CHEMICAL ENGINEERING TRANSACTIONS
VOL. 112, 2024
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Selena Sironi, Laura Capelli

ISBN 979-12-81206-13-7; **ISSN** 2283-9216
DOI: 10.3303/CET24112021

Machine Learning-Based Electronic Nose for Universal Mapping of Blood Odors and Diagnosis of Cancer

Ehab I. Mohamed[a,*], Israa A. Saleem[b], Amira A. Darwish[c], Mohamed S. Mshaly[d], Moustafa S. Mahmoud[d], Mohamed S. Turkey[e], Marwa A. Mohamed[f], Samy H. Darwish[g], Mohamed A. Abdou[h,i], Mohamed M. El Safwany[j]

[a]Medical Biophysics Department, Medical Research Institute, Alexandria University, Alexandria, Egypt
[b]Optometry Department, Technical Medical Institute, Erbil Polytechnic University, Erbil, Iraq
[c]Medical Laboratory Technology Department, Faculty of Applied Health Sciences, Pharos University, Alexandria, Egypt
[d]Physics Department, Faculty of Science, Alexandria University, Alexandria, Egypt
[e]Microbiology and Immunology Department, Faculty of Pharmacy, October 6 University, Sixth of October City, Giza, Egypt
[f]Chemical Pathology Department, Medical Research Institute, Alexandria University, Alexandria, Egypt
[g]Higher Institute of Engineering and Technology – King Mariout, Alexandria, Egypt
[h]Faculty of Computer Science and Engineering, Alalamein International University (AIU), Alexandria, Egypt
[i]Informatics Research Institute, City for Scientific Research and Technology Applications, Alexandria, Egypt
[j]Radiological Sciences and Medical Imaging Department, Faculty of Applied Health Sciences Technology, Pharos University, Alexandria, Egypt
eimohamed@yahoo.com; ehab.abdo@alexu.edu.eg

Electronic nose (E-Nose) technology is gaining prominence as a tool for cancer diagnostics due to its ability to detect volatile organic compounds (VOCs) in bodily fluids. There has been no comprehensive study of the methodologic hurdles and overall diagnostic accuracy of E-Noses for cancer detection in blood samples. The aim of this investigation was to standardize a procedure for using a machine learning-based E-Nose to accurately diagnose five different cancers. This prospective (diagnostic/prognostic) study included 1001 newly diagnosed adult male and female participants with blood, brain, breast, liver, and lung cancers, as well as healthy controls. Blood samples were collected from all participants for complete blood counts, specific tumour markers testing, and E-Nose measurements. Sensor response patterns at the plateau region were used for training, testing, and machine learning cross-validation. Out of the final 550 participants, the E-Nose accurately categorized 100 to CLL, 50 to GBM, 150 to IDC, 50 to HCC, 50 to AC, and 150 to HC in agreement with specific tumour markers data; there were no false-positives or false-negatives. With an average area under the curve (AUC) of 1.0, the support vector machine had 100% accuracy, sensitivity, and specificity. Thus, the E-Nose had high diagnostic accuracy, sensitivity, and specificity in cancer detection in blood samples.

1. Introduction

Cancer is characterized by the proliferation of aberrant cells that develop uncontrollably, infiltrate neighbouring tissues, and metastasize to distant organs, leading to death (Scheepers et al., 2022). In 2020, GLOBOCAN reported a total of 19,292,789 newly diagnosed cancer cases and 9,958,133 associated deaths, with an age-standardized rate (ASR) of 190/100,000 for all cancers worldwide (Sung et al., 2021). In 2014, the ASR in Egypt was 166.6/100,000, with liver, breast, and bladder cancers being the most common types (WHO, 2024; WCRFI, 2024). Cancer is predicted to double to 29-37 million new cases by 2040, with a significant incidence in underdeveloped and developing countries (Sung et al., 2021). Of the 15 million deaths at age 30-69 years in 2018, 4.5 million died of cancer, and 70% of these deaths also occurred in these countries (Ibrahim et al., 2014). In 80% of countries, premature cancer mortality trends hinder progress towards achieving the sustainable development goals (SDG) target of 3.4 million by 2030 (Shah et al., 2019). To determine the type and extent of cancer, a comprehensive physical examination is performed together with laboratory testing and imaging procedures such as endoscopy, mammography, US, CT, MRI, and PET scans (Merkow et al., 2017).

Paper Received: 5 March 2024; Revised: 22 April 2024; Accepted: 17 June 2024
Please cite this article as: Mohamed E.I., Saleem I.A., Darwish A.A., Mshaly M.S., Mahmoud M.S., Turkey M.S., Mohamed M.A., Darwish S.H., Abdou M.A., El Safwany M.M., 2024, Machine Learning-based Electronic Nose for Universal Mapping of Blood Odors and Diagnosis of Cancer, Chemical Engineering Transactions, 112, 121-126 DOI:10.3303/CET24112021

The analysis of volatolomic biomarkers in body fluids using the Electronic Nose (E-Nose) technology has revolutionized the diagnosis of blood, breast, head and neck, lung, ovarian, gynaecologic, colorectal, and other malignancies (Mohamed et al., 2014; 2017; 2019; Farraia et al., 2019; Scheepers et al., 2022), as well as tumour cell lines in vitro (Gendron et al., 2007). This method is founded on the premise that pathophysiological processes alter the body's metabolism, which is directly expressed as a unique change in the compendium of low molecular weight volatile organic compounds (VOCs) (Shirasu & Touhara, 2011). By combining E-Nose sensor responses and binary outcomes with machine learning (ML) algorithms, high levels of accuracy in disease classification and diagnosis could be achieved (Wojnowski & Kalinowska, 2022; Kokabi et al., 2023). Standardized validation studies to assess the E-Nose's accuracy in cancer detection were, however, emphasized by a growing number of oncologists (Malone et al., 2022; Scheepers et al., 2022).
The aim of the current investigation was to develop a global map of blood odours that can assist the oncologist in cancer screening and early diagnosis by using a ML-based E-Nose.

2. Methods

2.1 Study Design

A total cohort of 1001 male and female participants from the Medical Research Institute Hospital and Alexandria University Main Hospital were recruited in this prospective diagnostic/prognostic study. Newly diagnosed cancer patients were enrolled consecutively before any medical intervention from January 2019 to June 2023. The initial diagnosis began with taking a comprehensive history and thorough medical examinations. Blood samples were collected from the upper limbs of all participants using duplicate sterile vacutainer tubes on EDTA for routine lab investigations and E-Nose measurements. Cancers of the blood (n = 188), brain (n = 72), breast (n = 213), liver (n = 93), and lung (n = 158) were the study's main categories, while healthy controls (n = 277) were those with no evident signs of illness. The exclusion criteria were youngsters, steroid treatment, endocrinopathy, and the presence or suspected presence of infections or autoimmune diseases. The last step in diagnosis involved testing for specific tumour markers, bronchoscopy and thoracoscopy investigations, and imaging radiograms (e.g., X-rays, US, CT, MRI, and PET), yielding a homogenous cohort of only 550 participants. All participants had signed written informed consents and gave blood samples. The Ethics Committee of the Medical Research Institute, Alexandria University, reviewed and approved the study protocol.

2.2 Electronic Nose Measurements

A single researcher blindly processed the blood tube samples without knowing the subjects' clinical status or reference standard tumour marker values using a commercially available E-Nose (PEN3, Airsense Analytics GmbH, Schwerin, Germany), comprising a set of 10 chemically nonspecific metal oxide semiconductors sensors. The PEN3 device complies with the perspectives of the EC, according to Harmonised European Standards (EN292 Part1, Part2 / EN294 / EN61010-1 / EN1050 / EN60204-1 / EN55011 Group 1, Class B / EN50270 / EN61326), and was approved by the EC-Declaration of Conformity ID# 210601005E. The E-Nose was connected to each sealed vacutainer tube through a 3 mm Teflon tube ending with a size-20G long lure-lock needle. To allow ambient air into the tube, a second, shorter needle was inserted through the seal. Filtered, dry room air was used for delivering the VOCs in the headspace above blood samples to the E-Nose sensor array at a 400 ml/min flow rate. Each time a sample was connected, solenoid valves alternated between room air and headspace VOCs, thereby automatically recording the single changes in sensor electric resistance (R) and relative conductance (G/G_o) continuously for a duration of 60 s in an independent file. Prior to the subsequent sample measurement, sensors were flushed with filtered dry room air for 60 s and then zeroed out for 10 s to restore signals to their initial levels ($G/G_o = 1$). All measurements were repeated twice, and ML algorithms were used to extract and analyse 10-sensor stable response patterns in the plateau region at 50 s (Figure 1, right panels).

2.3 Machine Learning (ML)

ML employs unsupervised and supervised automatic algorithms to learn from data, improve performance, and make decisions and predictions (Wojnowski & Kalinowska, 2022; Kokabi et al., 2023). It detects patterns in datasets to create predictive models, where the accuracy of predicted outputs is directly proportional to the amount of data used for their training (Salama et al., 2021; Meshref et al., 2023). Principal component analysis (PCA) clusters and associates uncategorized datasets based on similarities, patterns, and differences, whereas support vector machine (SVM) uses categorized data for accurate medical diagnosis and prognosis (BasuMallick, 2024).
PCA is a non-parametric ML algorithm used for exploratory medical data analysis and predictive modelling for clustering and classification (Elhaik, 2022). It reduces data dimensionality by transforming potentially correlated

E-Nose sensor response time series into principal components (PCs), maximizing variance and data variability, and identifying variable correlations (Mohamed et al., 2014; 2017; 2019). The built-in PCA algorithm (WinMuster, Version 1.6.2.2, Airsense Analytics GmbH, Schwerin, Germany) was applied to transform the E-Nose time series from 10 dimensions to orthogonal *x*-*y* coordinates. It is the least-squares transform for the E-Nose time series, with PC #1 having the highest *x*-coordinate variance and PC #2 the second most *y*-coordinate variance. PCA displays similarity patterns in clusters, analysing and extracting essential details from the quantitatively dependent 10-dimensional times series variables. It helps identify and classify all investigated group blood samples, including blood, brain, breast, liver, and lung cancer groups, compared to healthy control participants. Vector-based SVM is a supervised ML technique for outlier detection, regression, and data classification (Akinnuwesi et al., 2023). It creates a hyperplane, the best decision boundary in the E-Nose 10-dimensional space, based on support vectors, to determine the optimal training parameter combination for model predictions (Kokabi et al., 2023). The LibLinear-built SVM algorithm uses a one-versus-all multiclass technique, providing maximal penalty and loss function flexibility, and effectively performs with limited and large data (Sarker, 2021). SVM was implemented using Spider 4.14, provided in Anaconda Navigator 1.9.12 and based on Python platform version 3.8.3, running on a core-i7 PC under Windows 11.

2.4 Input Dataset

The input dataset was comprised of NOS-formatted standard measurement files of E-Nose 10-sensor responses (60 s time series) from blood samples of participants with a confirmed diagnosis of chronic lymphocytic leukaemia (CLL), glioblastoma multiforme (GBM), invasive ductal carcinoma (IDC), hepatocellular carcinoma (HCC), adenocarcinoma (AC), and healthy controls (HC). The cancer input vector for blood samples was a 550 $\times$ 1100 matrix, describing 1100 attributes of 550 blood samples. The training phase employed 60% of the dataset, 20% for cross-validation, and the remaining 20% for testing.

2.5 Output Vectors

There were six study groups represented by the output vector: blood (CLL), brain (GBM), breast (IDC), liver (HCC), lung (AC) cancer, and HC. The input and output vectors were randomly divided into two sets, and the output of the linear SVM was calculated for varying values of the constant C. The best results were obtained at a C value of 200 and a number of folds equal to 5 for cross-validation.

2.6 Statistical Analysis

For each study group, the SVM model's performance metrics, including precision, accuracy, sensitivity, specificity, F1-score, and mean absolute error (MAE) were evaluated for the training, cross-validation, and testing output results and score curves. Graphical representations of error rates, the area under the curve (AUC) of the receiver operating characteristic (ROC) analysis, and the classification confusion matrix for each study groups were also provided.

3. Results and Discussion

Based on specific tumour markers for cancer diagnosis, the study was carried out on 550 participants with an age range of 49 to 70 years, finally categorized into CLL (n = 100), GBM (n = 50), IDC (n = 150), HCC (n = 50), AC (n = 50), and HC (n = 150). The mean age of cancer patients was 55.53 ± 10.71 years, 171 (42.75%) being males, which included 116 (67.83%) active smokers, while 229 (57.25%) were females, with only 16 (6.99%) active smokers. The HC mean age was 58.22 ± 7.53 years; 78 (52%) were males, of which 51 (34%) were active mild smokers or with a history of smoking. The average age of study participants falls within the middle-aged range, with a significant gender disparity (females were more than 50%) due to the higher number of IDC patients. IDC patients, who were all class 2 obese females (BMI > 35 kg/m^2), were significantly heavier than CLL and HCC patients, who were overweight (BMI > 25 kg/m^2), while GBM and AC patients were within the normal weight range (BMI < 25 kg/m^2).

Figure 1 right panels show the typical sensor responses, which were characteristic fingerprints to the VOCs emanated from blood samples of various cancers (CLL, GBM, IDC, HCC, and AC) and HC. The multidimensional PCA cluster plots are shown in Figure 1 left panel, where the correlation matrix showed that the first and second main PCs had a variance of 80.59% and 16.51%, respectively. The PCA total variance was 97.10%, with no false-positive or false-negative results, ensuring 100% sensitivity and 100% specificity. E-Noses were reported to have a 90% sensitivity and 87% specificity in diagnosing cancer using exhaled breath, based on a pooled analysis of 52 studies involving 3677 cancer patients (Scheepers et al., 2022). Earlier studies also confirmed the PCA algorithm's high sensitivity and specificity in accurately diagnosing major leukaemia types (i.e., ALL, AML, CLL, and CML), breast cancer, and lung cancer in various biological fluids, including blood (Mohamed et al., 2014; 2017; 2019; Liu et al., 2023).

Results from further assessments using the SVM model showed the accuracy during training, cross-validation, and testing were 100, 97.60, and 100%, respectively, with a total accuracy of 99.20% (Table 1). The SVM model successfully produced accurate predictions of true values that coincided with the actual cancer and HC categories, as evidenced by the error rate measurements showing a zero MAE (Table 1). With a scoring time of only 0.02 s, the SVM model was ideal for real-time applications of cancer diagnosis (Meshref et al., 2023). Moreover, the output vectors for all categories of blood samples were perfectly accurate with respect to sensitivity, specificity, recall, and F1-Score. The best cross-validation was achieved, as shown graphically in Figure 2A, which displays the accuracy score throughout the entire blood sample training set. Figure 2B shows that the error rate of the model converged to zero by the end of training, thus the SVM model successfully learned E-Nose patterns, minimizing its errors over time. The SVM model had an AUC of 1.00, which is the best possible performance on an ROC curve, meaning a sensitivity and specificity of 100%, by 5-fold cross-validation test (Figure 2C). As a result, the SVM model correctly classified all the blood samples from different cancers and HC, with an accuracy rate of 100% for true positives and 0% for false negatives in the confusion matrix of Figure 2D.

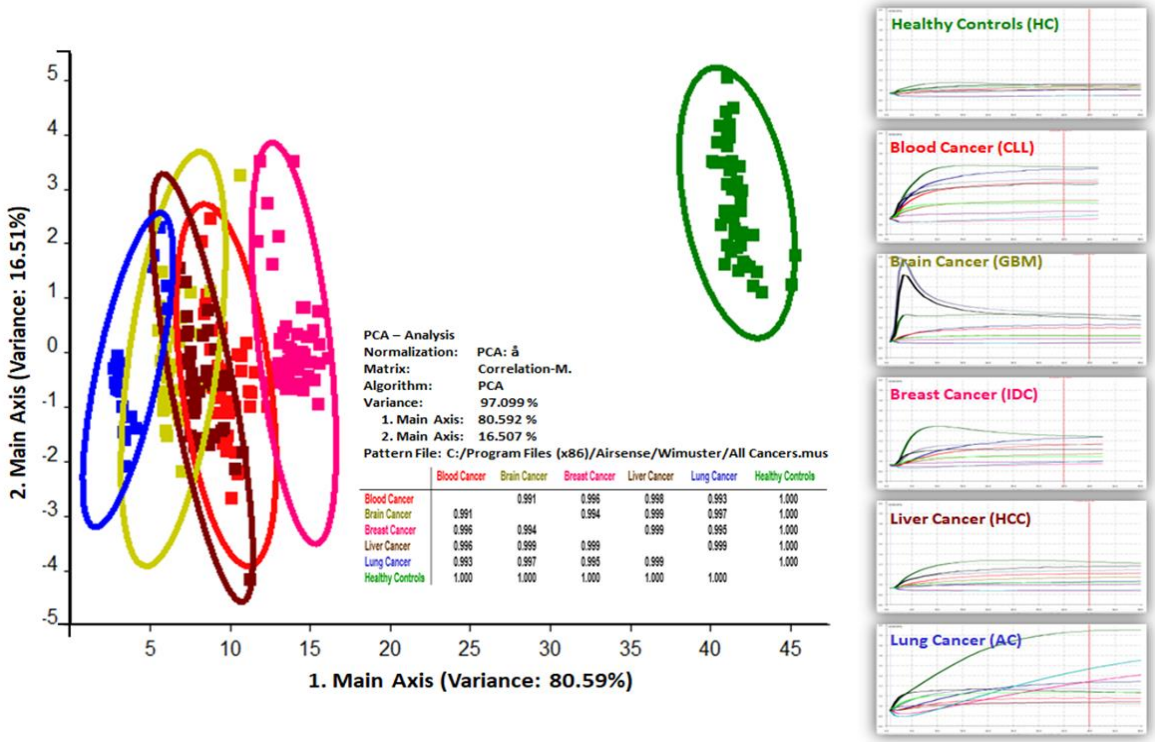

Figure 1: Typical sensor responses (right panels) and principal component analysis (PCA) (left clusters) for blood samples.

Table 1: Support vector machine (SVM) training, cross-validation, and testing output results.

	Output Result
Training Accuracy (%)	100
Cross-Validation Accuracy (%)	97.60
Testing Accuracy (%)	100
Total Accuracy (%)	99.20
Fit Time (s)	10.19
Score Time (s)	0.02
Mean Absolute Error (MAE)	0.00
Precision (%)	100
Sensitivity (%)	100
Specificity (%)	100
Recall (%)	100
F1-Score (%)	100

These results agree with our recent findings employing deep learning-based computer-aided systems to detect COVID-19 patients in chest X-ray images (Salama et al., 2021) and abdominal lymphadenopathy patients in CT scans (Meshref et al., 2023). Thus, this study demonstrates the potential of ML-based E-Noses for accurate blood-based cancer diagnosis, suggesting it as a reliable alternative or complimentary tool for early cancer detection and personalized medicine (Ferraia et al., 2019; Scheepers et al., 2022).

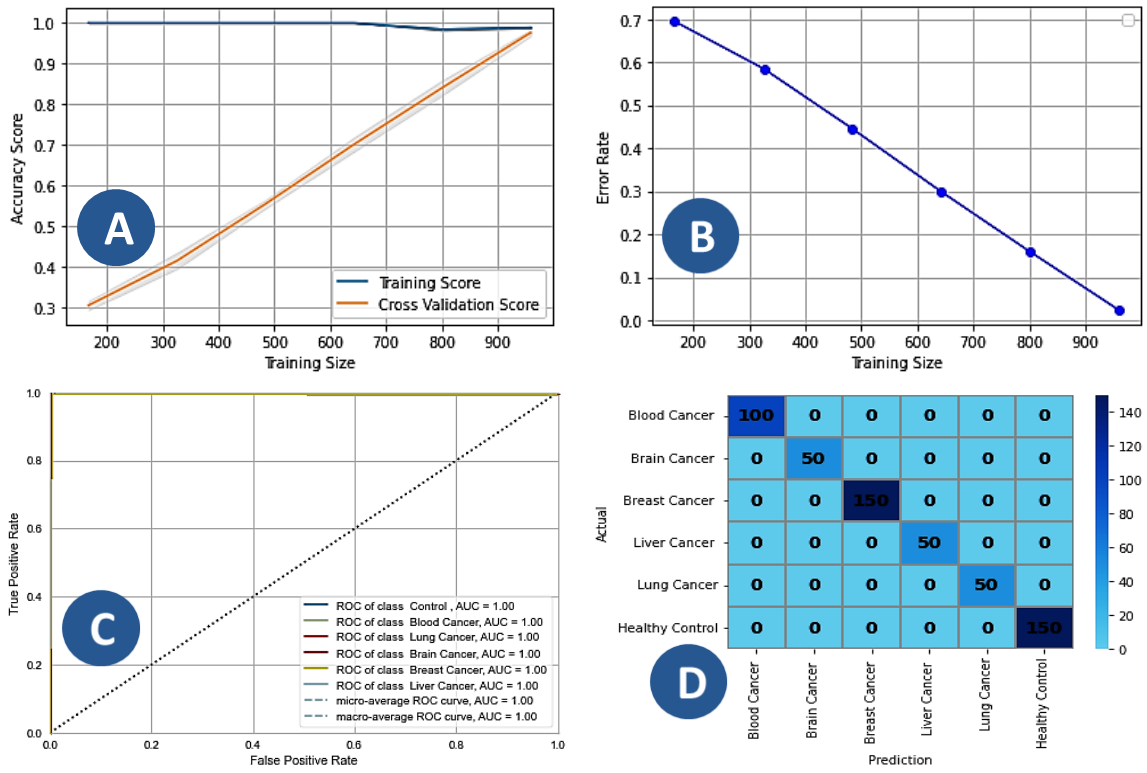

Figure 2: The SVM learning performance of the training and cross-validation score curves (A), error rate curve (B), receiver operating characteristic curve (ROC) (C), and the classification confusion matrix (D).

4. Conclusions

To the best of our knowledge, this study marks the first efficient implementation of the E-Nose to build a stereochemical global odour map of different cancers through the identification of specific VOCs in human underivatized blood samples. The E-Nose has achieved a sensitivity and specificity of 100% when using an SVM model to classify blood samples from patients with tumour markers-proven CLL, GBM, IDC, HCC, AC, and HC. Thus, we believe the global map of E-Nose cancer odours can assist the oncologist in cancer diagnostic decisions and prognosis.

Nomenclature

AC – Adenocarcinoma
ASR – Age-standardized rate
AUC – Area under the curve
BMI – Body mass index
CLL – Chronic lymphocytic leukemia
E-Nose – Electronic nose
GBM – Glioblastoma multiforme
HC – Healthy controls
HCC – Hepatocellular carcinoma
IDC – Invasive ductal carcinoma
MAE – Mean absolute error
ML – Machine learning
PCA – Principal component analysis
ROC – Receiver operating characteristic
SVM – Support vector machine
VOCs – Volatile organic compounds

Acknowledgments

This work was supported by a generous donation from Wolf Münchmeyer, Airsense Analytics GmbH, Schwerin, Germany. The authors would like to thank Prof. Mamdouh M. Shawki for valuable discussions and Dr. Hend N. Abdel-Reheim for assistance with liver cancer patients' clinical investigation.

References

Akinnuwesi B.A., Olayanju K.A., Aribisala B.S., Fashoto S.G., Mbunge E., Okpeku M., Owate P., 2023, Application of support vector machine algorithm for early differential diagnosis of prostate cancer. Data Sci Manag, 6(1), 1-12.

BasuMallick C. What is principal component analysis (PCA)? Meaning, working, and applications. Big Data. Spiceworks. Available at: https://www.spiceworks.com/tech/big-data/articles/what-is-principal-component-analysis/. (Last accessed: January 5, 2024).

Elhaik E., 2022, Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Sci Rep, 12, 14683.

Farraia M.V., Cavaleiro Rufo J., Paciência I., Mendes F., Delgado L., Moreira A., 2019. The electronic nose technology in clinical diagnosis: A systematic review. Porto Biomed J, 4(4), e42.

Gendron K.B., Hockstein N.G., Thaler E.R., Vachani A., Hanson C.W., 2007, In vitro discrimination of tumor cell lines with an electronic nose. Otolaryngol Head Neck Surg, 137(2), 269-73.

Ibrahim A.S., Khaled H.M., Mikhail N.N., Baraka H., Kamel H., 2014, Cancer incidence in Egypt: results of the national population-based cancer registry program. J Cancer Epidemiol, 2014, 437971.

Kokabi M., Tahir M.N., Singh D., Javanmard M., 2023, Advancing healthcare: Synergizing biosensors and machine learning for early cancer diagnosis. Biosensors, 13(9), 884.

Liu J., Chen H., Li Y., Fang Y., Guo Y., Li S., Xu J., Jia Z., Zou J., Liu G., Xu H., Wang T., Wang D., Jiang Y., Wang Y., Tang X., Qiao G., Zhou Y., Bai L., Zhou R., … Wang X., 2023, A novel non-invasive exhaled breath biopsy for the diagnosis and screening of breast cancer. J Hematol Oncol, 16, 63.

Malone E.R., Oliva M., Sabatini P.J.B., Stockley T.L., Siu L.L., 2020, Molecular profiling for precision cancer therapies. Genome Med, 12(8), 1-19.

Merkow R.P., Korenstein D., Yeahia R., Bach P.B., Baxi S.S., 2017, Quality of cancer surveillance clinical practice guidelines: Specificity and consistency of recommendations. JAMA Intern Med, 177(5), 701–709.

Meshref R.A., Elaskary S.A., Eldrieny A.M., Salama A.A., Saleem I.A., Mohamed E.I., 2023, Computer-aided system based on deep learning for lymph node lesions diagnosis in CT images from abdominal lymphadenopathy patients. Am J Biomed Sci & Res, 20(2), 138-145.

Mohamed E.I., Mahmoud G.N., El-Sharkawy R.M., Moro A.M., Abdel-Mageed S.M., Kotb M.A., 2014, Electronic nose for tracking different types of leukaemia: Future prospects in diagnosis. Hematol Oncol, 32(3), 165-7.

Mohamed E.I., Mohamed M.A., Abdel-Mageed S.M., Abdel-Mohdy T.S., Badawi M.I., Darwish S.H., 2019, Volatile organic compounds of biofluids for detecting lung cancer by an electronic nose based on artificial neural network. J Appl Biomed, 17(1), 67.

Mohamed E.I., Moustafa M.H., Mohamed M.A., Awad S.I., Maghraby H.K., Godeto T.W., Ross B.M., 2017, Qualitative and quantitative analysis of biological samples from non-metastatic breast cancer patients. Breast Cancer Rep, 4, 3.

Salama A.A., Darwish S.H., Abdel-Mageed S.M., Meshref R.A., Mohamed E.I., 2021, Deep convolutional neural networks for accurate diagnosis of COVID-19 patients using chest X-ray image databases from Italy, Canada, and the USA. Univ Louisville J Resp Infect, 5(1), 34.

Sarker I.H., 2021, Machine learning: Algorithms, real-world applications and research directions. SN Comput Sci, 2, 160.

Scheepers M.H.M.C., Al-Difaie Z., Brandts L., Peeters A., van Grinsven B., Bouvy N.D., 2022, Diagnostic performance of electronic noses in cancer diagnoses using exhaled breath: A systematic review and meta-analysis. JAMA Netw Open, 5(6), e2219372.

Shah S.C., Kayamba V., Peek R.M. Jr, Heimburger D., 2019, Cancer control in low- and middle-income countries: Is it time to consider screening? J Glob Oncol, 5, 1-8.

Shirasu M., Touhara K., 2011, The scent of disease: volatile organic compounds of the human body related to disease and disorder. J Biochem, 150(3), 257-66.

Sung H., Ferlay J., Siegel R.L., Laversanne M., Soerjomataram I., Jemal A., Bray F., 2021, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin, 71(3), 209-249.

Wojnowski W., Kalinowska K., 2022, Machine learning and electronic noses for medical diagnostics. In: Artificial intelligence in medicine. pp. 1203–1218.

World Cancer Research Fund International (WCRFI). Global cancer data by country. https://www.wcrf.org/cancer-trends/global-cancer-data-by-country/. (Last accessed: January 16, 2024).

World Health Organization (WHO). Newsroom, Fact Sheets, Detail: Cancer (February 3, 2022). https://www.who.int/news-room/fact-sheets/detail/cancer. (Last accessed: January 16, 2024).