ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.2263

Colorectal Cancer Detection in Sweat Samples Using Data Processing Techniques and the Cyranose 320 Electronic Nose

Abstract – Colorectal cancer (CRC) remains a leading cause of cancer-related mortality worldwide, underscoring the urgent need for non-invasive and cost-effective screening strategies. This study evaluates the feasibility of using the Cyranose 320 electronic nose to discriminate between CRC patients and healthy controls by analyzing volatile organic compounds (VOCs) in sweat samples. A total of 65 sweat samples (31 CRC, 34 controls) were analyzed. The data processing pipeline included Relative Difference (RD) feature extraction, Quantile Transformer scaling, Orthogonal Signal Correction (OSC), and Principal Component Analysis (PCA), followed by supervised machine learning classification. PCA revealed strong class separability, with the first three principal components explaining 95.72% of the total variance (PC1: 91.28%). Supervised classification using nested cross-validation demonstrated robust performance across seven algorithms. Random Forest achieved the best results, with 95.4% accuracy, 93.5% sensitivity, 97.1% specificity, and an AUC of 0.967. Decision Tree showed comparable performance, while all evaluated models exceeded an AUC of 0.90. Confusion matrix analysis confirmed high true positive rates and minimal false positives, particularly for tree-based ensemble methods. These findings demonstrate that sweat-derived VOC profiling using the Cyranose 320, combined with advanced data preprocessing and multivariate analysis, provides strong discriminative capability for CRC detection. The results support the potential of sweat-based electronic nose systems as a non-invasive, scalable, and patient-friendly screening approach, warranting validation in larger independent cohorts.

Keywords -- Electronic nose, Cyranose 320, Colorectal Cancer, Sweat analysis, Volatile organic compounds, Machine Learning

I. Introduction

Research has increasingly focused on developing non-invasive, cost-effective methods for early detection of various cancers, including colorectal cancer (CRC). In this context, particular emphasis has been placed on identifying biomarkers in biological fluids and on applying electronic biosensors for clinical screening [1], [2]. Conventional diagnostic methods for CRC, such as colonoscopy and tissue biopsy, are often invasive, expensive, and require specialized clinical infrastructure, which limits their accessibility and patient compliance [3], [4], [5]. Consequently, there is a growing demand for alternative diagnostic strategies that are affordable, minimally invasive, and suitable for large-scale screening.

Electronic biosensors have emerged as promising tools that translate molecular interactions into measurable electrical signals, providing sensitive, rapid, and low-cost platforms for disease detection [6], [7]. The use of non-invasive biological matrices significantly enhances this approach by simplifying sample collection and enabling repeated measurements. Among these matrices, sweat has gained increasing attention as a valuable biofluid for clinical diagnostics, due to its continuous secretion, ease of collection, and rich biochemical composition [8]. Unlike blood sampling, sweat can be collected noninvasively without causing discomfort, requiring trained personnel, or posing biohazard risks. It contains a wide range of metabolites, electrolytes, and VOCs that reflect systemic physiological and metabolic processes. Recent studies have demonstrated that alterations in metabolic pathways associated with cancer progression can be indirectly reflected in the composition of sweat VOCs, making sweat an attractive medium for early disease detection. Electronic nose systems promise diagnostic capabilities by capturing complex VOC patterns in biological matrices such as breath and fecal samples. These approaches have shown consistent discriminative performance, supporting the feasibility of VOC-based screening strategies as complementary tools to conventional diagnostic methods.

Despite these advances, most existing studies have focused on breath and stool analysis, while alternative biological matrices remain relatively underexplored. Sweat, in particular, is an attractive sampling medium due to its non-invasive collection, minimal patient discomfort, and potential to reflect systemic metabolic processes associated with disease states. Moreover, sweat-based sampling may offer practical advantages in terms of accessibility and ease of implementation in decentralized or resource-limited settings.

In this context, the present study investigates the potential of sweat-derived VOC profiles for colorectal cancer detection using the Cyranose® 320 electronic nose. By combining standardized sampling protocols with advanced signal processing and machine learning techniques, this work aims to contribute to the growing body of research on non-invasive diagnostic technologies and to expand the applicability of VOC analysis beyond traditionally studied biological matrices.

On the other hand, breath analysis using this technology has been widely explored; however, sweat-based analysis offers unique advantages for controlled sampling and signal reproducibility. Sweat samples are less influenced by environmental contamination and respiratory variability, which can affect breath measurements. Furthermore, sweat collection can be standardized using absorbent materials or wearable devices, enabling integration with point-of-care and continuous monitoring systems.

The concept of using emitted odors to diagnose disease dates back to Hippocrates, who associated body odors with hepatic disorders [9], [10]. Subsequent research identified specific VOCs such as acetone, dimethyl sulfide, and ketones as metabolic indicators of pathological states, reinforcing the

idea that volatile compounds can serve as non-invasive biomarkers of disease [11]. While early studies focused primarily on breath, the headspace above sweat samples has recently been recognized as a relevant source of diagnostically meaningful VOCs [12]. The term headspace refers to the gas phase immediately above a biological sample, containing volatile compounds released from the matrix. Analysis of sweat headspace enables indirect assessment of metabolic alterations without the need for direct chemical extraction.

Electronic nose systems mimic the human olfactory system by employing arrays of partially selective chemical sensors combined with pattern recognition algorithms. Exposure to VOC mixtures generates distinct electrical response patterns commonly referred to as olfactory fingerprints, which can be processed using multivariate analysis and machine learning techniques to discriminate between healthy and pathological conditions. Although analytical techniques such as gas chromatography–mass spectrometry (GC–MS) offer high chemical specificity, they are costly, time-consuming, and impractical for routine clinical screening. In contrast, eNose devices provide portable, rapid, and cost-effective solutions suitable for decentralized healthcare settings.

The Cyranose 320, developed through NASA-funded research at the Jet Propulsion Laboratory of the California Institute of Technology, is among the most extensively validated commercial electronic nose systems. Originally designed for air quality monitoring on spacecraft, it incorporates a proprietary NoseChip comprising 32 carbon-black–polymer sensors capable of detecting VOCs at parts-per-million levels. With nearly 500 peer-reviewed studies, the Cyranose 320 has demonstrated high accuracy in detecting bacterial infections and metabolic disorders, underscoring its potential for oncological diagnostics.

In this study, the primary objective is to evaluate the effectiveness of the Cyranose 320 in analyzing sweat samples obtained from colorectal cancer patients and healthy control subjects. Unlike previous works that primarily emphasize breath analysis, this research focuses on sweat as the main diagnostic matrix. Electrochemical sensor responses will be processed via signal normalization, feature extraction, and PCA-based dimensionality reduction. The composition of sweat VOCs is influenced by physiological and environmental factors, such as perspiration rate, temperature, and external contamination. To mitigate these effects, a standardized overnight collection protocol was implemented, including fasting conditions and controlled hygiene procedures. Additionally, headspace thermal conditioning (55–60 °C) was applied to stabilize VOC release and improve measurement reproducibility

Several machines learning algorithms, including Random Forest, Support Vector Machines, Logistic Regression, K-Nearest Neighbors, Naive Bayes, and Decision Trees, will be evaluated to identify the most robust classification model. By prioritizing sweat-based analysis, this study aims to contribute to the development of non-invasive, scalable, and patient-friendly diagnostic tools that facilitate early CRC detection and improve screening accessibility, particularly in high-risk and resource-limited populations.

II. MATERIALS AND METHODS

A. Study Design

This quantitative study evaluated the diagnostic potential of the Cyranose 320 electronic nose for colorectal cancer detection. Biological samples were collected from colorectal cancer patients and healthy controls at Hospital Erasmo Meoz in Cúcuta, Norte de Santander, Colombia. After collection, samples were transported to Mexico for analysis with the Cyranose 320 device. The study followed a controlled observational design, ensuring consistency in sample handling and analytical procedures.

The study comprised 65 sweat samples (34 controls and 31 cancer samples). All samples were properly labeled and stored in accordance with standard biosafety protocols prior to analysis. Sample anonymization was performed to protect patient confidentiality, and all procedures were conducted in accordance with ethical and institutional guidelines.

B. Instrumentation: Cyranose 320 Electronic Nose

The Cyranose 320 (Sensigent, Baldwin Park, CA, USA) is a portable electronic nose featuring a proprietary 32-channel polymer composite sensor array for volatile organic compound (VOC) detection. The device operates on the principle of chemiresistive transduction, in which each sensor comprises a unique conductive polymer mixed with carbon black particles. When VOCs interact with the sensor array, the polymeric films absorb varying amounts of vapor and swell, thereby altering electrical resistance, which is captured as digital patterns that represent unique “olfactory fingerprints” for different substances.

As illustrated in Figure 1, the compact, handheld design of the Cyranose 320 enables its use in clinical and laboratory settings, facilitating rapid, on-site analysis.

Fig. 1. Cyranose320 electronic nose.

The sensor array responds collectively to complex VOC mixtures, allowing pattern-based discrimination rather than single-compound identification. This makes the device particularly suitable for biomedical applications, in which disease-related metabolic changes are reflected as variations in global VOC profiles rather than by isolated biomarkers.

Technical specifications of the Cyranose 320 used in this study:

- Weight: <32 ounces (0.91 kg)
- Sensors: 32-channel polymer composite sensor array
- Dimensions: 10 × 22 × 5 cm
- Display: 320 × 200 graphic with LED backlight
- Response time: 10 seconds
- Sampling pump: 50-180 cc/min
- Operating temperature: 0 to 40°C (32 to 104°F)
- Humidity tolerance: 0-95%, non-condensing
- Built-in algorithms: K-Nearest Neighbors (KNN), K-means, Principal Component Analysis (PCA), Canonical Discriminant Analysis (CDA)
- Communication: RS-232 link, up to 57600 bps
- Battery: NiMH 4 AA battery pack or 4AA alkaline
- Battery life: 3 hours continuous operation
- Data storage capacity: 5 onboard methods, infinite on PC

The device's inlet probe consists of 2" and 4" interchangeable needles compatible with any standard female Luer Lock or slip adapter, facilitating consistent sample introduction across different biological matrices.

C. Sample Collection and Measurement Protocol

In this study, only sweat samples were analyzed to ensure methodological consistency and to focus on a single non-invasive biological matrix. Sweat collection was performed using a standardized protocol designed to minimize potential sources of contamination and ensure procedural reproducibility. Participants were instructed the day before sampling to fast for at least 8 hours and to avoid tobacco, alcohol, and certain foods, as well as perfumes, body lotions, or aerosol deodorants, due to their potential influence on the volatile organic compound (VOC) profile. Additionally, participants were advised to shower with neutral soap at least 2 hours prior to sample collection.

Each participant received a collection kit consisting of sterile gauze pads (5 cm × 5 cm), sterile water, microporous tape, sterile nitrile gloves, and 15 mL Falcon tubes. On the night prior to sampling, the lumbar region was cleaned with sterile water only and, once completely dry, two sterile gauze pads were placed and secured with microporous tape. The gauze pads remained in contact with the skin for approximately 8 hours, allowing spontaneous perspiration to be collected during sleep. The following morning, the gauze pads were carefully removed using the provided gloves, individually stored in Falcon tubes, and delivered to the laboratory.

For processing, previously frozen samples were thawed at 4 °C for 12 hours and equilibrated to room temperature for 15–20 minutes. The gauze pads were then transferred to airtight 15 mL vials sealed with a septum and heated at 55–60 °C for 5 minutes to promote VOC release into the headspace. Finally, the generated volatile compounds were directed to the Cyranose system for acquisition and analysis.

The exclusive use of sweat enabled improved reproducibility and reduced confounding from physiological differences across multiple biological fluids.

Each sweat sample was analyzed using the Cyranose 320 electronic nose, which recorded sensor resistance values (kΩ) in response to VOCs in the sample headspace. Prior to measurement, samples were equilibrated under controlled conditions to promote stable VOC release into the headspace. The device measured the collective response of the 32-sensor array, generating characteristic resistance patterns associated with the global VOC profile of each sweat sample.

All measurements were performed under controlled environmental conditions, with temperature and humidity maintained within the manufacturer-specified operating ranges. This approach minimized instrumental drift and reduced the influence of external environmental factors, thereby improving data reliability and ensuring comparability across all sweat samples analyzed in the study.

D. Data Preprocessing and Feature Extraction

Sweat samples were collected from patients using standardized protocols to ensure consistency and reproducibility. This study focused exclusively on sweat as the biological matrix for analysis, allowing for a controlled evaluation of a single non-invasive sample type. Each sweat sample was collected using the Cyranose 320 device, which measured sensor resistance in kilohms (kω) in response to volatile organic compounds (VOCs) in the sample's headspace. The measurement protocol maintained consistent environmental conditions within the device's operating specifications to minimize instrumental drift and environmental interference, thereby ensuring comparable sensor responses across all sweat samples.

• *Raw Data Structure and Preprocessing:* The Cyranose 320 generated CSV output files with 40 columns. The first two rows contained configuration metadata and were excluded from analysis. Of the remaining 40 columns, 8 contained system and environmental parameters (Time, Ambient Temperature, Thermistor Voltage, Battery status, Flag, and three unnamed monitoring variables) that were not directly relevant to VOC detection and were therefore discarded. Only the 32 sensor channels (S1–S32), representing resistance measurements in kilohms, were retained for analysis. Each sweat sample file contained between 700 and 1500 temporal measurements per

sensor, capturing the sensor array's dynamic response to headspace VOCs released from the samples.

• *Dataset Construction:* A unified dataset was created by consolidating all individual sweat sample files and assigning class labels based on clinical classification (Cancer or Control). This labeling enabled direct comparison of sweat-derived VOC profiles between colorectal cancer patients and healthy controls, producing a comprehensive data matrix of dimensions [65 samples × 32 sensors × temporal measurements] suitable for supervised machine learning analysis.

• *Feature Extraction Techniques:* To transform high-dimensional temporal sensor data into meaningful, computationally tractable features for classification, seven feature extraction techniques commonly used in electronic nose research were systematically applied to each sensor's time-series response. These techniques were selected to capture complementary aspects of sensor response dynamics, including signal amplitude, temporal evolution, variability, and relative changes during exposure to sweat-derived volatile organic compounds. Each method reduced the temporal measurements (700–1500 data points per sensor) to a single scalar value per sensor, yielding a final feature matrix of dimensions [65 samples × 32 features]. The mathematical formulation of each feature-extraction technique is presented below.

- DV (Delta Voltage): The difference between maximum and minimum resistance values.

$$DV = \max(R) - \min(R) \quad (1)$$

- DV2 (Delta Voltage 2): The difference between final and initial resistance values.

$$DV_2 = R_{\text{final}} - R_{\text{initial}} \quad (2)$$

- DV3 (Normalized Delta Voltage): The relative change from initial to final value.

$$DV_3 = \frac{R_{\text{final}} - R_{\text{initial}}}{R_{\text{initial}}} \quad (3)$$

- AUC (Area Under Curve): The integral of the sensor response over time using trapezoidal integration.

$$AUC = \int R(t)\, dt \quad (4)$$

- CV (Coefficient of Variation): The ratio of standard desviation to mean.

$$CV = \frac{\sigma}{\mu} \quad (5)$$

- GRAD (Average Gradient): The mean rate of changes of sensor resistance.

$$RAD = \text{mean}\left(\frac{dR}{dt}\right) \quad (6)$$

- RD (Relative Difference): The ratio between maximum and minimum resistance values.

$$RD = \frac{\max(R)}{\min(R)} \quad (7)$$

E. Data Scaling Techniques

To enable better comparison of sensor features and enhance the accuracy of dimensionality reduction and classification tasks, the QuantileTransformer with a uniform output distribution was applied to the feature matrix. This nonparametric method transforms original feature values to a uniform distribution via quantiles, thereby addressing non-Gaussian distributions and mitigating the effects of outliers commonly observed in chemiresistive sensor responses. Configured with 1000 quantiles, it provides a smooth transformation while maintaining the rank order of samples. This scaling method was chosen after comparing five normalization techniques, StandardScaler, MinMaxScaler, RobustScaler, PowerTransformer, and QuantileTransformer, and the latter showed the best classification results in preliminary tests. The five scaling methods are described below:

- Standard Scaler: Standardizes features by removing the mean and scaling to unit variance, assuming features follow a Gaussian distribution.

$$z = \frac{x - \mu}{\sigma} \quad (8)$$

- MinMax Scaler: Scales features to a fixed range [0, 1] through linear min-max normalization

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (9)$$

- Robust Scaler: Scales features using statistics robust to outliers, based on the median and interquartile range (IQR). Where IQR(x) = $Q_3(x)$ - $Q_1(x)$, with Q_1 and Q_3 representing the first and third quartiles.

$$x_{scaled} = \frac{x - \text{median}(x)}{\text{IQR}(x)} \quad (10)$$

$$\text{IQR}(x) = Q_3(x) - Q_1(x) \quad (11)$$

- Power Transformer: Applies a parametric power transformation (Yeo-Johnson) to make data more Gaussian-like and stabilize variance.

$$x_{transformed} = g(x, \lambda) \quad (12)$$

- Quantile Transformer (Uniform): Transforms features to follow a uniform distribution by mapping original values to their quantile ranks, effectively handling non-linear relationships and outliers while preserving rank order.

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \le x) \quad (13)$$

$$x_{scaled} = F^{-1}(U) \text{ where } U \sim \text{Uniform}(0,1) \quad (14)$$

Although several features were initially considered, the Relative Difference feature was used for the final classification because it produced the most stable class separation during exploratory analysis. Future studies should systematically compare individual and combined feature sets to quantify their contribution to classification performance.

F. Principal Component Analysis (PCA)

PCA was applied to the feature-extracted data as an unsupervised multivariate technique to reduce dimensionality while preserving maximum variance contained in the original dataset. PCA transforms the original 32 sensor features into a new set of orthogonal, uncorrelated principal components, ordered by the amount of variance they explain. This transformation facilitates the interpretation of complex sensor-array responses by projecting the data into a lower-dimensional space. As a result, PCA enables an effective visualization of class separability and highlights the most discriminative response patterns associated with sweat volatile organic compounds.

The variance explained by each principal component was calculated to assess each component's contribution to the overall data structure, and scatter plots of the first two principal components were generated to visualize the separation between the Cancer and Control groups and to explore underlying clustering patterns in the data.

G. Machine Learning Classification Algorithms

Seven supervised machine learning algorithms were implemented and evaluated for their ability to classify samples based on the extracted features. These algorithms were selected to represent different learning paradigms and decision strategies, enabling a comprehensive comparison of classification performance and robustness when analyzing sweat-based sensor data.

- Random Forest: An ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of individual trees. This method is robust to overfitting and provides feature importance rankings

$$\hat{y} = mode{h^1(x), h^2(x), \dots, h_n(x)} \quad (15)$$

- Support Vector Machine with RBF Kernel (SVM-RBF): A non-linear classification algorithm that maps input features into a higher-dimensional space using a radial basis function kernel, enabling separation of non-linearly separable classes.

$$K(x_i, x_j) = exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right) \quad (16)$$

$$The\ decision\ function\ is: f(x) = sign(\sum_i \alpha_i y_i K(x_i, x) + b) \quad (17)$$

- Support Vector Machine with Linear Kernel (SVM-Linear): A linear classification algorithm that finds the optimal hyperplane maximizing the margin between classes in the original feature space.

$$f(x) = sign(w \cdot x + b) \quad (18)$$

- Logistic Regression: A classical probabilistic binary classifier that models the probability of class membership using a logistic function, providing interpretable coefficients for each feature.

$$P(y = 1|x) = 1/\left(1 + exp\left(-(w \cdot x + b)\right)\right) \quad (19)$$

- K-Nearest Neighbors (KNN): A non-parametric algorithm that classifies samples based on the majority vote of their k nearest neighbors in the feature space, using Euclidean distance as the similarity metric.

$$\hat{y} = argmax \sum_i \in N_k(x) I(y_i = c) \quad (20)$$

- Naive Bayes: A probabilistic classifier based on Bayes' theorem with the assumption of conditional independence between features, providing fast training and prediction times.

$$P(y|x) = \frac{P(y)\prod_{m=1}^{3} P(x_m|y)}{P(x)} \quad (21)$$

- Decision Tree: A tree-structured classifier that recursively partitions the feature space based on feature thresholds, resulting in an interpretable model with clear decision rules.

$$Gini(D) = 1 - \sum_{k=1}^{2} p_k^2 \quad (22)$$

All analyses were performed in Python 3.8.10 using scikit-learn 1.0.2, NumPy 1.21.0, pandas 1.3.0, matplotlib 3.4.2, and seaborn 0.11.1. The main hyperparameters were as follows: Random Forest, 100 trees and maximum depth of 10; SVM, C = 1.0 with linear and RBF kernels; Logistic Regression, L2 regularization with C = 1.0; k-NN, k = 5 using Euclidean distance; Gaussian Naive Bayes with default variance smoothing; and Decision Tree with maximum depth of 10.

It should be clarified that to avoid data leakage, all preprocessing steps were embedded within the cross-validation loop. For each fold, normalization, OSC correction, and PCA projection were fitted exclusively on the training subset. The learned transformation parameters were then applied to the held-out validation subset. Thus, no information from the validation data was used during model fitting or dimensionality reduction.

H. Model Evaluation Metrics

Classification performance was evaluated using nested cross-validation and quantified through standard metrics commonly employed in binary classification tasks. The following performance measures were calculated for each machine learning algorithm: Accuracy, Precision, Recall (Sensitivity), and F1-score. Accuracy denotes the proportion of correctly classified samples; Precision reflects the reliability of positive cancer predictions; Recall quantifies the model's ability to correctly identify cancer cases; and the F1-score provides a balanced measure between Precision and Recall. These metrics were selected to comprehensively assess classification performance while accounting for potential class imbalance between cancer and control samples.

III. RESULTS

Following feature extraction and preprocessing, dimensionality reduction and classification algorithms were applied to evaluate the discriminative capability of VOC profiles derived from sweat samples. Feature extraction was performed using the Relative Difference metric, which captures the ratio of the maximum to the minimum sensor resistance, thereby emphasizing relative amplitude changes in sensor responses. Prior to multivariate analysis, the RD-based features were scaled using a QuantileTransformer to normalize feature distributions and mitigate skewness and outliers, thereby improving comparability across samples. The analysis focused on PCA for visualization and exploratory pattern recognition, followed by the evaluation of supervised machine learning algorithms to assess classification performance.

A. Data discrimination

To enhance class discrimination and remove systematic variation unrelated to the diagnostic outcome, principal component analysis was performed in conjunction with OSC. This method, introduced by Wold et al. [13] for calibration in near-infrared spectroscopy, removes variation orthogonal to the response variable while preserving predictive information. This method constructs latent factors that account for maximum variation in the predictor matrix while remaining orthogonal to the target variable, thereby improving the signal-to-noise ratio for subsequent multivariate analysis.

After OSC correction and Quantile Transformer scaling, PCA was applied to the RD-based feature matrix (n = 65; 31 cancers, 34 controls) to obtain a low-dimensional representation suitable for visualization and exploratory assessment of class separability. The first three principal components (PC1, PC2, and PC3) accounted for 95.72% of the total variance, with PC1 accounting for 91.28%, PC2 for 2.85%, and PC3 for 1.59%. As shown in Fig. 2, the separation between diagnostic groups is primarily driven by PC1.

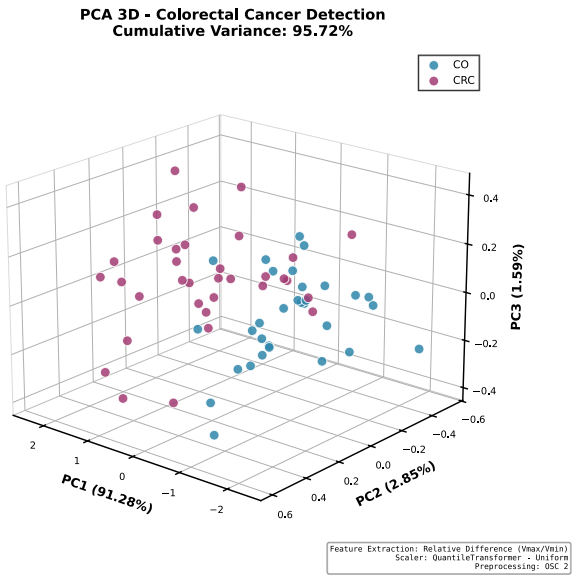

Fig. 2. PCA 3D Scatter, PC1 91.28%, PC2 2.85% and PC3 1.59% Cancer samples (purple) and Control Samples (blue).

Control samples (blue) are predominantly located at negative PC1 values, whereas cancer samples (magenta) cluster at positive PC1 values. PC2 and PC3 contribute marginally to class discrimination and mainly reflect residual variability within each group. This three-dimensional distribution indicates that the extracted features exhibit strong discriminative capability, with most class-relevant information concentrated along the first principal component, supporting their suitability for subsequent supervised classification.

B. Classification Performance

Seven machine learning algorithms were evaluated using nested cross-validation in the three-dimensional PCA space (PC1-PC3). Table I presents the comprehensive performance metrics for all algorithms.

TABLE I.
METRICS OBTAINED FROM MACHINE LEARNING MODELS

Classification Performance Metrics for Sweat – 3D				
Algorithm	***Accuracy***	***Precision***	***Recall***	***F1-score***
Random Forest	0.954	0.959	0.955	0.953
SVM (RBF)	0.862	0.876	0.867	0.860
SVM (Linear)	0.862	0.876	0.867	0.860
Logistic Regression	0.862	0.876	0.867	0.860
KNN	0.923	0.928	0.924	0.923
Naive Bayes	0.877	0.893	0.883	0.876
Decision Tree	0.954	0.955	0.955	0.954

The classification performance of seven supervised machine learning algorithms, evaluated in the three-

dimensional PCA space derived from sweat samples, varies across algorithms. Random Forest and Decision Tree achieved the highest accuracies of approximately 95.4%, with a precision of 95.9% and 95.5%, a recall of approximately 95.5%, and F1-scores of 95.3% and 95.4%, respectively. K-Nearest Neighbors (KNN) showed competitive performance with 92.3% accuracy. Support Vector Machines (both RBF and linear kernels) and Logistic Regression achieved moderate performance, with approximately 86.2% accuracy. Naive Bayes achieved 87.7% accuracy. These results indicate strong discrimination between Cancer and Control samples in the three-dimensional feature space, with tree-based ensemble methods demonstrating superior classification capability.

Confusion matrix analysis (see Fig. 3) revealed variable classification patterns across the seven algorithms in the three-dimensional PCA space. Random Forest demonstrated the strongest performance, achieving 33 true negatives (correctly identified controls) and 29 true positives (correctly identified cancer cases), along with 1 false positive and 2 false negatives, yielding 97.1% specificity and 93.5% sensitivity. The Decision Tree exhibited similar performance, with 32 true negatives, 30 true positives, 2 false positives, and 1 false negative. K-Nearest Neighbors yielded 32 true negatives, 28 true positives, 2 false positives, and 3 false negatives.

Fig. 3. Confusion matrices for seven classification algorithms using 5-fold cross-validation.

Support Vector Machines (both RBF and linear kernels) and Logistic Regression yielded identical confusion matrices: 27 true negatives, 29 true positives, 7 false positives, and 2 false negatives, resulting in moderate specificity (79.4%) while maintaining high sensitivity (93.5%). Naive Bayes achieved 28 true negatives, 29 true positives, 6 false positives, and 2 false negatives.

The variation in performance across algorithms indicates that, although all models maintain reasonable sensitivity for cancer detection, tree-based methods (Random Forest and Decision Tree) demonstrate superior specificity, thereby minimizing false alarms in control samples.

The high sensitivity maintained across most models (>90%) highlights their effectiveness in correctly identifying cancer cases, which is critical for screening applications where minimizing false negatives is paramount. Random Forest and Decision Tree's superior specificity (97.1% and 94.1%, respectively) ensures that healthy individuals are less likely to be incorrectly flagged as having cancer, thereby reducing unnecessary follow-up procedures and patient anxiety.

The performance variation across diverse machine learning paradigms from distance-based methods (KNN) to probabilistic classifiers (Naive Bayes) and ensemble techniques (Random Forest) indicates that algorithm selection significantly impacts classification outcomes in the three-dimensional feature space, with tree-based ensemble methods demonstrating optimal balance between sensitivity and specificity.

Receiver Operating Characteristic (ROC) curve analysis (see Fig. 4) further quantified the discriminative capability of each classifier in the three-dimensional feature space.

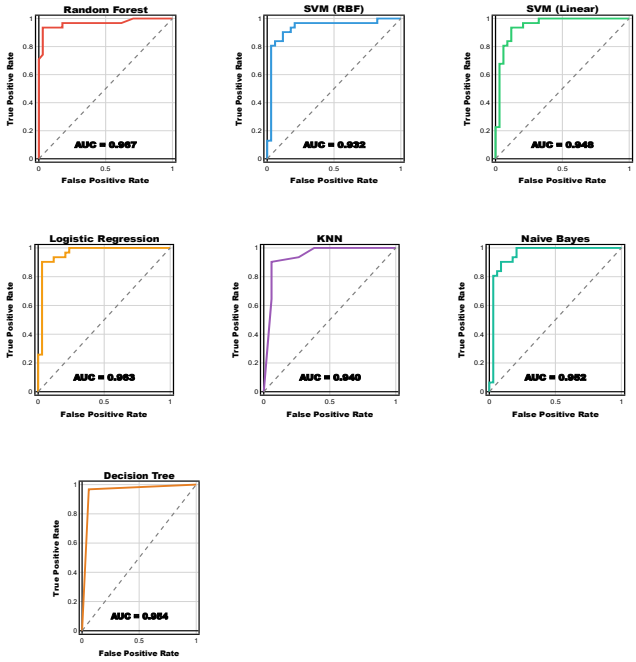

Fig. 4. ROC Curves for seven classification algorithms evaluated in the 3D PCA space using 5-fold cross-validation.

Random Forest, Decision Tree, and Logistic Regression achieved the highest area under the curve (AUC) values at 0.967, indicating excellent diagnostic performance. SVMs with linear and RBF kernels achieved strong performance, with AUC values of 0.948 and 0.932, respectively. K-Nearest Neighbors achieved an AUC of 0.940, while Naive Bayes

showed the lowest but still acceptable performance with an AUC of 0.952. All algorithms exceeded the AUC threshold of 0.90, demonstrating robust classification capability. The all-models comparison panel shows minimal variability in ROC curves across the algorithms, with all curves approaching the upper-left corner of the ROC space, indicating consistently high true-positive rates at low false-positive rates across the evaluated methods.

The high sensitivity maintained across most models (>90%) highlights their effectiveness in correctly identifying cancer cases, which is critical for screening applications where minimizing false negatives is paramount. The superior AUC values (>0.96) achieved by Random Forest, Decision Tree, and Logistic Regression, combined with Random Forest's exceptional specificity (97.1%), position tree-based ensemble methods as optimal choices for this diagnostic task. The strong ROC performance across diverse machine learning paradigms, ranging from distance-based methods (KNN) to probabilistic classifiers (Naive Bayes) and ensemble techniques (Random Forest), confirms that the three-dimensional PCA representation yields robust discriminative features for colorectal cancer detection. However, algorithm selection significantly impacts the balance between sensitivity and specificity, with Random Forest demonstrating the most favorable trade-off for clinical screening applications where both metrics are critical.

IV. CONCLUSIONS

This study demonstrated the feasibility of using the Cyranose 320 electronic nose to noninvasively discriminate between colorectal cancer patients and control subjects by analyzing volatile organic compounds in sweat samples. Feature extraction using the Relative Difference metric, combined with QuantileTransformer scaling and Orthogonal Signal Correction preprocessing, enabled measurable group differentiation.

Although the cohort size is moderate (n = 65), robust validation was ensured through stratified 5-fold cross-validation, minimizing overfitting and providing reliable performance estimation. Future work will focus on expanding the dataset across multiple institutions to enhance generalizability.

PCA revealed that three components accounted for 95.72% of the total variance (PC1: 91.28%, PC2: 2.85%, PC3: 1.59%), with control and cancer samples forming distinguishable spatial clusters primarily along PC1. Seven machine learning algorithms were evaluated using nested cross-validation in the three-dimensional space, achieving variable performance. Random Forest demonstrated optimal results with 95.4% accuracy, 93.5% sensitivity, 97.1% specificity, and 0.967 AUC, followed closely by Decision Tree. All algorithms achieved AUCs above 0.90, confirming robust discriminative capability. The superior performance of tree-based ensemble methods, particularly Random Forest's balance between sensitivity and specificity, positions them as optimal candidates for clinical screening applications. Thus, the performance metrics were averaged across folds, and variability was assessed to ensure stable classification results, reinforcing the robustness of the discriminative patterns.

These results confirm that the combination of Relative Difference extraction, QuantileTransformer scaling, and OSC preprocessing provides a robust analytical framework for non-invasive colorectal cancer detection using sweat-derived volatile organic compound profiles.

A limitation of this study was the absence of an independent external validation cohort. Although stratified 5-fold cross-validation was used to estimate model performance and reduce sampling bias, the results should be interpreted as preliminary and cohort-specific.

From a clinical perspective, the proposed system offers several advantages, including non-invasive sampling, low operational cost, and rapid analysis time. Compared to conventional screening methods such as colonoscopy, which are invasive and resource-intensive, the electronic nose approach could serve as a preliminary screening tool to identify high-risk individuals. However, regulatory approval, large-scale validation, and integration into clinical workflows remain necessary steps before implementation. Cost-effectiveness analyses and standardization of sampling protocols will be essential to support clinical adoption.

ACKNOWLEDGMENTS

This work was funded by Minciencias under Grant No. 1121-890-82391 and Contract No. 2022-0803. The authors sincerely acknowledge the valuable collaboration of the Erasmo Meoz University Hospital and the CEDIMOL Laboratory at the University of Pamplona for their support in sample collection and storage.

REFERENCES

[1] Vassilenko, V., Hernández, D., Alcaraz, M., & Pardo, A. (2023). Electrochemical biosensors for cancer biomarkers detection: Recent advances and perspectives. Biosensors and Bioelectronics, 230, 115282.

[2] Goldoni, M., Caglieri, A., Poli, D., & Apostoli, P. (2021). Salivary and urinary biomarkers in cancer diagnosis. Critical Reviews in Oncology/Hematology, 159, 103247.

[3] Lin, Y., et al. (2020). Noninvasive biosensors for cancer detection: Advances and future perspectives. Biosensors and Bioelectronics, 165, 112340.

[4] Kaczor-Urbanowicz, K. E., et al. (2016). Saliva diagnostics: Advances and emerging trends. Molecular Diagnosis & Therapy, 20(5), 437–446.

[5] Pathiraja, D., et al. (2020). Non-invasive cancer diagnostics using body fluids. Clinical Biochemistry, 84, 13–26.

[6] Fu, L., & Karimi-Maleh, H. (2024). Electrochemical sensors for early cancer diagnosis: Trends and challenges. Analytica Chimica Acta, 1276, 341560.

[7] Karimzadeh, S., et al. (2020). Emerging electrochemical biosensing strategies for cancer biomarker detection. TrAC Trends in Analytical Chemistry, 129, 115939.

[8] Singh, R., et al. (2025). Point-of-care diagnostics based on biosensors: Opportunities and challenges for non-invasive testing. Biosensors, 15(1), 45.

[9] van Keulen, K. E., et al. (2020). Breath analysis for detection of cancer in the clinical context. Lung Cancer, 146, 11–22.

[10] Shirasu, M., & Touhara, K. (2011). The scent of disease: Volatile organic compounds of the human body related to disease and disorder. Journal of Biochemistry, 150(3), 257–266.
[11] Tangerman, A. (2009). Volatile metabolites of bacteria and their role in diagnosis. Clinical Microbiology Reviews, 22(3), 451–474.
[12] Pavlou, A. K., et al. (2000). Detection of volatile organic compounds in biological samples using electronic nose technology. Annals of the New York Academy of Sciences, 945(1), 296–312.
[13] Fearn, T. (2000). On orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems, 50(1), 47-52. https://doi.org/10.1016/S0169-7439(99)00045-3