Employing Neural Networks to Analyze Human Sebum for Early Detection of Parkinson's Disease

Laura Ikic
FH Technikum
Vienna, Austria

Abstract—The detection of diseases through human odors, specifically via volatile organic compounds (VOCs), offers a promising pathway for non-invasive diagnostics. Parkinson's disease (PD), a neurodegenerative disorder lacking a definitive diagnostic test, is associated with distinct VOCs that can be detected in its early stages. This study presents a comparative analysis of existing research on the use of machine learning to analyze VOC samples from subjects for PD detection. One study employed thermal desorption-gas chromatography-mass spectrometry (TD-GC-MS), while the other utilized an electronic nose (E-nose) in conjunction with gas chromatography (GC). Both studies were evaluated based on their approaches to data analysis, including feature extraction processes, data modeling techniques, and model performance measured by Receiver Operating Characteristic (ROC) analysis and the area under the curve (AUC). The first study utilized Partial Least Squares Discriminant Analysis (PLS-DA) for modeling, while the second compared three machine learning models: Gradient Boosting Decision Trees (GBDT), Extreme Gradient Boosting (XGB), and Random Forest (RF). A comparison of the two approaches revealed that the TD-GC-MS-based method achieved a higher Correct Classification Rate (CCR), whereas the E-nose approach demonstrated superior AUC and ROC performance. These findings underscore the potential of machine learning models, particularly GBDT, in enhancing VOC-based early detection of Parkinson's disease, ultimately contributing to improved diagnostic accuracy and better clinical outcomes.

I. Introduction

A. Detecting diseases based on odors

The gases emitted by humans consist of approximately 30% dead space gas and 70% alveolar gas. Alveolar gas provides valuable information about metabolism and includes components such as nitrogen, oxygen, carbon dioxide, water vapor, rare gases, and various compounds produced during metabolic processes. This alveolar gas has already been utilized for detecting changes in biomarkers and disease detection [1]. Research has demonstrated that human scent can serve as an indicator for a range of diseases, including diabetes mellitus, tuberculosis, liver and kidney diseases, and cancer [2].

While most changes in odor are imperceptible to humans [3], there are notable exceptions. One such exception was reported by Joy Milne, a woman from Scotland with an enhanced sense of smell. Her heightened olfactory sensitivity enabled her to detect a distinct musky odor emanating from her husband, who was later diagnosed with Parkinson's disease (PD) [4]. PD is a neurodegenerative disorder that progresses over time and currently lacks a definitive diagnostic test [2]. Milne's discovery has significantly advanced the understanding of early-stage Parkinson's disease detection. In a series of controlled experiments, Milne demonstrated an exceptional ability to identify PD with 100% accuracy based solely on the scent of patients' T-shirts. This specific odor, which emerges in the early stages of PD, fades with treatment but reappears in later stages or when treatment is ineffective [4]. The presence and characteristics of this odor have become diagnostic indicators in the pursuit of earlier PD detection [5]. Further analysis of Milne's olfactory capabilities confirmed that she could detect the onset of Parkinson's disease before clinical symptoms were apparent. Subsequent research has supported her findings, indicating that individuals in the early stages of Parkinson's disease release unique volatile organic compounds (VOCs) detectable through scent. These VOCs serve as a potential diagnostic marker, suggesting that olfactory cues could play a pivotal role in the early detection of PD. This holds significant promise for non-invasive diagnostic approaches that may allow for earlier intervention and improve patient outcomes. Such advancements have the potential to revolutionize PD diagnosis and contribute to more effective clinical management of the disease [6].

Recent studies have identified four volatile organic compounds (VOCs) associated with the distinct odor differences between Parkinson's disease (PD) patients and healthy individuals: perillic aldehyde, hippuric acid, eicosane, and octadecanal. Of these, hippuric acid, eicosane, and octadecanal are considered primary contributors to the characteristic scent. The analysis of VOCs in PD patients is typically performed using sebum samples collected from the upper back, a region with a high concentration of sebaceous glands. These samples are exposed to an inert gas flow to volatilize the compounds, which are subsequently analyzed using gas chromatography (GC). Olfactory assessments can also be employed to detect these odors, though the complexity of the VOC mixture presents challenges, as the resulting "cocktail of odors" complicates precise detection [7]. Despite these limitations, the ability to identify odor changes holds diagnostic promise, and recent advances in electronic nose (E-nose) technology offer a viable alternative. E-noses provide a portable, cost-effective, and rapid means of VOC detection, representing a promising tool for early

PD diagnosis and complementing traditional diagnostic methods [8].

B. Electronic Nose

An E-Nose is a technological analog of the human olfactory system, designed to detect and identify odors using an array of sensors. While the human olfactory system relies on specialized receptors to detect odors, the E-Nose employs a gas sensor array to convert molecular gas signals into electrical signals, which are then processed by pattern recognition algorithms. The integration of Artificial Intelligence technologies, such as machine learning and pattern recognition, enables E-Nose systems to analyze and interpret complex scent data, enhancing their ability to identify a wide range of odors [9].

As illustrated in Figure 1, the human olfactory system involves the interaction of odor molecules with receptors in the nose. These receptors generate electrical signals that are processed in the brain, allowing us to recognize and interpret smells. In contrast, the E-Nose lacks the specialized receptors found in the human olfactory system but compensates for this by using a sensor array and advanced computational techniques. This allows the E-Nose to identify odors based on their unique "fingerprints"—distinct patterns of sensor responses—without relying on biological receptors [10].

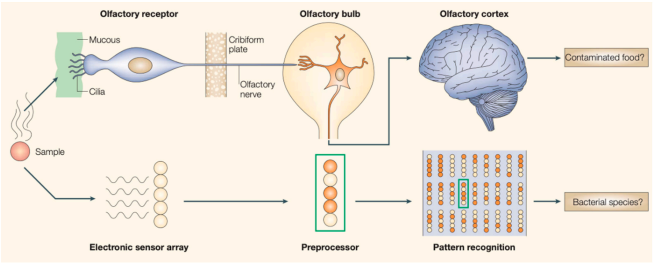

Fig. 1. Human and Electronic Model of the Olfactory System [10].

Although the E-Nose does not mimic the specific receptor-based detection process of the human nose, it can still distinguish different odors by recognizing and analyzing these odor patterns. Machine learning techniques further enhance its capabilities, allowing the E-Nose not only to identify known odors but also to predict future odors and adapt to new scent profiles. Thus, the E-Nose represents an advanced, technology-driven extension of the human sense of smell, with a wide range of practical applications, from environmental monitoring to food quality control [10].

A key component of the E-Nose is the artificial neural network (ANN), which is integral to feature extraction, modeling, and drift compensation. The system must be able to work with time-series data, which is essential for tracking VOC concentrations over time. As VOC levels fluctuate, capturing these temporal dynamics is crucial for accurate odor detection [11].

Time-series data presents a unique challenge in E-Nose systems, as the concentration of VOCs changes continuously. ANNs are well-suited for handling such data because they can model both spatial patterns (the sensor response at each time point) and temporal patterns (the evolution of responses over time). However, to ensure reliable results, efficient feature extraction techniques and noise reduction strategies are necessary to manage the large volumes of data generated by the sensors. Figure 2 illustrates how VOC concentrations fluctuate over time, emphasizing the need for time-series analysis in the detection process [10] [12].

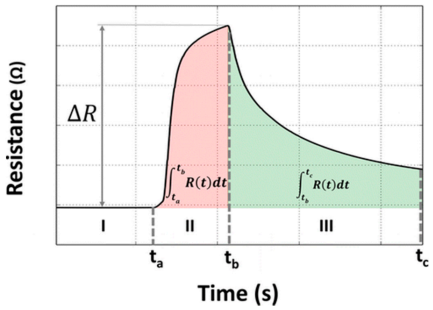

Fig. 2. Electrical Resistance measured over time [10].

Another challenge in E-Nose systems is sensor drift, which can result from factors such as aging or environmental conditions. Drift can degrade the accuracy of the system, making drift compensation techniques essential for maintaining long-term performance. Approaches such as adaptive learning and transfer learning are commonly employed to recalibrate models and compensate for sensor degradation. These methods help ensure that the E-Nose remains accurate over time, even in applications requiring continuous, long-term monitoring [13] [12].

II. Integrating Neural Networks into PD detection

The following section compares two innovative approaches for integrating neural networks into the prediction of PD. The first approach leverages machine learning to analyze sebum samples through Thermal Desorption-Gas Chromatography-Mass Spectrometry (TD-GC-MS). Unlike traditional Gas Chromatography (GC), which separates compounds in the gas phase, TD-GC-MS combines this separation process with mass spectrometry detection. This integration enhances sensitivity and specificity, allowing for detailed compound identification and precise quantification [14]. The second study explores the use of an E-nose, a cutting-edge technology for detecting volatile compounds, offering an alternative method for PD prediction [15].

A. Odor detection with TD-GC-MS

Sinclair et al. explored the use of TD-GC-MS for PD prediction by analyzing sebum samples from patients. The

samples were optained from the upper back, sealed in background-inert plastic bags, and subsequently cooled to appropriate storage temperatures until analysis [2].

1) Data Analysis: The data analysis was performed to account for variations in sebum production and ensure the integrity of the dataset before applying statistical modeling. Initially, the data were autoscaled to standardize the variables, ensuring that each had a mean of zero and a standard deviation of one. Missing values were addressed using spline interpolation, a technique for estimating and filling in gaps in the data. To control for differences in sebum production between participants, all samples were normalized to their respective Total Ion Count (TIC). This normalization step corrected for potential confounding factors related to varying amounts of sebum across samples. A further investigation of the relative intensities of common ions in high- and low-response samples from both the PD and control groups revealed no significant trends, indicating that sebum amount did not correlate with the overall VOC composition. These preprocessing steps ensured that the data were properly adjusted for analytical consistency before applying the more advanced modeling techniques [2].

2) Data Modeling: In this study, a two-pronged approach was employed for data modeling to classify and predict VOC signatures associated with PD. To mitigate the impact of class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was utilized, ensuring balanced representation between PD and control groups. The primary classification analysis was conducted using Partial Least-Squares Discriminant Analysis (PLS-DA) [2], a robust statistical method widely applied in high-dimensional datasets for predictive modeling, variable selection, and classification, particularly in fields such as food authentication, medical diagnostics, and forensic science [16].

PLS-DA was applied to identify key VOC features that effectively distinguish PD from control samples. Model validation was performed through resampling bootstrapping (n = 250), where the dataset was repeatedly partitioned into training and testing subsets to assess the stability and reliability of the resulting models. Receiver Operating Characteristic (ROC) analysis was employed to evaluate the classification performance, with the Area Under the Curve (AUC) serving as a quantitative metric for model accuracy. Additionally, Monte Carlo Cross Validation (MCCV), consisting of 30 random iterations of data splits, was carried out to assess model generalizability and calculate confidence intervals for the AUC, thereby further confirming the robustness and predictive reliability of the PLS-DA model in discriminating between PD and control groups [2].

The performance of the predictive model is summarized in the following chart, which includes key metrics such as True Positives (TP), False Positives (FP), False Negatives (FN), and the Correct Classification Rate (CCR), which is 84.4%. The model demonstrates a high accuracy of 92.4% in correctly predicting the presence of PD when it is truly present (TP). Additionally, the model shows a 64.5% accuracy in correctly identifying individuals without PD, resulting in a False Positive rate of 35.5%. The False Negative rate is 9.6%. These findings indicate that while the model is relatively effective in detecting PD, there is a notable incidence of both false positives and false negatives, suggesting potential areas for further refinement.

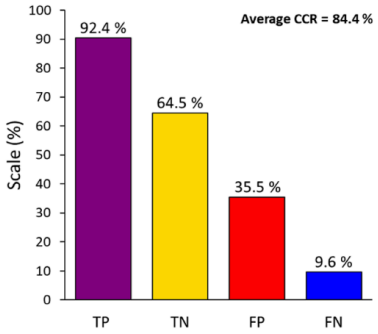

Fig. 3. TD-GC-MS Classification Results for Parkinson's Disease Detection [2].

The study also evaluated the model based on a ROC curve, which looked at the Variable importance in projection (VIP) scores which were determined for the PLS-DA model. The VIP combined subjects with subjects with PD and control subjects. Figure 4 shows the ROC curve and depicts that the addition of discriminatory compounds based on their VIP scores improves the classification model, but only up to a point (around 5-7 features). Beyond this point, further inclusion of features doesn't notably enhance the model's ability to differentiate PD from controls, and the model's predictive power plateaus. Each variable count (listed as VIP) is represented by a color in the ROC curve.

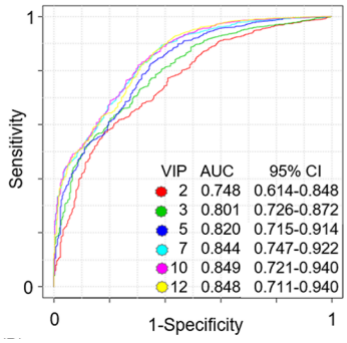

Fig. 4. TD-GC-MS ROC Curve Analysis for VIP Compound Biomarker Models [2].

B. Odor detection with E-Nose

In the second study, Cao et al. outlined a sample collection process similar to that of Sinclair et al., but with greater detail and stricter protocols. These included restrictions such as no showering, use of alcohol, perfume, or exercise for at least 12 hours prior to testing. The samples were securely sealed in background-inert plastic bottles. For analysis, the samples were examined using a surface acoustic wave (SAW) sensor in conjunction with GC [15].

1) Data Analysis: The features distinguishing PD from healthy controls (HC) were manually selected through a comprehensive analysis of VOC data. To identify the most relevant features for differentiating between the two groups, the data characteristics of PD patients and HC individuals were compared, with particular focus on time points and VOC frequencies. A univariate analysis using Spearman's rank correlation was conducted to assess the significance of various factors, including age, sex, BMI, and other clinical parameters. This analysis is a common step in manual feature selection, enabling the exclusion of features with no significant association to the target variable (PD vs. HC). Additionally, the Mann-Whitney U test was applied to identify statistical differences between the PD and HC groups, further refining the selection of key features. The significant time points, primarily between 5 and 12 seconds, based on VOC frequency, were identified as crucial for further investigation. This process of feature selection, driven by statistical significance, highlights a manual approach to identifying the most pertinent data points for distinguishing between PD and HC [15].

2) Data Modeling: The dataset, comprising 250 samples (121 PD patients and 129 HCs), was randomly split into training and testing subsets. The training dataset consisted of 200 samples (97 PD patients and 103 HCs), while the testing dataset comprised 50 samples (24 PD patients and 26 HCs). Three machine learning models—Random Forest (RF), Extreme Gradient Boosting (XGB), and Gradient Boosting Decision Tree (GBDT)—were trained on the training dataset using the significant features identified in the feature selection step. These models were designed to classify the samples into PD or HC groups, with the training process aimed at learning the underlying patterns in the data. Model performance was subsequently evaluated on the testing dataset [15].

Principal Component Analysis (PCA) was applied to reduce the dataset to three principal components, which are visualized in the 3D plot shown in Figure 5. The plot demonstrates some separation between the PD and HC groups, but also reveals significant overlap between them. This overlap suggests that the data's features may not be fully captured by linear relationships, implying that more advanced classification techniques could be necessary for better differentiation. In the plot, the first principal component score is represented along the X-axis, the second principal component score along the Y-axis, and the third principal component score along the Z-axis.

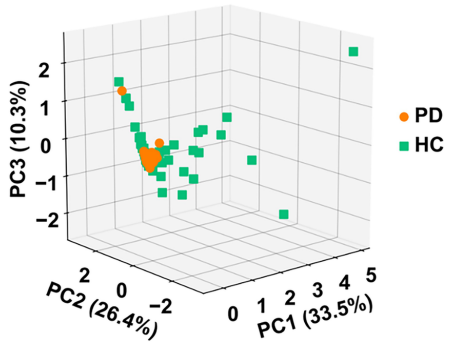

Fig. 5. 3D PCA plot showing the separation between PD and HC groups [15].

Three machine learning algorithms — Random Forest (RF), Extreme Gradient Boosting (XGB), and Gradient Boosting Decision Trees (GBDT) — were evaluated for their effectiveness in classifying individuals based on VOC data [15].

The results demonstrated that the GBDT model exhibited superior performance compared to the other models, achieving the highest sensitivity (84.00%) and specificity (83.33%). This indicates that GBDT was the most effective at accurately distinguishing between PD patients and HCs. Specifically, the true positive rate for GBDT was 83.33%, compared to 75.00% for XGB and 83.33% for RF. Similarly, the true negative rate for GBDT was 84.00%, while XGB achieved 84.00% and RF 72.00%. In terms of error rates, the false positive rate for GBDT was 16.00%, identical to that of XGB but lower than RF's 28.00%, and the false negative rate for GBDT was 16.67%, also matching XGB while outperforming RF, which had a false negative rate of 25.00% [15].

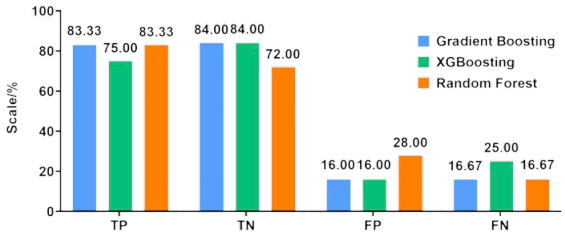

Fig. 6. E-Nose Classification Results for PD detection [15].

Receiver Operating Characteristic (ROC) curve analysis further validated these findings. The area under the curve (AUC) for GBDT was 0.893, outperforming XGB (0.870) and RF (0.852). These results confirm the robustness and reliability of GBDT in VOC-based classification tasks. The Correct Classification Rate for GBDT and XGB was 83.67% and for RF 73.50% [15].

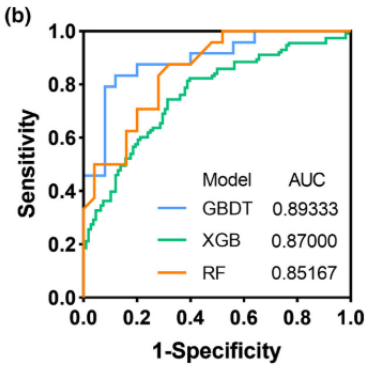

Fig. 7. E-Nose ROC Curve Analysis for VIP Compound Biomarker Models [15].

A notable limitation of the study was the small sample of drug-naive PD patients included, which prevented an analysis of how medication might affect the results in medicated PD patients.

III. Comparison of Approaches

This chapter provides a detailed comparative analysis of two advanced methodologies integrating neural networks for PD detection. The first employs TD-GC-MS to analyze sebum samples, while the second leverages an E-nose using a SAW sensor and gas chromatography. Both approaches represent state-of-the-art advancements in VOC-based diagnostics, combining sophisticated data processing and machine learning to achieve predictive insights.

A. Data Collection and Preprocessing

1) TD-GC-MS Approach: Sinclair et al. employed a rigorous protocol for collecting sebum samples from the upper back of participants. Samples were stored in background-inert plastic bags and cooled immediately to preserve integrity. Preprocessing involved several critical steps:

Autoscaling ensured that variables were standardized with a mean of zero and a unit variance, minimizing bias from large-scale differences. Spline Interpolation addressed missing data, providing smooth estimates for missing values. Normalization to Total Ion Count (TIC) adjusted for variations in sebum production between participants, ensuring data comparability. A detailed analysis confirmed that sebum quantity did not correlate with VOC composition, validating the normalization process.

2) E-Nose Approach: Cao et al. implemented stricter collection protocols, including restrictions on showering, alcohol, perfume, and physical activity before sampling. Samples were sealed in background-inert bottles for analysis. The analysis emphasized manual feature selection, relying on statistical methods such as Spearman's rank correlation and Mann-Whitney U tests to identify significant VOC features. Time points between 5 and 12 seconds were highlighted as critical based on VOC frequency, providing targeted insights into feature importance.

3) Comparison: The TD-GC-MS method exhibited a more automated and reproducible preprocessing pipeline, suitable for high-throughput analysis. In contrast, the E-nose approach relied on manual feature selection, offering flexibility but introducing potential subjectivity and variability. The stricter collection protocols in the E-nose study reduced external noise but required greater participant compliance, potentially limiting scalability.

B. Feature Extraction and Selection

1) TD-GC-MS Approach: Data preprocessing facilitated the identification of key VOC features for subsequent modeling. An analysis of variable importance in projection (VIP) scores identified discriminatory compounds, revealing that predictive performance plateaued after the inclusion of 5-7 features. This automated approach enabled high-throughput analysis with minimal user intervention.

2) E-Nose Approach: Feature selection was performed manually, guided by univariate analyses to exclude non-significant variables. The reliance on specific time points (5-12 seconds) underscored a more targeted approach but highlighted potential challenges in capturing broader VOC patterns.

3) Comparison: While both methods identified key VOC features, the TD-GC-MS approach offered a more comprehensive, data-driven framework for feature selection, reducing bias. The E-nose method's reliance on manual selection introduced the possibility of missing subtle but important patterns, though it allowed flexibility in tailoring feature selection to specific data characteristics.

C. Machine Learning Models and Predictive Performance

1) TD-GC-MS Approach: The study applied Partial Least-Squares Discriminant Analysis (PLS-DA), a robust statistical model designed for high-dimensional datasets. Techniques like SMOTE mitigated class imbalance, while validation methods such as bootstrapping and Monte Carlo Cross Validation ensured reliability.The model achieved an accuracy of 92.4%, with a Correct Classification Rate (CCR) of 84.4%. The false-positive rate (35.5%) and false-negative rate (9.6%) highlighted areas for refinement. AUC metrics demonstrated that adding discriminatory features improved performance until a plateau, suggesting a limit to the model's benefit from additional variables.

2) E-Nose Approach: Three machine learning models—Random Forest (RF), Extreme Gradient Boosting (XGB), and Gradient Boosting Decision Trees (GBDT)—were evaluated. Among these, GBDT emerged as the most effective: GBDT achieved a sensitivity of 84.00%, specificity of 83.33%, and an AUC of 0.893, outperforming XGB and RF. However, significant feature overlap in PCA results suggested that more sophisticated techniques might be needed to capture non-linear relationships fully.

3) Comparison: PLS-DA in the TD-GC-MS approach demonstrated superior statistical robustness, excelling in handling high-dimensional data with a focus on feature relevance. However, its relatively high false-positive rate could limit clinical applicability. The GBDT model in the E-nose approach balanced sensitivity and specificity effectively, achieving comparable AUC values while offering scalability for larger datasets. Nonetheless, the overlap in PCA results indicated potential limitations in the selected features' discriminatory power.

D. Practical Considerations and Limitations

1) TD-GC-MS Approach: The complex preprocessing and analytical pipeline, while ensuring high data quality, could limit accessibility and scalability in clinical settings. Additionally, the model's high false-positive rate suggests potential diagnostic challenges, particularly in distinguishing PD from other neurodegenerative disorders.

2) E-Nose Approach: The E-nose system offers faster analysis and is potentially more adaptable to clinical workflows. However, the reliance on manual feature selection and the small sample size, particularly the limited representation of drug-naive PD patients, highlight areas for improvement.

3) Comparison: TD-GC-MS excels in analytical rigor and precision, making it suitable for research applications requiring high accuracy. The E-nose approach offers a more pragmatic solution for real-world implementation but may require refinements in feature selection and model design to improve reliability and generalizability.

IV. Conclusion

This review highlights the potential of volatile organic compounds (VOCs) as biomarkers for the early detection of Parkinson's disease and emphasizes the effectiveness of machine learning models in analyzing VOC data. Based on the analysis of existing studies, Gradient Boosting Decision Trees (GBDT) consistently outperformed other models, achieving the highest sensitivity, specificity, and AUC values, thereby demonstrating its superiority in diagnostic applications. These findings underscore the promise of VOC analysis, combined with advanced machine learning approaches, as a robust and non-invasive diagnostic tool for Parkinson's disease. Future research should focus on refining electronic nose (E-Nose) technologies and integrating them with machine learning models to enhance diagnostic accuracy, portability, and scalability. Further exploration of deep learning techniques and temporal data analysis could also optimize the detection process, paving the way for more effective diagnostic strategies and improved patient outcomes.

References

[1] Y. Z. J. W. R. Y. Ying Li, Xiangyang Wei, "Research progress of electronic nose technology in exhaled breath disease analysis," Microsystems Nanoengineering, vol. 9, no. 129, pp. 1–22, Oct. 2023.

[2] D. S. K. A. H. J. M. S. H. L. T. K. A. M. R. R. M. A. d. B. M. S. D. K. T. P. B. Eleanor Sinclair, Caitlin Walton-Doyle, "Validating differential volatilome profiles in parkinson's disease," ACS Central Science, vol. 7, p. 300−306, Feb. 2021.

[3] N. S. Lee Sela, "Human olfaction: a constant state of change-blindness," Experimental Brain Research, vol. 1, pp. 13–29, Aug. 2010.

[4] J. Morgan, "Joy of super smeller: sebum clues for pd diagnostics," The Lancet Neurology, vol. 15, pp. 138–139, Feb. 2016.

[5] O. K. S. K. S. S. M. M.-R. Z. R. T. Adrina Habibzadeh, Vahid Reza Ostovan, "Volatile organic compounds analysis as promising biomarkers for parkinson's disease diagnosis: A systematic review and meta-analysis," Clinical Neurology and Neurosurgery, vol. 235, Dec. 2023.

[6] J. J. S. P. R. S. Shrinidhi Kulkarni, Neenu George Kalayil, "Detection of parkinson's disease through smell signatures," in Lean and Computing in Construction Congress, Jul. 2020.

[7] Q. Y. J. F. G. Z. Y. L. C. P. H. D. D. W. H. R. Y. G. Q. L. J. L. X. C. Wei Fu, Linxin Xu, "Artificial intelligent olfactory system for the diagnosis of parkinson's disease," ACS Omega, vol. 7, pp. 3836–4723, Jan. 2022.

[8] Y. X. D. S. C. W.-D. C. L. P. B. J. M. M. S. T. K. R. G. P. B. Drupad K. Trivedi, Eleanor Sinclair, "Discovery of volatile biomarkers of parkinson's disease from sebum," ACS Central Science, vol. 5, p. 599–606, Mar. 2019.

[9] S. Husam, "Odors detection and recognition based on intelligent e-nose," Journal of Electrical Systems, vol. 20, pp. 956–966, Mar. 2024.

[10] Q. L. Zhenyi Ye, Yuan Liumm, "Recent progress in smart electronic nose technologies enabled with machine learning methods," Sensors, 2021, vol. 21, no. 7620, pp. 1–22, Nov. 2021.

[11] R. J. Q. L. Zhenyi Ye, Yaonian Li, "Toward accurate odor identification and effective feature learning with an ai-empowered electronic nose," IEEE Internet of Things Journal, vol. 11, pp. 4735 – 4746, Feb. 2024.

[12] S.-W. C. K.-T. T. Syuan-He Wang, Ting-I Chou, "Using a hybrid deep neural network for gas classification," IEEE Sensors Journal, vol. 21, pp. 6401–6407, Mar. 2021.

[13] M. L. M.-T. P. S. D. V. F. F. G. D. F. P. D. V. E. Massera, B. Alfano, "Rapid parallel calibration for environmental bulky gas sensor systems," in Lecture Notes in Electrical Engineering, Feb. 2020.

[14] A. W. C. B. Andrea Marcillo, Juan C. Baca Cabrera, "A comparison between mobile and stationary gas chromatography–mass spectrometry devices for analysis of complex volatile profiles," Analytical and Bioanalytical Chemistry, vol. 415, p. 137–155, Nov. 2022.

[15] J. Z. Y. F. Q. L. W. F. J. Z. X. X. G. Z. D. K. X. C. J. F. Yina Cao, Lina Jiang, "A fast and non-invasive artificial intelligence olfactory-like system that aids diagnosis of parkinson's disease," European Journal of Neurology, vol. 31, pp. 1–10, Mar. 2024.

[16] A. A. J. Loong Chuen Lee, Choong-Yeun Liong, "Partial least squares-discriminant analysis (pls-da) for classification of high-dimensional (hd) data: a review of contemporary practice strategies and knowledge gaps," Analyst, vol. 143, pp. 3526–3539, Jun. 2018.