Journal of Food Measurement and Characterization
https://doi.org/10.1007/s11694-026-04063-w

ORIGINAL PAPER

A smart sensor system for gas fingerprint classification in *Cistus Ladanifer* essential oil

Ali Ahmad[1] · Francisco Javier Diaz[1] · Sandra Viciano-Tudela[1] · Lorena Parra[2] · Sandra Sendra[1] · Jaime Lloret[1] · Veronika Chaloupková[3,4] · Raquel Bados[4] · Luis Saul Esteban Pascual[4] · Irene Mediavilla[4] · Pedro V. Mauri[5] · Celia Cano Shaw[5] · Jose Plaza Ramos[5] · Barbara Stefanutti[5] · Alicia Sanchez-Gorostiaga[5]

Received: 29 July 2025 / Accepted: 9 January 2026
© The Author(s) 2026

Abstract
Volatile organic compounds (VOCs) are central to the aromatic and therapeutic properties of essential oils (EOs), with their profiles serving as reliable indicators of EO quality. In *Cistus ladanifer*, the synthesis and emission of VOCs—particularly terpenic hydrocarbons—are strongly influenced by environmental variables such as temperature, humidity, and solar radiation. Traditional EO quality assessment methods, including gas chromatography (GC), although highly accurate, are costly, labor-intensive, and destructive. This study proposes a smart sensor system that utilizes a low-cost array of MQ gas sensors combined with machine learning (ML) algorithms for non-destructive classification of VOC fingerprints in *Cistus ladanifer* EO. VOC data from 33 EO samples were collected using MQ sensors and paired with environmental datasets (Daily, 15-Day, and All) obtained from weather stations in the cultivation areas. Gas chromatography coupled with flame ionization detection and mass spectrometry (GC-FID/MS) was used as the reference method to quantify the concentrations of terpenic hydrocarbons. A total of 5,154 data points were used, with 75% used for training and validation and 25% reserved for independent testing. Model performance was evaluated using tenfold cross-validation within the training set. Several ML models—Neural Networks, Decision Trees, Logistic and Linear Regression, and Ensemble—were trained to predict hydrocarbon levels from VOC signals and environmental inputs. The "Daily" dataset produced the highest R^2 (92.68%), followed closely by the "All" dataset (R^2 = 92.67%). Minimal feature subsets were identified to optimize model performance: 8, 5, and 13 features for the Daily, 15-Day, and All datasets, respectively, using environmental variables alone. When VOC data were integrated, optimal feature sets increased to 12, 20, and 29. This study demonstrates the feasibility of a smart sensor-based framework for EO quality assessment via gas fingerprint classification. The integration of low-cost sensing technologies with ML models enables a scalable, rapid, and non-invasive approach to monitoring EO composition. While results were obtained under controlled laboratory conditions and real-world performance will require field validation to address environmental variability and sensor drift, the framework shows strong potential for deployment in precision agriculture in coastal areas and beyond, environmental monitoring, and the sustainable valorization of aromatic plant resources.

Keywords Precision agriculture · Volatile organic compounds · Gas sensors · Machine learning · Amount-of-substance · Integration of sensors

Abbreviations

DFA	Discriminant Function Analysis
EO	Essential Oils
Fuzzy ARTMAP	Fuzzy Adaptive Resonance Theory Mapping
GC-FID-MS	Gas Chromatography coupled with Flame Ionization Detection and Mass Spectrometry
GC-MS	Gas Chromatography–Mass Spectrometry
IoT	Internet of Things
HCA	Hierarchical Cluster Analysis

Extended author information available on the last page of the article

Published online: 14 March 2026

HPLC	High-Performance Liquid Chromatography
IDMAP	Intrinsic Dimensionality Mapping
LDA	Linear Discriminant Analysis
LR	Logistic Regression
ML	Machine Learning
MOS	Metal Oxide Semiconductor
MQ	MQ gas sensors
MRMR	Maximum Relevance–Minimum Redundancy
NN	Neural Networks
PCA	Principal Component Analysis
SLR	Simple Linear Regression
SVM	Support Vector Machine
VOC	Volatile Organic Compounds
e-Noses	Electronic Noses

Introduction

Essential oils (EOs) are complex substances obtained from plant materials, with multiple uses. The EOs can be composed of hundreds of compounds of various chemical classes that derive from the secondary metabolism [1, 2]. The chemical compounds are similar in their chemical structure, and specialized techniques are needed to identify and quantify them in EO samples. Identifying and quantifying these compounds in EOs is extremely important for different reasons, being the most important: (i) the assessment of the EO quality, (ii) the accomplishment of the required standards for EO commercialization, and (iii) the identification of potential adulteration processes. The quality of EOs is highly affected by the environmental conditions in which the harvested plants grow [3, 4] and the presence of certain stress [5, 6].

Numerous studies have explored the meteorological impact on EO production in various plant species, utilizing different statistical techniques [7, 8]. In fact, meteorological data have been used to rationalise trends in EO quantity and quality in Cinnamomum camphora [9] and even for predicting EO in *Curcuma longa* [10]. Additionally, Ložienė, et al. [11] observed that meteorological conditions influenced the accumulation of EOs in different *Thymus* species,. Similarly, Mehalaine and Chenchouni [12] observed that climatic factors affected the volatile oil yields of *Thymus algeriensis* and *Rosmarinus officinalis*. Another study reported that the seasonal variations, genetic properties, and developmental stages influenced rosemary oil composition, as analyzed through Gas Chromatography (GC) and GC coupled to Mass Spectrometry (GC–MS) [13].. It has been reported taht climate stress, high carbon dioxide and temperature, affects two aromatic plants (*Angelica glauca* and *Nardostachys jatamansi*) [8]. The therapeutic properties of *A. glauca* improved with increasing carbon dioxide and temperature, but monoterpenes decrease, while in *N. jatamansi*, sesquiterpenes increased.

The techniques used for the identification of the constituents of EO are mainly the chromatographic analyses including GC–MS and High-Performance Liquid Chromatography (HPLC) [14, 15]. These techniques, although being accurate, are expensive, cumbersome, and time consuming. The use of sensors to identify chemical compounds in EO represents a cost-effective alternative and faster method [16, 17]. The inherent characteristics of volatile organic compounds (VOCs) present in EOs poses challenges to the use of non-destructive identification techniques [16]. Recent advancements have led to the development of VOC sensors, initially designed to quantify specific compounds (e.g., ammonia, ethanol, benzene) [17], but now integrated with machine learning (ML) techniques to form electronic noses (e-Noses) [18]. In [17], the authors have used these sensors to differentiate among essential oils of two different species, including adulterated versions of these oils. These systems utilize multiple gas sensors and generate digital fingerprints that can differentiate odors. Low-cost metal oxide semiconductor (MOS) sensors, such as MQ sensors, are particularly notable for their role in e-Nose technology, despite criticisms regarding their specificity [19]. The integration of MOS sensors with ML models addresses these limitations, expanding their use in a range of applications from indoor air quality monitoring to the detection of adulterated commercial products [20–22].

In food analysis, e-Noses have been employed to analyze EOs, particularly for identifying their plant origins or detecting adulteration [23–25]. 7 MOS sensors (TGS822, TGS842, SP15A, SP32, SP53, TGS2610, TGS2620) achieved up to 99% accuracy using the ARTMAP algorithm to assess the quality of *Rosa damascena* oil [26]. As another example, an e-Nose with 6 MOS sensors (TGS815, TGS821, TGS822, TGS224, TGS825, TGS42) was used to identify compounds with potential antimicrobial activity in the resins of *Aucoumea klaineana* and *Canarium schweinfurthii* [27]. Similarly, an e-Nose using 9 metal oxide semiconductor gas sensors (MQ9, MQ4, MQ135, MQ8, TGS2620, MQ135, TGS813, TGS822, MQ3) coupled to pattern recognition techniques has been used to identify volatile compounds from a variety of edible plants, achieving accuracy rates between 96.7% and 100% [28]. Moreover, the detection of adulteration in lavender oil and *Cistus ladanifer* EO demonstrated high classification accuracy, with artificial neural networks (ANN) reducing MOS sensor requirements while maintaining accuracy levels exceeding 99% [17, 29]. In another similar study the authors proposed an EO characterization system combining ML techniques with

electronic noses, comprising 7 MQ sensors (MQ-2, MQ-3, MQ-4, MQ-5, MQ-6, MQ-7 and MQ-8) [30]. The samples used belong to different plants, which include the species used in this paper. Comparing five different algorithms, they found that the K-nearest neighbors method yielded precision, recall, and F1 scores of 1, 0.99, and 0.99, respectively. When using only two parameters for averaged data or 15 parameters for individual data, a precision of 100% was reported [30]. Nonetheless, in this study, the authors aimed to identify the species that produces the essential oil, rather than quantify the concentration of specific chemical compounds.

Previous studies demonstrated that low-cost MOS sensors can effectively detect a variety of compounds, often eliminating the need for individual sample processing. However, these reports did not consider the influence of meteorological parameters on EO yield and composition. In contrast, studies that incorporate meteorological factors typically rely on high-cost techniques, such as GC–MS, which involve complex sample processing in a destructive manner. To date, no research has shown the capability of MOS sensors combined with ML to correlate the production of specific EO compounds with environmental parameters. Therefore, this study aims to investigate the potential of cost-effective gas sensor technology in addressing this gap.

Our previous work demonstrated the feasibility of using MQ gas sensors combined with ML for the classification of *Cistus ladanifer* EOs based on their chemical composition, particularly terpenic hydrocarbons [31]. While that study marked the first successful application of metal oxide semiconductor (MOS) sensors to classify complex EO matrices with promising accuracy (up to 91.6%), it was limited to static VOC data and did not account for external environmental influences. Although the obtained accuracy was adequate, the methodology followed in [31] does not quantify the amount of chemical compounds; it only classifies, as it was based on Classification Learner tools. For multiple purposes, the simple classification cannot be adequate, and a regression is needed to quantify the concentration of given chemical compounds accurately. In contrast, the present study introduces a novel smart sensor system that integrates dynamic gas fingerprint data from MQ sensors (previously used in [31]) with environmental parameters—such as temperature, humidity, precipitation, and solar radiation—to predict EO concentrations. By incorporating meteorological variables and evaluating their temporal resolution, this work significantly advances beyond single-point VOC detection, allowing for more context-aware and accuracy EO quality monitoring. Furthermore, while the previous work categorized compound concentrations into discrete classes, the current approach uses regression-based ML models to predict compound values, thereby enhancing interpretability and applicability. This integrated framework bridges the methodological gap between low-cost sensing and high-cost chromatographic techniques, offering a non-destructive, real-time solution for EO quality assessment that is both scientifically robust and operationally practical for precision agriculture.

In this paper, we propose a smart sensor system that leverages ML algorithms to correlate historical meteorological data from *Cistus ladanifer* harvest locations with the concentration of terpenic hydrocarbons in their EO samples. Considering that some VOCs might have the same effect on MOS sensors, the objective is to provide additional data that might explain the different VOCs present, as indicated in some recent studies [8, 12, 31]. To evaluate the suitability of combining climatic data and MQ gathered data, a database was generated by gathering information from 33 samples of *C. ladanifer* EO. These samples were divided into a training dataset (4835 samples) and a test dataset (319 samples). The samples were introduced in a sensing system composed of 7 MQ sensors. Climatic data for the 360 days for the corresponding growth period was downloaded and analyzed. The results were used to propose regression models to predict the concentration of terpenic hydrocarbons for 33 samples of EOs.

Materials and methods

This section provides a detailed description of the EO sample collection, analysis, environmental data acquisition, and the gas sensor setup and configuration. It also outlines the data collection, preprocessing, and ML approaches employed in this study to analyse the relationship between EO composition and environmental parameters.

Sample material

EOs were extracted from wild *C. ladanifer* L. plants harvested at two distinct locations in the Sierra Norte de Guadalajara region, Central Spain: Bustares (30N 495423, 4,552,862) and Hiendelaencina (30N 499691, 4,546,683). The harvesting locations were selected due to their contrasting microclimates and ecological conditions, which allowed for the exploration of potential environmental influences on EO composition. The plant sampling methodology and EO extraction procedures adhered to the standard protocols detailed in our previous study [31].

The chemical characterization of the EO samples was performed using GC-FID-MS, following the guidelines established by Mediavilla, et al. [32]. This technique enabled the quantification and identification of individual VOCs present in the EO samples. A total of 33 EO samples

were collected during the study period, spanning February 28, 2021, to September 26, 2022. The samples were analyzed to determine their chemical profiles, with alpha-pinene, viridiflorol, and other terpenic hydrocarbons identified as the most abundant compounds. The concentration ranges for these compounds were determined as follows: alpha-pinene (15.81–61.07%), viridiflorol (2.76–18.37%), and remaining terpenic hydrocarbons (22.48–71.49%). This data description easily allows to differentiate between the three types of chemical compounds (Fig. 1). Among the terpenes analyzed, alpha-pinene and viridiflorol have been separated due to their higher concentration, and the rest have been grouped as “terpenic hydrocarbons”.

In light of the dominance of terpenic hydrocarbons along with their ecological and industrial significance, further analysis were only carried out with terpenic hydrocarbon compounds. Subsequent analyses aimed to identify correlations between the concentrations of these compounds and environmental parameters, using data captured by MQ gas sensors.

Environmental data

Meteorological data corresponding to the harvesting sites were obtained from the Jadraque weather station, the closest station to Bustares and Hiendelaencina. This data was sourced through the SIAR (Agroclimatic Information System for Irrigation) platform and cross-validated against records from the AEMET (State Meteorological Agency) [33].

Environmental parameters were organized into three datasets: "Daily," "15 Days," and "All" (Table 1). The "Daily" dataset provided high-resolution data on parameters such as temperature, humidity, and radiation, while the "15 Days" dataset offered averaged values over two-week periods to capture longer-term trends. The "All" dataset aggregated features from both "Daily" and "15 Days" datasets,

Table 1 Description of the environmental datasets employed in the study

		Features	
ID	**Dataset Name**	**Quantity**	**Description**
1	Daily	8	Maximum, minimum and average Humidity (%), Maximum, minimum, and average Temperature (°C), Precipitation (mm), and Radiation (MJm^2)
2	15 Days	8	Maximum, minimum and average Humidity (%), Maximum, minimum, and average Temperature (°C), Precipitation (mm), and Radiation (MJm^2)
3	All	16	Combined Daily and 15 Days

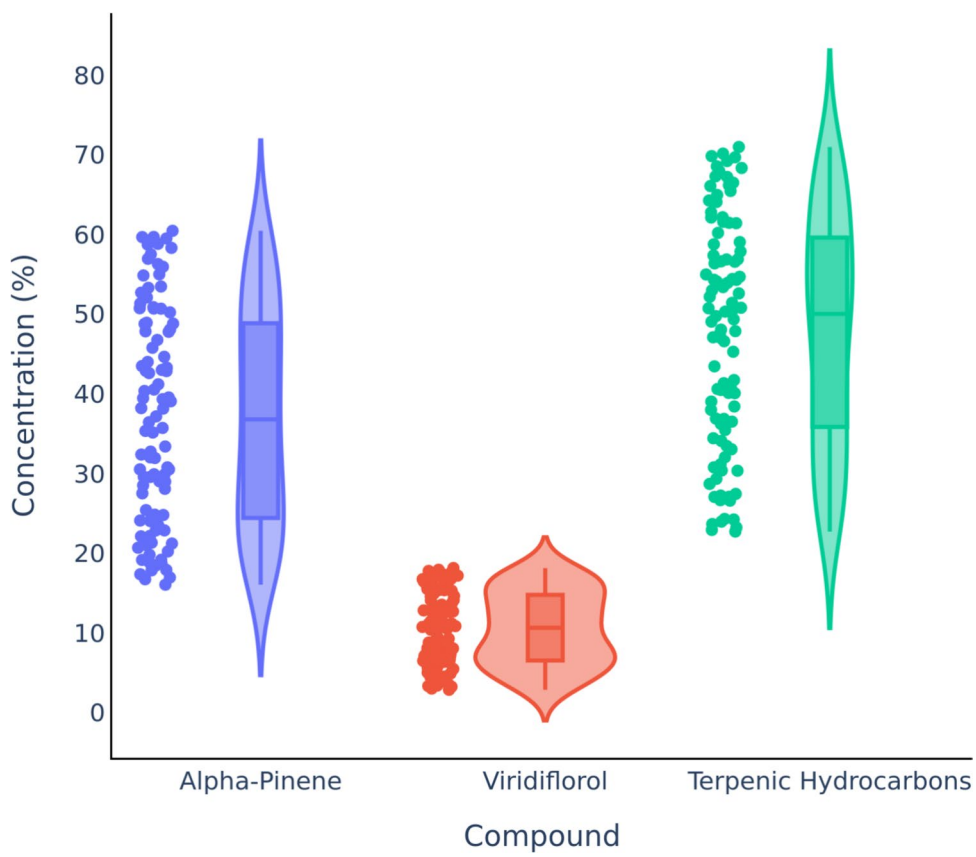

Fig. 1 Range of maximum and minimum concentrations for the organic compounds analyzed in *C. ladanifer* essential oils

enabling robust analysis by integrating detailed short-term and smoothed long-term environmental trends.

This multifaceted approach facilitated nuanced investigations into the relationships between environmental variables and terpenic hydrocarbon concentrations, while also addressing potential overfitting risks inherent in models with limited samples. The analysis was guided by the hypothesis that SnO_2 (MQ) sensor responses are proportional to the gaseous analyte concentrations, with sensor outputs reflecting the EO profiles.

Sensor description

The study employed MQ gas sensors to detect and quantify VOCs in EO samples. These sensors, based on MOS technology, were selected for their cost-effectiveness and versatility in detecting a wide range of chemical compounds. The working principle of MQ sensors involves changes in the resistance of the sensor material due to the adsorption and reaction of gaseous analytes on its surface. The MQ sensors used have an approximate price of €18 to €35, and the microcontroller costs approximately €37 to €53. In total, the device has an estimated cost of €55 to €88. Regarding energy consumption, the MQ sensors consume approximately 0.0063 kWh. The microcontroller, on the other hand, has an energy consumption of approximately 0.0096 kWh. In this study, a total of 33 biological samples were analyzed, yielding approximately 5,000 data points. The sensor and microcontroller system operated continuously for 24 h per biological sample. In total, the device's energy consumption in one day is 0.3816 kWh.

Each sensor includes a built-in heating element to maintain optimal operational conditions. The study utilized an array of MQ sensors tailored for detecting specific gases, including methane, LPG, carbon monoxide, and hydrogen. The sensor architecture comprised two circuits: a heating circuit and a resistance measurement circuit. The latter quantified resistance variations as a function of gas concentrations, enabling precise VOC detection. Detailed specifications of the sensors used are provided in Table 2.

Node configuration

The data acquisition system was architected around Arduino Mega 2560 microcontrollers, selected for their advantageous balance of performance, memory, and I/O expandability. Each Arduino Mega 2560 served as a dedicated sensor node controller due to its ATmega2560 microprocessor, which operates at 16 MHz and offers 54 digital I/O pins (15 of which can be used as PWM outputs), 16 analog inputs, and 4 UARTs for serial communication. These attributes rendered the platform particularly suitable for interfacing with multiple analog gas sensors and supporting modular expansion.

Each sensor node integrated seven gas sensors, including various MQ-series sensors (e.g., MQ-2, MQ-3, MQ-4, MQ-5, MQ-6, MQ-7, and MQ-8) (Fig. 2), which employ a SnO_2 sensing element, whose conductivity changes in the presence of target gases. MQ-series sensors are known for their high-power consumption, particularly due to their internal heating elements, which operate at 5.0V and draw significant current. The heater coils typically consume approximately 150mA to 200mA per sensor, leading to an aggregate current draw of up to 1.4A for all seven sensors in a single node configuration. This current requirement far exceeds the output capacity of the Arduino Mega 2560's onboard voltage regulator, which can typically supply only around 500mA total, including its own operational requirements.

To address this limitation, an external regulated 5 V DC power supply with a minimum current rating of 2 A was employed to power the MQ-series sensors independently of the Arduino's internal regulator. A dedicated power distribution module was used to split and regulate the voltage across all seven sensors, ensuring electrical stability and avoiding voltage drops that could affect sensor accuracy and longevity.

Table 2 Summary of compounds detected by each MQ sensor

Compound Detection	MQ2	MQ3	MQ4	MQ5	MQ6	MQ7	MQ8
Methane	X		X				
Butane	X				X		
Liquified Petroleum Gas (LPG)	X			X	X		
Smoke	X						
Alcohol		X					
Ethanol		X					
Compressed Natural Gas (CNG)		X					
Natural Gas			X				
Carbon Monoxide (CO)				X		X	
Hydrogen (H)							X

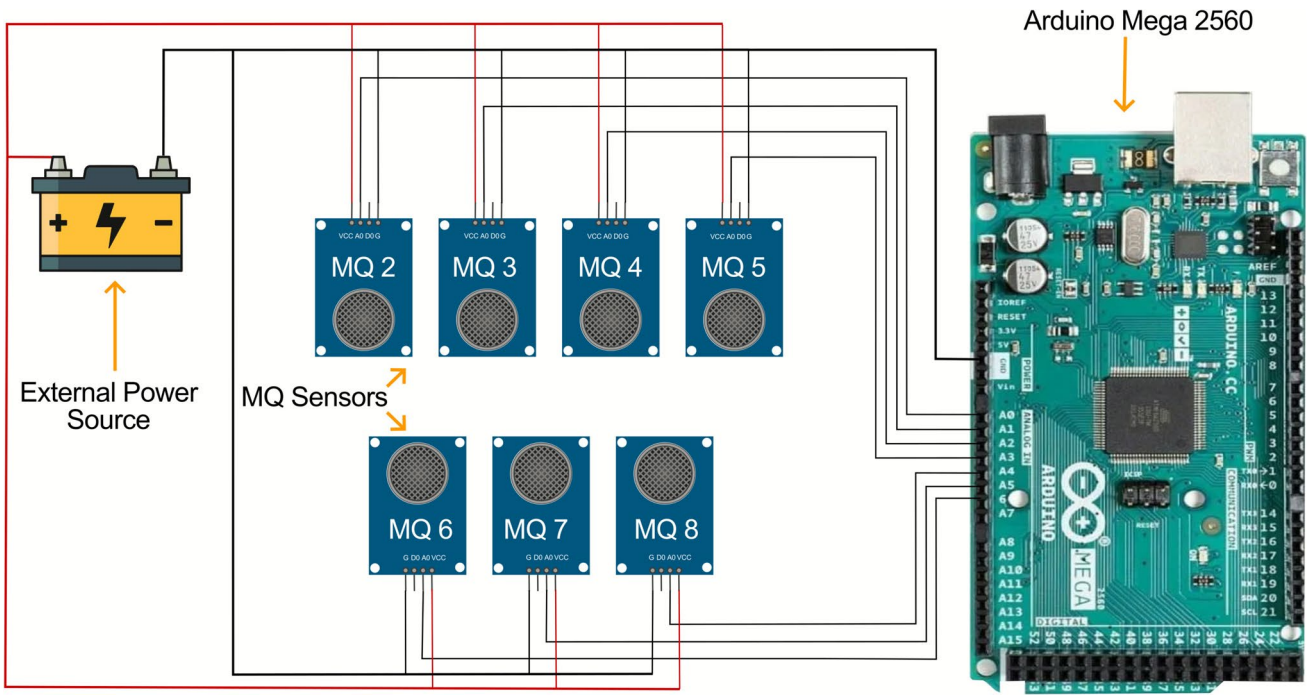

Fig. 2 Representation of node consisting of MQ sensors and Arduino Mega 2560

Each MQ sensor was connected as follows:

(i) VCC (5V): Connected to the external 5 V power supply via a distribution rail.
(ii) GND: Common ground was established between the external power supply and the Arduino to maintain a unified reference potential.
(iii) Analog Output (AO): Connected to one of the Arduino Mega 2560's analog input pins (A0–A6), enabling simultaneous sampling of all sensor outputs.

This setup ensured that the sensor heating elements were reliably powered, while the Arduino could safely handle the analog voltage signals without risk of thermal overload or unstable operation.

Communication architecture and data transmission

Each Arduino Mega 2560 node communicated serially with a Raspberry Pi 4 unit using UART (Universal Asynchronous Receiver/Transmitter) interface. The Raspberry Pi, equipped with a quad-core ARM Cortex-A72 processor and 4GB RAM, served as a local edge processing unit, responsible for temporal data aggregation, timestamping, and onward transmission.

Sensor readings were relayed from the Arduino to the Raspberry Pi in a structured JSON format at defined sampling intervals. The Raspberry Pi executed a Python-based acquisition script that parsed the incoming data stream and inserted the structured data into a MariaDB relational database hosted on a centralized local server. This local architecture was selected over cloud alternatives to reduce latency, increase resilience in the event of network failure, and preserve data privacy in field deployments.

The physical layout of each sensor node was modular, enclosed in weather-resistant casings with internal partitioning to prevent heat buildup from the MQ heaters influencing neighboring sensors. Each sensor node was powered through a centralized DC distribution board, with surge protection and current limiting circuitry. A schematic overview of the node architecture and wiring diagram is provided in Fig. 3, detailing the interconnections between the microcontroller, sensors, power supply, and the Raspberry Pi gateway.

Data acquisition system was built around Arduino Mega 2560 microcontrollers, chosen for their robust processing capabilities, ample memory, and extensive input/output pin availability. These microcontrollers were interfaced with Raspberry Pi 4 units, which served as data aggregation and transmission hubs. The configuration allowed seamless integration of MQ sensors with the data acquisition nodes.

The Raspberry Pi units transmitted sensor data to a MariaDB database hosted on a local server. This setup ensured efficient data storage and enabled real-time data retrieval through an access point. A schematic representation of the sensor and node configuration is shown in Fig. 3.

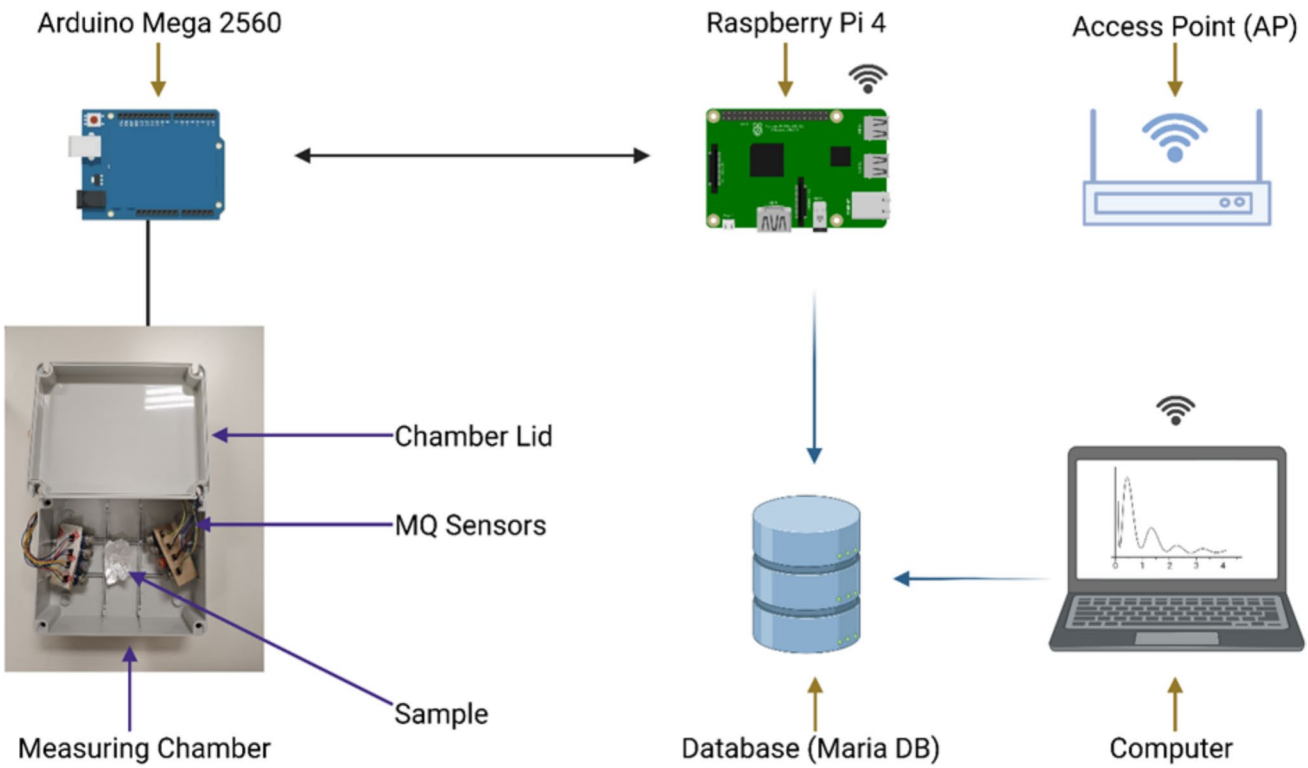

Fig. 3 Diagram of the sensing node, with the indication of each component of the system

Data collection and processing

Data collection protocol

Prior to data acquisition, a standardized calibration and drift-mitigation protocol was applied to all MQ sensors. Baseline correction was performed by recording the steady-state sensor output in clean air immediately before each measurement cycle [34]; these corrections for the baseline values were employed for all VOC readings to reduce drift and environmental noise. Sensor drift was further controlled through short preheating periods at the beginning of each sampling session and by discarding early unstable responses (<200 units). To ensure reproducibility over time, periodic calibration checks were conducted after every 10 measurement sessions. Together, these steps provided consistent sensor behavior, minimized long-term drift effects, and enhanced the reliability of VOC fingerprint classification.

EO samples were stored at 4 °C to preserve their integrity and were acclimatized to room temperature for 5–10 min before analysis. For each measurement, 0.05 mL of the EO sample was injected into a custom-designed closed chamber equipped with MQ sensors. The chamber was purged with ambient air between measurements to ensure atmospheric purity and sensor stability. Sensor signals were recorded at a sampling frequency of 1 Hz (one reading per second), resulting in a minimum of 1,000 time-resolved data points per sample over the 24-h monitoring period.

Data acquisition intervals of 2 h were implemented to capture comprehensive VOC profiles. Each sample was monitored over a 24-h period at ambient conditions (temperature: 25–34 °C; humidity: 22–38%). To enhance the robustness of the dataset, a minimum of 1,000 sensor readings per sample were recorded, capturing dynamic variations in VOC emissions. A stepwise illustration of data collection steps is presented in Fig. 4.

Data preprocessing and analysis

Initial data preprocessing involved discarding readings below 200 (to exclude unstable conditions) and above 600 (to cap outliers). Data were averaged every 10 readings to enhance homogeneity and mitigate sensor or electronic anomalies. Readings below 200 correspond to the well-documented initial

Fig. 4 Stepwise illustration of data acquisition through gas sensors process

instability phase of MQ sensors, during which the sensing element has not yet reached a steady resistive state. Values above 600 indicate signal saturation, where the sensor response is no longer proportional to VOC concentration and therefore does not reflect meaningful analytical information. These thresholds were applied uniformly across the dataset to ensure that only measurements within the stable operational range of the sensors were retained. Processed data were analyzed using MATLAB (R2022b) for regression tasks. ML models were developed to elucidate the relationships between VOC concentrations and environmental factors and sensor response.

Environmental datasets ("Daily," "15 Days," and "All") were first evaluated for completeness, and missing values were addressed using mean imputation [35] to maintain dataset integrity. During model development in MATLAB (R2022b), predictor normalization was performed automatically by enabling the *Standardize Predictors* option, which standardized all numeric variables to zero mean and unit variance. This ensured uniform scaling across predictors and supported more stable model convergence. Categorical temporal attributes (e.g., day-interval identifiers) were numerically encoded to enable their use in regression algorithms. Collectively, these preprocessing steps produced a consistent, well-scaled feature space that minimized the influence of variable magnitude differences and improved model interpretability.

A total of 5,154 data points were used, with 75% of the data allocated for training and validation, and 25% reserved for testing. Figure 5 demonstrates the detailed data processing procedures employed for this study. A thorough evaluation was performed utilizing 31 models, which were then compared to assessing their effectiveness. Initially, the models employed 16 input features, sourced from a dataset comprising readings from MQ sensors and temporal data. Subsequently, feature selection techniques, specifically the Maximum Relevance–Minimum Redundancy (MRMR) algorithms, were applied. These algorithms identified the most relevant variables by optimizing the values generated during the selection process. This analytical framework enabled the identification of key environmental predictors of EO composition and provided insights into the potential applications of MQ sensors in agri-food and ecological research.

Methodological rationale

This study employed a meticulously structured methodological framework to evaluate the temporal dynamics and sensor-retrieved VOCs influencing terpenic hydrocarbon production. By integrating advanced sensor technology, comprehensive data acquisition systems, and ML analytics, the approach offers a scalable and reproducible model for elucidating the interactions between environmental conditions and hydrocarbon biosynthesis.

Temporal data, encompassing diurnal and seasonal variations in temperature, humidity, precipitation, and light intensity, facilitates a dynamic analysis of environmental parameters. Cost-efficient MQ gas sensors deliver high-resolution VOCs data, ensuring continuous monitoring and enabling robust assessments of environmental influences. Despite their affordability, MQ sensors exhibit accuracy and reliability when paired with rigorous preprocessing and calibration, highlighting their potential for real-time monitoring in scientific contexts. Furthermore, correlation analyses, supported by statistical and ML-based feature selection, isolate key features impacting hydrocarbon production while mitigating the complexity of multivariate data. ML algorithms, including Random Forests, Gradient Boosting, and NN, are leveraged for predictive modeling of terpenic hydrocarbon yields. Model validation through tenfold cross-validation and independent test sets ensure reliability and generalizability.

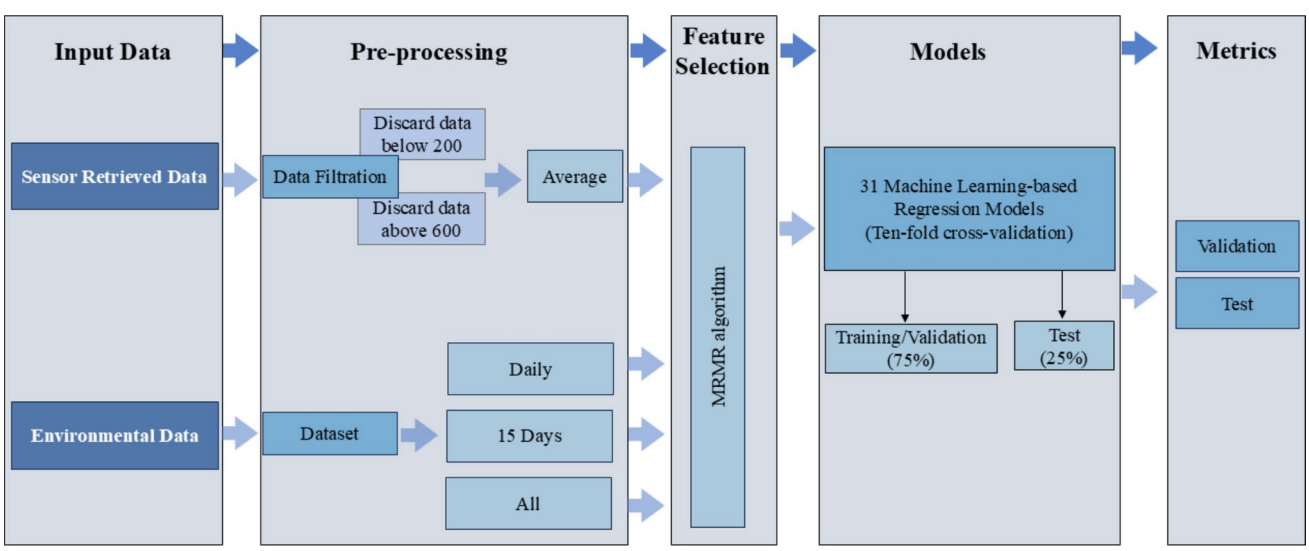

Fig. 5 Schematic illustration of the data processing procedure

The methodological approach is designed for reproducibility and scalability, utilizing open-source software to promote transparency and facilitate replication. This comprehensive framework bridges temporal, sensor-derived, and analytical data to optimize terpenic hydrocarbon production, offering significant applications in the agri-food sector and beyond.

Results

This section is organized in three subsections. First, the general results of all the tests that were conducted are summarized. Then, we present the details of the performance of the models generated for the assessment of terpenic hydrocarbons.

General results

In order to elucidate the relationship between environmental parameters and the production of terpenic hydrocarbons, we conducted a comprehensive analysis employing the three datasets and various feature sets. Our findings revealed that the highest accuracies in the test phase were attained using the daily dataset, the complete dataset, and the 15-days average dataset, with accuracies of 92.68%, 92.68%, and 90.10%, respectively. Corresponding training-validation accuracies were 94.75%, 94.78%, and 94.62%. Reducing the number of features to minimize resource usage did not yield any improvement in these accuracies (Table 3).

Table 3 Overview of the studied parameters and their highest recorded accuracy percentages for respective features selected. The (*) indicates the best test accuracy for the corresponding dataset

Studied Parameter	No. of features	R^2 Training-Validation accuracy (%)	R^2 Test accuracy (%)
Only Environmental Parameters with All Datasets			
Environmental Parameters	16 (all)	94.78	92.68
Environmental Parameters	15	94.78	92.68
Environmental Parameters	13	94.78	92.73*
Environmental Parameters	11	92.67	91.36
Environmental Parameters	8	92.68	92.48
Only Environmental Parameters with Daily Dataset			
Environmental Parameters	8 (all)	94.75	92.68*
Environmental Parameters	7	94.75	92.63
Environmental Parameters	6	94.75	92.57
Environmental Parameters	5	94.75	92.05
Environmental Parameters	4	92.43	91.82
Only Environmental Parameters with 15 Days Dataset			
Environmental Parameters	8 (all)	94.62	90.10
Environmental Parameters	7	94.62	90.10
Environmental Parameters	6	94.62	90.10
Environmental Parameters	5	94.62	90.10*
Environmental Parameters	4	93.39	89.96
All Parameters with All Datasets			
Data Gathered and Environmental Parameters	52 (all)	99.93	90.11
Data Gathered and Environmental Parameters	21	99.92	90.49
Data Gathered and Environmental Parameters	20	99.90*	90.56*
Data Gathered and Environmental Parameters	19	99.90	90.33
All Parameters with Daily Dataset			
Data Gathered and Environmental Parameters	44 (all)	96.59	85.16
Data Gathered and Environmental Parameters	30	99.92	86.68
Data Gathered and Environmental Parameters	15	99.84	85.80
Data Gathered and Environmental Parameters	13	99.77	83.58
Data Gathered and Environmental Parameters	12	99.80*	88.25*
Data Gathered and Environmental Parameters	11	96.37	86.64
All Parameters with 15 Days Dataset			
Data Gathered and Environmental Parameters	44 (all)	99.92	91.39
Data Gathered and Environmental Parameters	36	99.92	92.41
Data Gathered and Environmental Parameters	30	99.92	92.44
Data Gathered and Environmental Parameters	29	99.93*	93.15*
Data Gathered and Environmental Parameters	28	99.92	91.35
Data Gathered and Environmental Parameters	18	99.93	89.99

Table 4 Features contributing to the best performing machine learning models for each type of dataset

Number of Features	Feature Name	Assigned ID	ML Model
Only Environmental Parameters			
13 features for All Dataset	Minimum Temperature (°C) 15D	Min_Temp_15D	Exponential GPR (SLR)
	Maximum Humidity	Max_Hum	R^2 (Validation) = 94.78
	Average Humidity	Avg_Hum	R^2 (Test) = 92.73
	Average Temperature (°C) 15D	Avg_Temp_15D	
	Radiation (MJm^2) 15D	Radiation_15D	
	Precipitation (mm) 15D	Precipitation_15D	
	Average Temperature (°C)	Avg_Temp	
	Maximum Temperature (°C)	Max_Temp	
	Maximum Temperature (°C) 15D	Max_Temp_15D	
	Minimum Temperature (°C)	Min_Temp	
	Minimum Humidity 15D	Min_Temp_15D	
	Radiation (MJm^2)	Radiation	
	Average Humidity 15D	Avg_Hum_15D	
8 features for Daily Dataset	Maximum Humidity	Max_Hum	Interactions Linear (LR)
	Average Humidity	Avg_Hum	R^2 (Validation) = 94.75
	Average Temperature (°C)	Avg_Temp	R^2 (Test) = 92.68
	Maximum Temperature (°C)	Max_Temp	
	Minimum Temperature (°C)	Min_Temp	
	Radiation (MJm^2)	Radiation	
	Minimum Humidity	Min_Hum	
	Precipitation (mm)	Precipitation	
5 features for 15 Days Dataset	Minimum Temperature (°C) 15D	Min_Temp_15D	Coarse Tree (Tree)
	Average Temperature (°C) 15D	Avg_Temo_15D	R^2 (Validation) = 94.62
	Radiation (MJm^2) 15D	Radiation_15D	R^2 (Test) = 90.10
	Precipitation (mm) 15D	Precipitation_15D	
	Maximum Temperature (°C) 15D	Max_Temp_15D	
All Parameters with All Datasets			
20 features for All Dataset	MQ6_CH4	MQ6_CH4	Bagged Trees (Ensemble)
	MQ7_H2	MQ7_H2	R^2 (Validation) = 99.90
	MQ3_Hexane	MQ3_Hexane	R^2 (Test) = 90.56
	Maximum Humidity	Max_Hum	
	Maximum Temperature (°C) 15D	Max_Temp_15D	
	Average Temperature (°C)	Avg_Temp	
	MQ8_Alcohol	MQ8_Alcohol	
	Minimum Temperature (°C)	Min_Temp	
	Minimum Temperature (°C) 15D	Min_Temp_15D	
	Average Temperature (°C) 15D	Avg_Temp_15D	
	MQ6_H2	MQ6_H2	
	Average Humidity	Avg_Hum	
	Minimum Humidity 15D	Min_Hum_15D	
	Maximum Humidity 15D	Max_Hum_15D	
	Radiation (MJm^2) 15D	Radiation_15D	
	Precipitation (mm) 15D	Precipitation_15D	
	MQ2_LPG	MQ2_LPG	
	Average Temperature (°C)	Avg_Temp	
	Average Humidity 15D	Avg_Hum	
	Maximum Humidity	Max_Hum	
12 features for Daily Dataset	MQ6_CH4	MQ6_CH4	Bagged Trees (Ensemble)
	MQ7_H2	MQ7_H2	R^2 (Validation) = 99.80
	MQ3_Hexane	MQ3_Hexane	R^2 (Test) = 88.25
	Maximum Humidity	Max_Hum	
	Average Temperature (°C)	Avg_Temp	
	MQ8_Alcohol	MQ8_Alcohol	
	Minimum Temperature (°C)	Min_Temp	
	Average Humidity	Avg_Hum	
	Maximum Temperature (°C)	Max_Temp	
	Minimum Humidity	Min_Hum	
	MQ6_Alcohol	MQ6_Alcohol	
	Radiation (MJm^2)	Radiation	

Table 4 (continued)

Number of Features	Feature Name	Assigned ID	ML Model
29 features for 15 Days Dataset	MQ6_CH4	MQ6_CH4	Bagged Trees (Ensemble)
	MQ7_H2	MQ7_H2	R^2 (Validation) = 99.93
	MQ3_Hexane	MQ3_Hexane	R^2 (Test) = 93.15
	Maximum Temperature (°C) 15D	Max_Temp_15D	
	Average Temperature 15D	Avg_Temp_15D	
	Minimum Humidity 15D	Min_Hum_15D	
	MQ6_H2	MQ6_H2	
	Minimum Temperature (°C) 15D	Min_Temp_15D	
	Average Humidity 15D	Avg_Hum_15D	
	Radiation (MJm^2) 15D	Radiation_15D	
	Precipitation (mm) 15D	Precipitation_15D	
	MQ8_LPG	MQ8_LPG	
	Maximum Humidity 15D	Max_Hum_15D	
	MQ2_LPG	MQ2_LPG	
	MQ8_H2	MQ8_H2	
	MQ6_Alcohol	MQ6_Alcohol	
	MQ7_CO	MQ7_CO	
	MQ8_Alcohol	MQ8_Alcohol	
	MQ5_CO	MQ5_CO	
	MQ4_Alcohol	MQ4_Alcohol	
	MQ3_Alcohol	MQ3_Alcohol	
	MQ7_LPG	MQ7_LPG	
	MQ3_CO	MQ3_CO	
	MQ5_H2	MQ5_H2	
	MQ2_Alcohol	MQ2_Alcohol	
	MQ3_LPG	MQ3_LPG	
	MQ8_CH4	MQ8_CH4	
	MQ5_CH4	MQ5_CH4	
	MQ4_CO	MQ4_CO	

In contrast, an evaluation encompassing all 44 features—including sensor data, environmental parameters, and terpenic hydrocarbons from *C. ladanifer* EO—demonstrated an overall validation accuracy of 99.92% and a test accuracy of 91.39%. Interestingly, the accuracy metrics improved with fewer features selected to optimize resources. Specifically, utilizing only 29 features derived from a 15-day average dataset resulted in peak validation and test accuracies of 99.93% and 93.15%, respectively (Table 3).

Across all experimental configurations, the achieved predictive efficiencies reflect a coherent relationship between feature richness and model capability (Table 4). When using only environmental parameters, the Exponential GPR model demonstrated the strongest performance for the full dataset (R^2_val = 94.78; R^2_test = 92.73), indicating that temperature- and humidity-derived metrics—particularly multi-scale descriptors such as 15-day aggregates—encode sufficient environmental variability to delineate VOC fingerprint patterns. For the daily-scale subset, the Interactions Linear Regression model achieved comparably high accuracy (R^2_val = 94.75; R^2_test = 92.68), underscoring that linear interactions among core meteorological variables (temperature, humidity, radiation, precipitation) are already highly informative at finer temporal resolution. The 15-day subset, characterized by a reduced feature set dominated by aggregated thermal and radiation signals, yielded slightly lower performance with a Coarse Tree (R^2_val = 94.62; R^2_test = 90.10), consistent with the loss of short-term environmental variability.

Integrating sensor-derived VOC features with environmental parameters markedly enhanced model discriminative power. For the feature space (20 variables), Bagged Trees achieved near-perfect validation performance (R^2_val = 99.90) and strong generalization (R^2_test = 90.56), reflecting the complementary value of gas-specific responses (e.g., MQ6_CH_4, MQ7_H_2, MQ3_Hexane) combined with meteorological descriptors. The daily-scale fusion dataset achieved similarly high validation accuracy (R^2_val = 99.80) but a moderate decline in test performance (R^2_test = 88.25), suggesting greater day-to-day variability in sensor responses. Notably, the 15-day fusion dataset, enriched with 29 features capturing both long-term environmental signals and a wide spectrum of gas-sensitive channels, yielded the most robust generalization (R^2_val = 99.93; R^2_test = 93.15). This performance confirms that ensemble-based learners particularly benefit from high-dimensional feature spaces where both environmental trends and multi-gas fingerprints synergistically contribute to stable VOC classification. A comprehensive ranking of features for environmental and sensor data based on MRMR scores is provided in Figure S1-S2.

The error metrics of the best-performing minimal-feature models highlight clear differences in predictive efficiency across datasets and feature configurations (Table 5). When

Table 5 Performance metrics (RMSE, MSE, MAE) for the best-performing machine learning models across different datasets and feature configurations

Model Type	Preset	Dataset	No. of Features	Validation			Test		
				RMSE	**MSE**	**MAE**	**RMSE**	**MSE**	**MAE**
Only Environmental Parameters									
Gaussian Process Regression	Exponential GPR	All	16	3.09	9.55	1.49	3.39	11.50	2.83
Linear Regression	Interactions Linear	Daily	8	3.10	9.60	1.52	3.19	10.18	2.43
Tree	Coarse Tree	15-Days	5	3.14	9.84	1.58	3.71	13.77	3.06
Data Gathered and Environmental Parameters									
Ensemble	Bagged Trees	All	20	0.43	0.19	0.11	3.62	13.13	3.08
Ensemble	Bagged Trees	Daily	12	0.61	0.37	0.16	4.04	16.35	3.17
Ensemble	Bagged Trees	15-Days	29	0.36	0.13	0.07	3.09	9.53	2.57

using only environmental variables, Exponential Gaussian Process Regression (All dataset), Interactions Linear Regression (Daily dataset), and Coarse Tree (15-Days dataset) yielded comparable validation errors (RMSE = 3.09–3.14; MAE = 1.49–1.58), indicating moderate predictive precision under constrained feature sets. However, their test errors increased substantially (RMSE = 3.19–3.71; MAE = 2.43–3.06), revealing residual variability and limited generalization capacity when exposed to unseen conditions. By contrast, Bagged Trees models integrating environmental and gas-sensor predictors achieved markedly lower validation errors across all temporal resolutions, with the strongest performance observed for the 15-Days dataset (RMSE = 0.36; MAE = 0.07), reflecting excellent fit and minimal residual dispersion. Although test errors were higher than validation metrics, ensemble models maintained stable performance (RMSE = 3.09–4.04; MAE = 2.57–3.17), confirming robust generalization without error inflation. Notably, the 15-Days temporal aggregation consistently produced the lowest test errors among the ensemble configurations, demonstrating that extended observation windows enhance predictive stability and support more reliable VOC forecasting. Collectively, these results confirm that multi-source feature integration and ensemble learning substantially improve prediction accuracy over models restricted to environmental inputs and that temporal aggregation further strengthens model robustness.

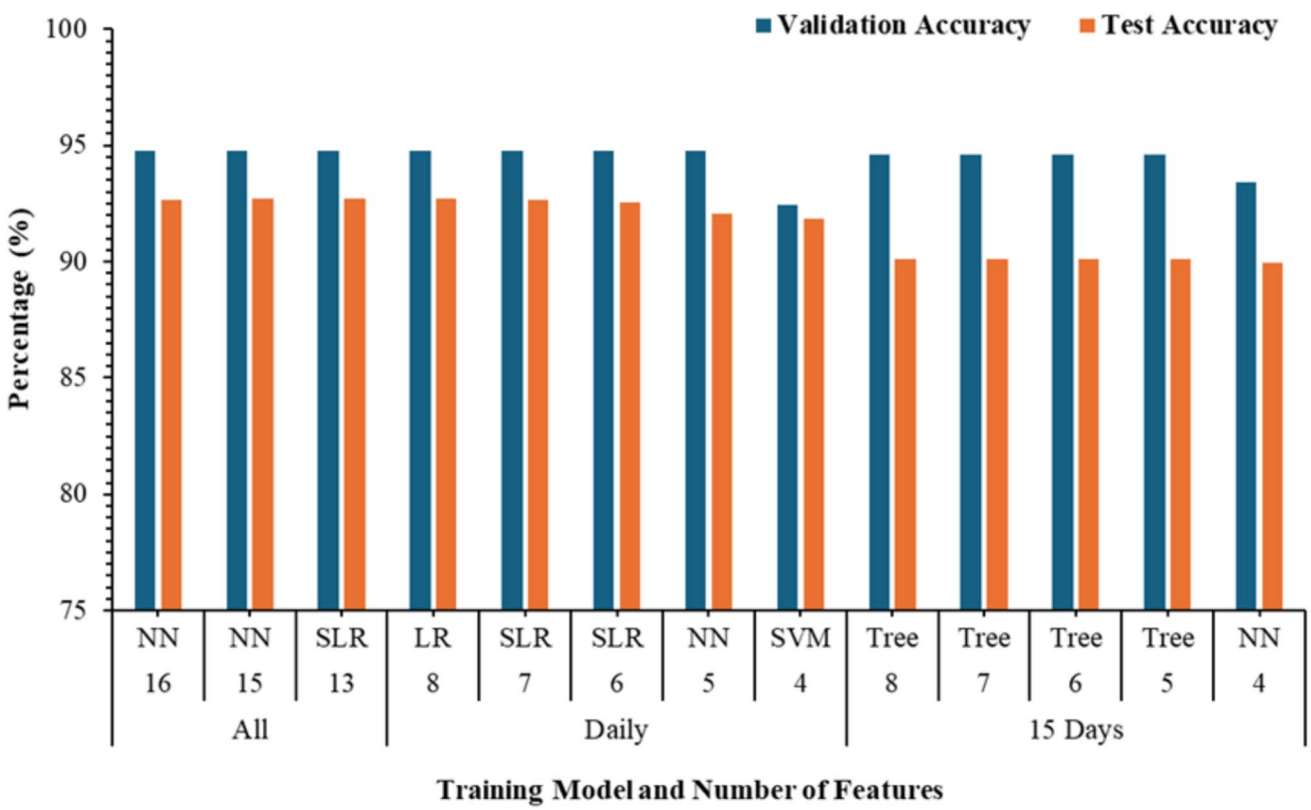

Fig. 6 Summary of the best performing machine learning models when different number of features were employed. All, Daily, and 15 Days represent three kinds of datasets. NN: Neural Network; SLR: Simple Linear Regression; LR: Linear Regression; SVM: Support Vector Machine

Environmental parameters

Best performing models

The environmental parameters were thoroughly analyzed to ascertain their influence on the concentration of terpenic hydrocarbon, as illustrated in Fig. 6. In this context, the application of all 16 features within a NN model resulted in a significant validation accuracy of 94.78%. Correspondingly, the test accuracy of this trained model exhibited a comparable value of 92.68%. Similar results were found when 15 features were employed. On selection of fewer features, the accuracy for test metrics declined except for 13 features where it was noted to be 92.73% -the highest achieved, with the corresponding validation of 94.78% for Stepwise Linear Regression model.

Conversely, when employing a Tree model for datasets spanning 15 days, an overall highest accuracy of 90.10%, along with a validation of 94.62%, was observed. Particularly, this accuracy remained consistent despite the reduction in the number of features until only 5 features were left, at which point the test accuracy began to drop below the 90% threshold. The models that demonstrated a subsequently higher accuracy for this dataset were Coarse Tree and Narrow Neural Network.

Focusing on the "daily" datasets, the Linear Regression (LR) model demonstrated a marginally higher test accuracy of 92.68%, achieved by utilizing all 8 available features. However, a systematic decline in accuracy was noted when fewer features were included in the analysis. The other models that demonstrated a subsequently higher accuracy were Stepwise Linear Regression, Stepwise Linear Regression, Neural Network, and SVM respectively with overall test accuracies of over 90%.

Prediction analysis

Based on the identification of best performing ML models, we further selected the two models with the highest test accuracy from each of the three datasets for an in-depth assessment of their predictive precision. The performance of these models is illustrated in Figs. 7, 8 and 9, which depicts the relationship between true and predicted responses.

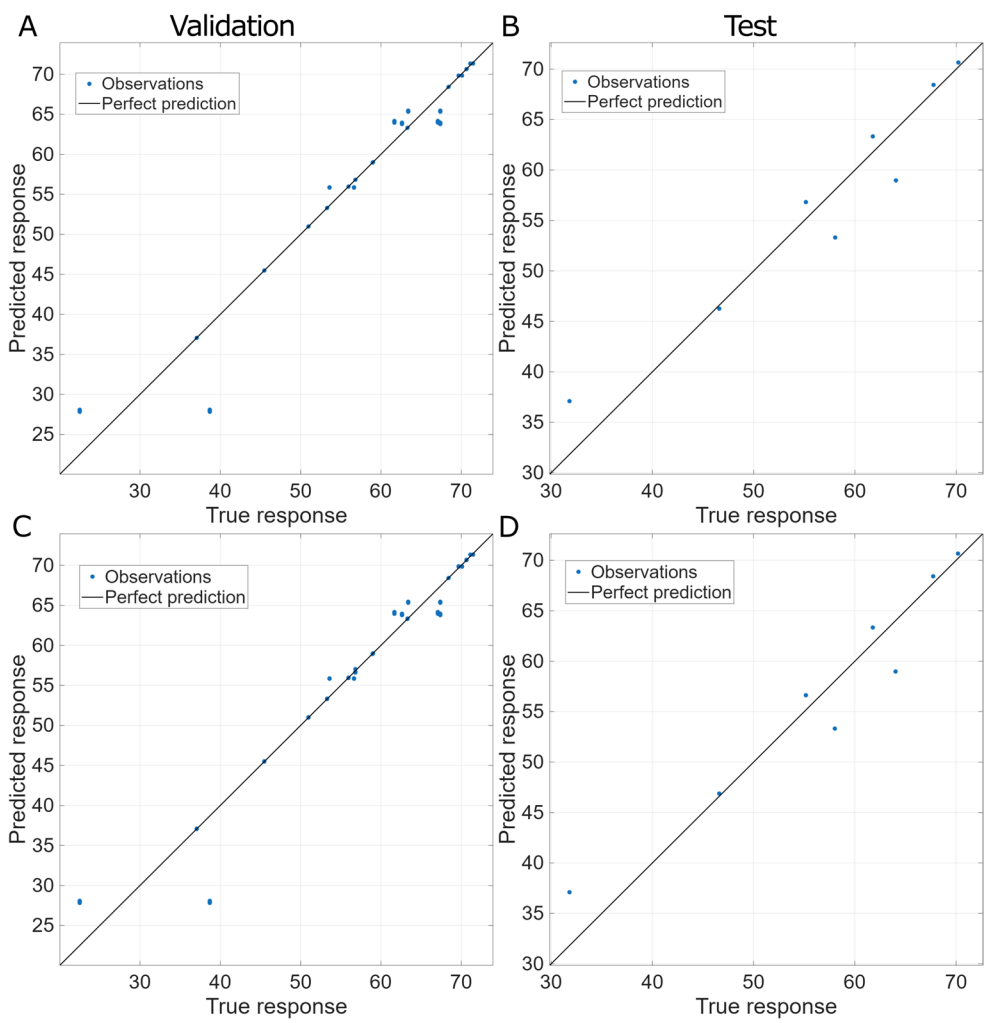

Fig. 7 Prediction analysis for "All" dataset for the production estimation of terpenic hydrocarbons based on environmental parameters. **A**) Validation for the Regression model using 16 features. **B**) Test for the Regression model using 16 features. **C**) Validation for the Regression model using 13 features. D) Test for the Regression model using 13 features

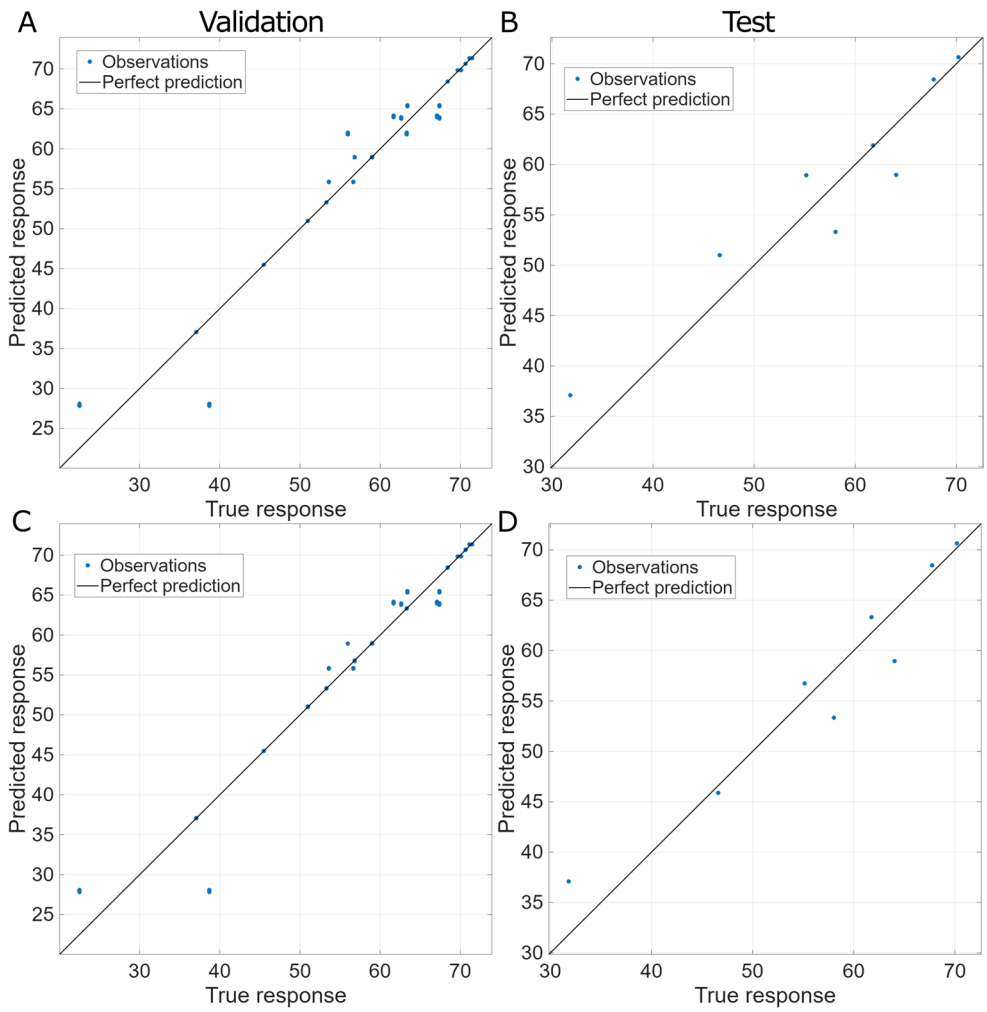

Fig. 8 Prediction analysis for Daily dataset for the production estimation of terpenic hydrocarbons based on environmental parameters. **A**) Validation for the Regression model using 8 features. **B**) Test for the Regression model using 8 features. **C**) Validation for the Regression model using 4 features. D) Test for the Regression model using 4 features

Precisely, when both the "Daily" and "15-day" datasets were combined, the models exhibited significantly improved predictive performance, as depicted in Fig. 7. This combined approach yielded superior predictive accuracy; a fact corroborated by the markedly reduced number of outliers in the predicted values. This improvement suggests a more accurate and reliable model response, likely due to the increased volume and diversity of data that enhances the learning process. The robustness of this combined dataset approach is further demonstrated by the enhanced generalizability and stability of the models' predictions.

In the case of the daily dataset, despite achieving noteworthy validation results, the test results revealed a considerable deviation from the expected range, as shown in Fig. 8. A large fraction of the data points fell outside the anticipated prediction intervals during testing, indicating substantial outlier issues. This discrepancy points to potential overfitting during the training phase, where the model performs well on the validation set but fails to generalize to unseen test data. The overall predictive accuracy of models trained on the daily dataset alone was thus inferior to those trained on the combined dataset. This outcome underscores the limitations of using a narrower dataset, which may lack the temporal diversity needed to capture more complex patterns and variations.

For the "15-day" dataset, both validation and test results exhibited lower accuracy relative to the other datasets, as illustrated in Fig. 9. A significant proportion of the predicted values deviated from the expected regression slope, suggesting that the models were less effective with this dataset. The reduced predictive performance could be attributed to the limited temporal resolution, which might not provide sufficient data variability for the models to learn effectively. The diminished accuracy and increased outliers in the predictions highlight the challenges associated with using less granular data.

SHAP analysis of environmental predictors

SHAP (SHapley Additive exPlanations) analysis was used to quantify the relative importance and directional effects of environmental predictors on VOC concentration estimates

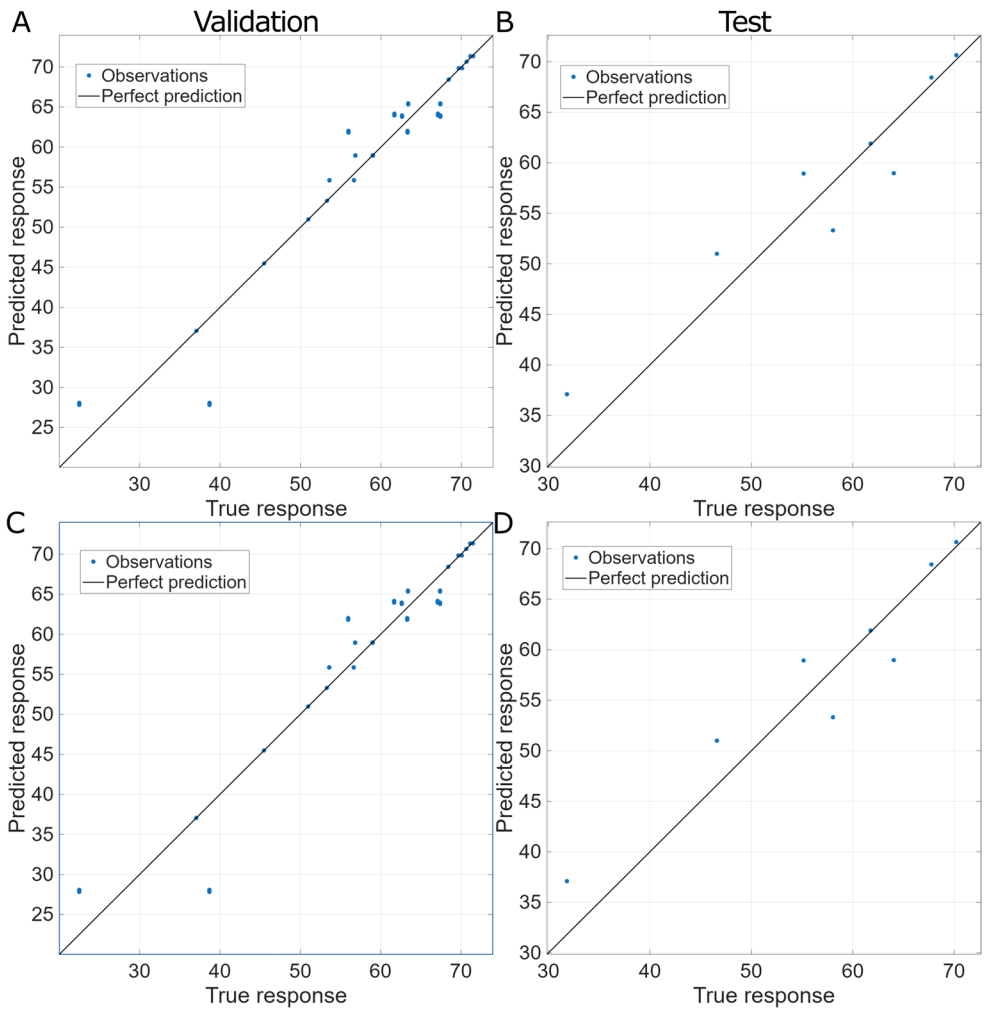

Fig. 9 Prediction analysis for 15 Days dataset for the production estimation of terpenic hydrocarbons based on environmental parameters. **A**) Validation for the Regression model using 8 features. **B**) Test for the Regression model using 8 features. **C**) Validation for the Regression model using 5 features. **D**) Test for the Regression model using 5 features

for the best-performing models associated with the All, Daily, and 15-Day datasets (Fig. 10). This approach provided an interpretable framework for understanding how individual meteorological parameters shaped VOC fingerprints from *C. ladanifer* EO.

In the All dataset (Stepwise Linear Regression; R^2 = 94.78% validation, 92.73% test), temperature-related variables exerted the strongest effects. The 15-day maximum and average temperatures, together with the daily maximum temperature, produced the largest positive SHAP contributions. Solar radiation and its 15-day aggregate also ranked highly, indicating that sustained thermal and light exposure were possibly the principal drivers of VOC fingerprints. Humidity effects were moderate and variable: low humidity tended to lower VOC signals, whereas higher humidity supported stabile contribution to model estimations. The 15-day precipitation variable showed only minor influence, consistent with its indirect linkage through soil moisture (Fig. 10A).

For the Daily dataset (Interactions Linear Regression; R^2 = 94.75% validation, 92.68% test), temperature again dominated the predictor hierarchy. Higher daily average, minimum, and maximum temperatures consistently were found greatly contributing to VOC estimates. Humidity exhibited bidirectional effects, with low values resulting in poor VOC predictions and higher values facilitating improved contribution in model prediction. Solar radiation continued to be an important positive driver, while daily precipitation contributed minimally (Fig. 10B).

In the 15-Day dataset (Coarse Tree; R^2 = 94.62% validation, 90.10% test), the 15-day maximum temperature produced the strongest positive SHAP values. This reflected the importance of sustained warm periods in enhancing volatilization and biosynthetic activity. The solar radiation also contributed strongly to the VOC fingerprints, possibly with its biological relevance in photosynthesis and secondary metabolism. Precipitation exhibited mixed effects, reflecting its dual influence with low contribution for higher humidity conditions. Multi-day minimum and average temperatures provided moderate positive contributions, likely supporting basic physiological functioning in terms of biological significance (Fig. 10C).

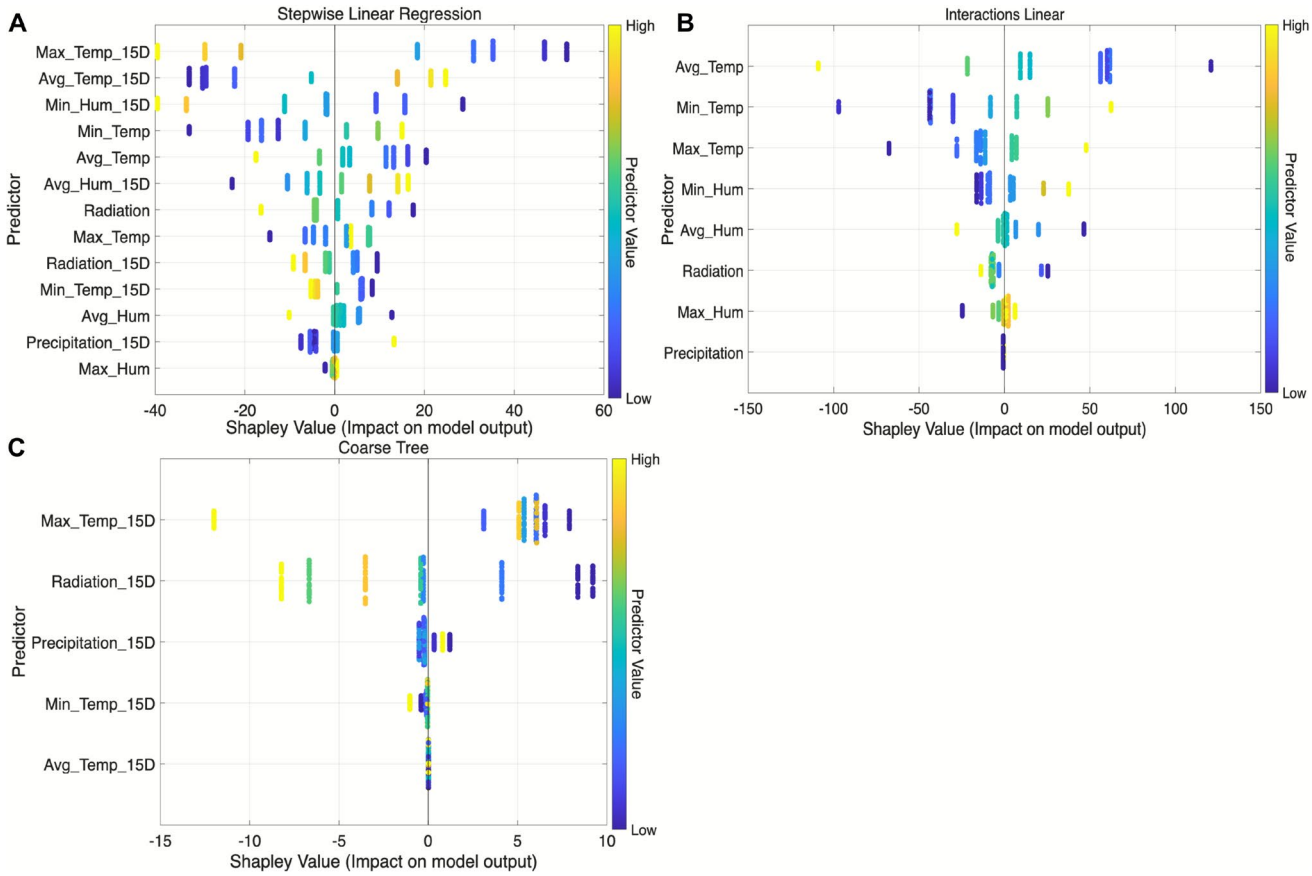

Fig. 10 SHAP analysis of environmental predictors for the best-performing machine learning models across the All (A), Daily (B), and 15-Day (C) datasets

Residual analyses for these models (Figures S3–S5) confirmed their robustness and lack of overfitting. Across validation and independent test sets, residuals clustered near zero, with minor localized deviations at extreme VOC concentrations. Test residuals generally exhibited narrower distributions than validation, indicating strong generalization to unseen data. No systematic trends were observed across the x-axis, supporting the assumption of homoscedasticity and the reliability of model predictions.

Collectively, the SHAP analyses showed that temperature and solar radiation were the primary drivers of VOC fingerprints, as used by ML modelling. Humidity acted as a moderating factor, while precipitation introduced only secondary variability. The combination of SHAP analysis and residual validation confirms that the models captured biologically relevant VOC–environment relationships with high accuracy and generalization.

sensors-based and environmental parameters

The data acquired through MQ sensors from *C. ladanifer* EO was evaluated along with the environmental parameters in order to achieve higher accuracy through the training of ML models. Once the environmental parameters were studied, they were related to parameters from sensors placed in the environment along with the values of terpene hydrocarbon concentrations.

Best performing ML models

The combination of all features, environmental, extracted from the sensors and terpene hydrocarbons, were analyzed and plotted in Fig. 11. As previously depicted, the two best models with the highest test accuracy were selected for each of the 3 datasets. The performance of each of these models is represented in Fig. 11, where the interaction between the true and predicted values is observed.

In the first case, all sensor data, all terpene hydrocarbon concentration values and all environmental parameter data were selected. When looking at the accuracy values of the models, it was observed that with all the features introduced (52 features in total); high Test accuracy values were detected, reaching up to 90.11%. To determine whether reducing the number of features would result in higher Test accuracy values, several modifications were carried out. After testing all possible combinations of features,

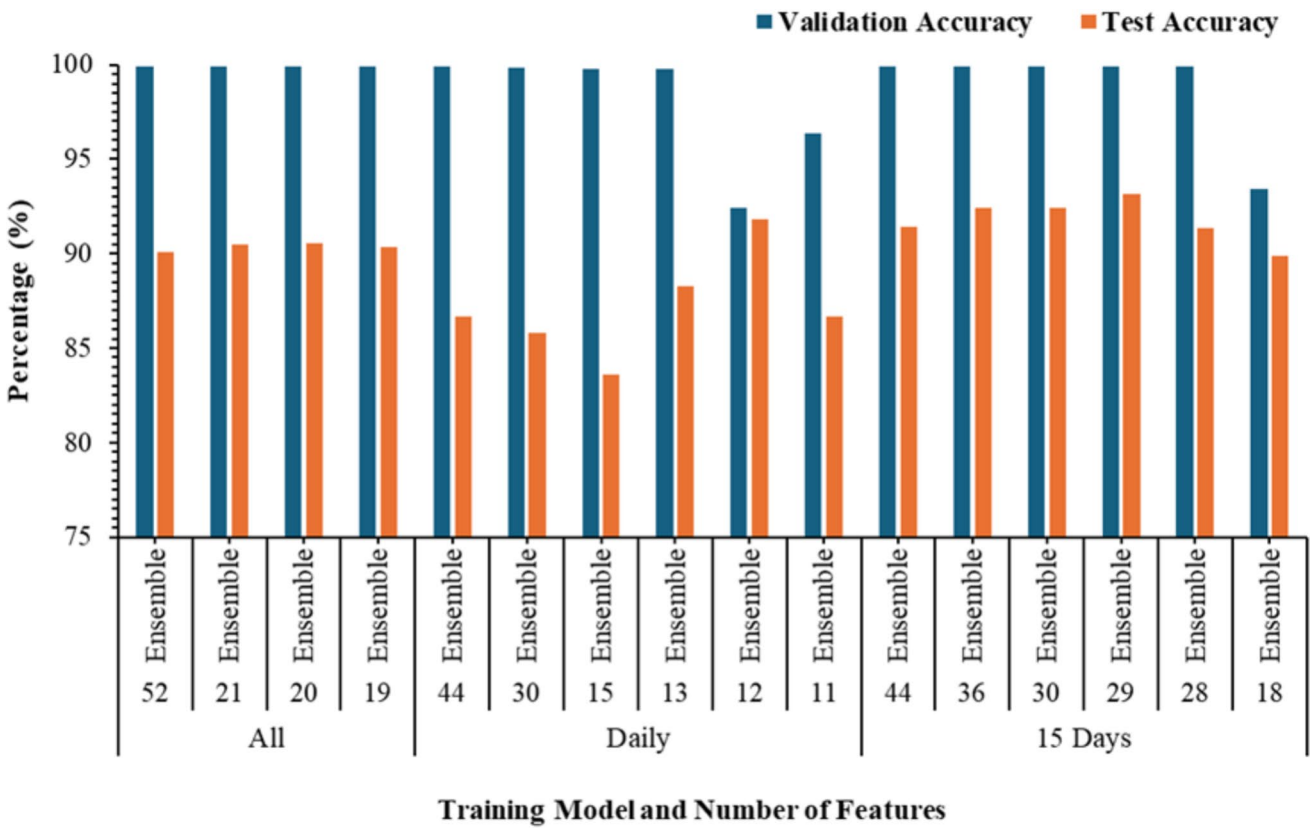

Fig. 11 Summary of best performing machine learning models when different number of features were used. "All", "Daily", and "15 Days" represent three kinds of datasets

it was determined that the highest Test accuracy value was achieved with 20 features, reaching up to 90.56%.

In the "Daily" data case, all sensor data, all terpenic hydrocarbon concentration values and only the daily environmental parameter data, were selected. When looking at the Test accuracy values of the models, it was observed that with all the features introduced (44 features in total), lower accuracy values were observed than in the previous case, where all the values were added. In this case, all the selected features achieved a Test accuracy of 85.16%. Again, to determine whether reducing the number of features would result in higher accuracy values, several modifications were carried out. After testing all possible combinations of features, the highest accuracy value was achieved with 12 features, reaching up to a maximum of 88.25% in Test accuracy values.

Finally, in the case of the "15-Days" mean data, all sensor data, all terpene hydrocarbon concentration values and only the "15-Days" mean data for environmental parameters were selected. When looking at the accuracy values of the models, it was observed that with all the features introduced (44 features in total); higher Test accuracy values were observed than in the two previous cases with a Test accuracy of 91.39%. Several modifications were made to assess whether reducing the number of features would improve accuracy. After testing all possible combinations of features, the highest accuracy value was achieved with 29 features, reaching 93.15% Test accuracy. Bagged Trees achieved the best Test accuracy when the entire dataset and the "15 Days" averages were used, both with all features and the selected features. Nevertheless, in the case of the "Daily" data, with all possible features, a Preset of Boosted Trees was obtained, while for the case of the features that obtained the best Test accuracy, was Bagged Trees.

Prediction analysis

After identifying the best-performing ML models, we selected the two models with the highest test accuracy from each of the three datasets for a detailed evaluation of their predictive accuracy. Figures 12 and 13 illustrate the performance of these models, in terms of validation and test, showing the relationship between true and predicted responses.

In the first model with all the 52 features included (Fig. 12A, B) the points distributed in the scatter plot are very similar. But, when only 20 features are selected (Fig. 12C), the dots appear closer to the prediction line

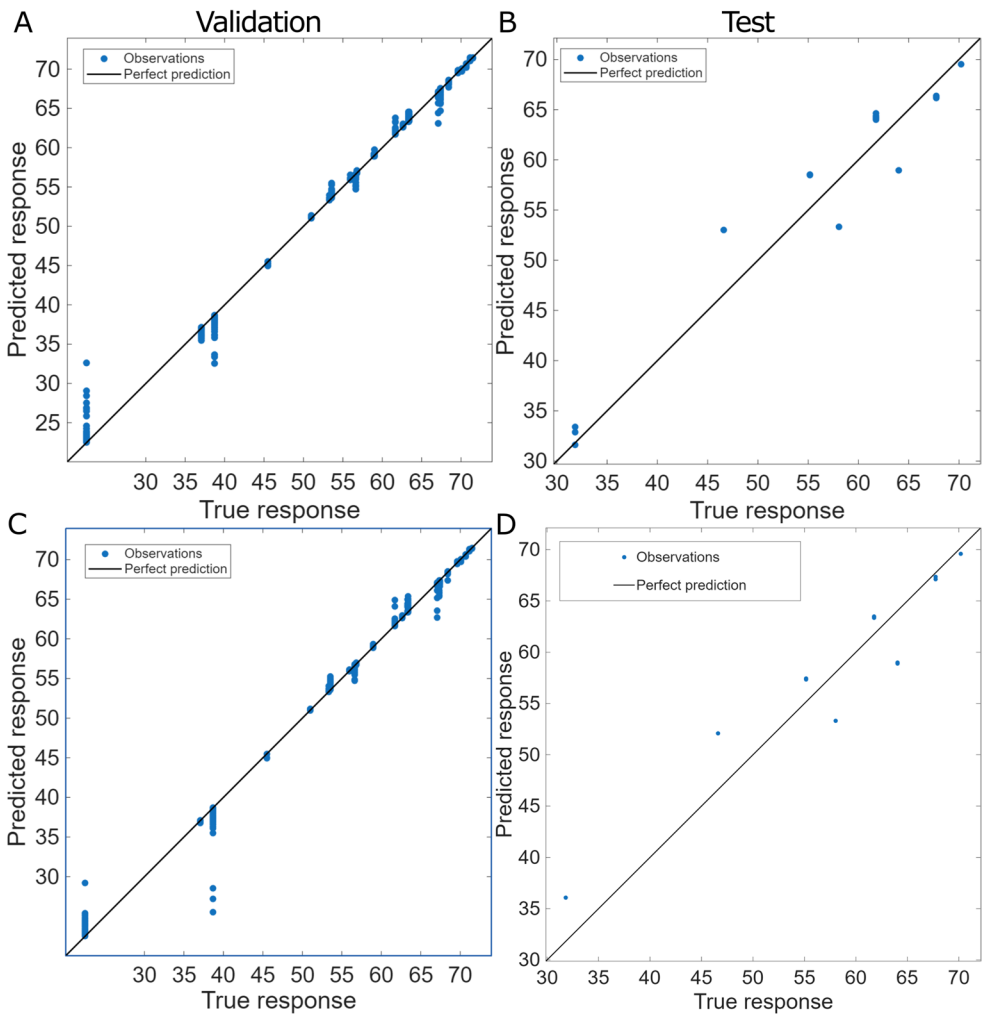

Fig. 12 Prediction analysis for "All" dataset for the production estimation of terpenic hydrocarbons based on environmental parameters and sensor retrieved data. **A**) Validation for the Regression model using 52 features. **B**) Test for the Regression model using 52 features. **C**) Validation for the Regression model using 20 features. **D**) Test for the Regression model using 20 features

(Fig. 12D). In this case, the coefficient of determination (R^2) values for the validation and test datasets, when utilizing all 52 features, were 0.99 and 0.90, respectively. Conversely, when the feature set was reduced to 20 features, similar R^2 values for the validation and test datasets were observed i.e., 0.99 and 0.90 respectively. This indicates that the model is more accurate with fewer number of features. However, although this accuracy value is high, it is still far from the graphical representation. This indicates that, although it is a valid model, it is not able to accurately indicate the expected results of the data collection.

In the "Daily" data case shown in Fig. 13, the points represented with all the features tend to deviate quite a bit from the perfect prediction line (Fig. 13A, B), unlike when only 12 features are selected (Fig. 13C, D). Even so, it is not possible to achieve that all points are completely close to the accuracy line. In this case, the precision values are lower as compared to regression analysis for "All" dataset. Precisely, in this case, the R^2 values for the validation and test datasets, when utilizing all 44 features, were 0.96 and 0.85, respectively. Conversely, when the feature set was reduced to 12 features, the R^2 values for the validation and test datasets were 0.99 and 0.88, respectively. However, it should be noted that the difference between the models is easily visible, since in the first one, the points clearly deviate from the perfect prediction slope, showing the difference between the two accuracy percentages.

Finally, in the case of the "15-day" mean data, depicted in Fig. 14, the points represented with all the features deviate slightly from the perfect prediction line (Fig. 14A, 14) as well as when the 29 features are selected (Fig. 14C, B). However, in the latter case, the values are much closer to the accuracy line. In this case, the R^2 values for the regression analysis for validation and test datasets, when utilizing all 44 features, were 0.99 and 0.91, respectively. Conversely, when the feature set was reduced to 29 features, the R^2 values for the validation and test datasets were 0.99 and 0.93, respectively. Therefore, the points close to the slope are very close to it, representing the difference in percentage of accuracy between this and the other models depicted in Figs. 11 and 12.

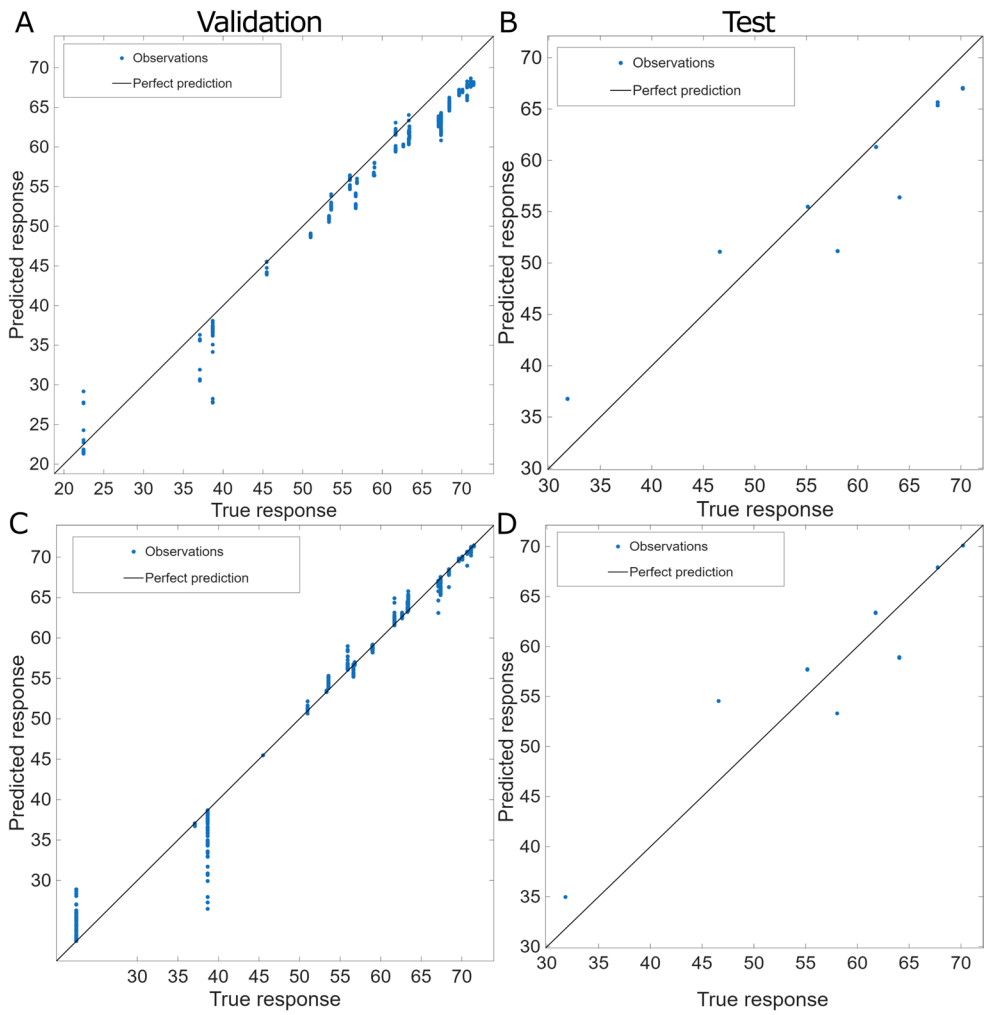

Fig. 13 Prediction analysis for "Daily" dataset for the production estimation of terpenic hydrocarbons based on environmental parameters and sensor retrieved data. **A**) Validation for the Regression model using 44 features. **B**) Test for the Regression model using 44 features. **C**) Validation for the Regression model using 12 features. **D**) Test for the Regression model using 12 features

SHAP analysis of environmental and sensor-based predictors

The SHAP analysis for the combined environmental and sensor-based dataset provides a detailed assessment of how meteorological and gas sensor variables jointly influence VOC predictions (Fig. 15). Across all temporal resolutions, environmental variables remain the dominant contributors, with sensor-derived VOC features offering complementary but secondary effects.

In the All dataset (Bagged Trees; $R^2 = 99.90\%$ validation, 90.56% test), multi-day warm periods and accumulated solar radiation produced the highest SHAP contributions. These sustained thermal and light conditions likely drove the core biosynthetic and volatilization processes of *Cistus ladanifer*. Short-term temperature indicators also contributed positively, consistent with rapid enzymatic responses. Humidity exerted moderate effects, generally higher VOC fingerprints related to higher moisture levels. Gas sensor features (MQ2, MQ3, MQ6, MQ7) ranked lower but added stability. However, MQ8contributed substantially in comparison to other sensor features. (Fig. 15A).

For the Daily dataset (Bagged Trees; $R^2 = 99.80\%$ validation, 88.25% test), daily radiation and daily maximum temperature emerged as the dominant predictors. Daily mean and minimum temperatures contributed additional positive effects to ML model prediction, reflecting their possible roles in maintaining metabolic activity. Daily humidity showed moderate influence, while gas sensor signals again provided small but consistent positive contributions that refined short-term predictions (Fig. 15B).

In the 15-Day dataset (Bagged Trees; $R^2 = 99.93\%$ validation, 93.15% test), maximum temperature and radiation predictors again dominated the SHAP ranking (Fig. 15C). These multi-day predictors captured sustained environmental influence on VOC fingerprints. Gas sensor predictors, especially MQ8, substantially contributed towards model prediction for this dataset. The humidity exerted moderate positive effects to model prediction, whereas gas sensor features offered steady but secondary enhancements through VOC fingerprinting.

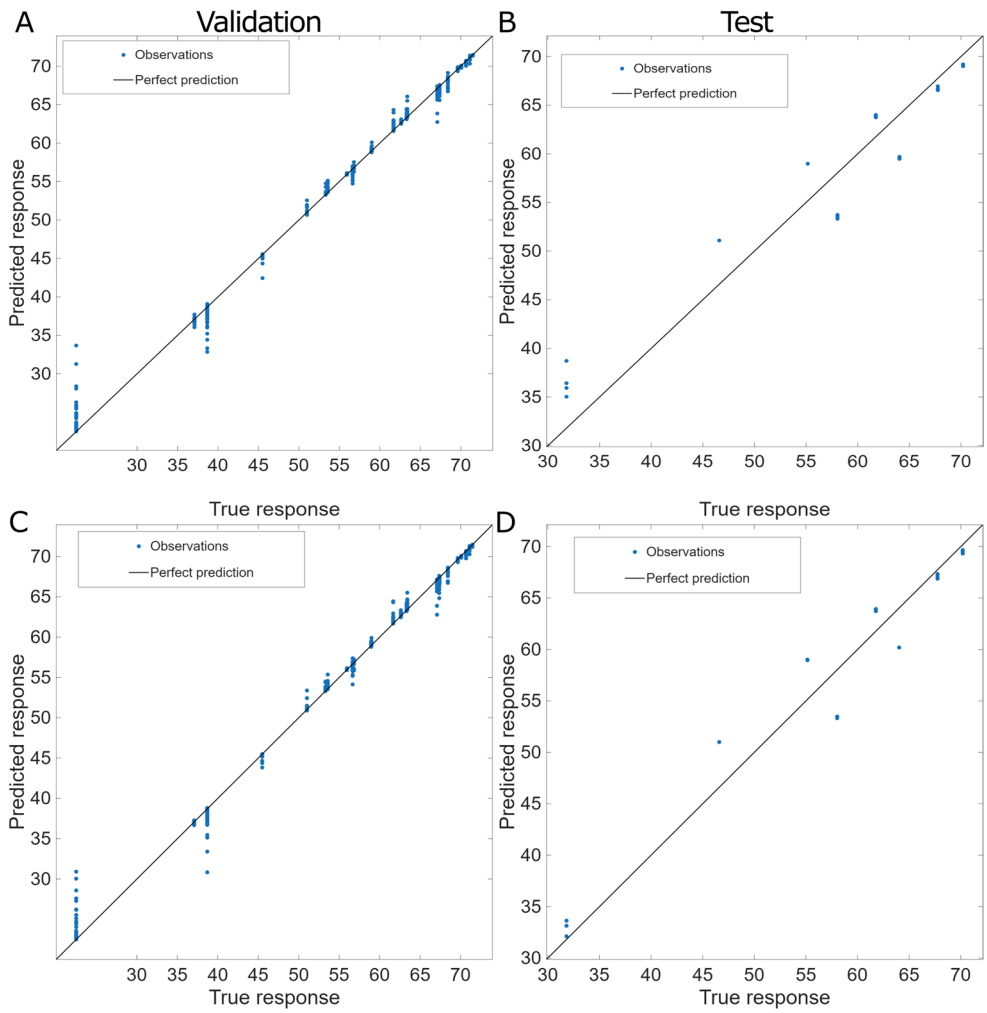

Fig. 14 Prediction analysis for 15 Days dataset for the production estimation of terpenic hydrocarbons based on environmental parameters and sensor retrieved data. **A**) Validation for the Regression model using 44 features. **B**) Test for the Regression model using 44 features. **C**) Validation for the Regression model using 29 features. **D**) Test for the Regression model using 29 features

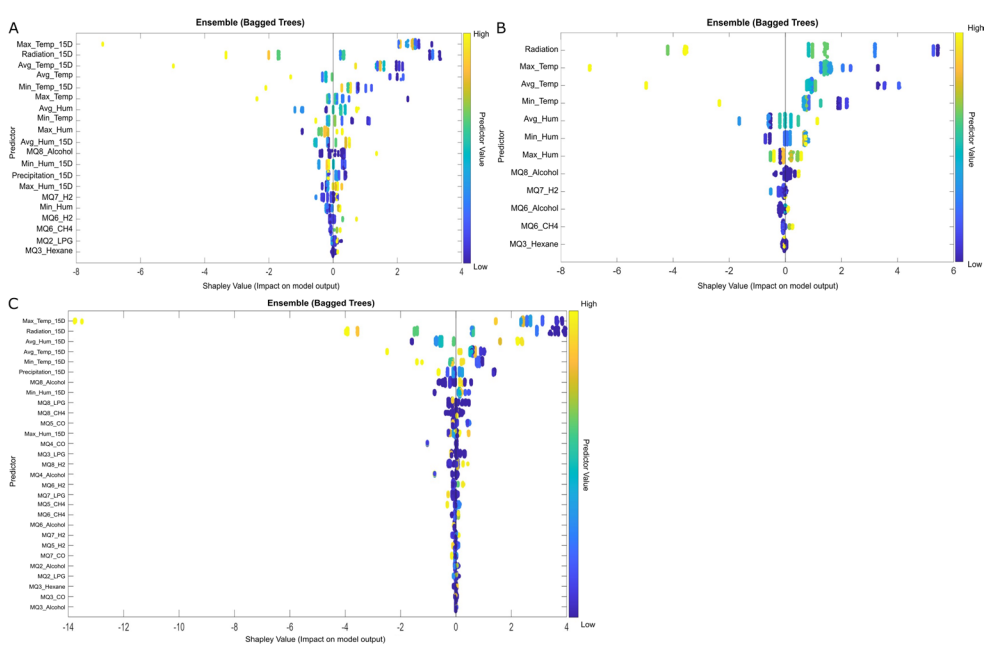

Fig. 15 SHAP analysis of combined environmental and gas sensor predictors for the best-performing machine learning models across the All (**A**), Daily (**B**), and 15-Day (**C**) datasets

Residual analyses across the three datasets (Figures S6–S8) confirmed the robustness and generalization of the Bagged Trees models, with no strong evidence of overfitting. Validation and test residuals clustered near zero across all datasets, with minor deviations at extreme VOC concentrations. The Daily dataset showed slightly larger residuals at high and low VOC values, yet errors were randomly distributed without directional bias. The 15-Day dataset exhibited the most balanced residual distribution, with reduced variability and improved test performance, indicating that temporal aggregation enhanced model stability.

Overall, SHAP results indicated that temperature and radiation, across both All and Daily datasets, were the primary determinants of the recorded VOC fingerprints, with humidity contributing moderately. The Bagged Trees model effectively leveraged these structured environmental signals, while gas sensor inputs improved its ability to resolve rapid VOC fluctuations, emphasizing the model's suitability for smart-sensor VOC fingerprinting in *C. ladanifer* EO systems.

Discussion

The integration of environmental parameters and gas sensors with internet of things (IoT) technologies is pivotal, in precision farming, to predict the concentration of organic compounds and yield of EOs from *C. ladanifer*. By combining MQ sensor data with meteorological conditions, preemptive agricultural strategies aimed for the optimization of crop management practices could be adopted. Such integrated approach could empower informed decision-making, mitigate agricultural risks, and enhance overall productivity.

The present study provides valuable insights into the optimal number of features required to maintain high accuracy in predicting terpenic hydrocarbon concentrations in the EO of *C. ladanifer*, thereby offering a robust framework for environmental monitoring and quality assessment. For a comprehensive overview of studies on environmental and gas sensor applications in precision agriculture and EO analysis, refer to Table 6.

Environmental parameters

The temporal resolution of environmental data plays a crucial role in the predictive capabilities of ML models. Temporal resolution refers to how frequently data is collected or updated over time. High temporal resolution data, such as hourly or daily measurements, capture short-term variations and enable ML models to respond to rapid environmental changes, while low temporal resolution data, such as monthly or yearly averages, offer a broader view and help identify long-term trends [36].

Environmental conditions impact on the physiology and phytochemical content (flavonoids and terpenes) of *C. ladanifer* has been reported through several studies. For example, Chaves, et al. [37] observed a significant increase in flavonoid secretion during summer, driven primarily by UV irradiation and synergistically enhanced by drought, to protect the plant from these environmental stresses. The secretion of more methylated, hydrophobic flavonoids under high temperatures and drought conditions suggests an adaptive defense mechanism of *C. ladanifer* against hydric stress. Similar evidence was reported in *C. monspeliensis*, where severe warming caused a significant decline in terpene emission, accompanied by increased antioxidant activity [38]. Likewise, meteorological parameters, temperature and precipitation, were found to provoke significant in situ physiological variation in *C. albidus* across different habitats in south-east Spain [39]. Precisely, authors identified photochemical efficiency as the primary source of variation, in response to differing environmental conditions, through PCA. A recent study demonstrated that *C. ladanifer* EOs exhibited significant variability in yield and chemical composition, particularly rich in oxygenated sesquiterpenes and monoterpenes, when PCA, CCA, and correlation analyses were undertaken [40]. Consequently, it can be deduced that the production of terpenes, in *Cistus sp.*, is substantially dependent on environmental conditions. Therefore, the identification of the optimal harvest time to obtain high-quality EOs would benefit from daily to bi-weekly observations of meteorological parameters.

The use of ML models has been reported effective for studies employing temporal data and agronomic features. For example, simulation of US maize yield under climate change scenarios was comprehensively assessed comparing process-based models, regression models, and ML algorithms [41]. While process-based models struggle with bias even with optimized parameters, both regression and ML models effectively replicated average yield patterns. Importantly, ML models were reported to outperform both methods in capturing yield variability and extremes, demonstrating its potential for robust probabilistic risk analysis of climate impacts on crop production, projecting a 13.5% decrease in US maize yield under this scenario.

High temporal resolution data (daily or hourly) has been found effective for ML models tasked with making short-term predictions. These models can learn and respond to diurnal patterns, seasonal variations, sudden environmental shifts, and energy demand estimation within a day. For instance, Gasch, et al. [42] demonstrated that temporal components significantly influence the alterability in all variables i.e., soil temperature, water content, and electrical conductivity, by introducing a framework for modeling dynamic soil properties in three dimensions and time, utilizing automated sensor networks.

Table 6 Summary of studies on environmental and gas sensor applications in precision agriculture and EO analysis

Reference	Methodology	Sensor Employed	Key Findings
[37–39]	PCA	N/A	Increase in flavonoid secretion during summer due to UV irradiation and drought Severe warming caused a decline in terpene emission and increased antioxidant activity in *Cistus monspeliensis* Photochemical efficiency as primary source of variation in *Cistus albidus* across different habitats
[40]	PCA, CCA, Correlation Analyses	N/A	Significant variability in yield and chemical composition of *C. ladanifer* EOs
[41]	Process-based models, Regression models, ML algorithms	N/A	ML models outperformed other methods in capturing yield variability and extremes for US maize yield under climate change scenarios (13.5% decrease in yield)
[42]	Framework for modeling dynamic soil properties	Automated sensor networks	Temporal components influence variability in soil temperature, water content, and electrical conductivity
[43]	Extensive time series crop yield monitoring data, ML algorithms	N/A	High concordance correlation coefficients (0.89–0.92) for wheat, barley, and canola yields
[54, 55]	Various ML models	Gas sensors, electronic noses (E-Noses)	High accuracy in detecting ripening, growth stages of fruits, and product origin
[56]	Sensing array, fluorescent polymers	36 fluorescent polymers	Differentiated 30 analytes and a control, but performance decreased with compound pairs
[57]	MQ-based, TGS-based sensors	MQ, TGS sensors	Differentiated chemical compounds
[58]	Array of 9 functionalized reduced graphene oxide	Reduced graphene oxide	Identified individual organic volatile compounds
[26]	Fuzzy ARTMAP, LDA	Various gas sensors	Classified EOs from *Rosa damascena* with accuracies of 82–99%
[59]	LDA, SVM	Various gas sensors	Classified tree EOs with 95–99% accuracy
[60]	PCA, IDMAP	Custom carbon nanocomposites	High coefficient of determination ($R^2 = 0.99$) using custom sensors
[27]	PCA, DFA, HCA	TGS815, TGS821, TGS822, TGS224, TGS825, TGS42	Differentiated between EOs from *Aucoumea klaineana* and *Canarium schweinfurthii*
[28]	4 ML algorithms	9 gas sensors	Classified EOs with accuracies of 96.7–100%
[61]	N/A	8 gas sensors	Differentiated EO from distilled water extracts with accuracies of 85.6–100%
[62]	N-way k-shot meta-learning	24 printable chemiresistive odor sensors	Recognized and classified 30 different aromas with 98.70% accuracy
[29]	KNN, NB, DA	MQ3, MQ4, MQ6, MQ135	Classified commercial lavender EOs with 65% accuracy
[63]	PCA	9 gas sensors	Detected adulteration of patchouli oil with candlenut oil
[17, 30]	ANN, NB, KNN, DA, SVM	Various gas sensors	Identified *C. ladanifer* EO adulterated with Pinus pinaster EO (99.97% accuracy). Classified herbs and fruits EOs (88.74–99.82% accuracy)
[64]	1-D convolutional neural network, ArcLoss	MQ3, MQ4, MQ7, MQ9	Discriminated odors from EOs, coffee, and whiskey with highest accuracy, precision, and recall scores (94.61%, 96.1%, 93.45%)

On the other hand, lower temporal resolution data, such as weekly averages, are beneficial for identifying long-term trends and gradual changes. These trends can be crucial for understanding and predicting phenomena like climate change, vegetation growth, and annual average temperature shifts. A direct correlation between the dataset and the training model precision was reported by Filippi, et al. [43]. In this study, authors leveraged extensive time

series and crop yield monitoring data alongside ancillary datasets like soil conductivity and environmental parameters, thus generating predictive models for wheat, barley, and canola yields. They observed high Lin's concordance correlation coefficients of 0.89 to 0.92 at field resolution. Whereas, combining high and low temporal resolution datasets can enhance the predictive capabilities of ML models by balancing short-term accuracy with long-term trend identification. Such hybrid approaches allow models to leverage the detailed dynamic patterns captured by high-resolution data while also considering the broader trends from low-resolution data [36].

Considering the results obtained in our study, the comparative analysis across these datasets revealed that the combined "Daily" and "15-Day" dataset approach significantly enhances the predictive capabilities of the models (Fig. 6). This method benefits from a broader temporal context, which likely provides a more comprehensive representation of the underlying patterns. The superior performance of models trained on the combined dataset, demonstrating an R^2 test accuracy of 92.73%, suggests that incorporating a wider range of data frequencies can mitigate issues related to overfitting and outliers, thereby enhancing the robustness and accuracy of the predictions. Whereas a comparative lower R^2 test accuracy, of 90.10%, suggests its suitability for identifying long-term trends and gradual changes, which could be more suitable for monitoring purposes.

Therefore, it can be deduced that the findings of our study underscore the critical role of temporal resolution in the performance of ML models for predictive analysis, which are in accordance with the previous investigations where temporal data was employed. While the daily datasets offer fine-grained insights, they are susceptible to overfitting and the influence of outliers due to insufficient variability. In contrast, the 15-Day dataset provides broader temporal insights but compromises on accuracy because of reduced granularity. The combined dataset approach successfully navigates this trade-off, harnessing the advantages of both datasets to achieve superior predictive accuracy and enhanced model robustness. These results emphasize the importance of strategic dataset composition in the development of reliable and accurate ML models for predictive analytics for precision agriculture.

Best performing ML models

The superior classification performance achieved in this study could be attributed to the intrinsic capabilities of the selected ML models and their alignment with the structural complexity of the datasets. Neural Networks demonstrated outstanding results, exceeding 94% validation accuracy and 92% test accuracy, largely because of their capacity to model highly non-linear, multivariate interactions between environmental parameters and terpenic hydrocarbon concentrations. This is consistent with recent evidence showing that deep learning architectures—particularly feedforward, CNN-based, and LSTM-enhanced hybrids—excel at capturing dynamic dependencies in environmental systems and outperform conventional statistical approaches in both robustness and predictive precision [44–46].

Moreover, Stepwise Linear Regression performed exceptionally well when dimensionality reduction was employed, particularly with 13 features. Its strength lies in its systematic removal of collinearity and retention of only the most influential predictors, thereby reducing noise while improving interpretability. Prior studies in environmental and chemical modeling similarly highlight its reliability for identifying dominant variables in complex multivariate systems [47, 48].

For datasets combining sensor, environmental, and chemical inputs, ensemble tree-based models, especially Bagged and Boosted Trees, consistently yielded high accuracy. Bagged Trees reached 93.15% test accuracy for the 15-Day mean dataset, benefitting from bootstrap aggregation that reduces overfitting and enhances generalization. In contrast, Boosted Trees excelled in daily datasets by iteratively correcting residual errors and adapting to temporal variability, a behavior widely documented in ecological prediction tasks [49–51].

Collectively, these results illustrate that optimal model selection hinges on matching algorithmic strengths to data complexity. Neural Networks are well suited for high-dimensional, interaction-rich environments, while ensemble methods dominate when feature heterogeneity and non-linear patterns are prominent. These conclusions align with current advances recommending hybrid and ensemble-deep learning strategies for scalable, interpretable agro-environmental prediction [52, 53].

Sensor-based and environmental parameters

Most studies utilizing gas sensors and electronic noses have focused on specific applications, such as detecting ripening, growth stages of fruits, or determining product origin. While these studies have demonstrated high accuracy in their results, they do not provide quantification of the chemical compounds present [54, 55].

In terms of detecting and identifying individual compounds, several studies have been conducted. For instance, Zhao, et al. [56] evaluated the response of a sensing array composed of 36 fluorescent polymers to differentiate 30 analytes and a control. The results were promising, but the performance decreased when compounds were tested in

pairs. Klouček, et al. [57] used MQ-based and TGS-based gas sensors to differentiate chemical compounds, and Liu, et al. [58] employed an array of 9 functionalized reduced graphene oxide to identify individual organic volatile compounds. However, none of these studies quantified the concentration of the given compounds using regression or classification models.

Some studies have focused on using gas sensors to classify different products based on their features [26, 59]. For example, EOs from *Rosa damascena* were classified using Fuzzy ARTMAP and LDA, achieving accuracies ranging from 82 to 99% [26]. However, the authors did not provide information about the quantity of selected chemical compounds for the classification nor give clear data about the range in each class. Similarly, EOs from shrubs (*Rosa damascene*) were classified using ML algorithms, but the authors did not provide information about the achieved accuracies nor the detection quantification of specific chemical components of the EOs [59]. Nevertheless, the authors employed LDA and SVM as ML algorithms for multiclass classification, achieving accuracies between 95 and 99% in the validation dataset.

In 2020, two studies utilized gas sensors to analyze EOs from tree sources (*Syzygium aromaticum*). Graboski, et al. [60] developed custom gas sensors based on carbon nanocomposites, creating six distinct sensors. These sensors were used to train and validate a model using principal component analysis (PCA) and independent component analysis (IDMAP). The authors achieved a high coefficient of determination (R^2) of 0.99, indicating excellent performance. However, this was achieved using custom sensors that are not commercially available due to their high cost. In contrast, Aghoutane, et al. [27] employed low-cost sensors (TGS815, TGS821, TGS822, TGS224, TGS825, TGS42) and three ML algorithms (PCA, DFA, HCA) to differentiate between EOs from *A. klaineana* and *C. schweinfurthii*. Although the authors did not report the achieved accuracies or quantify specific chemical compounds in the EOs, this study demonstrates the potential of gas sensors for tree-derived EO analysis.

Recent studies have also utilized gas sensors to classify EOs from herbs and fruits. For instance, Rasekh, et al. [28] used 9 gas sensors and 4 ML algorithms to classify EOs from various sources, achieving accuracies ranging from 96.7 to 100%. However, no tests were conducted, and the authors did not identify or quantify chemical compounds. Likewise, Zorpeykar, et al. [61] used 8 sensors to differentiate EO from distilled water extracts, achieving averaged accuracies from validation and test ranging from 85.6 to 100%. However, the authors did not quantify chemical compounds, and the averaged data precluded correct comparison. In 2022, Liu, et al. [62]. proposed the application of gas sensors for the recognition and classification of aroma oils. The authors employed 24 printable chemiresistive odor sensors to measure up to 30 different aromas. The data was analyzed using N-way k-shot meta-learning, resulting in an accuracy of 98.70%. However, the test dataset was not provided.

In 2023, Zlatev and Draganova [29] utilized gas sensors to classify the quality of commercial lavender EOs. The authors used four MOS sensors (MQ3, MQ4, MQ6, MQ135) and three ML algorithms (KNN, NB, DA), achieving an accuracy of 65% in the validation dataset. Notably, no quantification of the chemical compounds in the tested EOs was conducted. Similarly, Sudarmaji, et al. [63] used gas sensors to detect adulteration of patchouli oil with candlenut oil. The authors selected nine gas sensors and analyzed the data using PCA, achieving the best results with six of the nine sensors. However, the authors did not provide accuracy in the paper, and no quantification of chemical compounds was conducted. Likewise, Viciano-Tudela, et al. [17] used gas sensors to identify *C. ladanifer* EO adulterated with *Pinus pinaster* EO. The authors achieved an accuracy of 99.97% using ANN. The same sensors were used to recognize different types of EOs from herbs and fruits [30]. In this case, the used ML models were higher, including ANN, NB, KNN, DA, and SVM, achieving accuracies from 88.74 to 99.82%. Nonetheless, in those papers, the authors have not used gas sensors to quantify the content of any chemical compound. However, in these papers, the authors did not use gas sensors to quantify the content of any chemical compounds. Another study by Ye, et al. [64] addressed the challenge of odor detection and identification through an advanced E-Nose system, which employed 4 MQ gas sensors (MQ3, MQ4, MQ7, MQ9). The study emphasized the importance of feature quality extracted from E-Nose data; a facet often overlooked in favor of ML discriminative enhancements. The authors constructed a novel E-Nose with automated data processing to discriminate odors from EOs, coffee, and whiskey. They employed a 1-D convolutional neural network coupled with ArcLoss, a technique inspired by face recognition technology, to improve feature distinctiveness. Authors reported a highest accuracy, precision, and recall score of 94.61% (Random Forest), 96.1% (Random Forest), and 93.45% (Logistic Regression) for the classification of EO.

The empirical outcomes from previous studies utilizing MQ sensors substantiate our current findings, demonstrating that MQ sensors exhibit a heightened capability to accurately discern the EOs of *C. ladanifer*, especially when leveraged by selective ML features. Although gas sensors have been extensively employed in various studies for the classification of EOs, the precise quantification of the chemical constituents has not traditionally been the primary

focus. The precision achieved in both the training and validation datasets in this study aligns closely with the accuracy reported by other researchers in the field. However, it is imperative to highlight that many existing studies fail to disclose test accuracy metrics. In contrast, our study presents a comprehensive evaluation, demonstrating significant accuracy for the selected ML models and their predictive capabilities.

Limitations and future prospects

Despite the robust findings and methodological advancements presented in this study, several limitations must be acknowledged. Firstly, the reliance on MQ sensors, while effective in detecting and classifying VOCs, is inherently limited by the sensor's sensitivity and specificity to the diverse chemical profiles present in *C. ladanifer* EO. Although the MQ sensors demonstrated a significant accuracy, there remains a need for more sophisticated sensor technology that could offer enhanced sensitivity and broader detection capabilities to detect individual terpenic hydrocarbons and related analytes.

Furthermore, the study's focus on temporal resolution and environmental parameters segmented into "Daily", "15 Days", and "All" datasets, while informative, may not fully capture the complex interactions and temporal dynamics of environmental factors influencing EO composition. In addition, the environmental parameters were limited to a specific set, potentially omitting other influential factors such as soil composition, microbial activity, and plant phenological stages, which could significantly impact EO profiles. Future studies should aim to integrate a broader spectrum of environmental and biological data to enhance the predictive models' accuracy and robustness and the intricate biological relationships.

Future studies should include a systematic characterization of MQ sensor response ranges under controlled VOC exposures to empirically validate stability and saturation thresholds. Incorporating exploratory data analysis and distribution-based thresholding will allow data-driven filtering criteria, improving reproducibility and ensuring that preprocessing decisions are fully supported by quantitative evidence.

Although measurements were performed under controlled laboratory conditions (temperature: 25–34 °C; humidity: 22–38%), MQ sensors inherently present limitations that may influence classification accuracy. Notably, humidity fluctuations can modify sensor resistance and distort response profiles, while cross-sensitivity to non-target VOCs can introduce additional noise into the gas fingerprint patterns. Despite the implementation of baseline correction and drift-mitigation procedures, these factors cannot be fully eliminated. Future work should incorporate more robust calibration strategies, including multi-point calibration across varying humidity ranges, algorithmic humidity compensation, and hybrid sensor arrays combined with machine-learning–based feature disentanglement, to reduce cross-sensitivity effects and further improve reproducibility and specificity.

The ML models employed in this study, although rigorously validated, could benefit from the exploration of more advanced algorithms that may offer superior predictive performance and generalizability. Additionally, the current study's scope was confined to *C. ladanifer*, and thus, the generalizability of the findings to other plant species and EO compositions remains to be explored.

Likewise, future work should expand the sampling design by incorporating additional sites, seasons, and a larger number of EO samples to improve statistical representativeness and capture broader environmental and chemotypic variability. A more extensive dataset will further strengthen model generalization, allow for robust external validation, and support the development of transferable predictive frameworks for VOC monitoring across diverse *C. ladanifer* populations and environmental conditions.

In terms of future prospects, the integration of next-generation sensor technologies, such as nano-sensors or biosensors, with advanced ML algorithms holds significant promise for the field of precision agriculture. These technologies could offer unprecedented accuracy and resolution in monitoring VOCs and environmental parameters, thereby facilitating more precise and real-time decision-making processes. Moreover, expanding the research framework to include other economically important plants and their EOs could provide valuable insights into the broader applicability of the developed models and methodologies. In a similar manner, longitudinal studies capturing seasonal and interannual variations in EO composition and environmental conditions could further elucidate the temporal dynamics and resilience of agricultural systems.

This study marks a significant step forward in the application of gas sensors and ML models for EO analysis and environmental monitoring, along with highlighting the necessity for ongoing innovation and refinement in sensor technology, data integration, and model development. The pursuit of these advancements will undoubtedly enhance our understanding and management of agricultural ecosystems, contributing to the sustainability and efficiency of precision agriculture practices. Furthermore, regarding the sensor's applicability, the resulting device can be used in cases where oils have been mixed. This allows the producer to monitor the oil's quality and whether it has been adulterated with other substances during the extraction process. It

is an inexpensive option that provides a preliminary assessment of the oil's condition. However, for the pharmaceutical industry, which requires pure oils, this option would not be entirely viable, although it could be useful for a first approximation before processing with more expensive techniques such as HPLC, if necessary. It is worth highlighting the speed and the fact that highly qualified personnel are not required to carry out the analysis with the developed MQ sensor system. Comparing the proposed system with traditional chromatographic analysis techniques (GC-FID-MS), the price per sample ranges from €150 to €250. The main difference, besides the price, is that MQ sensors can be reused for a considerable number of samples, requiring only a single investment (€55-€88). Furthermore, the sample is not destroyed, making it a sustainable method.

Conclusions

This study demonstrates the potential of a smart sensor system integrating low-cost MQ gas sensors and environmental data within ML frameworks to predict terpenic hydrocarbon concentrations in *Cistus ladanifer* EO. By capturing the gas fingerprint of VOC emissions, the system enables a non-destructive, rapid, and cost-effective alternative to traditional analytical techniques. Environmental data were segmented into Daily, 15-Day, and All datasets and paired with VOC measurements to evaluate the impact of temporal resolution on model performance. The Daily dataset yielded the highest predictive R^2 (90.10%) using environmental parameters alone, with all tested ML models—including Neural Networks, Linear and Logistic Regression, Decision Trees, and Ensemble methods—achieving over 90% accuracy across datasets. Feature selection revealed that minimal yet informative subsets (8, 5, and 13 features for Daily, 15-Day, and All datasets, respectively) were sufficient for accurate prediction using environmental inputs. The integration of VOC signals from MQ sensors enhanced model robustness, though it required a greater number of features (12, 20, and 29, respectively). These findings highlight the importance of dataset granularity and optimized feature selection in enhancing ML-based predictive systems. While the proposed smart sensor system demonstrates strong performance in controlled laboratory settings, its real-time applicability has yet to be validated under field conditions. Practical deployment challenges include maintaining stable sensor calibration under fluctuating temperature and humidity, mitigating cross-sensitivity effects, ensuring power efficiency for continuous monitoring, and enabling reliable data transmission in outdoor environments. Future work will address these limitations through field trials, integration with IoT communication platforms, and the development of adaptive, humidity-compensated calibration strategies to enhance robustness and long-term stability. In conclusion, the developed smart sensor system offers a promising approach for EO quality assessment and environmental monitoring, with potential integration into precision agriculture frameworks following successful field validation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11694-026-04063-w.

Authors' contributions Conceptualization, J.L., S. S., and A. A.; methodology, L.P., A.A., and F.J.D.B.; software, A.A., F.J.D.B., S.V.T., and L.P.; validation, A.A., F.J.D.B., I.M., and A.S.G.; formal analysis, A.A.; F.J.D.B.; S.V.T., and L.P.; investigation A.A., F.J.D.B., S.V.T., L.P., V.C., R.B., L.S.E.P., P.M., C.C.S., J.P.R., and B.S.; resources, J.L., S.S., A.S.G., and P.V.M.; data curation, A.A. and F.J.D.B.; writing—original draft preparation, A.A.; F.J.D.B.; S.V.T., L.P., and A.S.G.; writing—review and editing, A.A., F.J.D.B., S.V.T., S.S., J. L.,V.C., R.B., L.S.E.P., P.M., C.C.S., J.P.R., B.S., and A.S.G.; visualization, A.A., J.F.D.B.; supervision, S.S. and J.L.; project administration, P.V.M. and J.L.; funding acquisition, S.S., P.V.M., and J.L. All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work is partially funded by the "Programa Estatal de I + D + i Orientada a los Retos de la Sociedad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2017–2020" through the projects Project PID2020-114467RR-C31/C32/C33 funded by MICIU/AEI/https://doi.org/10.13039/501100011033; "Proyectos Estratégicos Orientados a la Transición Ecológica y a la Transición Digital" project TED2021-131040B-C31; Grant PRE2021-100809 funded by MICIU/AEI/https://doi.org/10.13039/501100011033 and by European Social Fund Plus (ESF +), and through the project FP24-ESSENZIALI + D funded by IMIDRA. B. Stefanutti is the beneficiary of a predoctoral R&D&I contract within the framework of the project 'FP24 ESSENZIALI + D,' funded by IMIDRA.

Declarations

Institutional review board statement Not applicable.

Conflicts of interest All authors have read the manuscript and declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. L. Nahar, H.R. El-Seedi, S.A. Khalifa, M. Mohammadhosseini, S.D. Sarker, Ruta essential oils: composition and bioactivities. Molecules **26**, 4766 (2021)
2. C. Pérez-Izquierdo, P. Serrano-Pérez, M. del Carmen Rodríguez-Molina, Chemical composition, antifungal and phytotoxic activities of *Cistus ladanifer* L. essential oil and hydrolate. Biocatal. Agric. Biotechnol. **45**, 102527 (2022)
3. A. Laftouhi, N. Eloutassi, E. Ech-Chihbi, Z. Rais, A. Abdellaoui, A. Taleb, M. Beniken, H.-A. Nafidi, A.M. Salamatullah, M. Bourhia, The impact of environmental stress on the secondary metabolites and the chemical compositions of the essential oils from some medicinal plants used as food supplements. Sustainability **15**, 7842 (2023)
4. A. Laftouhi, N. Eloutassi, S. Drioua, E. Ech-Chihbi, Z. Rais, A. Abdellaoui, A. Taleb, M. Beniken, M. Taleb, Impact of water stress and temperature on metabolites and essential oil of *Rosmarinus officinalis* (phytochemical screening, extraction, and gas chromatography). J. Ecol. Eng. (2023). https://doi.org/10.12911/22998993/161760
5. M.M. Qaderi, A.B. Martel, C.A. Strugnell, Environmental factors regulate plant secondary metabolites. Plants **12**, 447 (2023)
6. S.S. Bashir, A. Hussain, S.J. Hussain, O.A. Wani, S. Zahid Nabi, N.A. Dar, F.S. Baloch, S. Mansoor, Plant drought stress tolerance: understanding its physiological, biochemical and molecular mechanisms. Biotechnol. & Biotechnol. Equip. **35**, 1912–1925 (2021)
7. E.D. Cruz, L.D. Barros, B.D. Guimarães, R.H. Mourão, J.G. Maia, W.N. Setzer, J.K. da Silva, P.L. Figueiredo, Seasonal Variation in Essential Oil Composition and Antioxidant Capacity of Aniba canelilla (Lauraceae): A Reliable Source of 1-Nitro-2-phenylethane. Molecules **28**, 7573 (2023)
8. P. Dobhal, V.K. Purohit, S. Chandra, S. Rawat, P. Prasad, U. Bhandari, V.L. Trivedi, M. Nautiyal, Climate-induced changes in essential oil production and terpene composition in Alpine aromatic plants. Plant Stress **12**, 100445 (2024)
9. T. Zhang, Y. Zheng, C. Fu, H. Yang, X. Liu, F. Qiu, X. Wang, Z. Wang, Chemical variation and environmental influence on essential oil of *Cinnamomum camphora*. Molecules **28**, 973 (2023)
10. A. Akbar, A. Kuanar, J. Patnaik, A. Mishra, S. Nayak, Application of artificial neural network modeling for optimization and prediction of essential oil yield in turmeric (*Curcuma longa* L.). Comput. Electron. Agric. **148**, 160–178 (2018)
11. K. Ložienė, V. Vaičiulytė, R. Maždžierienė, Influence of meteorological conditions on essential oil composition in geraniol-bearing *Thymus pulegioides* and *Thymus* hybrid. Acta Physiol. Plant. **43**, 1–9 (2021)
12. S. Mehalaine, H. Chenchouni, Quantifying how climatic factors influence essential oil yield in wild-growing plants. Arab. J. Geosci. **14**, 1257 (2021)
13. A. Ben Arfa, H. Gouja, H. Hannachi, H. Isoda, M. Neffati, H. Najjaa, Seasonal changes in rosemary species: A chemotaxonomic assessment of two varieties based on essential oil compounds, antioxidant and antibacterial activities. PLoS ONE **17**, e0273367 (2022)
14. A.S. Mohareb, M.A. Elashmawy, M.E. Nawar, A.K. Abdelrahman, F.M. Ahmed, A.E. Hassona, M.Z. Salem, Chemical compositions and antifungal activity of *Corymbia citriodora*, *Cupressus macrocarpa*, and *Syzygium cumini* extracts: GC–MS and HPLC analysis of essential oils and phenolic compounds. Biomass Convers. Biorefin. (2023). https://doi.org/10.1007/s13399-023-05106-8
15. Z. Wang, L. Tian, Y. Xiao, M. Zhao, Y. Chang, Y. Zhou, S. Liu, H. Zhao, Y. Xiu, Quantitative and differential analysis between *Bupleurum chinense* DC. and *Bupleurum scorzonerifolium* Willd. using HPLC-MS and GC-MS coupled with multivariate statistical analysis. Molecules **28**, 5630 (2023)
16. P.K. Khatri, M. Paolini, R. Larcher, L. Ziller, D.A. Magdas, O. Marincas, A. Roncone, L. Bontempo, Botanical characterization and authentication of lavender essential oil using its volatile organic compounds and compound-specific carbon and hydrogen isotope ratio analysis. Food Control **154**, 110002 (2023)
17. S. Viciano-Tudela, S. Sendra, L. Parra, J.M. Jimenez, J. Lloret, Proposal of a gas sensor-based device for detecting adulteration in essential oil of *Cistus ladanifer*. Sustainability **15**, 3357 (2023)
18. S.I. Sabilla, R. Sarno, J. Siswantoro, Estimating gas concentration using artificial neural network for electronic nose. Procedia Comput. Sci. **124**, 181–188 (2017)
19. A. Ajiboye, J. Opadiji, A. Yusuf, J. Popoola, Analytical determination of load resistance value for MQ-series gas sensors: MQ-6 as case study. TELKOMNIKA **19**, 575–582 (2021)
20. K. Gupta, G.G. Krishna, T. Anjali, An IoT based system for domestic air quality monitoring and cooking gas leak detection for a safer home. In Proceedings of the 2020 International Conference on Communication and Signal Processing (ICCSP), pp. 0705–0709 (2020)
21. A. Onyan, An electronic device for detecting adulteration of petroleum products at point-of-sale terminals. FUPRE J. Sci Ind Res **6**, 67–78 (2022)
22. H.J. Khadim, F.K. Obaed, Z.T. Abd Ali, Application of mq-sensors to indoor air quality monitoring in lab based on iot. In Proceedings of the 2021 International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE) 1–5. (2021)
23. M. Rasekh, H. Karami, A.D. Wilson, M. Gancarz, Performance analysis of MAU-9 electronic-nose MOS sensor array components and ANN classification methods for discrimination of herb and fruit essential oils. Chemosensors **9**, 243 (2021)
24. P.P. Lal, A.A. Prakash, A.A. Chand, K.A. Prasad, U. Mehta, M.H. Assaf, F.S. Mani, K.A. Mamun, Iot integrated fuzzy classification analysis for detecting adulterants in cow milk. Sens Bio-Sensing Res **36**, 100486 (2022)
25. S. Natarajan, V. Ponnusamy, Development of a Smart Sensor Array for Adulteration Detection in Black Pepper Seeds using Machine Learning. Int J Adv Comput Sci Appl 14 (2023)
26. A. Gorji-Chakespari, A.M. Nikbakht, F. Sefidkon, M. Ghasemi-Varnamkhasti, J. Brezmes, E. Llobet, Performance comparison of fuzzy ARTMAP and LDA in qualitative classification of Iranian rosa damascena essential oils by an electronic nose. Sensors **16**, 636 (2016)
27. Y. Aghoutane, M. Moufid, S. Motia, G.S. Padzys, L.P. Omouendze, E. Llobet, B. Bouchikhi, N. El Bari, Characterization and analysis of Okoume and aiele essential oils from Gabon by gc-ms, electronic nose, and their antibacterial activity assessment. Sensors **20**, 6750 (2020)
28. M. Rasekh, H. Karami, A.D. Wilson, M. Gancarz, Classification and identification of essential oils from herbs and fruits based on a MOS electronic-nose technology. Chemosensors **9**, 142 (2021)
29. Z. Zlatev, A. Draganova, Development and research of a measuring device for gas and optical features of essential oils. In Proceedings of the AIP Conference Proceedings (2023)
30. S. Viciano-Tudela, L. Parra, P. Navarro-Garcia, S. Sendra, J. Lloret, Proposal of a new system for essential oil classification based on low-cost gas sensor and machine learning techniques. Sensors **23**, 5812 (2023)
31. F.J.D. Blasco, S. Viciano-Tudela, L. Parra, A. Ahmad, V. Chaloupková, R. Bados, L.S.E. Pascual, I. Mediavilla, S. Sendra, J. Lloret, Employment of MQ gas sensors for the classification of *Cistus ladanifer* essential oils. Microchem. J. **206**, 111585 (2024)

32. I. Mediavilla, M.A. Blázquez, A. Ruiz, L.S. Esteban, Influence of the storage of *Cistus ladanifer* L. bales from mechanised harvesting on the essential oil yield and qualitative composition. Molecules **26**, 2379 (2021)
33. SIAR. EMap. Available online: https://eportal.mapa.gob.es/websiar/SeleccionParametrosMap.aspx?dst=1 (accessed on April 20).
34. A. Miquel-Ibarz, J. Burgués, S. Marco, Global calibration models for temperature-modulated metal oxide gas sensors: a strategy to reduce calibration costs. Sensors Actuators B Chem. **350**, 130769 (2022)
35. A. Jadhav, D. Pramod, K. Ramanathan, Comparison of performance of data imputation methods for numeric dataset. Appl. Artif. Intell. **33**, 913–933 (2019)
36. T.P. Kharel, A.J. Ashworth, P.R. Owens, M. Buser, Spatially and temporally disparate data in systems agriculture: issues and prospective solutions. Agron. J. **112**, 4498–4510 (2020)
37. N. Chaves, J.C. Escudero, C. Gutierrez-Merino, Role of ecological variables in the seasonal variation of flavonoid content of *Cistus ladanifer* exudate. J. Chem. Ecol. **23**, 579–603 (1997)
38. I. Nogués, M. Medori, C. Calfapietra, Limitations of monoterpene emissions and their antioxidant role in *Cistus* sp. under mild and severe treatments of drought and warming. Environ. Exp. Bot. **119**, 76–86 (2015)
39. O.M. Grant, L.D. Incoll, Photochemical efficiency is an important component of ecophysiological variation of *Cistus albidus* between habitats in south-east Spain. Funct. Plant Biol. **32**, 107–115 (2005)
40. C. Pérez-Izquierdo, M.J. Jordán Bueso, M. del Carmen Rodríguez-Molina, F. Pulido, Spatial variation in yield, chemical composition, and phytotoxic activity of *Cistus ladanifer* essential oils. Chem. Biodivers. **20**, e202300995 (2023)
41. G. Leng, J.W. Hall, Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. Environ Res Lett **15**, 044027 (2020)
42. C.K. Gasch, T. Hengl, B. Gräler, H. Meyer, T.S. Magney, D.J. Brown, Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D+ T: the Cook Agronomy Farm data set. Spat. Stat. **14**, 70–90 (2015)
43. P. Filippi, E.J. Jones, N.S. Wimalathunge, P.D. Somarathna, L.E. Pozza, S.U. Ugbaje, T.G. Jephcott, S.E. Paterson, B.M. Whelan, T.F. Bishop, An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precis. Agric. **20**, 1015–1029 (2019)
44. S. Liu, K. Liu, Z. Wang, Y. Liu, B. Bai, R. Zhao, Investigation of a transformer-based hybrid artificial neural networks for climate data prediction and analysis. Front. Environ. Sci. **12**, 1464241 (2025)
45. X. Sun, T. Lu, Z. Wang, Comparative analysis of different deep learning algorithms for the prediction of marine environmental parameters based on Cmems products. Acta Geodyn. Geomater. **21**, 343–363 (2024)
46. M. Wesselkamp, N. Moser, M. Kalweit, J. Boedecker, C.F. Dormann, Process-informed neural networks: a hybrid modelling approach to improve predictive performance and inference of neural networks in ecology and beyond. Ecol. Lett. **27**, e70012 (2024)
47. J. Fan, Y. Zhao, D. Wang, X. Zhou, Y. Li, W. Zhang, F. Xu, S. Wei, A stepwise multifactor regression analysis of the interactive effects of multiple climate factors on the response of vegetation recovery to drought. Atmosphere **15**, 1094 (2024)
48. G. Del Serrone, L. Moretti, A stepwise regression to identify relevant variables affecting the environmental impacts of clinker production. J. Clean. Prod. **398**, 136564 (2023)
49. G. Spanakis, G. Weiss, A. Roefs, Bagged boosted trees for classification of ecological momentary assessment data. https://arxiv.org/abs/1607.01582https://arxiv.org/abs/1607.01582 (2016)
50. L. Nespoli, V. Medici, Multivariate boosted trees and applications to forecasting and control. J. Mach. Learn. Res. **23**, 1–47 (2022)
51. G. De'Ath, Boosted trees for ecological modeling and prediction. Ecology **88**, 243–251 (2007)
52. R. Majumder, J. Pollard, M.S. Salek, D. Werth, G. Comert, A. Gale, S.M. Khan, S. Darko, M. Chowdhury, Development and evaluation of ensemble learning-based environmental methane detection and intensity prediction models. Environ. Health Insights **18**, 11786302241227307 (2024)
53. Y. Jiang, R. Zhang, H. Jiang, B. Zhang, K. Chen, J. Lv, J. Chen, Y. Song, Comparative performance analysis of heterogeneous ensemble learning models for multi-satellite fusion GNSS-IR soil moisture retrieval. Land **14**, 1716 (2025)
54. J. Xu, K. Liu, C. Zhang, Electronic nose for volatile organic compounds analysis in rice aging. Trends Food Sci. Technol. **109**, 83–93 (2021)
55. M.M. Ali, N. Hashim, S. Abd Aziz, O. Lasekan, Principles and recent advances in electronic nose for quality inspection of agricultural and food products. Trends Food Sci. Technol. **99**, 1–10 (2020)
56. P. Zhao, Y. Wu, C. Feng, L. Wang, Y. Ding, A. Hu, Conjugated polymer nanoparticles based fluorescent electronic nose for the identification of volatile compounds. Anal. Chem. **90**, 4815–4822 (2018)
57. P. Klouček, M. Malina, J. Havlík, P. Maršík, L. Maršán, Identification and quantification of plant volatiles by low-cost electronic nose. The 2nd IMEKOFOODS, 110–114 (**2016**)
58. B. Liu, Y. Huang, K.W. Kam, W.-F. Cheung, N. Zhao, B. Zheng, Functionalized graphene-based chemiresistive electronic nose for discrimination of disease-related volatile organic compounds. Biosens Bioelectron X **1**, 100016 (2019)
59. A. Gorji-Chakespari, A.M. Nikbakht, F. Sefidkon, M. Ghasemi-Varnamkhasti, E.L. Valero, Classification of essential oil composition in *Rosa damascena* Mill. genotypes using an electronic nose. J. Appl. Res. Med. Aromat. Plants **4**, 27–34 (2017)
60. A.M. Graboski, C.A. Zakrzevski, F.M. Shimizu, R.T. Paschoalin, A.C. Soares, J. Steffens, N. Paroul, C. Steffens, Electronic nose based on carbon nanocomposite sensors for clove essential oil detection. ACS Sens. **5**, 1814–1821 (2020)
61. S. Zorpeykar, E. Mirzaee-Ghaleh, H. Karami, Z. Ramedani, A.D. Wilson, Electronic nose analysis and statistical methods for investigating volatile organic compounds and yield of mint essential oils obtained by hydrodistillation. Chemosensors **10**, 486 (2022)
62. C. Liu, H. Miyauchi, K. Hayashi, Deepsniffer: a meta-learning-based chemiresistive odor sensor for recognition and classification of aroma oils. Sensors Actuators B Chem. **351**, 130960 (2022)
63. A. Sudarmaji, A. Margiwiyatno, S.B. Sulistyo, Characteristics of array MOS gas sensors in detection of adulteration on patchouli oil with candlenut oil. In Proceedings of the AIP Conference Proceedings, (2023)
64. Z. Ye, Y. Li, R. Jin, Q. Li, Towards accurate odor identification and effective feature learning with an AI-empowered electronic nose. IEEE Internet of Things Journal (2023).

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

Authors and Affiliations

Ali Ahmad[1] · Francisco Javier Diaz[1] · Sandra Viciano-Tudela[1] · Lorena Parra[2] · Sandra Sendra[1] · Jaime Lloret[1] · Veronika Chaloupková[3,4] · Raquel Bados[4] · Luis Saul Esteban Pascual[4] · Irene Mediavilla[4] · Pedro V. Mauri[5] · Celia Cano Shaw[5] · Jose Plaza Ramos[5] · Barbara Stefanutti[5] · Alicia Sanchez-Gorostiaga[5]

✉ Jaime Lloret
jlloret@dcom.upv.es

Ali Ahmad
aahmad1@upv.es

Francisco Javier Diaz
fjdiabla@doctor.upv.es

Sandra Viciano-Tudela
svictud@upv.es

Lorena Parra
lorena.parra@upm.es

Sandra Sendra
sansenco@upv.es

Veronika Chaloupková
veronika.chaloupkova@ciemat.es

Raquel Bados
raquel.bados@ciemat.es

Luis Saul Esteban Pascual
luis.esteban@ciemat.es

Irene Mediavilla
irene.mediavilla@ciemat.es

Pedro V. Mauri
pedro.mauri@madrid.org

Celia Cano Shaw
celia.cano.shaw@madrid.org

Jose Plaza Ramos
jose.plaza.ramos@madrid.org

Barbara Stefanutti
barbara.stefanutti@madrid.org

Alicia Sanchez-Gorostiaga
alicia.sanchez.gorostiaga@madrid.org

1 Instituto de Investigación Para La Gestión Integrada de Zonas Costeras, Universitat Politècnica de València, Gandía C/Paranimf, 1, 46730 Grao de Gandia, Spain

2 Departamento de Producción Agraria, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid, 28040 Madrid, Spain

3 Department of Agricultural Production, School of Agricultural, Food and Biosystems Engineering (ETSIAAB), Universidad Politécnica de Madrid (UPM), Madrid, Spain

4 Centro de Desarrollo de Energías Renovables (CEDER-CIEMAT), Autovía de Navarra A15, Salida 56, Lubia, 42290 Soria, Spain

5 Instituto Madrileño de Investigación y Desarrollo Rural, Agrario y Alimentario (IMIDRA), Finca El Encín, Autovía A-2. Km 38, 200, 28805 Alcalá de Henares, Spain