Multi-Window Feature Extraction for SVM-Based Electronic Nose Classification of Four Herbal Essential Oils

Bambang Heru Iswanto[1*], *Fajar* Hardoyono[2], and *Haris* Suhendar[1]

[1]Department of Physics, Universitas Negeri Jakarta, Jl. Rawamangun Muka, Jakarta 13120, Indonesia
[2]Department of Environmental Science, UIN Prof KH Saifuddin Zuhri, Purwokerto 53126, Indonesia

Abstract. Rapid authentication of herbal essential oils requires sensor models that are accurate, fast, and chemically interpretable. This paper presents an electronic-nose workflow for four oils, namely red ginger, white turmeric, turmeric, and lemongrass, using multi-window feature extraction and support vector machine classification. Signals were collected from a ten-sensor metal-oxide-semiconductor array during an exposure period from 10 to 70 seconds after baseline correction. Features were extracted from short, medium, and long windows corresponding to onset, peak development, and tail dynamics, then classified using radial-basis-function support vector machines under nested cross-validation. Gas chromatography-mass spectrometry profiles were used as class-level chemical anchors rather than direct regression targets. Early windows already contained strong class information, especially for citral-rich lemongrass, whereas mid and late windows improved interpretation for slower terpene-rich oils. The compact window triad produced competitive macro-averaged F1-score performance while preserving an auditable link between sensor features and volatile kinetics. The workflow supports rapid screening of herbal essential oils with a parsimonious and chemically interpretable feature set.

1 Introduction

Herbal essential oils contain complex volatile mixtures that support food, pharmaceutical, cosmetic, and medicinal applications. Red ginger, white turmeric, turmeric, and lemongrass are important Indonesian herbal commodities because their aroma profiles are associated with bioactive terpenes and oxygenated compounds. In routine quality control, gas chromatography-mass spectrometry (GC-MS) provides molecular confirmation, but it is relatively slow and resource-intensive. An electronic nose (e-nose) is an artificial olfaction instrument that combines an array of partially selective gas sensors, signal acquisition, and pattern-recognition algorithms to generate volatile fingerprints for sample recognition. E-noses based on metal-oxide-semiconductor (MOS) sensor arrays offer a complementary screening tool because they can record volatile fingerprints rapidly and non-destructively [1–3].

* Corresponding author: bhi@unj.ac.id

A persistent challenge in MOS e-nose modeling is that sensor signals are transient, drift-sensitive, and highly dependent on the chosen feature-extraction interval. Conventional full-window features may ignore useful early information, whereas overly short windows may miss slower adsorption and desorption phenomena. Recent studies on food e-nose systems emphasize that signal acquisition, data processing, feature construction, and environmental drift control jointly determine recognition performance [3–5]. Time-resolved gas-sensor studies also show that volatile release can be tracked dynamically and summarized using multivariate representations such as principal component analysis [4].

This paper focuses on a compact multi-window representation for support vector machine (SVM)-based classification of four herbal essential oils. In this work, SVM refers to a supervised maximum-margin classifier, radial basis function (RBF) refers to the Gaussian nonlinear kernel used to model nonlinear class boundaries, and radial basis function support vector machine (RBF-SVM) refers to an SVM classifier that uses the RBF kernel. Time windows are treated as chemically meaningful feature-engineering units: the short window captures rapid volatile onsets, the medium window captures peak organization, and the long window captures slower tails. RBF-SVM was selected because SVM-based chemometric models remain effective for small-sample, high-dimensional e-nose data [6]. Model selection and evaluation were separated through nested cross-validation to reduce optimistic bias, while uncertainty was summarized using bootstrap intervals, and macro-F1 was used as the main balanced multi-class performance metric [7].

The contribution of this work is threefold. First, in contrast to broad essential-oil e-nose classification studies that mainly report whole-response fingerprints or global feature sets [1] and food e-nose reviews that call for more interpretable, drift-aware, and field-ready validation [2,3,5]. This study uses time windows as the primary representation, so that features correspond to response onset, peak development, and tail behavior. This provides a more auditable feature set for herbal essential oils than a single full-exposure summary. Second, whereas rapid-detection and time-resolved e-nose studies have emphasized early non-steady-state signals for efficiency [4,8] or feature fusion in other food matrices [6], this work evaluates cumulative windows and a compact triad under the same nested cross-validation protocol, directly quantifying the trade-off between decision time and balanced four-class performance. Third, unlike studies that use chemical analysis only as separate confirmation, this work uses GC-MS profiles as class-level chemical anchors, allowing RBF-SVM separation to be discussed in relation to volatile chemistry rather than accuracy alone [9–12]. These differences provide an interpretable rapid-screening workflow tailored to four Indonesian herbal essential oils.

2 Materials and methods

2.1 Essential oil samples and acquisition protocol

The study used four herbal essential oils: red ginger, white turmeric, turmeric, and lemongrass. Each class consisted of 20 independent e-nose acquisitions, giving 80 observations in total. Aroma responses were measured using an e-nose datalogger equipped with 10 MOS chemiresistor channels (S1-S10) and auxiliary temperature and relative humidity channels. Each acquisition consisted of a 0-10 s clean-air baseline, a 10-70 s exposure phase, and a 70 -100 s purge phase. Signals were sampled synchronously at 10 Hz.

Before feature extraction, each channel was baseline-corrected by subtracting the mean signal during the 0-10 s interval. Analysis was then restricted to the 10-70 s exposure period. Temperature and humidity were retained as audit covariates because environmental variation is known to affect MOS responses; humidity-aware rapid-detection strategies are therefore

important for practical deployment [8]. Figure 1 illustrates the baseline and exposure response used for subsequent windowing. The complete acquisition and feature-engineering design is summarized in Table 1.

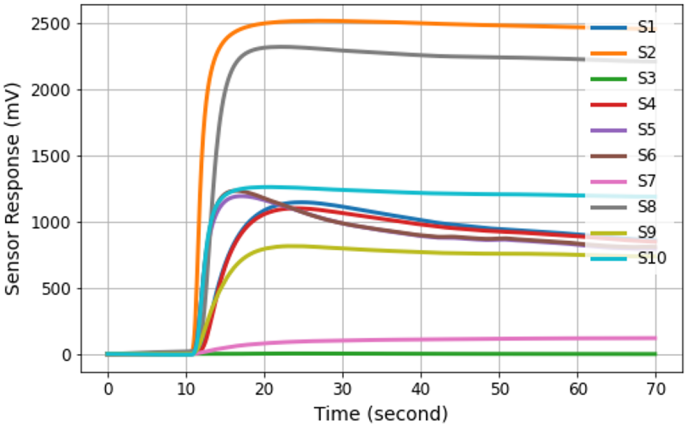


Fig. 1. Representative baseline-corrected responses of ten e-nose sensors during the 0-70 s acquisition, showing rapid response after the exposure begins at 10 s.

Table 1. E-nose acquisition and feature-engineering design.

Component	Design used in this study
Classes	Red ginger, white turmeric, turmeric, and lemongrass essential oils
Measurements	20 acquisitions per class; 80 e-nose observations in total
Sensor system	Ten MOS channels (S1-S10) with temperature and relative humidity
Timing	0-10 s baseline; 10-70 s exposure; post-exposure purge
Windows	Short 10-20 s; medium 20-40 s; long 40-70 s
Descriptors	Maximum amplitude, area under the curve, and slope for each sensor channel
Classifier	RBF-SVM with C and gamma tuned in the inner cross-validation loop

2.2 Multi-window feature extraction and validation

The core representation used three non-overlapping windows: 10-20 s, 20-40 s, and 40-70 s. These windows were chosen to capture three kinetic regimes of MOS response. The first window represents fast onset, which is expected to be informative for highly volatile oxygenated monoterpenes. The second window represents peak formation and stabilization. The third window captures slower tails associated with less-volatile, higher-molecular-weight compounds. This design is consistent with recent e-nose studies that emphasize dynamic volatile organic compound release and time-resolved signal processing [3,4,8]. Figure 2 shows how the triad is mapped onto the exposure period.

For every sensor channel and window, three descriptors were computed: maximum amplitude, area under the curve using trapezoidal integration, and local slope estimated from the initial part of the window using the same rule for all samples. The triad therefore produced 90 features (3 windows × 10 sensors × 3 descriptors). For comparison, cumulative single windows [10-T] were also evaluated, where T was 20, 30, 40, 50, 60, or 70 s. This comparison allowed the time-accuracy trade-off of the compact multi-window design to be assessed against simpler baselines.

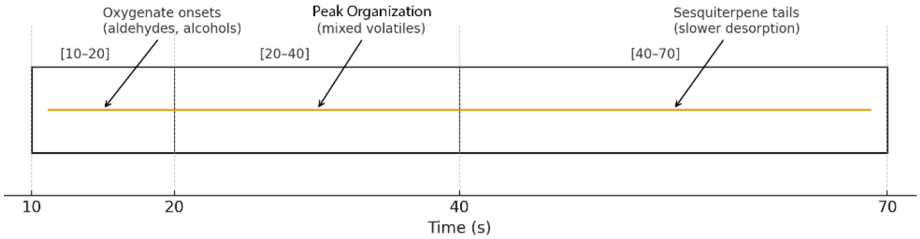


Fig. 2. Multi-window triad showing short [10-20], medium [20-40], and long [40-70] s segments for onset, peak organization, and slower tail dynamics.

All feature standardization was performed within the training folds to avoid information leakage. Classification used the RBF-SVM defined above. The regularization parameter C and kernel parameter gamma were tuned in an inner three-fold cross-validation loop, while generalization was estimated using an outer five-fold stratified cross-validation loop. Nested cross-validation was used to separate model selection from model assessment, following current recommendations for time-series classification and machine-learning prediction studies [13,14]. Performance was reported using accuracy and macro-F1; macro-F1 was prioritized because it gives equal weight to all classes in multi-class evaluation [7].

GC-MS was used as a class-level interpretive anchor, not as a direct sensor-to-concentration regression target. Lemongrass was characterized by a citral-rich composition, especially neral and geranial [9]. Turmeric was interpreted through turmerone-type constituents [10]. White turmeric was associated with Curcuma zedoaria-type sesquiterpenes, including germacrone, curzerene, and curdione [11]. Red ginger was interpreted through major volatile constituents reported for Zingiber officinale extracts [12].

3 Results and discussion

3.1 Chemical rationale for windowed sensor features

The GC-MS summary supported the use of different time windows. Lemongrass contained a high citral fraction, with neral and geranial contributing approximately half of the area in the original dataset. Such oxygenated monoterpenes are expected to produce strong early transients in MOS sensors [9]. In contrast, turmeric and ginger contained more slowly expressed terpene patterns, and white turmeric showed a mixed profile with oxygenates, sesquiterpenes, and non-terpene alkane peaks that require careful interpretation [10,12].

These chemical differences provide a rationale for the multi-window design. If the classifier uses only the earliest segment, it may separate lemongrass effectively but can confuse oils with slower or more complex kinetic profiles. Adding medium and long windows gives the model access to peak shape and tail information, which are more relevant for discriminating red ginger, turmeric, and white turmeric. The multi-window representation therefore functions not merely as a larger feature vector but as a chemically motivated encoding of volatile response stages.

3.2 Classification performance and decision time

Table 2 compares cumulative single-window features with the three-window triad. The cumulative [10-20] window already produced a macro-F1 of 0.872, showing that the first seconds of exposure contained strong discriminative information. The highest point estimate was obtained with the full [10-70] single-window (macro-F1 = 0.900), but the multi-window

triad remained competitive (macro-F1 = 0.862) and provided a more interpretable decomposition of early-, mid-, and late-signal behavior. The bootstrap interval for the triad overlapped with the high-performing single-window intervals, indicating that differences should be interpreted with caution in this small dataset.

Table 2. Classification performance of selected single-window and multi-window configurations.

Window Set	End Time (s)	#Features	Accuracy (mean ± SD)	Macro-F1 (mean ± SD)	Macro-F1 95% CI
[10 - 20]	20	30	0.875 ± 0.1250	0.8717 ± 0.1271	[0.7636, 0.9702]
[10 - 30]	30	30	0.875 ± 0.1768	0.8619 ± 0.2007	[0.7601, 0.9680]
[10 - 40]	40	30	0.875 ± 0.1768	0.8619 ± 0.2007	[0.7601, 0.9680]
[10 - 50]	50	30	0.825 ± 0.2092	0.8233 ± 0.2087	[0.6957, 0.9266]
[10 - 60]	60	30	0.875 ± 0.1768	0.8619 ± 0.2007	[0.7601, 0.9680]
[10-70]	70	30	0.9 ± 0.1369	0.9 ± 0.1369	[0.7522, 0.9699]
[10-20] + [20-40] + [40-70]	70	90	0.875 ± 0.1768	0.8619 ± 0.2007	[0.7910, 0.9782]

The T90 decision threshold, defined here as the earliest window end time reaching 90% of the best macro-F1 value, was reached early because 90% of the best macro-F1 value (0.900) equals 0.810, and all tested windows exceeded that level. This suggests that much of the information needed for four-class discrimination appears before the end of the 70 s exposure. Figure 3 visualizes both the time-accuracy curve and the associated class-level error structure.

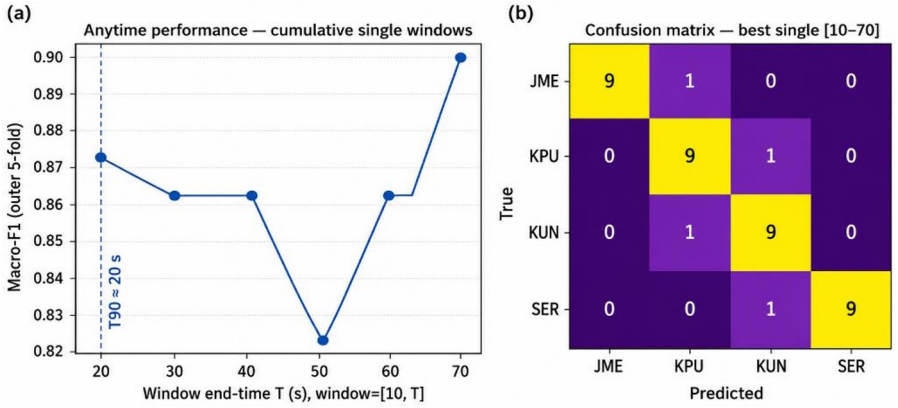
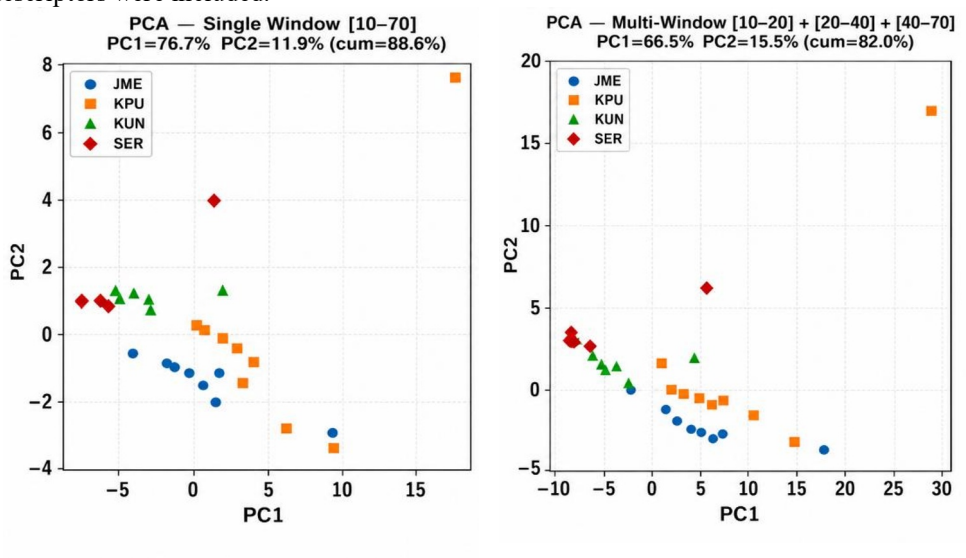


Fig. 3. (a) Anytime performance of cumulative single windows expressed as macro-F1 versus window end-time; (b) confusion matrix for the best single-window configuration, showing dominant diagonal classification with limited inter-class confusion.

3.3 Chemical interpretation, PCA visualization, and robustness

The strongest and most stable class separation was observed for lemongrass. This agrees with the GC-MS profile, as citral-rich oils should produce rapid, high-amplitude early responses. The short window, therefore, carries a strong lemongrass signature, explaining why early classification was already effective. Turmeric and red ginger required more information from medium- and longer-window periods because their discriminating compounds are associated with slower-response components.

Principal component analysis (PCA) was used as an unsupervised diagnostic visualization for the engineered feature space. PCA and related multivariate representations are commonly used in e-nose studies to check whether sensor features preserve meaningful sample structure [1–4]. Figure 4 compares the single-window and multi-window feature spaces. In the original analysis, lemongrass formed a compact cluster along the early-response axis, whereas red ginger, white turmeric, and turmeric were more clearly separated when mid- and late-window descriptors were included.



(a) (b)

Fig. 4. PCA score plots for (a) cumulative single-window [10-70] features and (b) multi-window [10-20]+[20-40]+[40-70] features. JME, KPU, KUN, and SER denote red ginger, white turmeric, turmeric, and lemongrass, respectively.

The main benefit of the triad is therefore interpretability and auditability. Early-window slopes and maxima can be associated with volatile onsets, while mid- and late-window integrals capture broader peak morphology and tail behavior. In small datasets, however, adding windows increases dimensionality, so the number of windows should remain parsimonious. The chosen triad kept a direct chemical meaning while avoiding excessive fragmentation.

Robustness remains an important limitation. The dataset contained only 80 acquisitions, so confidence intervals are wide and the reported ranking among window configurations should not be overinterpreted. Environmental effects were addressed through baseline correction, fold-wise standardization, and temperature-humidity audit variables, but independent validation across instruments, days, and batches is still needed. Future work should also test adaptive window selection and per-sample GC-MS linkage to connect sensor predictions more directly to chemical variation.

4 Conclusion

This study presents an electronic-nose workflow for classifying four herbal essential oils using multi-window feature extraction and support vector machine classification. The proposed window triad, consisting of 10-20 s, 20-40 s, and 40-70 s intervals, provides a compact way to encode fast-onset, peak-development, and slower tail dynamics. The approach produced high and interpretable classification performance under nested cross-

validation, with macro-F1 values comparable to strong single-window baselines and confidence intervals that reflect the uncertainty of the small-sample setting.

The main value of the multi-window design lies in its interpretability and auditability. Lemongrass was separable from early windows because of its citral-rich volatile profile, while turmeric, red ginger, and white turmeric benefited from information distributed across later response segments. GC-MS anchors helped relate SVM separation to plausible chemical kinetics. The method therefore supports rapid screening of herbal essential oils while preserving a clear connection between sensor features, classifier behavior, and chemical composition.

This publication was funded by the Indonesia Endowment Fund for Education (LPDP) on behalf of the Indonesian Ministry of Higher Education, Science and Technology, and managed under the EQUITY Program (Contract Number 4308/B3/DT.03.08/2025 and B/284/UN39/HK.07.00/2025).

References

1. M. Rasekh, H. Karami, A.D. Wilson, M. Gancarz, Chemosensors 9, 142 (2021). https://doi.org/10.3390/chemosensors9060142
2. T. Sanislav, G.D. Mois, S. Zeadally, et al., Sensors 25, 4437 (2025). https://doi.org/10.3390/s25144437
3. M. Wang, Y. Chen, Eur. Food Res. Technol. 250, 21 (2024). https://doi.org/10.1007/s00217-023-04381-z
4. Z. Han, W. Ahmad, Y. Rong, et al., Foods 13, 1721 (2024). https://doi.org/10.3390/foods13111721
5. L. Ren, G. Cheng, W. Chen, et al., Sens. Rev. 44, 733 (2024). https://doi.org/10.1108/SR-06-2024-0554
6. S. Gu, Z. Wang, W. Chen, J. Wang, J. Sci. Food Agric. 101, 4220 (2021). https://doi.org/10.1002/jsfa.11061
7. O. Rainio, J. Teuho, R. Klén, Sci. Rep. 14, 6086 (2024). https://doi.org/10.1038/s41598-024-56706-x
8. M. Cai, S. Xu, X. Zhou, H. Lu, Sensors 24, 5881 (2024). https://doi.org/10.3390/s24185881
9. D.N. Do, D.P. Nguyen, V.-D. Phung, et al., Processes 9, 593 (2021). https://doi.org/10.3390/pr9040593
10. A.M. Orellana-Paucar, Molecules 29, 4210 (2024). https://doi.org/10.3390/molecules29174210
11. D.K. Poudel, P.K. Ojha, A. Rokaya, et al., Plants 11, 1932 (2022). https://doi.org/10.3390/plants11151932
12. A. Spyrou, M.G.F. Batista, M.L. Corazza, et al., Molecules 29, 871 (2024). https://doi.org/10.3390/molecules29040871
13. M. Tahkola, Z. Guangrong, IEEE Access 10, 39299 (2022). https://doi.org/10.1109/ACCESS.2022.3166525
14. C.L.A. Navarro, J.A.A. Damen, T. Takada, et al., BMJ 375, n2281 (2021). https://doi.org/10.1136/bmj.n2281