Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmujournal.com.my/index.php/araset/index
ISSN: 2462-1943

A Robust Support Vector Machine Model for Monitoring Methane Emission Levels in Paddy Ecosystems using Electronic Nose Technology

Mohd Muzamir Othman[1,*], Muhamad Khairul Ali Hassan[1], Sukhairi Sudin[1], Fathinul Syahir Ahmad Saad[1], Khairul Salleh Basaruddin[2], Muhammad Juhairi Aziz Safar[1], Shafriza Nisha Basah[1], Haniza Yazid[3], Mohd Hanafi Mat Som[3]

[1] Faculty of Electrical Engineering & Technology, Universiti Malaysia Perlis, Kampus Pauh Putra, Arau, Perlis, 02600, Malaysia
[2] Faculty of Mechanical Engineering & Technology, Universiti Malaysia Perlis, Kampus Pauh Putra, Arau, Perlis, 02600, Malaysia
[3] Faculty of Electronic Engineering & Technology, Universiti Malaysia Perlis, Kampus Pauh Putra, Arau, Perlis, 02600, Malaysia

ARTICLE INFO

Article history:
Received 27 April 2026
Received in revised form 23 May 2026
Accepted 30 May 2026
Available online 3 June 2026

Keywords:
Carbon Emission; Paddy Farming; Electronic Noise; Greenhouse Gas Emissions

ABSTRACT

Paddy farming is a significant contributor to global greenhouse gas emissions, particularly methane (CH_4) and carbon dioxide (CO_2). While accurate monitoring is vital for carbon management, traditional analytical methods are often cost-prohibitive and technically complex. This study develops a robust Support Vector Machine (SVM) model integrated with Electronic Nose (E-Nose) technology to predict carbon emission levels by analyzing multi-dimensional environmental and gas parameters. A comprehensive dataset was collected across three field locations (water inlet, mid-field, and water outlet), comprising CH_4 and CO_2 concentrations, ambient temperature, humidity, and rice growth stages. The data underwent a rigorous pre-processing pipeline, including median imputation for missing values and outlier removal using the Interquartile Range (IQR) method, which refined the dataset from 1,499 to 1,486 instances. Feature selection via the SelectKBest method identified CH_4 and CO_2 concentrations, growth stages, and gas fluxes as the most significant predictors. The SVM model utilizing a Radial Basis Function (RBF) kernel was evaluated using two configurations: "All Features" and "Selected Features". The "Selected Features" model demonstrated superior predictive performance on the testing set, achieving a coefficient of determination R^2 of 0.9975, a Mean Absolute Error (MAE) of 0.0067, and a Root Mean Square Error (RMSE) of 0.0085. In comparison, the "All Features" model yielded a slightly lower R^2 of 0.9918, with a higher MAE of 0.0122 and RMSE of 0.0153. This indicates that isolating key environmental drivers not only reduces computational redundancy but also enhances the model's accuracy and generalizability. The results demonstrate that an optimized SVM model, supported by strategic feature selection, provides a highly reliable and cost-effective tool for monitoring carbon footprints in paddy ecosystems. This framework facilitates the transition toward data-driven sustainable farming and precision emission management.

---

[*] *Corresponding author.*
E-mail address: syahmir90@gmail.com (Mohd Muzamir Othman)

https://doi.org/10.37934/araset.60.5.3350

1. Introduction

Agricultural sectors are significant contributors to global climate change, with rice paddies identified as one of the largest anthropogenic sources of methane (CH_4) and carbon dioxide (CO_2) emissions. Modeling emissions from rice paddies is critical due to the complex and dynamic nature of the ecosystem [1]. Methane is primarily produced through methanogenesis, a metabolic byproduct of archaeal microorganisms in anaerobic, flooded soil conditions, while carbon dioxide emissions are heavily influenced by the balance between plant photosynthesis, respiration, and microbial decomposition of organic matter as shown in Figure 1 [2,25]. These biochemical pathways are highly sensitive to fluctuating environmental factors such as soil temperature, moisture levels, and nutrient availability. Consequently, as future climate pressures intensify, the unpredictability of these emissions poses a significant challenge to global environmental stability and food security.

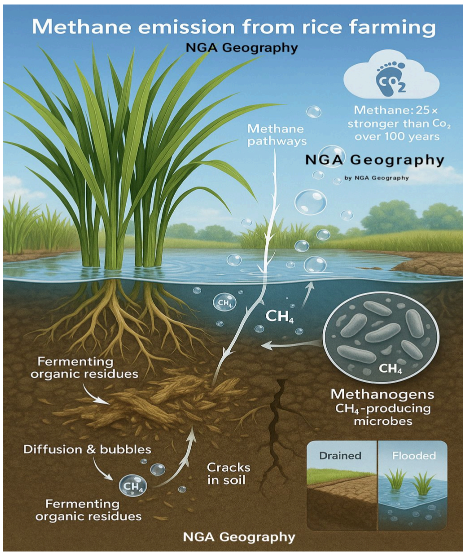


Fig. 1. Mechanisms of Methane (CH_4) and Carbon Dioxide (CO_2) Emission Pathways in a Paddy Ecosystem [25]

Furthermore, the global shift toward sustainable rice production has transformed emission monitoring from a purely scientific endeavor into a socio-economic necessity. Assessing greenhouse gas (GHG) emissions is now a core priority for international carbon credit management and the development of robust environmental mitigation strategies [4,5]. For nations heavily reliant on rice cultivation, such as Malaysia and other Southeast Asian countries, the ability to quantify carbon footprints accurately is essential for participating in emerging carbon markets and complying with ESG (Environmental, Social, and Governance) standards. This paradigm shift necessitates the transition from generalized emission estimates to high-resolution, site-specific data acquisition. By integrating advanced monitoring technologies, stakeholders can implement "low-carbon" farming techniques, such as alternate wetting and drying (AWD), thereby optimizing the trade-off between crop yield and environmental preservation.

Although various simulation models and automated sensing frameworks have been developed to monitor agricultural gases, a critical research gap remains in the real-time, non-destructive, and high-resolution quantification of concurrent (CH_4) and Carbon Dioxide (CO_2) fluxes specifically utilizing electronic nose (E-Nose) technology coupled with machine learning in tropical paddy ecosystems. Most existing frameworks either rely on costly, specialized optical sensors or lack the localized microclimatic integration needed to predict non-linear emissions accurately. To address this gap, this study aims to develop and validate a robust Support Vector Machine (SVM)-based prediction model integrated with an optimized E-Nose data acquisition pipeline for high-frequency greenhouse gas monitoring. The significance of this study lies in providing a cost-effective, site-specific digital tool that enables precision carbon footprint auditing for smallholder farmers, thereby supporting active participation in emerging carbon markets and accelerating the adoption of low-carbon agricultural policies in Southeast Asia.

1.1 Problem Statement

Current methods for monitoring greenhouse gas (GHG) emissions in agricultural systems, particularly in paddy ecosystems, suffer from several critical limitations that hinder accurate and timely environmental assessment. Traditional measurement techniques such as manual gas sampling using static chambers and laboratory-based analysis are widely used; however, these approaches are inherently limited in temporal resolution. They typically rely on periodic or episodic sampling intervals, which means that data collection is sparse and unable to continuously represent dynamic field conditions. As a result, important short-term emission fluctuations such as sudden methane release events triggered by temperature changes, irrigation cycles, or soil disturbance are often missed, leading to incomplete datasets and significant observational gaps [6].

In addition, these conventional methods are labour-intensive, time-consuming, and often require specialised equipment and trained personnel, making them less practical for large-scale or real-time monitoring applications. The delay between sample collection and laboratory analysis further limits their usefulness for immediate decision-making at the farm level.

Moreover, emission patterns in paddy ecosystems are highly complex, non-linear, and influenced by a wide range of interacting environmental factors. Variables such as ambient temperature, soil moisture, humidity, redox potential, microbial activity, and microclimatic variations all contribute to the variability of methane (CH_4) and carbon dioxide (CO_2) fluxes. This environmental heterogeneity introduces significant uncertainty in modelling and prediction tasks [7,1]. Without continuous, high-frequency monitoring systems, it becomes extremely difficult to accurately characterise emission behaviour or establish reliable predictive relationships between environmental conditions and gas output.

1.2 Objectives

The primary objective of this study is to develop a robust SVM-based monitoring model to accurately classify and predict carbon emission levels in paddy ecosystems using E-Nose technology. The specific objectives are:

i) To develop an E-Nose data acquisition framework that captures CH_4, CO_2, temperature (surrounding and chamber), and humidity across various locations and diurnal sessions.

ii) To implement a data pre-processing and feature selection pipeline to identify the most significant input parameters affecting carbon emission flux.

iii) To evaluate the performance and robustness of the SVM model in predicting carbon emission outputs, ensuring high accuracy and reliability for real-world agricultural carbon management.

2. Literature Review

2.1 Methane Dynamics and Environmental Drivers in Paddy Ecosystems

The production and emission of methane (CH_4) in paddy ecosystems are primarily driven by methanogenesis processes occurring under anaerobic soil conditions. Flooded rice fields create oxygen-limited environments that support the activity of methanogenic archaea, which decompose organic substrates such as acetate, hydrogen, and carbon dioxide into methane [2,3]. The rate of methane production is strongly influenced by soil characteristics and the availability of organic matter, including plant residues, root exudates, and organic amendments. Higher levels of labile carbon increase substrate availability for methanogens, thereby enhancing methane emissions [3,8].

In addition to soil and microbial factors, methane emissions are significantly affected by environmental and climatic conditions. Temperature plays a key role, as higher soil temperatures accelerate microbial metabolism and organic matter decomposition, leading to increased methane production [1]. Hydrological conditions such as flooding duration also regulate soil redox potential, influencing the balance between methane generation and oxidation. Furthermore, rice plants contribute to methane transport through aerenchyma tissues, which act as conduits for methane release from the soil to the atmosphere. These interacting factors make methane dynamics in paddy ecosystems highly variable and challenging to predict accurately.

2.2 Evolution of Gas Sensing Technologies and Materials

Gas sensing technologies have undergone significant evolution from conventional laboratory-based analytical systems to compact and portable Electronic Nose (E-Nose) platforms. Early gas detection systems were typically bulky, expensive, and required controlled laboratory conditions, limiting their application in field environments. The development of E-Nose systems has enabled the replication of olfactory sensing using arrays of gas sensors combined with pattern recognition techniques, allowing for real-time monitoring of gaseous compounds in various environments, including agricultural ecosystems [6,1]. In parallel, the integration of Internet of Things (IoT) technology has further enhanced these systems by enabling continuous data acquisition, remote monitoring, and cloud-based analytics for large-scale environmental applications [9].

Recent advancements in sensing materials have significantly improved the sensitivity, selectivity, and stability of gas detection systems. Electrochemical sensors have been widely adopted due to their high responsiveness and suitability for detecting specific gas species, while optoresistive and optical-based sensing platforms offer improved accuracy through light-matter interaction mechanisms [10,11]. These developments have expanded the capability of gas sensors to operate under diverse environmental conditions with enhanced reliability, making them suitable for real-world applications such as environmental monitoring and precision agriculture.

In addition, there is a growing focus on the development of sustainable and multifunctional sensing materials. Carbon-based materials, including graphene and carbon nanotubes, have gained attention due to their excellent electrical conductivity and large surface area, which enhance gas adsorption and detection performance. Fluorescent-based sensing systems have also been introduced as an alternative approach for highly sensitive and selective gas detection [12,13]. Beyond environmental monitoring, these advanced sensing technologies are increasingly being applied in plant health assessment and agricultural diagnostics, where they assist in detecting stress conditions and monitoring ecosystem changes in real time [14,15].

2.3 Machine Learning and Chemometrics in Gas Analysis

Electronic Nose (E-Nose) systems generate high-dimensional datasets due to the simultaneous response of multiple gas sensors under varying environmental conditions. These datasets are often complex, noisy, and highly non-linear, making traditional statistical methods insufficient for effective interpretation. Chemometrics has therefore become an essential tool in gas analysis, enabling feature extraction, dimensionality reduction, and pattern recognition to convert raw sensor signals into meaningful chemical information. This improves the reliability and interpretability of gas sensing data, particularly in complex environmental applications [16].

Machine learning techniques have further enhanced gas classification and prediction capabilities in E-Nose systems. Methods such as Artificial Neural Networks (ANN) and Random Forest (RF) are commonly used; however, they may struggle with highly non-linear and dynamic environmental datasets, especially under fluctuating conditions such as temperature and humidity variations [17,1]. To overcome these limitations, Support Vector Machines (SVM) are widely applied due to their strong ability to handle non-linear data using kernel functions, allowing better class separation in complex gas mixture analysis [18]. In addition, recent developments in high-accuracy computational algorithms have further improved classification performance and prediction reliability in real-time gas sensing applications [19].

2.4 Agricultural Sustainability and Integrated Carbon Monitoring

Modern agriculture increasingly integrates advanced sensing technologies to support environmental monitoring and sustainability objectives. Sensor-based systems enable continuous tracking of emissions and environmental conditions in agricultural ecosystems, particularly in relation to greenhouse gas dynamics [20]. Automated monitoring platforms can also capture microbial activity and odor-related characteristics in real time, which are important indicators of organic matter decomposition processes [7,3]. These measurements are crucial for understanding methane production pathways and evaluating the degradation of organic substrates in paddy environments, where anaerobic conditions strongly influence gas emissions [8].

In addition to emission monitoring, carbon tracking systems are increasingly linked to broader sustainability strategies, including operational efficiency and cost reduction in environmental management [5]. The versatility of sensing technologies extends beyond agriculture, with applications in corrosion detection and non-destructive testing, demonstrating their robustness across different engineering domains [21,22]. Furthermore, reliable real-time environmental data is essential for Environmental, Social, and Governance (ESG) compliance and reporting frameworks, where accurate carbon accounting is required [4]. The integration of IoT-based soil and environmental monitoring systems further strengthens data accuracy and supports comprehensive analysis of agricultural ecosystems, enabling more informed decision-making for sustainable farming practices [23].

2.5 Research Gaps in Tropical Carbon Flux Monitoring

Despite significant advancements in gas sensing and Electronic Nose (E-Nose) technologies, several critical gaps remain in their application for tropical carbon flux monitoring. Most existing studies are primarily focused on air pollution assessment or general odor detection applications rather than real-time prediction of greenhouse gas emissions such as methane in agricultural ecosystems [1,24]. In tropical environments, the complexity of carbon flux dynamics is further intensified by high humidity, temperature variability, and active microbial processes. However, these environmental noise factors and microclimate fluctuations are often insufficiently addressed in current sensing and modelling approaches, particularly in field-deployed systems [7,25].

In addition, there is a limited integration of temporal variability, such as diurnal and seasonal data patterns, in model training for carbon emission prediction. This leads to reduced generalisation performance and lower prediction accuracy under real-world conditions. Furthermore, many existing approaches lack robust machine learning frameworks capable of handling non-linear and dynamic environmental datasets, where techniques such as Support Vector Machines (SVM) are underutilised for structured classification and prediction tasks in tropical agriculture [4]. Therefore, this study aims to address these limitations by developing a robust E-Nose-based monitoring system tailored for tropical agricultural environments, incorporating advanced sensing and data-driven modelling approaches to improve real-time carbon flux estimation and environmental decision-making [6].

3. Methodology

3.1 System Architechture

The E-Nose system was engineered as a customized multi-sensor platform designed for the rugged conditions of a paddy field. The core architecture consists of a sensor array comprising specific metal-oxide semiconductor (MOS) sensors calibrated to detect Methane (CH_4) and Carbon Dioxide (CO_2). These sensors were integrated with an ESP32 microcontroller, chosen for its low power consumption and built-in Wi-Fi/Bluetooth capabilities, which facilitate seamless data transmission. To ensure data integrity, a secondary data logger with an SD card module was incorporated to prevent data loss during network fluctuations in remote farming areas.

The system is housed in a weather-resistant enclosure with a controlled ventilation system that draws air into a dedicated sensing chamber. In this study, the sensing chamber was linked to a static gas chamber placed over the paddy plants. This setup allows the E-Nose to monitor the accumulation of gases over time, providing a high-resolution "digital smell" profile of the plant’s emission cycle.

The system incorporates a diverse array of Metal Oxide Semiconductor (MOS) sensors. Specifically, the TGS-2611 was calibrated for Methane (CH_4) detection, while the MG-811 was utilized for Carbon Dioxide (CO_2) monitoring. These sensors operate based on the change in conductivity of the sensing layer upon contact with gas molecules. The schematic diagram for E-Nose system is shown in Figure 2.

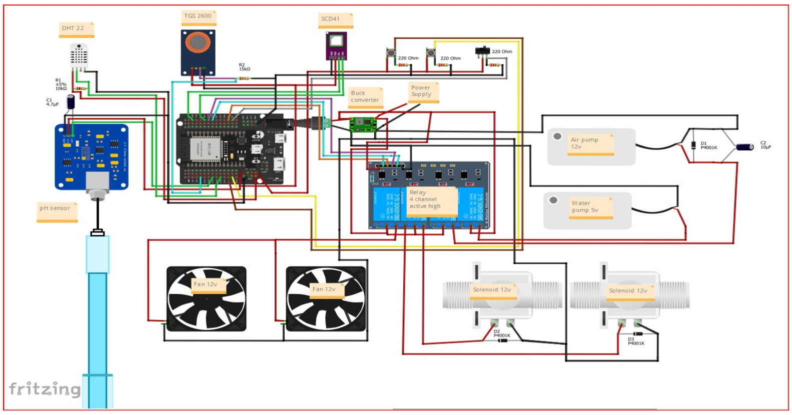


Fig. 2. Schematic diagram of the E-Nose system showing sensor connections, ESP32 control unit, power regulation, relay module, and actuation components

3.2 Data Acquisition

Data collection was conducted on-site at the paddy fields to capture the authentic variance of the agricultural ecosystem. The acquisition process focused on eight primary input parameters: CH4 concentration, CO_2 concentration, surrounding temperature, surrounding humidity, gas chamber temperature, gas chamber humidity, location coordinates, and diurnal sessions (morning, afternoon, and evening). The "session" parameter is vital as it accounts for the metabolic changes in the paddy plants and soil microbes driven by solar intensity and diurnal cycles.

The sampling procedure followed a structured protocol where the gas chamber was sealed for a specific duration (e.g., 30 minutes) to allow gas accumulation, with the E-Nose recording data at 1-minute intervals. This high-frequency sampling ensures that the "flux", the rate of gas change over time is captured accurately. By recording both ambient and internal chamber conditions, the study accounts for the "micro-climate" effect, which is often a source of noise in traditional sensing method.

Micro-climatic variables, specifically temperature and relative humidity, were simultaneously monitored by deploying high-precision DHT22 sensors both inside the gas chamber and at the surrounding ambient level

3.3 Field Data Acquisition and Experimental Design

The data acquisition protocol was designed to capture the high temporal and spatial variability of gas flux in tropical paddy ecosystems. Samples were taken from diverse locations within the paddy field to account for soil heterogeneity. Each location was geo-tagged using a GPS module to include spatial coordinates as a potential input variable.

Observations were categorized into three distinct diurnal sessions: Morning (08:00–10:00), Afternoon (12:00–14:00), and Evening (16:00–18:00). This captures the peak of methanogenic activity typically stimulated by high soil temperatures during the afternoon, as well as the carbon respiration cycle of the rice plants.

A customized static gas chamber (0.5m times 0.5m times 1.0m) was deployed as shown in Figure 3. Once the chamber is sealed, the E-Nose monitors the accumulation of gas concentration over a 30-minute window. This accumulation rate, or slope, is later used to calculate the Gas Flux, which serves as the ground truth for our carbon emission output.

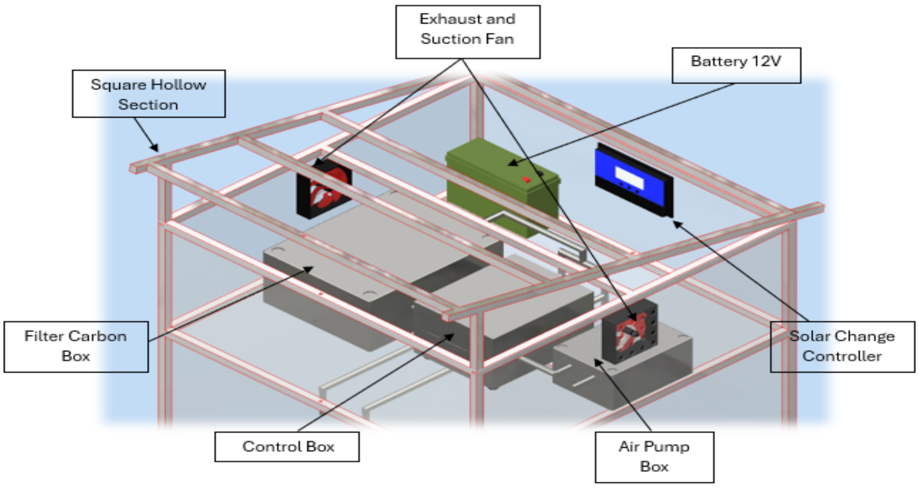


Fig. 3. E-Nose main chamber

3.4 Data Pre-processing

This section describes the rigorous data pre-processing pipeline established to ensure data integrity and optimize the dataset for machine learning model development. Pre-processing is a critical phase in E-Nose applications to mitigate sensor noise, handle environmental fluctuations, and standardize multi-modal datasets for improved predictive accuracy. Figure 4 shows data pre-processing pipeline.

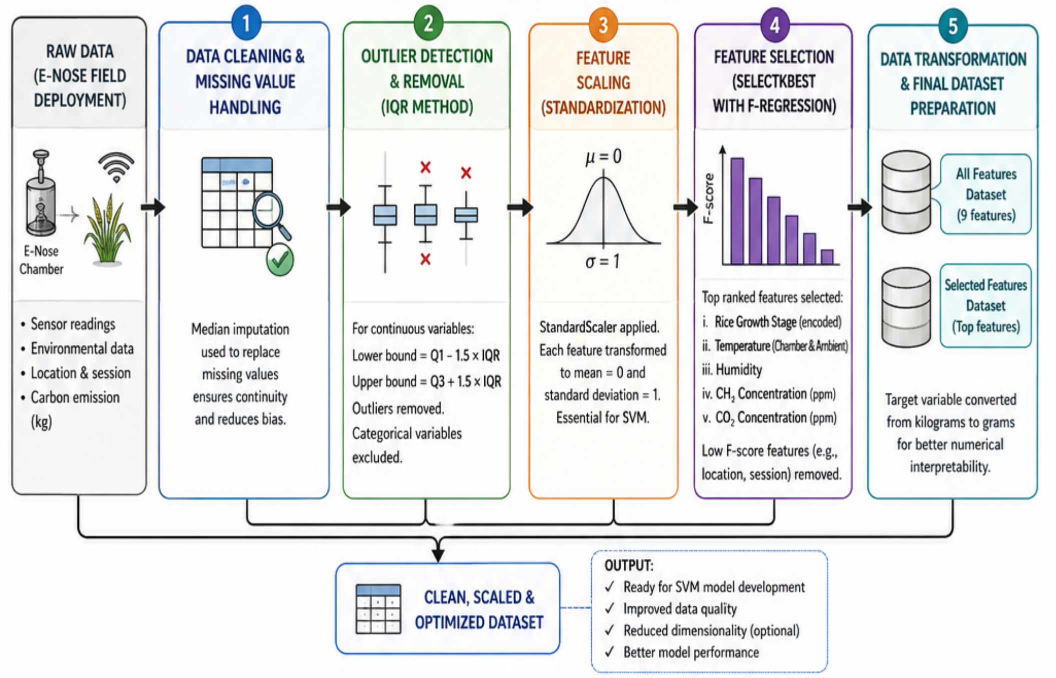


Fig. 4. Data Pre-Processing Pipeline

3.4.1 Data Cleaning and Missing Value Handling

The raw dataset obtained from field deployments often contains missing entries resulting from sensor latency, wireless communication intermittent errors, or physical disturbances in the paddy ecosystem. To maintain dataset continuity without introducing significant bias, median imputation was employed. The median value of each specific feature was used to replace null entries, as it provides a robust measure of central tendency that is less susceptible to extreme outliers compared to the mean, a factor particularly important for environmental sensors exposed to volatile outdoor conditions.

3.4.2 Outlier Detection and Removal

To eliminate anomalies caused by abnormal sensor spikes or temporary hardware disturbances, the Interquartile Range (IQR) method was implemented. Outliers were statistically identified and removed based on the following boundary conditions as shown on Eq. 1 and Eq.2.

$$Lower\ bound = Q1 - 1.5\ X\ IQR \tag{1}$$

$$Upper\ bound = Q3 + 1.5\ X\ IQR \tag{2}$$

This process was applied exclusively to continuous variables, including ambient and chamber temperature, humidity levels, gas concentrations (CH_4 and CO_2), and calculated carbon flux. Categorical variables, such as location and session, were excluded from this statistical filtering to preserve the integrity of the encoded experimental design.

3.4.3 Feature Scaling

Given that the inputs consist of vastly different units, ranging from parts per million (ppm) for gas concentrations to degrees Celsius for temperature and encoded categorical values, Standardization (StandardScaler) was performed. Each feature was transformed to achieve a mean of zero ($\mu = 0$) and a standard deviation of one ($\delta = 1$). This step is mathematically essential for the Support Vector Machine (SVM) algorithm, as SVM relies on the calculation of distances between data points in the feature space; features with larger magnitudes would otherwise disproportionately influence the model's decision boundary.

3.4.4 Feature Selection

To identify the most influential drivers of carbon emission, feature selection was conducted using the SelectKBest method integrated with the fregression scoring function. This approach evaluates the linear dependency between each input feature and the target variable by calculating the F-statistic. Based on the statistical ranking, the following features were prioritized:

i. Rice Growth Stage (encoded);
ii. Temperature (Chamber and Ambient);
iii. Humidity;
iv. Methane Concentration (CH_4 ppm);
v. Carbon Dioxide Concentration (CO_2 ppm);

Features such as location coordinates and diurnal sessions, which exhibited lower F-scores, were excluded in the "Selected Features" model to reduce dimensionality and focus the model on the most significant biochemical and physical drivers of emission.

3.4.5 Data Transformation and Final Dataset Preparation

The final pre-processing stage involved transforming the dataset into two distinct configurations to test the study's hypothesis regarding model robustness:

- All Features Dataset: Consisting of all nine initial input variables.
- Selected Features Dataset: Consisting only of the top-ranked variables.

Furthermore, the target variable (carbon emission) was scaled from kilograms to grams. This transformation improves numerical interpretability and ensures that the machine learning algorithm can more effectively process the variations in emission levels without being hindered by extremely small floating-point values.

3.5 SVM Model Development

The development of the Support Vector Machine (SVM) model was centered on its capability to handle the inherent non-linearity and high dimensionality of the E-Nose sensor data. Given the complex biochemical interactions in paddy ecosystems, the Radial Basis Function (RBF) was selected as the kernel for this study. The RBF kernel is particularly effective in agricultural gas sensing because it can map non-linear input features such as the fluctuating concentrations of CH_4 and CO_2 relative to humidity and temperature, into a higher-dimensional feature space. This transformation allows the algorithm to establish an optimal hyperplane that separates different emission levels, which would be mathematically inseparable in a standard two-dimensional linear space.

To ensure the model's robustness and prevent overfitting, a rigorous Parameter Tuning process was implemented using a grid-search strategy with 5-fold cross-validation. Two critical hyperparameters were optimized: the Regularization parameter (C) and the Kernel coefficient (δ). The C parameter was tuned to balance the trade-off between maximizing the margin of the decision boundary and minimizing classification errors on the training set. Meanwhile, the δ parameter was optimized to define the reach of the influence of a single training example; a low δ value considers distant points in the boundary calculation, while a high δ focuses on points closer to the decision line. By identifying the optimal (C, δ) pair, the SVM model could accurately generalize carbon emission patterns across different locations and diurnal sessions, even when faced with the "noisy" signals typical of field-deployed MOS sensors.

3.6 Performance Evaluation Metrics

The predictive performance and reliability of the SVM model were assessed using a comprehensive suite of statistical metrics designed for regression tasks. Accuracy was prioritized to provide a general overview of the model's predictive ability. Specifically, accuracy was quantified by calculating standard regression metrics that evaluate the magnitude of prediction errors, such as Mean Absolute Error and Root Mean Square Error. The model’s effectiveness in explaining the variance within the continuous dataset was determined using the Coefficient of Determination. Robustness of the regressor was ensured by maintaining consistent performance regardless of fluctuating emission levels, validated through a visual summarization of residual analysis. Collectively, these statistical evaluation methods allow for a transparent comparison between the All Features and Selected Features datasets, ultimately validating the E-Nose as a precise tool for data-driven sustainable farming.

4. Results and Discussions

4.1 Dataset Preparation and Sensor Response Analysis

The dataset was constructed from environmental and gas emission parameters collected via the E-nose system across three critical paddy field locations: the water inlet, mid-field, and water outlet. These locations provided a realistic representation of the environmental variations inherent in rice cultivation. The initial dataset comprised 1,499 instances of CH_4 concentration, CO_2 concentration, temperature, humidity, and rice growth stage data.

Analysis of the raw sensor signals revealed a characteristic "signal signature," where resistance decreased in the presence of reducing gases. However, the response was highly sensitive to environmental fluctuations, particularly temperature. Sensors exhibited a higher baseline sensitivity during afternoon sessions due to elevated chamber temperatures enhancing surface reaction kinetics. This non-linear relationship justified the use of a non-linear kernel in the subsequent machine learning phase.

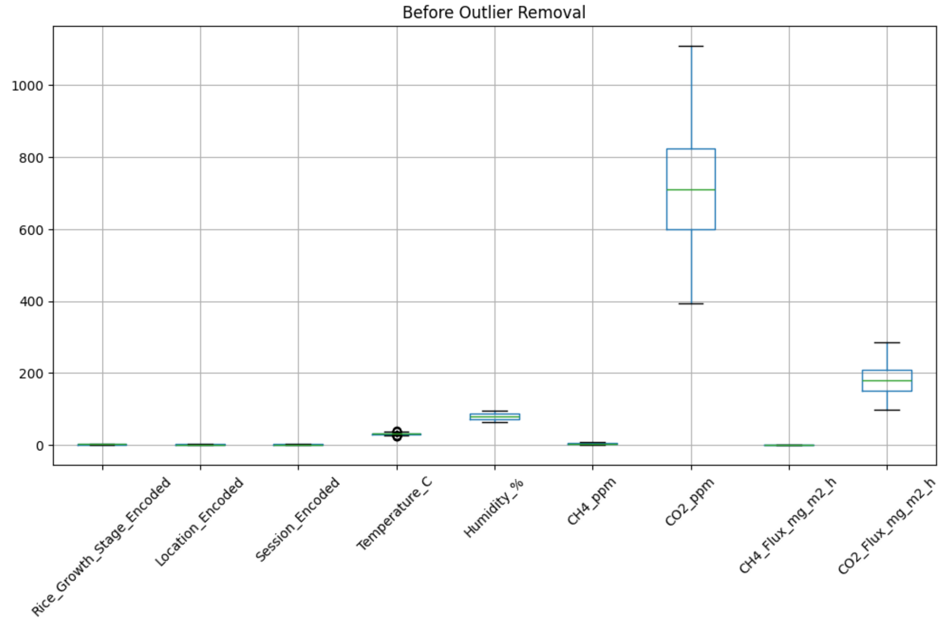
4.2 Data Pre-processing and Feature Importance

To ensure the integrity and robustness of the predictive model, a rigorous data pre-processing pipeline was implemented across the raw dataset. This stage was critical for mitigating sensor noise and addressing the inconsistencies inherent in field-acquired data. Missing values, which may have occurred due to sensor latency or communication intermittent errors, were systematically addressed using median imputation. This method was specifically chosen because it preserves the overall data distribution and provides a higher resistance against extreme outliers compared to mean imputation, thereby maintaining a consistent and reliable dataset for subsequent analysisTo ensure robustness, a rigorous pre-processing pipeline was implemented. Missing values were addressed using median imputation, preserving the overall data distribution.

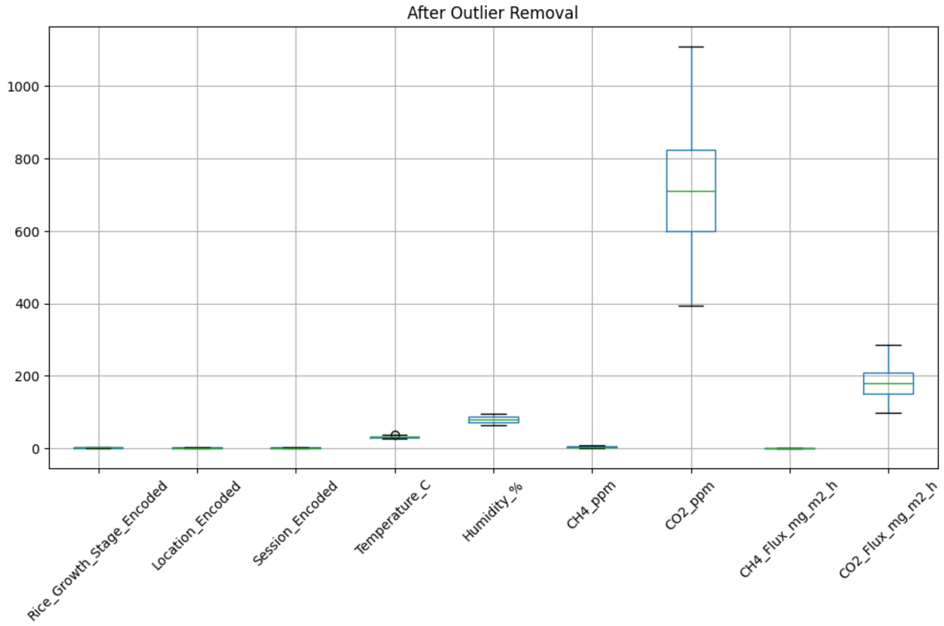
4.2.1 Outlier Removal

To ensure the stability of the predictive model and eliminate potential training bias, the Interquartile Range (IQR) method was systematically applied to the dataset. This statistical technique was essential for identifying and purging "spikes" or anomalies, such as sudden surges in temperature or methane concentration that likely resulted from transient sensor interference rather than authentic environmental flux.

The implementation of this filtering process refined the dataset from an initial 1,499 instances to a final count of 1,486 instances as shown in Figure 5. By removing these 13 outliers, the variance in the training data was significantly reduced, allowing the SVM algorithm to converge more efficiently on the true biochemical patterns of the paddy ecosystem. The transition from the raw to the cleaned data distribution is visualized in the boxplots below, which clearly illustrate the removal of extreme data points across variables such as CH_4 and CO_2 ppm.



(a)



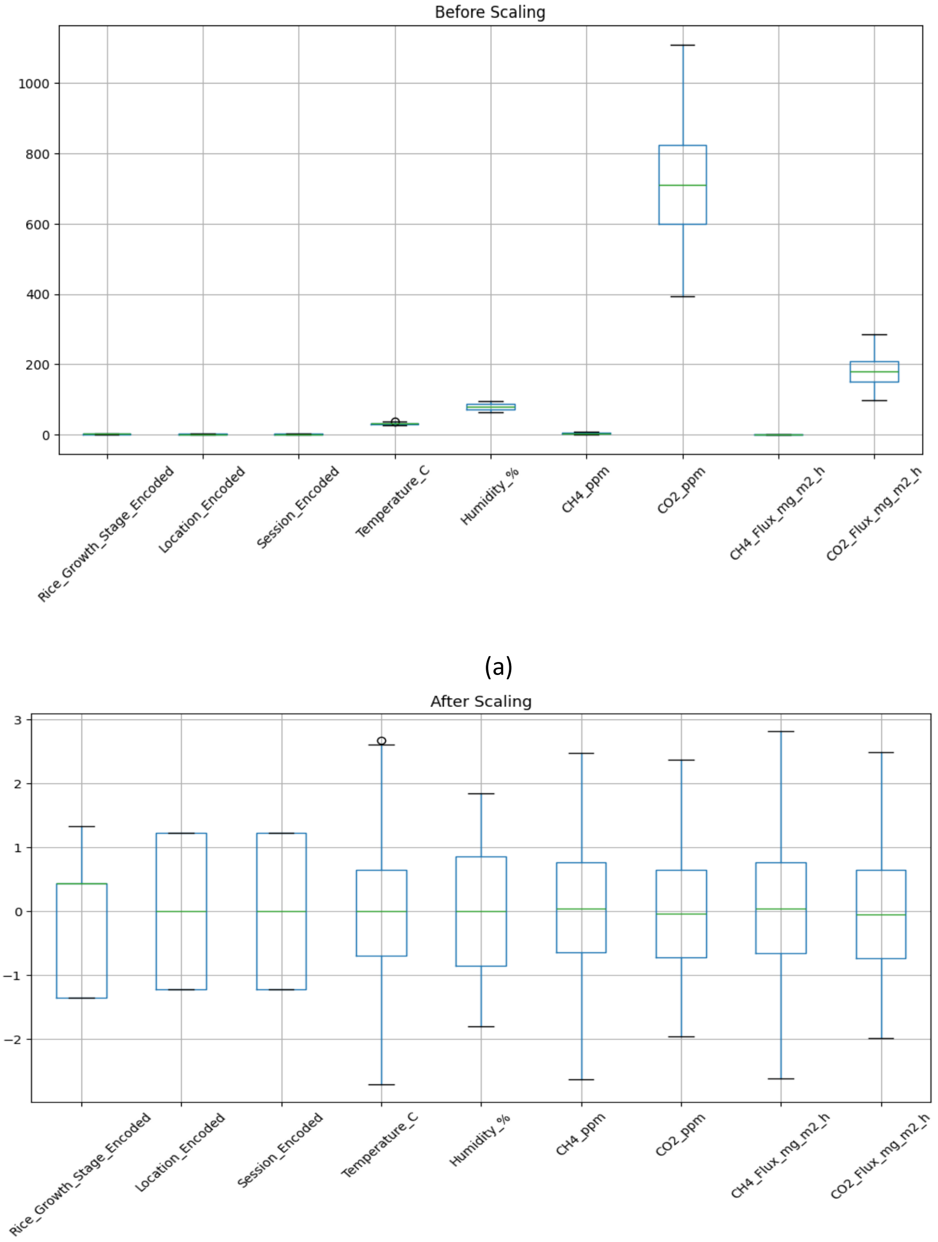
(b)

Fig. 5. Boxplot Analysis of Feature Distribution (a) Before Outlier Removal and (b) After Outlier Removal

4.2.2 Feature Scaling and Selection

To optimize the training environment for the Support Vector Machine (SVM) algorithm, standardization was performed on the refined dataset. This step was mathematically critical because SVM models are highly sensitive to the scale of input features; without standardization, variables with larger numerical ranges, such as CO_2 ppm, would disproportionately dominate the model's decision-making process. By applying a StandardScaler, each feature was transformed to achieve a mean of zero and a unit variance, ensuring that every environmental and gas parameter contributed equally to the objective function. The impact of this transformation is clearly evidenced in the comparative boxplots, where all input variables were brought to a common scale between approximately -3 and +3. The result before and after scaling process is shown in Figure 6.

Following standardization, feature selection was executed using the SelectKBest method integrated with the f_regression scoring function. This analytical stage aimed to identify the most potent predictors of carbon emissions while discarding "noisy" or redundant variables that could lead to overfitting. The statistical results identified six primary drivers that exhibited the strongest correlation with the target variable: 'Rice_Growth_Stage_Encoded', 'CH4_ppm', 'CO2_ppm', 'CH4_Flux_mg_m2_h', ‘Temperature-C’ and 'CO2_Flux_mg_m2_h' as shown in Figure 7. Notably, the feature importance analysis confirmed that CH_4 and CO_2 concentrations, alongside their respective fluxes, were the most significant predictors of the ecosystem’s carbon output.



(a)

(b)

Fig. 6. Boxplot Analysis of (a) Before Scaling and (b) After Scaling Process.

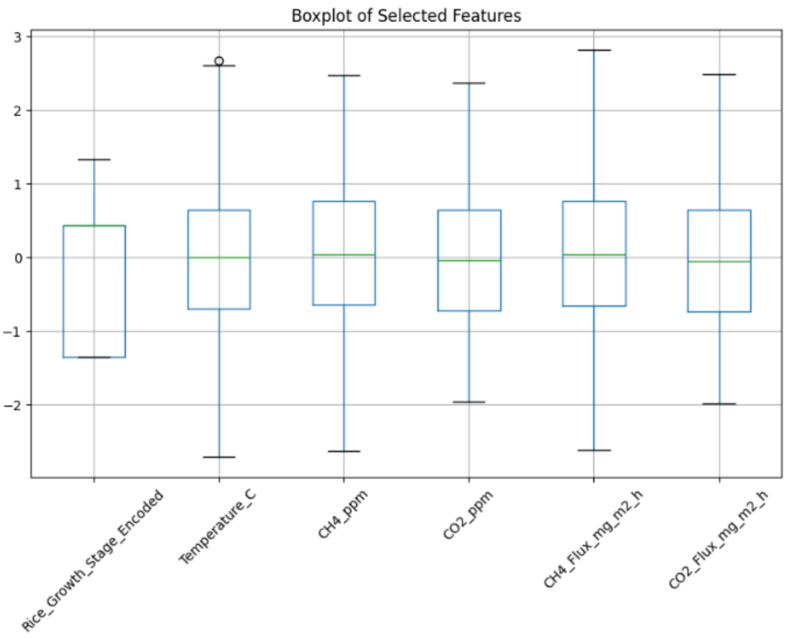


Fig. 7. Feature Importance Ranking Derived via SelectKBest.

4.3 Model Performance and Comparative Analysis

The Support Vector Machine (SVM) model, utilizing a Radial Basis Function (RBF) kernel, was rigorously trained and evaluated across two distinct configurations to determine the impact of dimensionality reduction on predictive accuracy. The first configuration, the "All Features" model, incorporated all nine initial input variables, including encoded location and session data. The second configuration, the "Selected Features" model, utilized only the six most significant predictors identified during the feature selection process.

4.3.1 Performance Metrics

To provide a holistic assessment of the model's reliability, performance was quantified using three standard statistical metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). These metrics collectively evaluate the magnitude of prediction errors and the model's ability to explain the variance within the carbon emission dataset.

Table 1
Comparative Performance of SVM Configurations (Testing Set)

Configuration	MAE	RMSE	R^2
Selected Features	0.006717	0.008468	0.997492
All Features	0.012188	0.015319	0.991794

The comparative analysis reveals that the "Selected Features" configuration yielded superior predictive performance. While the "All Features" model achieved a lower MAE (0.001985) and RMSE (0.002625) during the training phase , its performance on the testing set showed higher error rates compared to the optimized model. Specifically, the "Selected Features" model achieved an R^2 score of 0.9975, indicating that the model can explain over 99.7% of the variance in carbon emissions with minimal error.This high correlation is visually confirmed in Figure 8, which plots the actual measured carbon emissions against the values predicted by the optimized model.

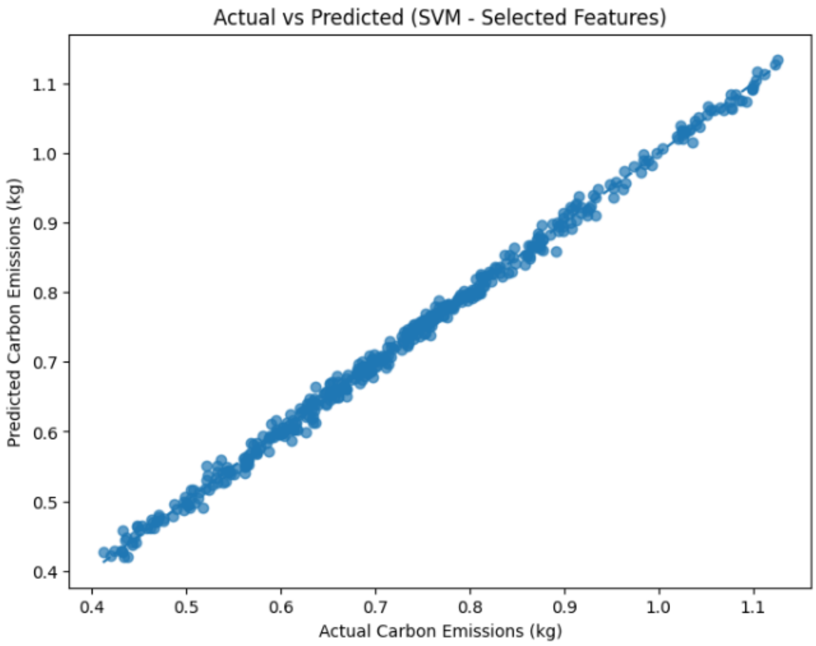


Fig. 8. Actual vs. predicted carbon emission levels using the Support Vector Machine (SVM) model with selected features

This enhancement in performance suggests that the exclusion of low-ranking features, such as location and session encoded data, effectively mitigated "dimensionality noise". By focusing on the primary biochemical drivers, CH_4 and CO_2 concentrations and fluxes, the SVM model achieved greater generalizability and precision, which is critical for accurate reporting in agricultural carbon management.

4.4 Robustness and Practical Implications

To confirm the reliability of the system, the stability of the SVM model was rigorously validated through a detailed residual analysis. This analysis involved examining the distribution of residuals—the differences between the actual carbon emission values and the values predicted by the model. The resulting residual plot demonstrated that prediction errors remained consistently low and randomly distributed across the entire data range, without exhibiting any distinct patterns. This lack of heteroscedasticity indicates that the model is robust and maintains high predictive accuracy regardless of the magnitude of the carbon emission levels being measured. This lack of heteroscedasticity is confirmed in Figure 9, where the residuals are randomly scattered around the zero horizontal line, indicating constant variance in prediction errors across all predicted values.

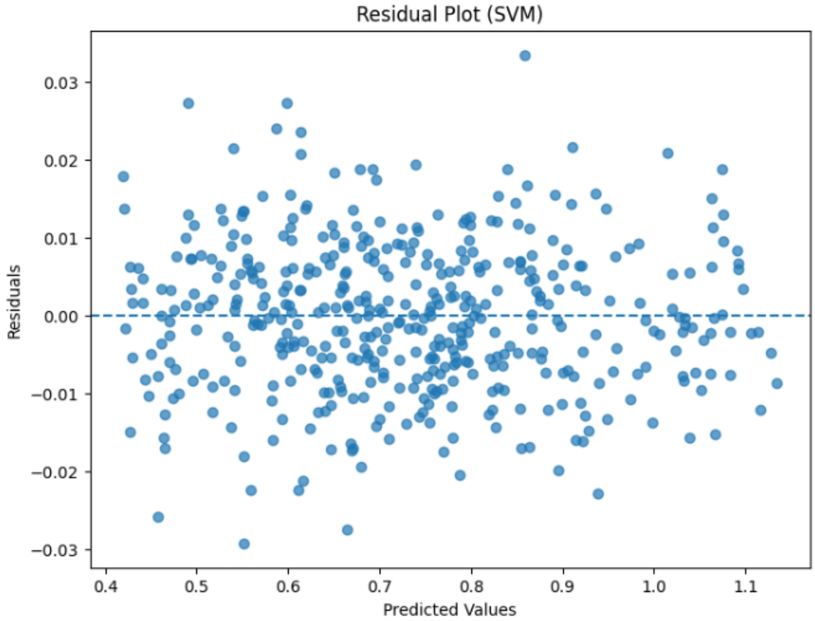


Fig. 9. Residual Plot of the SVM Model (Selected Features).

The integration of a high R^2 score of 0.9975 and low error values (MAE: 0.0067 and RMSE: 0.0085) provides empirical proof that the SVM-E-nose framework can reliably predict carbon footprints within the complex paddy ecosystem. From a practical standpoint, this high degree of precision enables the transition from delayed laboratory analysis to real-time monitoring of methane flux. By providing immediate data on emission surges, the system allows farmers and stakeholders to implement "Smart Farming" feedback loops. For instance, real-time alerts can trigger the use of Alternate Wetting and Drying (AWD), a water management technique that aerates the soil to suppress methanogenesis, effectively reducing greenhouse gas emissions while maintaining crop productivity.

5. Conclusion

This study successfully developed and validated a robust carbon emission monitoring framework for paddy ecosystems by integrating Electronic Nose (E-Nose) technology with a Support Vector Machine (SVM) regressor. By analyzing multi-dimensional environmental and gas parameters, the system addressed the crucial need for cost-effective, high-resolution, site-specific emission data in tropical agriculture.

The rigorous multi-stage pre-processing pipeline established in Chapter 3 proved essential for managing the inherent noise and volatility of field-acquired MOS sensor data. Median imputation maintained dataset continuity, while the statistical IQR method effectively refined the dataset from 1,499 to 1,486 instances by purging abnormal spikes and temporary hardware disturbances. The scaling process standardized variables with vastly different magnitudes allowing for unbiased model training. Furthermore, feature selection derived via SelectKBest provided critical biochemical insights, identifying six primary drivers: rice growth stage, concentrations of CH4 and CO2, their respective fluxes, and temperature, as the most influential predictors of the ecosystem's carbon output.

Comparative analysis demonstrated that dimensionality reduction significantly enhances model robustness. While incorporate redundant input sets increased dimensionality noise, the optimized "Selected Features" SVM model utilizing a Radial Basis Function (RBF) kernel achieved superior precision on the testing set. The final model yielded a low Mean Absolute Error (MAE) of 0.0067, a low Root Mean Square Error (RMSE) of 0.0085, and an exceptionally high coefficient of determination (R^2) of 0.9975. This empirical validation proves that the optimized SVM-E-Nose framework can explain over 99.7% of the variance in carbon emissions within a complex paddy ecosystem.

A comprehensive stability assessment through residual analysis confirmed the model's reliability across the entire data range. The resulting residual plot demonstrated that prediction errors remained consistently low, randomly scattered centered at zero, and free of heteroscedasticity. This statistical validation indicates that the SVM model is robust against environmental fluctuations and maintains consistent high predictive accuracy, regardless of whether the system is detecting low baseline emissions or peak methane surges, which typically occur during afternoon sessions stimulated by high microbial metabolism.

References

[1] Seesaard, Thara, Kamonrat Kamjornkittikoon, and Chatchawal Wongchoosuk. "A Comprehensive Review on Advancements in Sensors for Air Pollution Applications." *Science of the Total Environment* 951 (2024): 175696. https://doi.org/10.1016/j.scitotenv.2024.175696.

[2] Jiang, Kunyu, Zhaowen Cheng, Ziyang Lou, et al. "Chemical and Olfactive Impacts of Organic Matters on Odor Emission Patterns from Simulated Waste." *Journal of Environmental Sciences* 103 (2021): 196–206. https://doi.org/10.1016/j.jes.2020.10.016.

[3] Gao, Wei, Xiaofang Yang, Xinmeng Zhu, et al. "The Variation of Odor Characteristics of Wastewater Sludge Treated by Advanced Anaerobic Digestion." *Science of the Total Environment* 840 (2022): 156722. https://doi.org/10.1016/j.scitotenv.2022.156722.

[4] Dal Magro, Adele, Daniela Lovarelli, Federica Borgonovo, et al. "Can Coffee Silverskin Be a Suitable Substrate for Biogas Production and Odour Impact Mitigation of Digestate?" *Journal of Cleaner Production* 547 (2026): 147814. https://doi.org/10.1016/j.jclepro.2026.147814.

[5] Kellerman, Hunter, Jess C. Mace, Kara Y. Detwiller, et al. "Travel Costs and Carbon Savings Associated With Telemedicine in a Tertiary Care Rhinology Center." *International Forum of Allergy and Rhinology* 15, no. 7 (2025): 692–698. https://doi.org/10.1002/alr.23540.

[6] Doty, Anna C., A. Dan Wilson, Lisa B. Forse, and Thomas S. Risch. "Assessment of the Portable C-320 Electronic Nose for Discrimination of Nine Insectivorous Bat Species." *Biosensors* 10, no. 2 (2020): 12. https://doi.org/10.3390/bios10020012.

[7] Sun, Yurui, Guilin Shan, Wilfried Berchtold, et al. "An Automated Experimental System for Tracking Multiple Factors of Aerobic Deterioration in Silage." *Computers and Electronics in Agriculture* 239 (2025): 110902. https://doi.org/10.1016/j.compag.2025.110902.

[8] Lee-Rangel, Héctor Aarón, José Alejandro Roque-Jiménez, Rubén Oswaldo Cifuentes-López, et al. "Evaluation of Marine Algae on Degradability and Methane Emissions." *Fermentation* 8, no. 10 (2022): 511. https://doi.org/10.3390/fermentation8100511.

[9] Ismail, Fadzil A., Muhamad Khairul Ali Hassan, Fathinul Syahir Ahmad Saad, et al. "An IoT Agricultural System for Harumanis Farm." *Journal of Physics: Conference Series* 2107 (2021): 012038. https://doi.org/10.1088/1742-6596/2107/1/012038.

[10] Banga, Ivneet, Anirban Paul, Sriram Muthukumar, and Shalini Prasad. "ZENose Platform for Noninvasive Ammonia Detection." *ACS Applied Materials and Interfaces* 13, no. 14 (2021): 16155–16165. https://doi.org/10.1021/acsami.1c02283.

[11] Santos, José Pedro, Carlos Sánchez-Vicente, Isabel Sayago, and Jesús Lozano. "Optoresistive Measurement System for Metal Oxide Nanosensors." *Lecture Notes in Electrical Engineering* 1334 (2025): 81–87. https://doi.org/10.1007/978-3-031-82076-2_12.

[12] Thongsai, Nichaphat, Aphinya Chuiduang, Sumana Kladsomboon, et al. "Fluorescent Carbon Dots from Papaya Waste for Sensing Applications." *Journal of Metals, Materials and Minerals* 33, no. 3 (2023). https://doi.org/10.55713/jmmm.v33i3.1684.

[13] Shahdost-Fard, Faezeh, Arafeh Bigdeli, and Mohammad Reza Hormozi-Nezhad. "Smartphone-Based Fluorescent Electronic Tongue for Dopamine Detection." *ACS Chemical Neuroscience* 12, no. 17 (2021): 3157–3166. https://doi.org/10.1021/acschemneuro.1c00160.

[14] Takshi, Arash, and Mohammad Shakhawat Hossain. "Electronic-Nose for Plant Health Monitoring in Controlled Environment." *Proceedings of SPIE* 12210 (2022): 1221002. https://doi.org/10.1117/12.2632828.

[15] Li, Xiaofang, Qi Liu, Yunpeng Liu, et al. "Flexible Multifunctional Humidity-Strain Sensors Based on Perovskite Filaments." *Advanced Fiber Materials* 7, no. 3 (2025): 762–773. https://doi.org/10.1007/s42765-025-00518-9.

[16] Bouchikhi, Benachir, Mohammed Moufid, and Nezha El Bari. "Role of Chemometrics in Wastewater Treatment." In *Current Advancements in Nanomaterials for Wastewater Remediation*, 363–381. 2025. https://doi.org/10.1016/B978-0-443-16112-4.00010-4.

[17] Seilov, Shakhmaran, Marat Baydeldinov, Dias Abildinov, et al. "Electronic Nose and Neural Network-Based Monitoring of Vegetable Freshness." *Journal of Robotics and Control* 6, no. 6 (2025): 2834–2843. https://doi.org/10.18196/jrc.v6i6.28505.

[18] Zhang, Jiade, Mingzhi Jiao, Liangsong Duan, et al. "Gas Classification System Based on Hybrid Waveform Modulation Technology." *Sensors and Actuators B: Chemical* 435 (2025): 137637. https://doi.org/10.1016/j.snb.2025.137637.

[19] Feng, Ruibo, Zhenyu Liu, Wenchao Gan, et al. "Fast Algorithm for Quantitative Analysis of Harmful Gas Mixtures." *Measurement* 273 (2026): 121159. https://doi.org/10.1016/j.measurement.2026.121159.

[20] Carotti, I., D. R. Billson, D. A. Hutchins, et al. "Detection of Rust Corrosion Using Electronic Nose." *IEEE Sensors Journal* 25, no. 12 (2025): 22573–22579. https://doi.org/10.1109/JSEN.2025.3562033.

[21] Schlag, Mareike, Kai Brune, Hauke Brüning, et al. "Extended Non-Destructive Testing for Surface Quality Assessment." In *Adhesive Bonding of Aircraft Composite Structures*, 119–222. 2021. https://doi.org/10.1007/978-3-319-92810-4_3.

[22] Hopanna, Mamatha, and Brian Balchunas. "Odor Management at Annapolis WRF: Data-Driven Solutions." *WEFTEC Proceedings* (2025).

[23] Zaini, Muhammad Faiz Aiman Mohamed, Muhamad Khairul Ali Hassan, Fathinul Syahir Ahmad Saad, et al. "An IoT-Based System to Monitor Soil Parameter of Harumanis Tree." In *Lecture Notes in Mechanical Engineering*. 2024. https://doi.org/10.1007/978-981-97-0169-8_18.

[24] Chu, Ming, Ruoxi Wang, Xiaoyuan Jing, et al. "Subacute Inhalation Toxicity of E-Cigarette Aerosols." *Ecotoxicology and Environmental Safety* 271 (2024): 116002. https://doi.org/10.1016/j.ecoenv.2024.116002.

[25] NGA AgroClimate Tech Service. "Infographic on Methane Cycle and Greenhouse Gas Emissions in Paddy Fields." 2023. https://web.facebook.com/photo/?fbid=715725647809468.