically functionalized gold nanoparticles and the GC–MC technique (GC–MS) that distinguishes the breath patterns of different cancers. Figure 1 summarises various eNose descriptions found in bibliography.

Aims of our study.

The medical community has long accepted the fact that BC produces metabolic changes in human physiology[27–41]. More specifically, Vignoli et al. emphasise on a "BC metabolomic signature in breast tissue, blood, serum/plasma and urine", detected through Nuclear Magnetic Resonance spectroscopy (NMR)[42]. A metanalysis by Dent et al.[33] reviews the findings from several authors from 2003 up to 2012 and concludes that the "VOC fingerprint" differs significantly among publications. In conclusion, the key to successfully identifying cancer is focusing on the proportion between VOCs rather than on VOCs themselves.

The aim of this study is therefore to obtain multidimensional data related to urine smell and analyze it using statistical algorithms. Like the human nose, the implemented software will respond in concert to a given set of odors—a pattern or *smellprint*—which will be analyzed, compared with stored patterns, and recognized. The device under study is based on the principle that BC causes certain inflammatory processes, resulting in specific

Published eNose approaches				
Sensor type Metal oxide semiconductors	Gold nanoparticles	Quartz microbalance	Colorimetry	Cyranose 320
References [44-46]	[33, 34]	[35, 36]	[47]	[48, 49]
Electronic fundament 2 metal oxide semiconductors and an array of 10 metal oxide semiconductor field-effect transistor sensors.	14 GNP electrodes with changing resistance when in contact with VOCs.	8-sensor array of oscillating quartz crystals coated with varied metalloporphyrins that adsorb VOCs.	36 spots impregnated with chemically sensitive compounds, forming an array on a disposable cartridge.	32 built-in carbonblack polymer composite chemoresistors array and a processor.
Physical principle VOC-dependant resistance that outputs a voltage decay proportional to the sensed combination of VOCs.	Change in physical properties depending on their size and shape (altered by contact with VOCs).	Adsorption of VOCs by a thin quartz crystal area results in a variation of its mass, which induces a variation of the oscillation frequency.	Adsorption of VOCs to the dots of metalloporphyrins causes them to change in color. Change is quantified and denotes VOC composition.	Sensor swells upon VOCs exposure causing an increase in electrical resistance.
Cancer type, sample Bladder cancer, headspace urine sample. Prostate cancer, cell medium	Breast, lung, colorectal, prostate cancer, breath samples	Lung cancer, breath samples (headspace)	Lung cancer, breath samples	Lung cancer, pleural cancer; breath samples (solid-phase)
Specifications Convenient due to its small size, no need for high voltage, and high sensitivity (i.e. no complex pre-processing needed).	Performs irrespective of age, gender... Separates early from the advanced stage.	Good sensitivity for alcohols and ketones and for breath affections apart from lung cancer.	Results are not affected by the subjects' demography, smoking history, or stage of cancer.	Might as well apply to the detection of bacterial pathogens, blood glucose level, bronchogenesis....
Performance 70% accuracy, n=89; 94% accuracy, n=27	88% accuracy, n=72	94% accuracy, n=44; 85% sensitivity, n=98	73.3% specificity, n=143	71.4% sensitivity, n=76; 90% accuracy, n=30

Figure 1. Summary of published eNose approaches to portable low-cost screening solutions: Metal oxide semiconductors[43–45], gold nanoparticles[34,35], quartz microbalance[36,37], colorimetry[46] and the *Cyranose 320*[47,48].

circulating metabolites. These subsequently interact with the excretory system, which translates into the urine of BC patients containing the decomposition products of these metabolites.

Results

Software considerations: classic biostatistics approach.
As depicted in Fig. 2, the first approach consisted of building a first model for sample classification based on classic biostatistics. Before classifying the data, it was pre-processed to ensure that the classification would, later on, rely on the characteristic of cancer instead of on irrelevant artifacts. Pre-processing steps are described in Fig. 3. Later on, a Principal Components Analysis (PCA) was performed, and it was used to prove the hypothesis mentioned below:

Urine samples from breast cancer patients are significantly different from control samples.

The PCA was firstly conducted using GC–MS data. A visual inspection of this classification is displayed in Fig. 4A,B. Once observed that it was possible to discriminate control samples from BC samples, the same procedure was repeated by analyzing the same data with the eNose prototype. This time, the classification algorithm was simplified so that it could run on the edge—e.g., in the microcontroller inside the eNose. It was thus a simplified PCA, consisting of only 2 Principal Components (PCs). The projection of urine samples against PC1 and PC2 is shown in Fig. 4C,D.

Hence, the scientific principle for BC eNose was proved. Below we compare the *smellprint*, obtained from two data sources: GC–MS-acquired data (Fig. 4A,B), versus the eNose-captured signal (Fig. 4C,D).

Software considerations: a machine learning approach.
Once the null hypothesis has been tested, an AI-based algorithm has been implemented to assert the classification capacity of a convolutional neural network (from now on, CNN). This neural network consisted of four convolutional filters of size 32, 64, 128 and linear activation. The output layer of the neural network presented two cells (two outputs) with a softmax activation function. Dropout was implemented to avoid model overfitting. Figure 5 is a sketch of how the model structure was designed. In our first approach to the CNN (ConvNet) model, GC–MS data was used: 90 urine samples from control and BC subjects. This model was trained with 50 epochs of batch size 10. Out of 90 samples, 65 (72.3%) samples were used for training, 12 (13.3%) for testing, and 13 (14.4%) for cross-validation. The achieved training accuracy was 98.20%, and the training loss was 7.70%. As far as validation is concerned, the output of the model
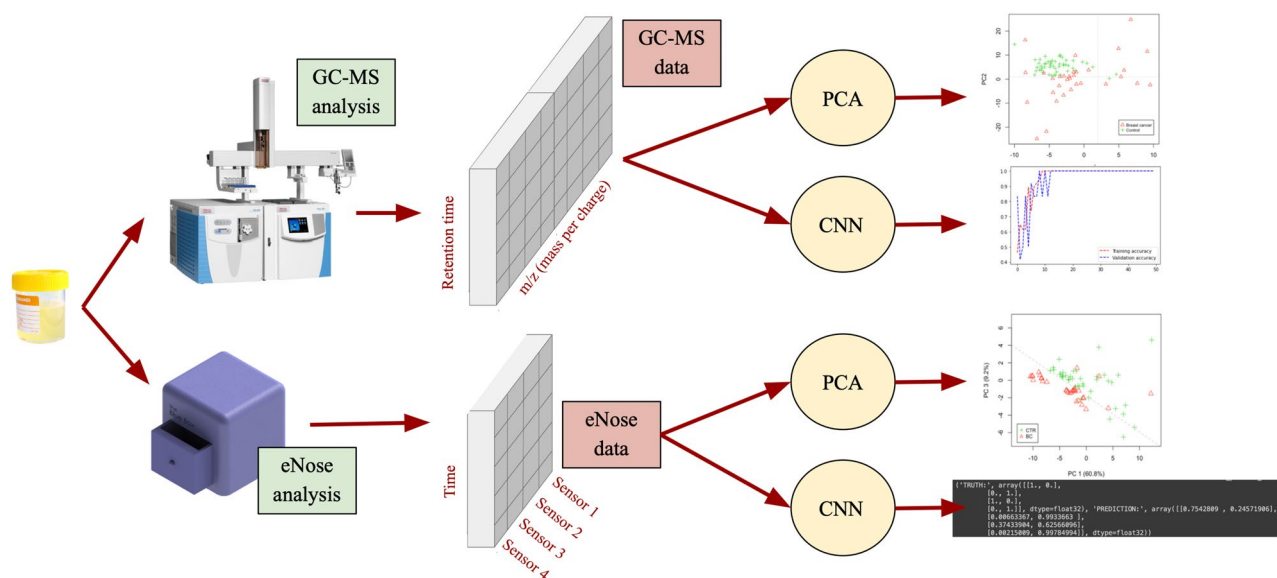
Figure 2. Design of the project. Upon collection, urine samples were analyzed using both an eNose and a GC–MS. Data resulting from the GC–MS analysis contain information relative to the m/z of the VOCs found in the sample per unit of time (retention time). The output of the eNose analysis was a 2D array that was informative of the intensity of each sensors' reaction as a function of time. Finally, both datasets were input to a biostatistics-based classification (PCA) as well as an AI-based classification (CNN).

(with GC–MS data) was 6 true positives, 6 true negatives, 0 false negatives and 1 false positive, presented in a confusion matrix in Table 1. Hence, the *validation accuracy was 92.31%*. As displayed in Table 3, the model has a *sensitivity of 100% and a specificity of 85.71%*.

Similar to the previous section, the model was subsequently trained and tested with eNose-obtained data. Out of 44 samples, 36 (81.8%) samples were used for training, 4 (9.1%) for testing, and 4 (9.1%) for validation. The structure of the model was the same as presented in Fig. 5. The achieved training accuracy was 93.3% and the training loss was 14.72%. As far as validation is concerned, the output of this model consisted on 2 true positives, 1 true negative, 1 false positive and 0 false negatives. These results are presented in a confusion matrix in Table 2. Validation accuracy was 75.00%. As displayed in Table 3, the model has a sensitivity of 100.00% and specificity of 50.00%.

Accuracy, sensitivity and specificity. Upon collection, urine samples were first analyzed using a gas chromatographer–mass spectrometer (GC-MS), which provided an insight into the intensity in which every type of molecule was found at every retention time. In other words, the output consisted of a report for every retention time, which indicated the number of molecules found at that given time for every molecular mass, relative to their ionized charge. Figure 2 shows the shape of the GC–MS data. Later on, this data was processed by a PCA and CNN, which will be detailed as follows. Secondly, urine samples were also analyzed utilizing the eNose prototype that we developed. Please refer to "Methods" section for a detailed description of the prototype. The output of this second analysis was a 2D matrix containing the response of the 4 sensors at every instant of time. As one can observe in Fig. 2, this data was classified by a PCA and a CNN as well. Table 3 displays the accuracy of the two models that were described before: the biostatistics model and the AI-based model respectively.

As one can observe, the introduction of neural networks plays a critical role in enhancing the classification capability of the system. Additionally, it should also be noted that the model classification rate is highly dependent on the sample space size. Therefore we consider it necessary to conduct further studies to keep training the algorithm and achieve a sensitivity and specificity that are acceptable for an oncology screening.

Discussion

As these results denote, machine learning—in this case, CNN—causes a significant impact on sample classification: In the presented case, the accuracy of the model rises from 58.30 to 75.00% (when using eNose data). Additionally, it should be noted that the eNose data set has a smaller dimensionality. Further, because this technology is intended for screening, special attention should be placed on sensitivity rather than specificity. In our case, if we compare our AI-powered model to mammography, our sensitivity is notably higher (100%). Mammography has a sensitivity of 86.9% and its performance is highly dependent on age and tissue density[42]. On the other hand, our highest specificity value is achieved when implementing CNN with GC–MS data.

In conclusion, the experiments carried out in this paper indicate that the implementation of AI in the medical field can yield new approaches and discoveries that classic biostatistics could not reach. This paper also suggests the feasibility of a future potential medical device for in-home and non-irradiating BC screening. Additional studies are yet to be conducted to better identify the classification capacity of the technology. The

6-step pre-processing protocol

STEP 1 — Interpolation along time and m/z axis. Each 2D sample reading informs of the intensity at which VOCs of each size (mass/charge, i.e. m/z) are found in urine per instant of time. Because all samples have been analyzed the same number of times but not at the same specific times, time interpolation is needed. *interpolate_t()* function takes as input the evolution of intensity along time and uses this data to predict the behavior of the function at the desired times. The same logic is used to interpolate over m/z with *interpolate_mz()*. Mass spectra have been predicted as if sampled 1000 equispaced times within 30-300 m/z and 3000 times between 2-20 min
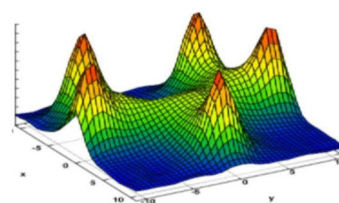
STEP 2 — Chromatogram design. Thanks to interpolation, data can now be integrated along the m/z axis and therefore convert a 3D signal (intensity v.s. time v.s. m/z) into a 2D signal: The evolution of the intensity of all masses along time. From this point onwards, a subject is described by a single vector, the total ion chromatogram (TIC). The resulting TIC is the blue signal presented below.
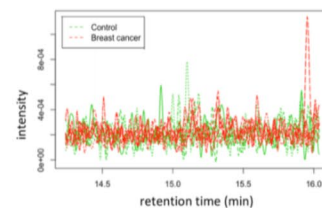
STEP 3 — Baseline removal. GC-MS records the signal together with added noise that increases over time. This baseline low-frequency noise and drift modifications have been removed utilizing *baseline_removal()*, which does not subtract from the sample when intensity increase is not due to noise but an actual chromatogram peak.
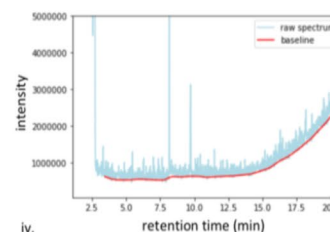
STEP 4 — Savitzky-Golay filter. *SG()* takes each point of the chromatogram together with their 30 surrounding neighbors and estimates the parameters of the 11-order polynomial that best fits this data. This polynomial is then used to set the intensity of the central point to the value the polynomial takes at this place and reduce additive high-frequency noise. An optimal SG filter removes high-frequency contributions to the signal caused by noise but ignores low frequencies.
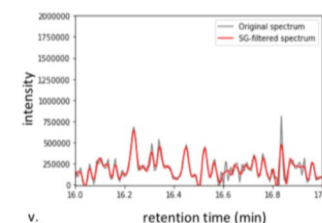
STEP 5 — Normalization. It prepares the feature vector for the next pattern analysis steps and ensures subjects are comparable among them and classifications do not rely on water consumption per day or diuresis rate i.e. classification is not quantitative but qualitative. However, the area under the curve of large peaks contributes largely to that of the TIC, leading to over-classified samples. By not considering the area under peaks higher than 106, the enhanced *normalize_avoidPeak()* does not over-normalize.

STEP 6 — 3-fold sample averaging. To seek repeatability and disregard features related to GC-MS performance, each sample has been analyzed three times, pre-processed, and then averaged. Samples are 4600-dimensional vectors with an intensity value per instant of time (2 - 20 min).
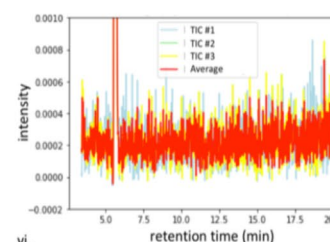
Figure 3. Summary of the pre-processing steps from GC–MS raw data reading up to total ion chromatogram (TIC) design. Sample 151 (belonging to a breast cancer subject) has been used to illustrate those steps.

results presented above were achieved with a sample size of 90 patients. The authors believe that if the study was continued and a bigger sample size was achieved (of a magnitude of 300–500 patients), better classification results might be obtained.

In addition to the latter, the authors also conclude that there is also room for improvement considering both hardware and sensorics of the presented technology: The current technology solely uses commercially-available sensors, which are sensitive to a wide range of VOCs (very sensitive but poorly specific). As a result, future steps to improve the detectability of our system (e.g., building new sensors specifically designed for VOCs that are breast cancer biomarkers) might result in a better classification rate. In fact, the sensor's response is not unequivocally correlated to the concentration of a single VOC but rather a *smellprint* consisting of a wide range of volatile BC biomarkers.
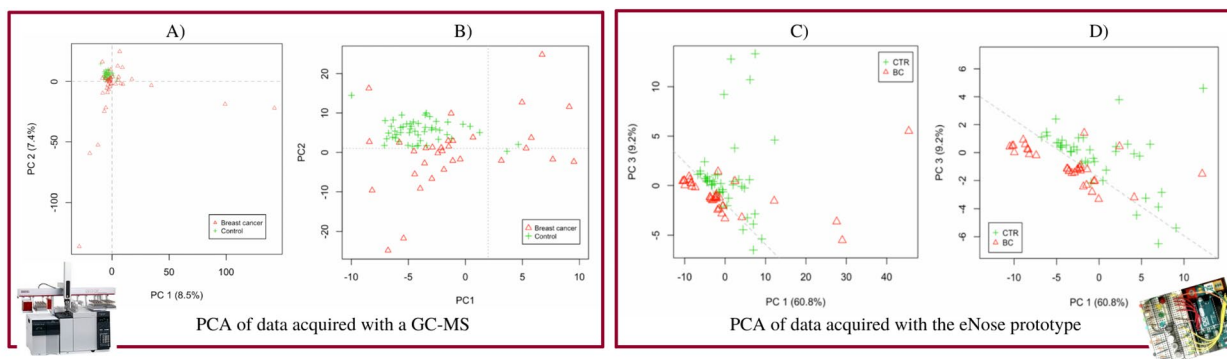
Figure 4. (**A**) PCA performed on data obtained by analyzing control (green) and BC (red) human urine with a GC–MS. (**B**) Figure on the right is a zoom of the figure on the left. (**C**) PCA performed on data obtained by analyzing control (green) and BC (red) human urine with the eNose prototype. (**D**) Figure on the right is a zoom of the figure on the left.
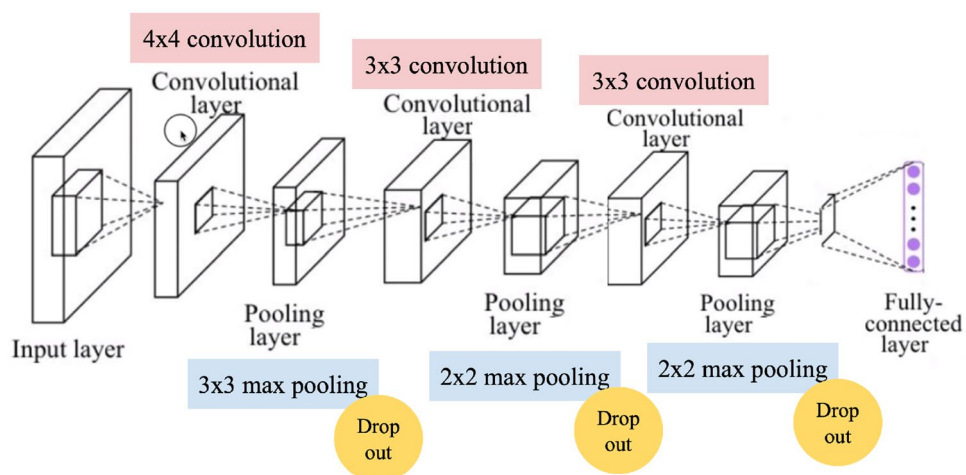

Figure 5. Convolutional neural network used for sample classification, consisting of convolutional filters of size 32, 64, 128 and linear activation. The output layer of the neural network presents two cells (two outputs) with a softmax activation function.

	(+) predicted	(−) predicted
(+) ground truth	6	0
(−) ground truth	1	6

Table 1. Confusion matrix of the CNN model using GC–MS data. Out of 90 samples, 65 (72.3%) samples were used for training, 12 (13.3%) for testing and 13 (14.4%) for validation.

	(+) predicted	(−) predicted
(+) ground truth	2	0
(−) ground truth	1	1

Table 2. Confusion matrix of the CNN model using data acquired by the eNose prototype. 36 (81.8%) samples were used for training, 4 (9.1%) for testing and 4 (9.1%) for validation.

	Technique	Accuracy (%)	Sensitivity (%)	Specificity (%)	Sample size
PCA	GC–MS data	77.11	75.05	68.33	N = 90
	eNose data	58.30	75.00	45.00	N = 44
cNN	GC–MS data	92.31	100.00	85.71	N = 90
	eNose data	75.00	100.00	50.00	N = 44
Mammography	–	> 91.00[8]	86.90[42]	74.00–98.00[8]	–

Table 3. Sample classification using biostatistics (PCA) and machine learning (cNN) vs reported performance of mammography.
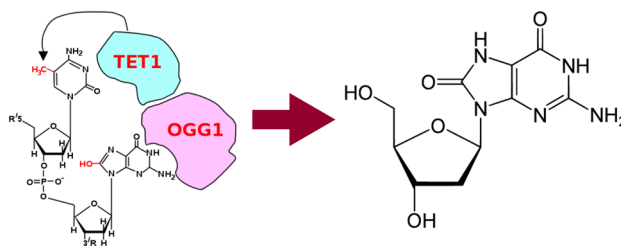

Figure 6. Formation of 8-oxodG out of guanine from a DNA sequence with intervention of TET1 and OGG1. 8-oxodG is a sensitive marker of DNA damage due to hydroxyl radical attack at the C8 of guanine. This damage is mutagenic and promotes cancer.

The study of urine smellprint and its correlation to breast cancer has been under study for a long time. However, a variety of electronic noses described in bibliography are subject to urine or exhaled alveolar breath and commonly apply to a wide variety of diseases, but none of them are designed specifically for BC. Urine metabolome is highly dependent on diet, environment and lifestyle, and consequently has more daily variability than serum or plasma[42]. It is for this reason that the metabolic fingerprint of BC in urine has been less studied than that of serum, plasma and breast tissue, but Vignoli et al. highlight the potential of this becoming a new tool for early-stage BC screening and report the findings from 4 different studies that coincide on the following: BC is associated with downregulation of acetate, alanine, creatinine, dimethylamine, glutamine, glycine, guanidoacetate, hippurate, isoleucine, lactate, leucine, succinate, taurine, threonine, trimethylamine n-oxide and valine[42]. The authors therefore believe that further research on urine characterization is yet another opportunity for improvement. Relevant knowledge related to urine characterization to date is presented as follows.

8-oxodG as a breast cancer volatile biomarker. When it comes to BC, publications on its specific biomarkers are currently rare yet consistent. Some articles point out that an increased presence in urine of the volatile 8-oxo-7,8-dihydro-2'-deoxyguanosine, shown in Fig. 6, (from now on, 8-oxodG) denotes cancer. Guo et al.[27] found a significant increase of 8-oxodG in patients with early-stage BC (p < 0.001) by ultraperformance liquid chromatography-electrospray ionization tandem mass spectrometry combined with a solid-phase microextraction (n = 184). According to another publication by Guo et al., reactive oxygen species (ROS) are produced by endogenous oxygen metabolism, as well as after exposure to ionizing radiation and chemical carcinogens[49]. The enzyme TET1 catalyzes a reaction with a guanine. For this reaction to happen, the guanine needs to have been oxidized to 8-hydroxy-2′-deoxyguanosine (8-OHdG or its tautomer 8-oxodG) due to the presence of a ROS. This oxidation allows enzyme OGG1 to bind to 8-oxodG and thus recruits TET1, which oxidizes the molecule, see Fig. 6 Increased ROS can cause oxidative base modifications and thus lesions in DNA. Since guanine exhibits the lowest oxidation potential, it is more vulnerable to free a radical, leading to the formation of 8-oxodG, commonly conceived as a biomarker of oxidative damage to DNA and a mutagen that contributes to carcinogenesis. Those features imply that a urine-based screening for 8-oxodG would be non-invasive, the patient would not get irradiated and it would not result in false positives, with the consequent economic savings[22,27,50].

Although being one of the most reliable BC biomarkers, 8-oxodG has a significant size (average molecular weight of 283.2407), and is therefore poorly volatile. For this reason, the authors suspect that when samples were heated up, its concentration in urine headspace was very low. This is an additional opportunity for improvement: designing a system that can heat the sample more, or designing a sensor that is more sensitive to this molecule.

Benzoic acid absence as a breast cancer volatile biomarker. Benzaldehyde is the only downregulated component by BC. It is absorbed via the gastrointestinal tract, skin and lungs, then distributed—especially in the blood and kidneys, and finally excreted very rapidly almost exclusively via urine. During the process, benzaldehyde gets oxidized to benzoic acid[51]. Hence, if Lavra et al.[29] reported a lack of benzaldehyde in BC cells mediums, a lack of benzoic acid should be expected in BC patients' urine. In fact, the PubChem database[52] con-

	Bening breast tissue (MCF10A)	Cancerous breast tissue (BT474)	p-value
2-nonanone	Very low	Very high	$6.5.10^{-21}$
4-methil-2-heptanone	Very low	Very high	$5.8.10^{-11}$
Isobutyric acid allyl ester	Very low	Very high	$3.6.10^{-10}$
1,3-dis-ter-butylbenzene	Very low	Very high	$1.8.10^{-9}$
Benzaldehyde	Very high	Very low	0.013

Table 4. Intensity of concentration of various BC biomarkers encountered in cell media: control, benign breast tissue (MCF10A), and cancerous breast tissue (BT474).

firms benzoic acid presence in control urine within the range 350–630 nmol/mmol creatinine. Therefore, some of the sensors inside the analysis chamber have been specifically selected to react to benzenes and benzoic acid.

Theoretical fundament for 2-nonanone presence in breast cancer urine samples. Lavra et al.[29] suggests 2-nonanone, 4-methil-2-heptanone, isobutyric acid allyl ester, 1,3-dis-ter-butylbenzene and benzaldehyde among others as cancer biomarkers encountered directly in BC cell mediums. As depicted in Table 4, such components exhibit a significant difference in concentration between control cell medium and primary tumor cell medium (BT474) with respective p-values of $6.5.10^{-21}$, $5.8.10^{-11}$, $3.6.10^{-10}$, $1.8.10^{-9}$ and 0.013. Table 1 shows a summary of the different VOCs found in human physiology according to many authors.

Acetone and 2-butanone as potential confounders in control urine. In the last years, the urge to deeply characterize the VOCs present in BC patients' urine has led to the need for a deeper understanding of control urine volatiles as well. Because VOCs present in urine are numerous, any analysis searching for a specific VOC will result in an extremely noisy and superposed signal with several VOC signals to be filtered out. It is long well-known that control urine consists on a mixture of water (91–96%), urea (9.3 g/dL), creatinine (0.670 g/L), sodium (1.17 g/L), potassium (0.750 g/L), chloride (1.87 g/L) and several VOCs[53]. Acetone and 2-butanone are the two predominant control VOCs in urine. Mochalski et al.[54] performed a selective reagent ionization time of flight mass spectrometry and gas chromatography and headspace solid-phase microextraction to determine VOCs in human urine (n = 19). A total of 16 VOCs exhibiting high incidence rates were quantified in urine. Amongst them, there were ten ketones, three volatile sulfur compounds and three heterocyclic compounds (furan, 2-methylfuran, 3-methylfuran). According to this study, acetone ($C_3H_6O \cdot NO^+$) has a parent ion m/z of 88.04, and when analyzed with a GC–MS, it appears at retention time Rt = 16.08 min, very early, at an intensity of 3.0–52,000 nmol/L, i.e. extremely variable among samples. 2-Butanone ($C_4H_8O.NO^+$) has a parent ion m/z of 102.06, and when analyzed with a GC–MS, it appears at retention time Rt = 22.22 min, at an intensity of 0.9–637 nmol/l, i.e. its peak is less intense than acetone's. Figure 7 summarizes the main findings of VOCs in urine according to most authors. Further investigation in this direction might allow the technology under study to capture a less noisy signal and therefore achieve a better classification.

Further directions. The results presented in this paper indicate that a bigger sample size would need to be analyzed and fed to a ML algorithm before one can assume that an algorithm can undoubtedly discriminate between breast cancer and control samples. Our observations indicate that the (a) urine contains enough information to allow discrimination between early-stage breast cancer and control class; and (b) if our technology was further improved, we might be able to capture more of the information contained in the sample. In conclusion, the next directions of this project include further training our AI model with a greater sample size, designing new sensors to acquire more information from the sample, and building a more robust system that performs with higher repeatability and robustness. The expected outcome upon completion of these future steps is a new version of the device that performs with higher accuracy and that can potentially be applied to breast cancer screening in the future.

Methods
Overall system design.
System design consisted of three steps: identifying the targeted VOC biomarkers; collect urine samples to gather data; and building the final prototype based on previous observations.

1st prototype: proof of concept. The proposed eNose is based on a metal oxide sensor array that scans the olfactory imprinting of the initial part of the urine stream. Metal oxide semiconductors are widely used when designing an eNose since they are easily available on the market, non-expensive, and notably well-performing. As depicted in Fig. 8, their surface conductivity changes upon adsorption and subsequent reaction of gases with the already-adsorbed oxygen. Since this reaction is an oxidation or reduction one—in the case of SnO or ZnO sensors-, the concentration of electrons available for conduction gets altered and causes this material to be an optimal semiconductor[45].

This proof of concept was built to prove whether the analysis of human urine could be performed in a portable and low-cost way without significantly decreasing the amount of information contained in the captured *smellprint*. It consisted of a breadboard with 4 VOC-specific sensors able to capture the *smellprint* of urine as well as a microprocessor to control the system. Data was analyzed on edge, and results were displayed through
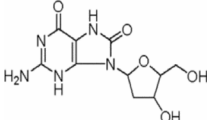
Breast cancer volatile biomarkers			
Urine	**Cell medium**		
Guo et al. [38]	Lavra et al. [29]	Silva et al. [41]	Huang et al. [39]
8-oxo-7,8-dihydro-2'-deoxyguanosine (8-oxodG)	**Benzaldehyde** (downregulated)	2-Pentanone	**1,4-dimethyl-benzene** (redox product)
	Prop-2-enyl 2-methyl propanoate	2-heptanone	Cyclohaxanol
IUPAC name:	**2-nonanone**	3-methyl-3-buten-1-ol	2-ethylhexan-1-ol
2-amino-9-[(2R,4S,5R)-4-hydroxy-5-(hydroxymethyl)oxolan-2-yl]-3,7-dihydropurine-6,8-dione	4-methil-2-heptanone	ethyl acetate	**2,4-dimethyl-benzaldehyde** (highly upregulated)
	1,3-dis-ter-butylbenzene	ethyl propanoate	
		2-methyl butanoate	

	Exhaled breath		
Chemical formula: $C_{10}H_{13}N_5O_5$	Brooks et al. [51]	Phillips et al. [25]	Li et al. [42]
	5-nonane	Alkanes derivatives	Hexanal
Structure:	6-ethyl-3-octyl ester 2-trifluoromethyl benzoic acid	**Benzene derivatives**	Heptanal
	Nonane	Pentane	Octanal
	propane, 2-methyl		**Nonanal**
	nonadecane, 3-methyl		

Figure 7. Summary of some VOCs encountered in urine, cell medium and expired breath, according to Guo et al.[27], Lavra et al.[29], Silva et al.[40], Huang *et al.*[38], Brooks et al.[50], Phillips et al.[25] and Li et al.[41]. Those VOCs whose evidence was found to be more consistent and their derivatives have been highlighted.
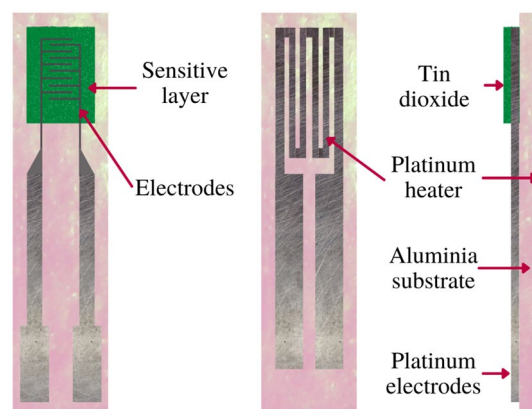

Figure 8. Structure of a metal oxide sensor. The sensitive layer on the top reacts with the presence of certain VOCs and thus sensors change in conductivity. The platinum heater placed on the bottom of the sensor consists of a dissipating resistance that outputs heat. By heating the sensor a higher VOC specificity is reached.

a color-coded LED-based user interface shown in Fig. 9A. The four gas sensors on the breadboard are the key components of the electronic system thanks to the fact that the SnO_2 layer (Fig. 8) absorbs the VOCs present in the sample and thus changes in conductivity[45]. However, this phenomenon is only possible at a given temperature, which is why 5 V need to be continuously supplied to the sensor's heater plate, which keeps the sensitive layer warm. The heater is an 83 Ω resistance working at 42 mA. Finally, metal oxide sensors are poorly selective in ambient temperature with a high presence of water vapor[55], and thus reach optimal performance at 68F, 65% HR. Arduino board also outputs voltage and ground to feed the electronic components of the system. In this case, the 5 V and GND outputs were used (Fig. 9A, orange and green wires). This is the prototype that was used to acquire the 44 urine *smellprints* mentioned in the "Results" section.

eNose final prototype. Based on the proof of concept mentioned above, this second prototype was aimed at better capturing the signal (i.e., to profile the smell of the sample more accurately) and allowing for a more complex algorithm: The new prototype's connectivity workflow makes it possible to host the algorithm on the cloud,
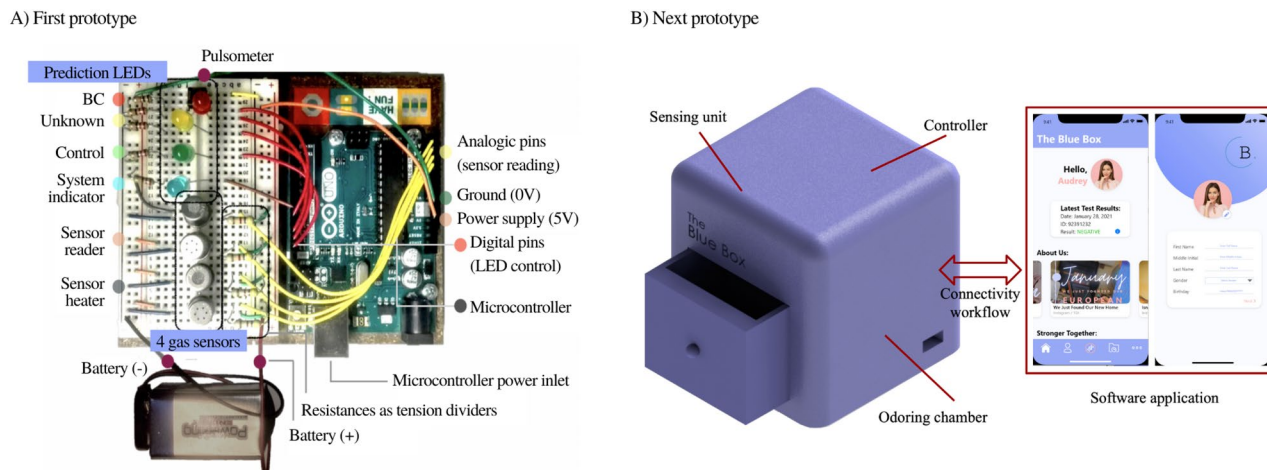
Figure 9. (**A**) First prototype: Main electronic components that integrate the first eNose prototype. The 4 gas sensors sniff the sample and send a VOC-dependent 0–5 V signal to the Arduinos analogic pins. Their combined signal is then reduced in dimensionality and undergoes sample classification. Depending on the prediction, output digital pins supply current to the corresponding LED, alerting of "possibly BC" "prediction no strong enough"(Blank) or "possibly control"(CTR). (**B**) Final prototype: Structure of the device and software application that we mentioned above. The 3D-printed structure contains all parts involved in the urine analysis. The software application is the interface to trigger the analysis and receive the results.
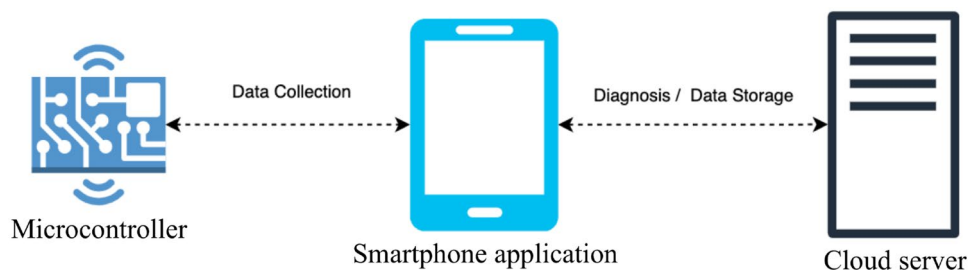

Figure 10. Connectivity workflow to gather data *on edge* and perform sample classification on the cloud.

allowing it to use more computational power, i.e., being more complex. This prototype, shown in Fig. 9B, was not tested in a clinical setting. It consists of a 3D-printed structure that contains the following:

- *Odoring chamber* Physical structure where to collect the smell of urine. The chamber consists of an irradiating wall that dissipates heat over time in a controlled fashion. By doing so, it is easier to assess which VOC is evaporated at every instant of time, depending on their boiling points.
- *Sensing unit* Array of sensors that acquire the *smellprint*.
- *Controller* Microprocessor that gathers all data from the sensors and executes the embedded code.
- *Connectivity workflow* Antenna and microchip that send the captured information via Bluetooth Low Energy to a software application in a smartphone, which will, in turn, send it via the internet to the cloud-based server.
- *Cool & clean system* Set of 2 fans and air conduits through the 3D-printed case creating an airflow through the device to keep the electronics from overheating and cleans the Sensing unit once the analysis has finished.
- *Software application* Mobile application installed on the user's smartphone to collect her demographic data, merge it with sensor data and forward it to the cloud.
- *Cloud-based server* Remote server hosting an AI-based classification algorithm that outputs the probability of a user having BC.

As detailed in Fig. 10, the connectivity workflow of the aforementioned technology mentioned above works as follows: When a urine sample is introduced inside the device, the sensors perform a change in voltage depending on the nature of the chemical compounds present in the urine sample. Once the microprocessor has acquired the sensors' signal, the BLE module sends the sample's *smellprint* to the user's phone, where the software application is installed, for 30 min. The app then sends this signal to the cloud via WiFi, where the AI-based classification AI algorithm is allocated.The algorithm classifies the sample with 95% accuracy. The signal is now not processed on edge (as in the previous prototype) but on the cloud to allow better computational power and thus better model accuracy.

Sample collection.   Urine samples were collected at the Southern Catalonia Institute of Oncology (Reus, Spain). Patient selection was carried out by a medical oncologist with experience in BC. Patients had locally advanced or metastatic BC (stages III and IV), thus presenting a high tumor burden and therefore potentially an increased concentration of excreted metabolites in urine. Control subjects were invited to participate in the study as well. These samples were obtained from a random population segment. Urine samples were stored at 4 °C (39.2°F) for no longer than 48 h before analysis so that 80% of the intensity of their VOCs was preserved. There is no evidence of freezing the sample harming its odor pattern[56,57].

Rovira i Virgili Institute of Medical Research's ethics committee approved the research. All research was performed following the provided good practice guidelines. Informed consent was obtained from all participants and/or their legal guardians. Only non-identifiable patient data was used.

Human urine is highly dependent on diet, lifestyles, ethnic background and other demographic features[42]. On the other hand, PCA works based on sample variability, taking the assumption that variability is merely based on class. To ensure this consideration does not affect the performance of the model, this assumption has been made: Human urine sample variability performs regardless of subject age. The control urine sample sub-space has a size of 51, and s0amples were collected from women aged 18–78 years (29.7 on average, with a standard deviation of 15.8). The breast cancer urine sample sub-space has a size of 39, and samples were collected from women aged 29–75 years (54.9 on average, with a standard deviation of 12.0).

Additionally, control subjects that reported being on a vegan or vegetarian diet were expected to follow a different fashion when it comes to sample prediction because a vegan diet can result in a decrease of inflammatory processes in the body and thus alters its physiology. A urine sample was taken from a 19-year-old control woman who had previously overcome thyroid cancer. In the PCA scoreplot, this sample lies in the vicinity of the BC-CTR threshold, which falls within the expected area because women who have previously overcome cancer and are currently healthy might present some cancer-like alterations in urine as well[25]. This sample has not been used for model training as specified in the patient inclusion requirements. Similar to that, one participant from the breast cancer patient group who was subject to a ketogenic diet also lies near the boundaries between cancer and control. This might be because a ketogenic diet causes pathogenic cells to become more sensitive to adjuvant cancer treatments[58].

Experimental validation.   The following samples were obtained from recruited subjects:

- 49 urine samples from *control subjects*.
- 37 urine samples from *BC patients*.

Urine analysis using GC–MS.   As shown in Fig. 2, urine samples have been firstly experimentally tested using a gas chromatography-mass spectroscopy. This procedure aimed at determining whether the GC–MS was sensible enough to discriminate BC samples from control samples—thus obtaining evidence that there existed enough difference between the two classes. Hence, the first step of sample classification was performed with GC–MS data.

The following protocol has been applied to the 90 human urine samples: We used a Focus DSQ II GC–MS and Triplus autosampler by Thermofisher Scientific (Fabrication number 12550090). Xcalibur and DSQ Tune II softwares were used to identify VOCs within the output signal and to visualize it. VOCs circulated along a 30 m × 0.25 mm 0.26 ţm HP-5MS column. Helium was used as the carrier gas (1.5 mL/min flow).

GC–MS vacuum setup and machine calibration have been performed according to the following parameters: Vacuum has been applied for 24 h before calibration. MS injector temperature was 220 °C; MS transfer line temperature was 200 °C; MS ion source temperature was 250 °C; MS column flow was 1.5 mL He/min; Acquisition rate was 20 Hz; GC oven temperature was 200 °C; GC Internal ambient temperature was 23 °C; GC RF generator temperature was 28 °C; GC slope was set to stay at 70 °C for 0.2 min, heat up at + 10 °C/min up to 270 °C, and stay at 270 °C for 10 min. Once calibrated, sample analysis has been performed according to the protocol that follows.

1. Aliquot a sample in three 11-mL vials for triple analysis.
2. Heat the sample in the oven at 150 °C for 20 min.
3. Inject the micro-syringe through the vial silicon septum and extract 6 ţL headspace gas.
4. Quickly introduce the 6 ţL of the sample headspace into the GC–MS inlet.
5. After 30.20 min the analysis will have finished.
6. Analyse a blank vial (filled with distilled water alone) for every 4–5 samples.

Data availability
The data that support the findings of this study are available from the corresponding author, J.G.B., upon reasonable request.

References

1. US Preventive Services Task Force. Screening for breast cancer: US Preventive Services Task Force recommendation statement. *Ann. Intern. Med.* **151**(10), 716–236 (2009).
2. Nelson, H. D. *et al.* Screening for breast cancer: an update for the US Preventive Services Task Force. *Ann. Intern. Med.* **151**(10), 727–737 (2009).
3. Center for Disease Control (CDC). *Health, United States, 2019, Table 33*.
4. Feldstein, A. C. *et al.* Patient barriers to mammography identified during a reminder program. *J. Womens Health* **20**(3), 421–428 (2011).
5. Jemal, Ahmedin, Ward, Elizabeth & Thun, Michael J. Recent trends in breast cancer incidence rates by age and tumor characteristics among US women. *Breast Cancer Res.* **9**(3), 1–6 (2007).
6. Broeders, M. *et al.* The impact of mammographic screening on breast cancer mortality in Europe: A review of observational studies. *J. Med. Screen.* **19**(1 suppl), 14–25 (2012).
7. Blumen, H., Fitch, K. & Polkus, V. Comparison of treatment costs for breast cancer, by tumor stage and type of service. *Am. Health Drug Benefits* **9**(1), 23 (2016).
8. Barlow, W. E. *et al.* Accuracy of screening mammography interpretation by characteristics of radiologists. *J. Natl Cancer Inst.* **96**(24), 1840–1850 (2004).
9. Desreux, J. A. Breast cancer screening in young women. *Eur. J. Obstetr. Gynecol. Reprod. Biol.* **230**, 208–211 (2018).
10. Siu, A. L. Screening for breast cancer: US Preventive Services Task Force recommendation statement. *Ann. Intern. Med.* **164**(4), 279–296 (2016).
11. Sechopoulos, I., Suryanarayanan, S., Vedantham, S., D'Orsi, C. J. & Karellas, A. Radiation dose to organs and tissues from mammography: Monte Carlo and phantom study. *Radiology* **246**(2), 434–443 (2008).
12. Gøtzsche, P. C. The debate on breast cancer screening with mammography is important. *J. Am. Coll. Radiol.* **1**(1), 8–14 (2004).
13. World Health Organization. *WHO Position Paper on Mammography Screening* (WHO, 2014).
14. World Health Organization. *Disease Burden and Mortality Estimates* (WHO, 2018).
15. Apantaku, L. M. Breast cancer diagnosis and screening. *Am. Fam. Phys.* **62**(3), 596–602 (2000).
16. Podgornova, Y. A. & Sadykov, S. S. Detection of malignant breast tumors on the background of fibrocystic breast disease. In *CEUR Workshop Proceedings*, Vol. 2210, 177 (2018).
17. Malherbe, K. & Fatima, S. *Fibrocystic Breast Disease* (2019).
18. Warner, E. *et al.* Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination. *JAMA* **292**(11), 1317–1325 (2004).
19. Mango, V. L., Goel, A., Mema, E., Kwak, E. & Ha, R. Breast MRI screening for averagerisk women: A Monte Carlo simulation costbenefit analysis. *J. Magn. Reson. Imaging* **49**(7), e216–e221 (2019).
20. Persaud, Krishna & Dodd, George. Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature* **299**(5881), 352–355 (1982).
21. Asimakopoulos, A. D. *et al.* Prostate cancer diagnosis through electronic nose in the urine headspace setting: A pilot study. *Prostate Cancer Prostatic Dis.* **17**(2), 206–211 (2014).
22. Guerrero-Flores, H. *et al.* A non-invasive tool for detecting cervical cancer odor by trainedscent dogs. *BMC Cancer* **17**(1), 79 (2017).
23. Buszewski, B. *et al.* Identification of volatile lung cancer markers by gas chromatography-mass spectrometry: Comparison with discrimination by canines. *Anal. Bioanal. Chem.* **404**(1), 141–146 (2012).
24. Blatt, R., Bonarini, A. & Matteuci, M. *Pattern Classification Techniques for Lung Cancer Diagnosis by an Electronic Nose* 397–423 (Springer, 2010).
25. Phillips, M. *et al.* Rapid point-of-care breath test for biomarkers of breast cancer and abnormal mammograms. *PLoS ONE* **9**(3), e90226 (2014).
26. Burton, C. & Ma, Y. Current trends in cancer biomarker discovery using urinary metabolomics: Achievements and new challenges. *Curr. Med. Chem.* **24**, 5–28 (2017).
27. Guo, C. *et al.* Discriminating patients with early-stage breast cancer from benign lesions by detection of oxidative DNA damage biomarker in urine. *Oncotarget* **8**(32), 53100 (2017).
28. Horvath, G., Järverud, G., Järverud, S. & Horváth, I. Human ovarian carcinomas detected by specific odor. *Integr. Cancer Ther.* **7**(2), 76–80 (2008).
29. Lavra, L. *et al.* Investigation of VOCs associated with different characteristics of breast cancer cells. *Sci. Rep.* **5**(1), 13246 (2015).
30. van Keulen, K. E., Jansen, M. E., Schrauwen, R. W., Kolkman, J. J. & Siersema, P. D. Volatile organic compounds in breath can serve as a noninvasive diagnostic biomarker for the detection of advanced adenomas and colorectal cancer. *Aliment. Pharmacol. Therap.* **51**(3), 334–346 (2020).
31. Antoce, A. O. & Namolosanu, I. O. A. N. Rapid and precise discrimination of wines by means of an electronic nose based on gaschromatography. *Rev. Chim.* **62**(6), 593–595 (2011).
32. Kishimoto, N. & Kashiwagi, A. Evaluation of filtration on volatile compounds in virgin olive oils using an electronic nose. In *2019 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*, 1–3 (IEEE, 2019).
33. Dent, A., Sutedja, T. & Zimmerman, P. Exhaled breath analysis for lung cancer. *J. Thorac. Dis.* **63**(2), 164–168 (2013).
34. Peng, G. *et al.* Detection of lung, breast, colorectal and prostate cancers from exhaled breath using a single array of nanosensors. *Br. J. Cancer* **103**(4), 542–551 (2010).
35. Peled, N. *et al.* Non-invasive breath analysis of pulmonary nodules. *J. Thorac. Oncol.* **7**(10), 1528–1533 (2012).
36. Di Natale, C. *et al.* Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors. *Biosens. Bioelectron.* **18**(10), 1209–1218 (2003).
37. D'Amico, A. *et al.* An investigation on electronic nose diagnosis of lung cancer. *Lung Cancer* **68**(2), 170–176 (2010).
38. Huang, Y., Li, Y., Luo, Z. & Duan, Y. Investigation of biomarkers for discriminating breast cancer cell lines from normal mammary cell lines based on VOCs analysis and metabolomics. *R. Soc. Chem. Adv.* **6**(48), 41816–41824 (2016).
39. Wang, C. *et al.* Volatile organic metabolites identify patients with breast cancer, cyclomastopathy, and mammary gland fibroma. *Sci. Rep.* **4**(1), 5383 (2014).
40. Silva, C., Perestrelo, R., Silva, P., Tomás, H. & Câmara, J. Volatile metabolomic signature of human breast cancer cell lines. *Sci. Rep.* **7**, 43969 (2017).
41. Li, J, *et al.* Investigation of potential breath biomarkers for the early diagnosis of breast cancer using gas chromatography-mass spectrometry. *Clin. Chim. Acta* **436**, 59–67 (2014).
42. Vignoli, A. *et al.* Precision oncology via NMR-based metabolomics: A review on breast cancer. *Int. J. Mol. Sci.* **22**(9), 4687 (2021).
43. Weber, C. *et al.* Evaluation of a gas sensor array and pattern recognition for the identification of bladder cancer from urine headspace. *Analyst* **136**(2), 359–364 (2011).
44. Roine, A. *et al.* Detection of smell print differences between nonmalignant and malignant prostate cells with an electronic nose. *Future Oncol.* **8**(9), 1157–1165 (2012).
45. Watson, J. The tin oxide gas sensor and its applications. *Sens. Actuators* **5**(1), 29–42 (1984).

46. Mazzone, P. *et al.* Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array. *Thorax* **62**(7), 565–568 (2007).
47. Machado, R. *et al.* Detection of lung cancer by sensor array analyses of exhaled breath. *Am. J. Respir. Crit. Care Med.* **171**(11), 1286–1291 (2005).
48. Dragonieri, S. *et al.* An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD. *Lung Cancer* **64**(2), 166–170 (2009).
49. O'Donovan, P. *et al.* Azathioprine and UVA light generate mutagenic oxidative DNA damage. *Science* **309**, 1871–1874 (2005).
50. Brooks, S., Moore, D., Marzouk, E., Glenn, F. & Hallock, R. Canine olfaction and electronic nose detection of volatile organic compounds in the detection of cancer: A review. *Cancer Investig.* **33**(9), 411–419 (2015).
51. *List of MAK and BAT Values 2017: Permanent Senate Commission for the Investigation of Health Hazards of Chemical Compounds in the Work Area. Report 53*, Vol. 17, 1st ed, 14–36 (WILEYVCH Verlag GmbH and Co. KGaA, 2017).
52. "The PubChem Project". Pubchem.ncbi.nlm.nih.gov. https://pubchem.ncbi.nlm.nih.gov (2018) (Accessed 10 March 2018).
53. "Human Metabolome Database", HMDB. http://www.hmdb.ca (2018) (Accessed 30 May 2018).
54. Mochalski, P. & Unterkofler, K. Quantification of selected volatile organic compounds in human urine by gas chromatography selective reagent ionization time of flight mass spectrometry (GC-SRI-TOF-MS) coupled with head-space solid-phase microextraction (HS-SPME). *Analyst* **141**(15), 4796–4803 (2016).
55. Brsan, N. & Weimar, U. Understanding the fundamental principles of metal oxide based gas sensors; the example of CO sensing with SnO2 sensors in the presence of humidity. *J. Phys. Condens. Matter* **15**(20), 813–839 (2003).
56. Henriksen, T., Hillestrom, P., Poulsen, H. & Weimann, A. Automated method for the direct analysis of 8-oxo-guanosine and 8-oxo-2'-deoxyguanosine in human urine using ultraperformance liquid chromatography and tandem mass spectrometry. *Free Radic. Biol. Med.* **47**(5), 629–635 (2009).
57. Smith, S. *et al.* A comparative study of the analysis of human urine headspace using gas chromatography–mass spectrometry. *J. Breath Res.* **2**, 037022 (2008).
58. Allen, B. *et al.* Ketogenic diets as an adjuvant cancer therapy: History and potential mechanism. *Redox Biol.* **2**, 963–970 (2014).

Author contributions

J.G.P. selected the patients and designed the sample collection protocol; J.G.B. and J.G.P. collected the 90 urine samples; J.G.B. and A.P.M. conceived the first experiment (biostatistics classification); J.G.B., M.S. and F.K. conceived the second experiment (ML classification); J.G.B. conducted the experiments; J.G.B., A.P.M. and J.G.P. analyzed the results of the first experiment; J.G.B., M.S., F.K. and M.K. analyzed the results of the second experiment. All authors reviewed the manuscript.

Competing interests

J.G.B. owns stock in The Blue Box Biomedical Solutions, a company owning intellectual property related to this publication. All other authors declare no potential conflict of interest and no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.G.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.