Enhancing Gas Detection Accuracy Using Stacking Ensemble Learning

D.Kesavan
CSBS Department
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
kesavand@student.tce.edu

T.Siva
CSBS Department
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
tsca@tce.edu

M.Kishore Kumar
CSBS Department
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
kishorekumarm@student.tce.edu

E.B.R.Nagaesvara
CSBS Department
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
nagaesvara@student.tce.edu

***Abstract*— In industrial and safety applications, efficient gas detection is essential to ensuring environmental safety. Conventional gas detection systems, which use chemical sensors, frequently produce erroneous readings due to sensor drift, background noise, and cross-sensitivity to different gases. This study uses machine learning approaches to create a reliable gas classification model utilizing sensor data in order to get around these restrictions. This study aims to improve gas detection systems in industrial and safety applications by utilizing machine learning techniques. The approach involves a combination of individual classifiers like Random Forest, k-Nearest Neighbors, and Support Vector Machine, along with ensemble learning techniques like stacking, bagging, and voting. The dataset is preprocessed using Tukey's method, feature standardization via StandardScaler, and label encoding for categorical target transformation. The dataset is divided using stratified sampling for balanced class distribution. Feature extraction techniques like Principal Component Analysis and Recursive Feature Elimination are explored. Ensemble learning significantly enhances classification accuracy, with the proposed model achieving up to 99% accuracy. This research presents a scalable and reliable solution for early gas detection, reducing health hazards, improving workplace safety, and mitigating environmental risks.**

Keywords*— *Gas classification, Machine learning, Ensemble learning, Sensor data, Industrial safety, Feature extraction

I. Introduction

In today's world, environmental safety and monitoring are essential, especially in fields where exposure to hazardous gases can pose major risks to one's health and well-being. In many industrial, environmental, and safety applications, gas detection is essential because precise identification of dangerous gases can avert possible health and environmental problems. Chemical sensors are the basis of conventional gas detection techniques, however sensor drift, ambient noise, and cross-sensitivity to different gases frequently restrict their efficacy. Machine learning (ML) approaches have become a potent tool for gas classification in response to these issues, allowing for automatic, precise, and effective detection. Early and precise identification of dangerous gases is essential to preventing accidents and ensuring a safe working environment.

The traditional gas detection method relies on chemical sensors to generate values for various gas compositions, but these methods are often limited by manual monitoring and rule-based systems. The development of data analytics and machine learning offers a promising alternative, enabling more accurate, automated, and scalable gas detection systems. Data preparation is crucial for ensuring accuracy and consistency, using techniques like Tukey's method (IQR-based filtering), StandardScaler, and label encoding. Exploratory data analysis (EDA) is used for extensive feature selection and data validation, with the dataset split into training, validation, and test sets at a 60:20:20 ratio. Ensemble approaches like bagging and stacking are used to reduce bias and variation. Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are used to decrease dimensionality and increase computing efficiency without sacrificing performance.

The objective of this work is to use advanced machine learning models and rigorous preprocessing techniques to deliver a reliable and scalable gas detection solution. In industrial settings, the system can be used to identify hazardous gases early, enhancing safety protocols, reducing health risks, and reducing environmental hazards.

II. Literature Review

Asmita Varma et al. proposed IoT-based gas leak detection system is the method that uses predictive analytics and intelligent alerting to improve on conventional detectors [1]. It detects chemicals such as H_2, LPG, CH_4, CO, alcohol, smoke, and propane using a MQ-2 sensor with an Arduino UNO, a

GSM module, and an Ethernet shield. It then sets off alarms and sends out notifications by text, email, and phone calls. When gas levels reach danger, the system can turn off the power source and send sensor data to the cloud for analysis. It is an affordable substitute for current detectors with additional IoT features, costing about $50. Big data analytics, a mobile app for improved system control, and the integration of temperature and humidity sensors are examples of future improvements.

Pai Peng et al. developed the Gas Net model with a high accuracy of 95.2%, that divides the gases into the four designated categories (carbon monoxide, methane, hydrogen, and ethylene) [2]. This is accomplished by using the deep convolutional neural network architecture to handle the electronic nose's raw sensor input, automatically extracting and learning pertinent features for classification. The model's performance demonstrates the promise of deep learning in gas classification challenges by outperforming more conventional techniques like SVM and MLP.

Syuan-He Wang et al. introduces SimResNet-10_X_MLP, a hybrid deep neural network for gas classification that incorporates environmental compensation to increase accuracy [3]. For improved CNN processing, it makes use of a novel data pre-processing technique (Image data-3) that maintains gas reaction waveforms. The model outperforms classical machine learning (KNN: 67.08%, SVM: 80.16%) by classifying 10 chemical gases, including carbon monoxide (CO), methane (CH_4), ethanol, ammonia (NH_3), acetone, and hydrogen (H_2), with a 95% accuracy rate using an open-source gas database. The hybrid method reduces sensor drift brought on by outside variables such as temperature, wind speed, and sensor location. The potential of hybrid deep learning models for accurate gas categorization in practical settings is demonstrated by this study.

Javier G. Monroy et al. proposed the effect of mobility on gas classification with a mobile robot equipped with an electronic nose (e-nose) [4]. Motion affects sensor response and reduces classification accuracy by up to 30% when trained on static data, according to experiments using ethanol and acetone as classified gases in a controlled indoor corridor. SVM and Naive Bayes, among other classification algorithms, produced the best accuracy results. Accuracy was greatly increased by using motion data to train classifiers; mixed-speed training produced the best results. The study emphasizes how motion effects must be taken into consideration in practical applications such as robotics and pollution monitoring.

Muhammad Hussan et al. developed a technique using a variety of metal oxide sensors for gas identification known as the Binary Decision Tree (BDT)-based classifier [5]. In contrast to conventional techniques, it improves accuracy even when gas concentrations fluctuate by using changes in sensor sensitivity as classification criteria. The BDT classifier outperformed KNN, MLP, and PCA-based methods with an accuracy of 95.5% on variable concentrations when tested on five gases (CO, CO_2, H_2, NH_3, and C_3H_8). The technique is scalable, reliable, and appropriate for real-world settings with varying gas concentrations.

Xiaojin Zhao et al. suggested a one-dimensional deep convolutional neural network (1D-DCNN) for mixture gas classification [6]. To precisely identify specific gases in binary mixtures (ethylene, CO, and methane), the model uses multi-label classification. With an accuracy of 96.30%, it surpasses more conventional techniques like SVM, ANN, KNN, and RF. By automatically extracting features from gas sensor data, the 1D-DCNN increases the effectiveness of categorization. For gas monitoring applications in industry, the environment, and healthcare, this method works quite well.

III. Design And Methodology

A. Dataset Description and Dataset Pre-Processing

In order to overcome the shortcomings of low-cost gas sensors, Multimodal Gas Data is a novel dataset for gas classification and detection that makes use of a thermal camera and gas sensors. There are 6400 samples in the collection, which are divided into four categories: perfume, smoke, neutral atmosphere, and smoke and perfume. With gas dispersion intervals varying from 15 to 45 seconds, data was gathered over a 90-minute period from two gas sources: smoke and scent. This multimodal approach helps develop and train AI models for electronic nose systems and gas detection in a variety of applications by improving detection accuracy even at low concentrations.

The dataset was organized using the Pandas package for efficient data handling. YAML files were used to manage file locations, preprocessing parameters, and dataset configurations. The dataset was read into a Pandas DataFrame for feature and label manipulation. Relevant predictor variables were retained after eliminating unnecessary columns. Integer-based columns were changed to int 32 for data integrity and uniformity. Assertions were implemented to detect discrepancies in data quality, and numerical columns for sensors were verified to ensure their values fell within expected operational ranges. The dataset was divided into training, validation, and test subsets for effective model training and evaluation. Outliers were identified and

eliminated using Tukey's Interquartile Range (IQR) technique..

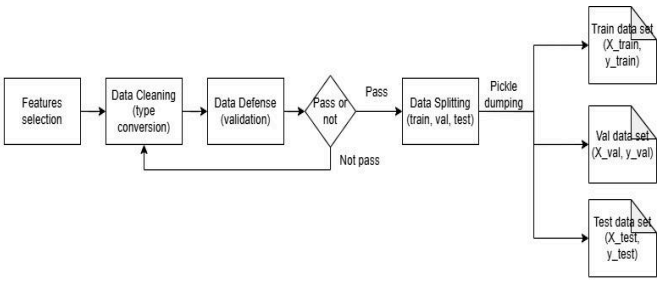

Fig. 1. Dataset Preparation

Additionally, StandardScaler was used to apply feature scaling to the predictor variables, which improved machine learning algorithm convergence by putting the data into a uniform distribution. LabelEncoder was used to numerically encode categorical target labels, preserving class information while enabling compatibility with predictive models. To ensure consistency in upcoming transformations, the encoding scheme was saved. Lastly, pickle format was used to store all processed datasets, including feature-engineered versions and transformations like scaled features and encoded labels. In the end, this procedure produced a reliable and optimized dataset for machine learning applications by guaranteeing effective data retrieval and consistency across several model building phases.

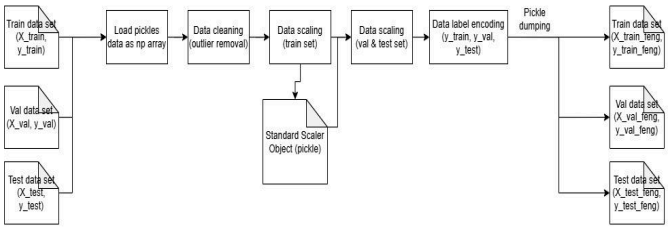

Fig. 2. Dataset Pre-processing

The study analyzed sensor data distribution using KDE Density Plot, revealing patterns like skewness and multimodal distributions. Conventional parametric models were deemed insufficient due to the complex data distribution. Machine learning algorithms like decision trees, random forests, k-nearest neighbors, and neural networks were chosen. Boxplots displayed Tukey's Interquartile Range (IQR) approach for outlier detection. Statistical data analysis confirmed dataset integrity. Pie charts and bar plots revealed no class imbalance. A heatmap for correlation analysis revealed strong connections between sensor features, potentially due to environmental factors or overlapping sensitivity. Understanding these connections provided insight into possible model biases or feature redundancy. The mutual information method was applied to assess each sensor's contribution to predicting gas types by measuring the dependency between sensor readings and gas classifications. It quantifies how much information a feature YY provides about the gas type XX using the formula:

$$I(X,Y) = \sum x \in X \sum y \in Y \; P(x,y) \; log(P(x,y)/P(x) * P(y)) \quad (1)$$

where P(x, y) represents the joint probability of XX and YY, while P(x) and P(y) denote their respective marginal probabilities. Higher mutual information values indicate stronger correlations, making a feature more relevant for classification. The analysis revealed that MQ3, MQ135, and MQ8 were the most influential sensors, but due to the moderate dataset size and manageable feature count, feature selection was not deemed necessary unless overfitting emerged during model training.

B. Methodology

A structured machine learning pipeline is used in the study to classify gasses according to sensor data. Data is gathered from several MQ-series sensors, preprocessed using the Chi-Square test, and normalized using conventional scale. Model performance is enhanced through the use of Grid Search Cross-Validation for hyperparameter adjustment. Using certain datasets, machine learning models are developed and evaluated. To assess the model's efficacy, performance metrics such as F1-score, recall, accuracy, and precision are employed. To increase classification accuracy and generalization across various gas types, ensemble learning techniques are used.

C. One-level Classification

In the gas classification framework, the one-level classifier serves as the basis for the independent training of three machine learning models: Random Forest (RF), k-Nearest Neighbors (kNN), and Support Vector Machine (SVM) to categorize gas samples from sensor data. SVM is strong at identifying intricate, non-linear patterns, Random Forest is excellent at feature importance analysis, and kNN manages local classification based on similarity, all of which have their own advantages.

Optimized with n_estimators=200 and max_depth=20, the Random Forest Classifier builds several decision trees and combines their results to better categorization. By identifying the most important gas sensor characteristics, it improves the interpretability of the model. Using a 'distance'-based weighting function and n_neighbors=10, the kNN Classifier classifies gas samples by assessing how close they are to other labeled data points. It is useful for capturing local fluctuations, but for large datasets, it can be computationally costly. By mapping input data into higher-dimensional space with an RBF kernel with C=10 and gamma='scale, the SVM Classifier improves the separation of overlapping gas types.

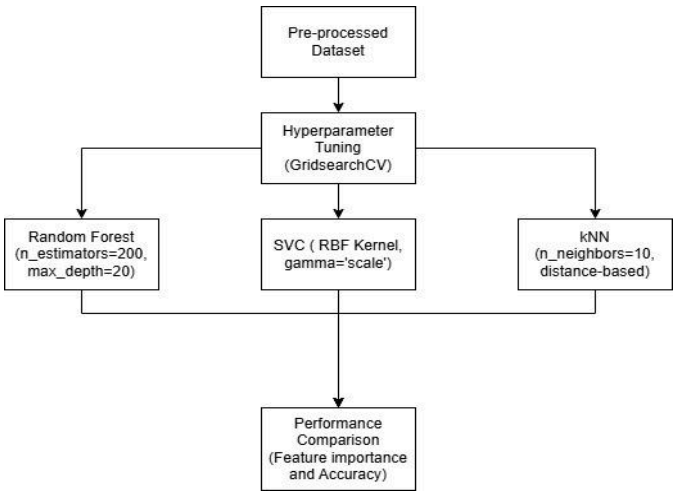

Fig. 3. One-level classification model training

The processed dataset is used to train each model, and a validation set is used to assess it. Based on feature importance analysis and classification accuracy, their performances are contrasted. Although these models provide accurate classification, they all have drawbacks: SVM demands careful hyperparameter tuning, Random Forest may have trouble generalizing, and kNN are likely to be sensitive to high-dimensional data. By combining the advantages of several models, a two-level classifier that uses ensemble learning techniques is presented to address these issues and improve classification resilience and accuracy.

D. Two-level Classification

A two-level classification framework is used to apply ensemble learning techniques to improve classification accuracy beyond individual models. By using stacking, voting, and bagging approaches to incorporate numerous base classifiers, this method improves generalization and robustness in gas classification. The Stacking Classifier uses an XGBoost classifier as the meta-model to provide final predictions, combining predictions from Random Forest, kNN, and SVC models. By utilizing the various strengths of individual classifiers, the stacking technique enables the model to increase overall classification accuracy.

Moreover, a soft voting technique is used to create a Voting Classifier, whereby predictions from RF, kNN, and SVM are combined and models with superior individual performance are given larger weights (RF=3, kNN=1, SVM=2). This guarantees that the stronger classifiers have a greater influence on the final decision, producing a more dependable result. Additionally, a Random Forest model with 100 estimators is employed as the base learner in a Bagging Classifier. Bagging lowers overfitting and improves prediction stability by training several models with random feature subsets (max_features=0.8).

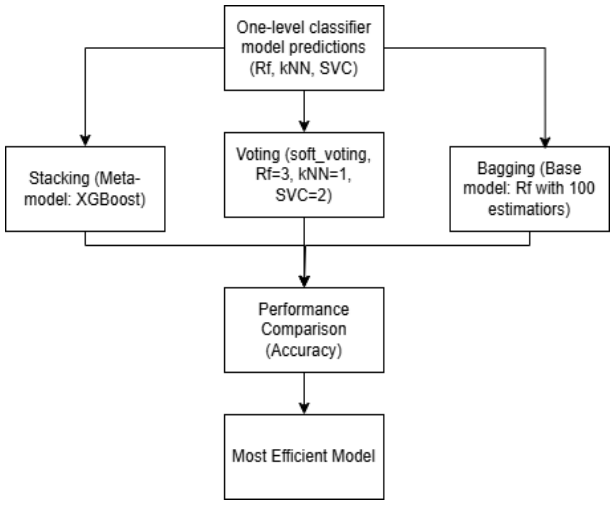

Fig. 4. Two-level classification model training

By effectively integrating the advantages of several classifiers, the two-level classifier framework efficiently increases accuracy over one-level models. The experimental findings show that ensemble methods work better than single classifiers, resulting in a gas detection system that is more reliable and accurate.

IV. Results and Discussion

In this section, we present a comprehensive evaluation of both one-level and two-level classification models, comparing their effectiveness in predictive performance. The assessment is based primarily on accuracy, which serves as a key metric for quantifying the correctness of model predictions. Additionally, we examine the strengths and limitations of each classification approach, providing insights into their real-world applicability.

The one-level classification models consist of **individual machine learning algorithms** applied directly to the dataset

without integrating multiple predictions. These models are evaluated based on their accuracy, as shown in **Table 1**:

TABLE 1

ONE-LEVEL CLASSIFICATION ACCURACY

Algorithm	One Level Classification
Random forest	97.4%
KNN Classifier	97.2%
SVC	94.4%

Among the one-level models, Random Forest (97.4%) achieved the highest accuracy, closely followed by k-Nearest Neighbors (kNN) (97.2%), while Support Vector Classifier (SVC) (94.4%) produced the lowest accuracy. Random Forest exhibited superior performance due to its ensemble nature, which helps reduce variance and overfitting by combining multiple decision trees. The model's ability to perform internal feature selection and handle non-linearity contributed to its high accuracy. kNN, although competitive, is more sensitive to local variations and noisy data, which can slightly impact performance. The choice of the optimal k-value plays a crucial role in determining its accuracy. SVC showed lower accuracy, potentially due to its dependence on hyperparameter tuning and sensitivity to high-dimensional feature spaces. The performance of SVC can be further improved by optimizing kernel functions and regularization parameters. The Visualization plot for one level classification can be represented as follows:

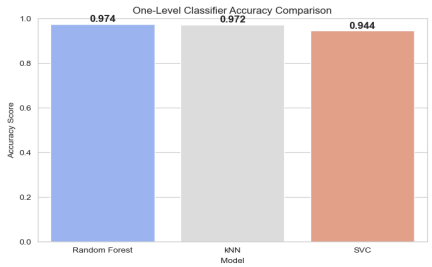

Fig. 5. One-level models accuracy comparison

The two-level classification models employ ensemble learning techniques to combine predictions from multiple base models, enhancing generalization and reducing overfitting. These methods leverage the strengths of different algorithms to achieve higher classification accuracy. The accuracy results for these ensemble techniques are shown in Table 2.

TABLE 2

TWO-LEVEL CLASSIFICATION ACCURACY

Algorithm	Two Level Classification
Stacking	98.28%
Voting	96.64%
Bagging	95.7%

The Stacking classifier, with a 98.28% accuracy rate, outperformed other two-level models due to its ability to combine predictions from heterogeneous models using a meta-learner. This enhances decision-making and reduces individual model weaknesses, leading to improved accuracy. Voting, with a 96.64% accuracy rate, combined predictions from multiple models using a soft voting strategy. However, it lagged behind Stacking as it does not dynamically learn optimal model weights, making it less adaptive to feature variations. Bagging, with a 95.7% accuracy rate, reduced variance by training multiple instances of the same base estimator on different data subsets but slightly lower accuracy compared to Stacking and Voting. Ensemble techniques, particularly heterogeneous model combinations like Stacking, significantly enhance classification accuracy compared to individual models. The Visualization plot for two level classification can be represented as follows:

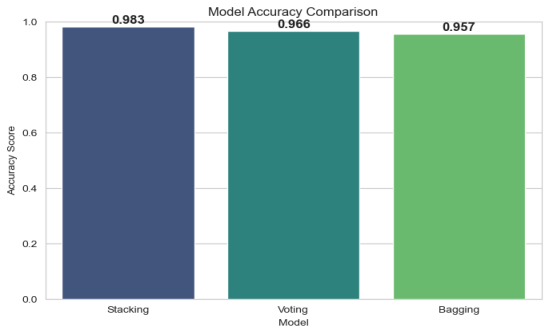

Fig. 6. Two-level models accuracy comparison

One-level models, such as Random Forest, k-Nearest Neighbors (kNN), and Support Vector Classifier (SVC), apply individual algorithms directly to the dataset. Among these, Random Forest achieved the highest accuracy (97.4%), benefiting from its ensemble nature that reduces variance. However, despite their simplicity and efficiency, one-level models are limited by their reliance on a single algorithm, making them more susceptible to overfitting or underfitting depending on data characteristics.

In contrast, two-level classification models, including Stacking, Voting, and Bagging, combine multiple base models to enhance predictive accuracy. The Stacking classifier demonstrated the best performance (98.28%) by leveraging a meta-learner to optimize predictions from diverse models. Voting (96.64%) and Bagging (95.7%) also showed improvements over single models by reducing variance and bias, though their effectiveness was constrained by fixed aggregation strategies. The comparative analysis reveals that two-level models consistently outperform one-level models due to their ability to integrate diverse learning paradigms, mitigate individual model weaknesses, and refine decision-making through meta-learning. While Random Forest remains the strongest one-level classifier, ensemble methods provide a systematic advantage by enhancing model generalization and robustness in complex datasets.

V. Conclusion

The study presents a gas detection framework that uses one-level and two-level classification models to improve gas classification accuracy, reliability, and robustness. This approach is crucial in industries like chemical manufacturing, environmental monitoring, and hazardous gas detection, which require high-precision classification systems to prevent accidents, ensure workplace safety, and monitor air quality. Ensemble learning is a practical and scalable approach for improving gas detection systems, making them more resilient to changes in sensor readings and environmental conditions. The study also suggests that future research could explore deep-learning-based methodologies, such as convolutional neural networks and recurrent neural networks, to capture more complex patterns in gas sensor data. Real-time implementation of gas detection models using edge computing or IoT frameworks could also enhance the practicality and deployment of such systems in dynamic environments. Adaptive learning mechanisms, where models continuously update themselves based on new sensor data, could further improve system efficiency and responsiveness.

This research establishes that ensemble-based classification techniques offer a significant improvement over individual models for gas detection applications. The findings provide a strong foundation for further exploration into more advanced, intelligent, and automated gas detection systems, ensuring increased reliability and safety in various industrial and environmental contexts.

References

[1] Varma, A., Prabhakar, S., & Jayavel, K. (2017, February). Gas leakage detection and smart alerting and prediction using IoT. In *2017 2nd International Conference on Computing and Communications Technologies (ICCCT)* (pp. 327-333). IEEE..

[2] Peng, P., Zhao, X., Pan, X., & Ye, W. (2018). Gas classification using deep convolutional neural networks. *Sensors*, *18*(1), 157.

[3] Wang, S. H., Chou, T. I., Chiu, S. W., & Tang, K. T. (2020). Using a hybrid deep neural network for gas classification. *IEEE Sensors Journal*, *21*(5), 6401-6407.

[4] Monroy, J. G., & Gonzalez-Jimenez, J. (2017). Gas classification in motion: An experimental analysis. *Sensors and Actuators B: Chemical*, *240*, 1205-1215.

[5] Hassan, M., & Bermak, A. (2014, June). Gas classification using binary decision tree classifier. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 2579-2582). IEEE.

[6] Zhao, X., Wen, Z., Pan, X., Ye, W., & Bermak, A. (2019). Mixture gases classification based on multi-label one-dimensional deep convolutional neural network. *IEEE access*, *7*, 12630-12637.

[7] Siva, T., & Merline, A. (2024). Optimizing network lifetime: ERBS-REE for resilient object detection and tracking in resource-constrained WSN environments. *Signal, Image and Video Processing*, 1-13.

[8] Han, L., Yu, C., Xiao, K., & Zhao, X. (2019). A new method of mixed gas identification based on a convolutional neural network for time series classification. *Sensors*, *19*(9), 1960.

[9] Brahim-Belhouari, S., Bermak, A., Shi, M., & Chan, P. C. (2005). Fast and robust gas identification system using an integrated gas sensor technology and Gaussian mixture models. *IEEE Sensors Journal*, *5*(6), 1433-1444.

[10] Romain, A. C., & Nicolas, J. (2010). Long term stability of metal oxide-based gas sensors for e-nose environmental applications: An overview. *Sensors and Actuators B: Chemical*, *146*(2), 502-506.'

[11] Smulko, J. M., Ionescu, R., Granqvist, C. G., & Kish, L. B. (2015). Determination of gas mixture components using fluctuation enhanced sensing and the LS-SVM regression algorithm. *Metrology and Measurement Systems*, *22*(3), 341-350.

[12] Jain, P. C., & Kushwaha, R. (2012, December). Wireless gas sensor network for detection and monitoring of harmful gases in utility areas and industries. In *2012 Sixth International Conference on Sensing Technology (ICST)* (pp. 642-646). IEEE.

[13] Shi, M., Bermak, A., Chandrasekaran, S., Amira, A., & Brahim-Belhouari, S. (2008). A committee machine gas identification system based on dynamically reconfigurable FPGA. *IEEE Sensors Journal*, *8*(4), 403-414.

[14] Liu, Q., Hu, X., Ye, M., Cheng, X., & Li, F. (2015). Gas recognition under sensor drift by using deep learning. *International Journal of Intelligent Systems*, *30*(8), 907-922.

[15] Aishima, T. (1991). Aroma discrimination by pattern recognition analysis of responses from semiconductor gas sensor array. *Journal of Agricultural and food chemistry*, *39*(4), 752-756.