Publishing Excellence
Inspiring Progress

ISSN: 2997-9587 https://aristonpubs.com/journal/chemsci-advances https://doi.org/10.69626/csa.2026.0166

RESEARCH ARTICLE

Multi–Model Evaluation and Feature–Driven Analysis for Predicting Carbon Monoxide and Ethylene Gas Concentrations

I. S. Jamadar [1], S. S. Potdar [1,*], N. K. Yadav [1], H. S. Tajanpure [1], A. B. Deore [1], K. S. Pakhare [2]

ABSTRACT: Predicting gas concentrations from metal oxide sensors is challenging. This is due to traditional methods of using extracted features such as sensitivity percentage, response time and recovery time. This study explores the effectiveness of using raw sensor resistance data to predict gas concentrations. We applied three different machine learning models: Random Forest, TabNet and AutoGluon to predict gas concentrations using an open source sensing array dataset provided by the University of California Irvine. The data contained responses from sixteen commercial sensors to CO and Ethylene mixtures. We applied feature importance through correlation to select the best subset of sensors. We evaluated our models using the coefficient of determination, mean absolute error and root mean squared error. AutoGluon performed best when predicting gas concentrations with all subsets of features showing the benefits of using a fully automated ensemble method. TabNet also performed well when more features were used. This is due to the attention mechanisms within the model. Random Forest provides stable results. Feature importance analysis found that at least six to ten most important features are required to ensure accuracy throughout all models. Our results show that raw sensor resistance data can be used to predict gas concentrations using a simplified, scalable and generalisable approach. This study shows progress towards efficient machine learning solutions.

Keywords: Gas concentration prediction, Metal oxide sensors, Machine learning, Feature selection, Ensemble learning, Sensor resistance data

Received: 24 May 2026; Revised: 20 July 2026; Accepted: 29 July 2026; Published Online: 19 August 2026

1. INTRODUCTION

Accurately detecting and quantifying gases in our environment is becoming increasingly important with many applications ranging from industrial safety, environmental monitoring and medical diagnostics. Metal Oxide Semiconductor sensors are now the most prevalent type of gas sensing technology due to their low cost, high sensitivity and quick response time [1]. Operationally they work by detecting a change in resistance, caused by interaction of the gas molecule with the sensor surface. This allows for detection of various gases such as Carbon Monoxide and Ethylene. Due to the complicated and non-linear nature of sensor outputs, alongside cross sensitivity and environmental interferences, accurate prediction is difficult [2]. Traditionally models have been based on calculated values such as sensitivity percentage, response times and recovery times.

Machine learning techniques have been used for the analysis of sensor data to enable automated feature extraction and advanced prediction tools. Studies have shown the usefulness of neural networks for classifying and predicting gas concentrations [3, 4]. More recently, deep learning has been used with convolutional neural networks and deep belief networks to improve the accuracy of gas recognition [5, 6]. Ensembling techniques have also been found to be useful for improving prediction using different models [7, 8]. However, many studies still use traditional regression techniques or require extensive preprocessing. This requires

[1] School of Computing, MIT Art, Design and Technology University, Pune, Maharashtra-412201, India
[2] Anandibai Raorane Arts, Commerce and Science College, Vaibhavwadi, Sindhudurg-416810, India

* Author to whom correspondence should be addressed: sachin.potdar@mituniversity.edu.in (S. S. Potdar)

© Ariston Publications 2026. All rights reserved.

a high level of domain knowledge and feature engineering. Additionally, features such as response and recovery time will be heavily affected by external factors such as temperature, humidity levels and ageing of the sensors [9].

This study focuses on overcoming the above issues by using raw resistance data to predict gas concentration. This means we are not having to calculate any new features and therefore simplifying our model. We compared three different machine learning algorithms to see which worked best. These were Random Forest, a well-known ensemble technique known for its robustness and interpretation [10]. Next was TabNet, a deep learning algorithm that is built to work with tabular data using attention mechanisms [11]. Lastly was AutoGluon, a machine learning algorithm that builds multiple layers of ensembles automatically [12]. We used the UCI gas sensor array data set which contains readings from sixteen commercial metal oxide sensors when exposed to mixtures of Ethylene and Carbon Monoxide at various concentrations [13]. The data set contains high time resolution data (in milliseconds) and a wide variety of concentration changes meaning it would be ideal for building a predictive model.

Feature selection is an important part of model creation. Correlation based feature ranking was used in this case to find the most useful set of sensors for predicting gas concentration. Overall correlations were found for each sensor using the average of correlation coefficients for each sensor with the two target gases. Features were then ordered in descending order based on their overall correlation. A sequential feature selection method was used to find how many features should be used in each of the models. This allows a balance to be struck between the complexity of the model and its accuracy.

This work is novel because we only use the raw sensor resistance data to predict carbon monoxide and ethylene concentrations at the same time. This overcomes the issues associated with using derived values to predict concentrations as previously done in many studies. Our model takes advantage of the information in the raw resistance data from multiple sensors [14], whereas many studies will use derived values such as response time and sensitivity to predict concentrations. These values can be highly dependent on the environment and sensor drift. By comparing three different models (traditional ensemble model, dedicated deep learning model and automated machine learning), we can see how each of them has benefits and drawbacks. Correlation based feature selection is combined with each of the models to give us a systematic way of finding small sets of features which are still enough to keep the accuracy high.

In this paper, we report the development and evaluation of machine learning models for predicting carbon monoxide and ethylene concentrations from raw sensor resistance data. The specific objectives are threefold: first, to demonstrate that raw resistance values alone can achieve high predictive accuracy without requiring derived metrics; second, to compare the performance of Random Forest, TabNet, and AutoGluon across varying feature subset sizes; and third, to identify the optimal feature selection strategy that balances model simplicity with predictive performance. The findings contribute to the broader effort of developing efficient, scalable, and generalizable solutions for gas sensing applications that can be readily deployed in real-world monitoring scenarios.

2. RELATED WORK

Gas concentration prediction has gained a lot of interest in the fields of industrial automation, environmental safety and agricultural efficiency. Fundamental research into the electro-physical, thermodynamic and structural properties of metal oxides helps to determine their suitability for gas sensing applications [1]. The chemical make-up, temperature and humidity are all key factors which affect the sensitivity of semiconducting metal oxide sensors [2]. This highlights the complex nature of metal oxide gas sensing and the difficulties in predicting the gas concentrations.

A variety of different uses for machine learning in gas sensing have been reported. Neural networks are widely used for gas classification [3] and concentration prediction [4]. Deep learning methods have been widely explored using convolutional neural networks for gas and gas mixtures [5-7]. Ensemble recurrent networks have also shown great potential for gas monitoring in IoT [8]. A combination of unsupervised and supervised learning techniques have been used for harmful gases [9]. Gradient boosting methods have also been explored using systems such as XGBoost for rapid gas mixture identification in an electronic nose system [10]. Recent work has seen the use of TabNet ensemble networks to combat gas sensor drift. Using the attention mechanism inherent in TabNet the performance of support vector machines was improved [11]. In general the ability to provide greater selectivity and stability to gas sensing using machine learning has been well reviewed [12, 13].

In many cases feature selection is an important aspect of sensor modeling. Using correlation based feature ranking algorithms can increase the discriminative aspects of a feature vector extracted from a sensor array by calculating the relevance of each feature within the sensor array [14]. Recursive feature elimination using correlation analysis has shown that careful feature selection can greatly increase regression accuracy and reduce computation [15]. This shows that there is a non-linear relationship between the feature dimension and model accuracy. There will always be an optimal feature subset.

The unique aspect of this work is that it only uses the raw resistance data from the sensors to predict the concentration of Carbon Monoxide and Ethylene. This differs from many studies which have used calculated features such as sensitivity percentages, response times and recovery times. This paper shows that using the raw resistance data is sufficient for accurate concentration predictions. By comparing a basic ensemble model to an attention based deep learning model with feature selection and an automated ensemble model, we can understand the

© Ariston Publications 2026. All rights reserved.

pros and cons of these models when applied to sensor data from metal oxide sensors. This work contributes to the development of methods for analyzing sensor data by demonstrating techniques which can be scaled, be quick and produce accurate results for predicting gas concentrations.

3. DATASET DESCRIPTION AND PREPROCESSING

3.1. Dataset Overview

The data used in this project is an open source dataset with a collection of sensors provided by the University of California Irvine. The dataset contains the dynamic gas response data of 16 commercial sensors situated within ethylene or CO mixtures (in air) at different concentrations. This dataset provides a full and dynamic dataset useful for detecting and modelling gas concentrations.

The dataset contains response data from 16 sensors which are used to detect changes in electron flow caused by the interaction between the chemical compounds of the targeted gases. The concentrations of CO and Ethylene are expressed in parts per million (ppm). A unique feature of this dataset is that it can provide response data every millisecond providing high frequency data. However, the high frequency of this dataset causes an increase in the overall size of the data and the computational complexity involved meaning smart preprocessing techniques are needed.

3.2. Preprocessing

To fulfill the study's aim of discernment of sensor responses exclusively to CO and ethylene, a number of preprocessing operations were performed. The initial dataset contained ten-second averages of the sensor readings and gas concentrations which were registered at a millisecond time scale. In order to control the size of the dataset and facilitate the analysis of temporal variations, these readings were accumulated into one-second time intervals. This averaging not only greatly reduced the high-frequency noise but also preserved the gas concentrations and sensor response dynamics to a considerable degree. In addition, the analysis was restricted to the gases of interest by retaining only those data points where either CO or ethylene was present in the gas mixture, thus eliminating irrelevant data where none of the target gases was present and making the analysis specific to the gases of interest.

The processed dataset finally comprised the mean resistance values of all 16 sensors accompanied by CO and ethylene concentrations. This structured and denormalized dataset not only enabled rapid modeling and analysis but also comprised the essential features for accurate predicting. Preprocessing was the critically important step in the dataset preparation process since it trans- formed a raw, high-dimensional dataset into an optimized format that was suitable for predictive modeling and detailed analysis.

4. METHODOLOGY

4.1. Data Loading and Processing

The initial processing stage involved importing the dataset into the Python environment using the pandas library, which provides efficient and flexible data manipulation capabilities. The input features comprising sensor resistance values were subsequently separated from the output variables representing carbon monoxide and ethylene concentrations. Standardization was applied to all input features using the StandardScaler transform to ensure compatibility with scale-sensitive algorithms and to prevent features with larger magnitudes from dominating the learning process [21]. A correlation analysis was then performed to rank the features according to their aggregate correlation with the output variables, thereby establishing a systematic basis for feature selection. This preliminary ranking served as the foundation for the iterative feature selection procedure described in the following subsection.

4.2. Feature Selection

Iterative feature selection was conducted to identify the most relevant sensor groups for accurate concentration prediction. Combined correlation scores were computed by averaging the individual correlation coefficients of each feature with both carbon monoxide and ethylene concentrations. Features were subsequently arranged in descending order according to these combined scores, and feature sets were constructed containing the top N features for systematic evaluation. The sequential feature selection strategy was applied uniformly across all three modeling frameworks to determine the optimal feature count for each approach [22]. This consistent methodology ensured that performance comparisons across models reflected genuine algorithmic differences rather than disparities in feature availability.

4.3. Model Frameworks

Three distinct modeling frameworks were evaluated in this study, each representing a different paradigm within the machine learning landscape. Random Forest was selected as a representative of classical ensemble learning, TabNet as a specialized deep learning architecture for tabular data, and AutoGluon as an automated machine learning framework capable of constructing sophisticated stacked ensembles.

4.3.1. Method 1: Random Forest

The Random Forest model was employed as the first modeling approach. As an ensemble technique that aggregates predictions from multiple decision trees, Random Forest was selected for its capacity to capture non-linear relationships while maintaining robustness against

 © Ariston Publications 2026. All rights reserved.

overfitting in sensor data applications [23]. The model was configured with one hundred trees, and a seventy-thirty train-test split was used for training and validation. Performance was quantified using the coefficient of determination, mean absolute error, and root mean squared error for both target variables across all feature subsets. This model served as a strong baseline against which the more sophisticated frameworks were compared.

Figure 1 presents the coefficient of determination scores obtained by the Random Forest model for both carbon monoxide and ethylene across feature subsets ranging from sixteen down to two features. The scores remained consistently high for both gases when using larger feature sets, with carbon monoxide achieving values above 0.75 and ethylene exceeding 0.85 for feature counts between sixteen and ten. A gradual decline was observed as the feature count decreased, with a pronounced drop below four features for both gases. Ethylene consistently maintained higher scores than carbon monoxide across the full range of feature subsets, suggesting that the sensor array captures ethylene concentration variations more effectively than carbon monoxide variations.

4.3.2. Method 2: TabNet

TabNet, a neural network architecture specifically designed for tabular data, was employed to predict gas concentrations. Its sequential attention mechanism dynamically selects relevant features during training, enabling the model to focus on the most informative sensor responses at each decision step while reducing the need for manual feature engineering [11]. Separate TabNet models were constructed for carbon monoxide and ethylene, each trained for a maximum of two hundred epochs with early stopping implemented to mitigate overfitting. Input features were formatted as NumPy arrays to ensure compatibility with the neural network framework. The attention masks generated during training additionally provided insight into the relative importance of individual sensors for concentration prediction.

Figure 2 illustrates the coefficient of determination scores achieved by TabNet for both target gases. The model demonstrated stable performance with larger feature sets, maintaining scores above 0.85 for carbon monoxide and above 0.75 for ethylene when using twelve or more features. Performance declined gradually as the feature count decreased, with a sharp deterioration observed below four features. The attention mechanism enabled TabNet to sustain reasonable accuracy even with moderately reduced feature sets, though the model exhibited greater sensitivity to feature reduction than Random Forest in the lower feature count regime.

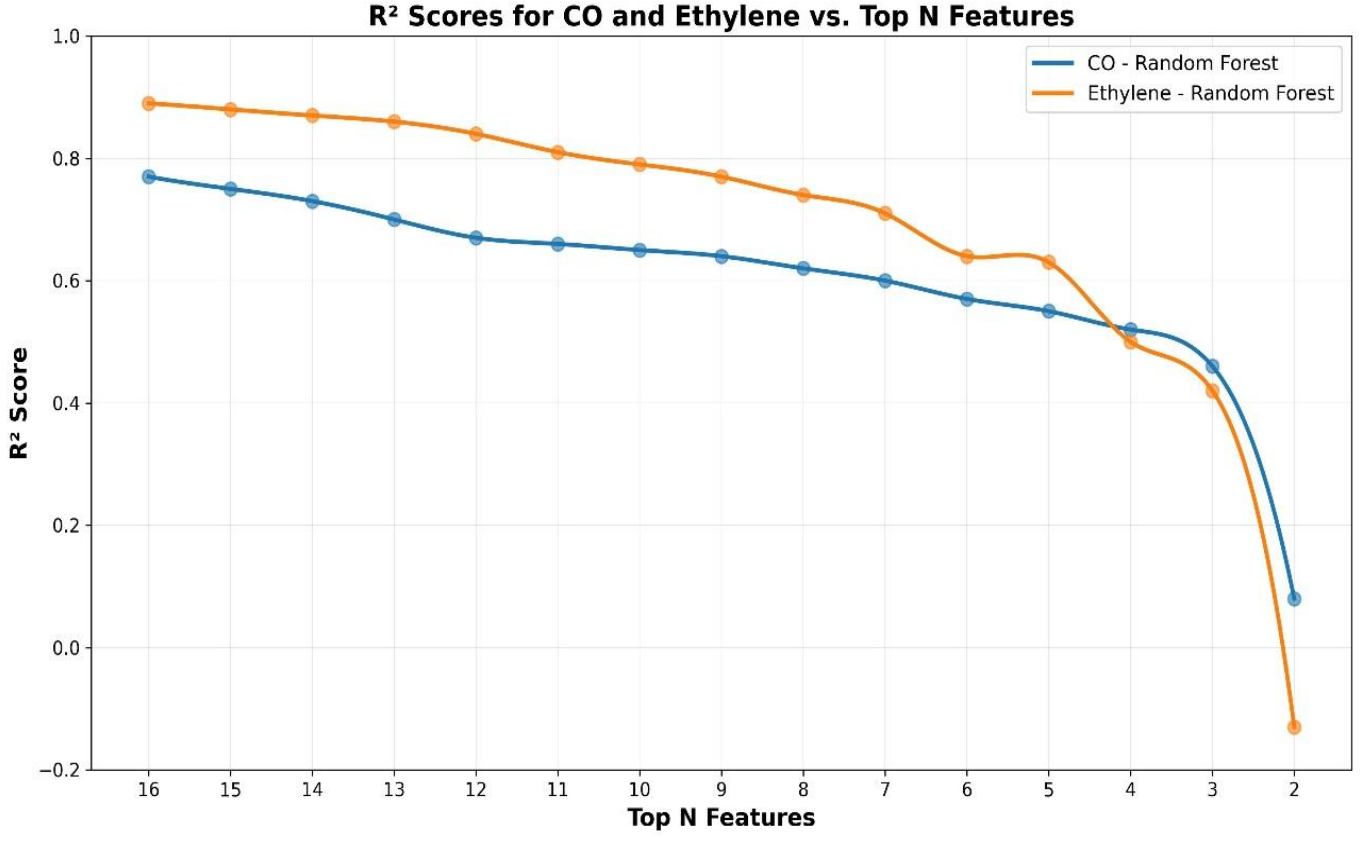


Fig. 1. Coefficient of determination scores for carbon monoxide and ethylene prediction using Random Forest across varying numbers of top-ranked features.

© Ariston Publications 2026. All rights reserved.

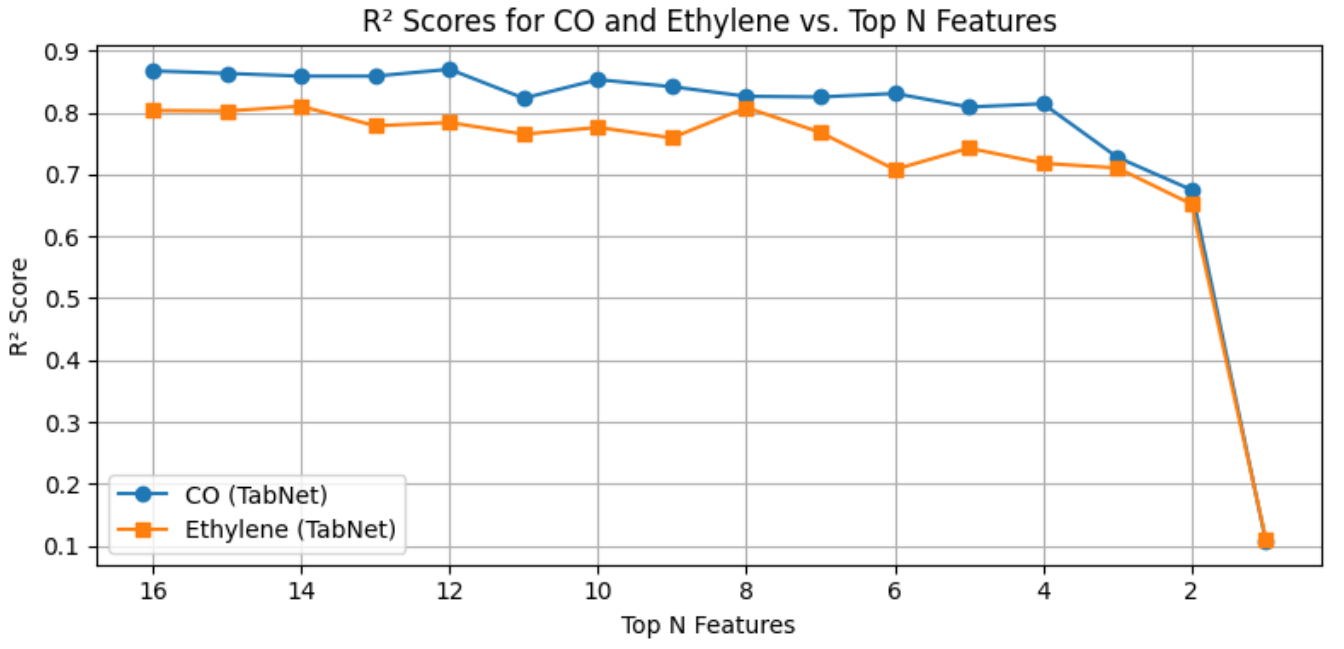


Fig. 2. Coefficient of determination scores for carbon monoxide and ethylene prediction using TabNet across varying numbers of top-ranked features.

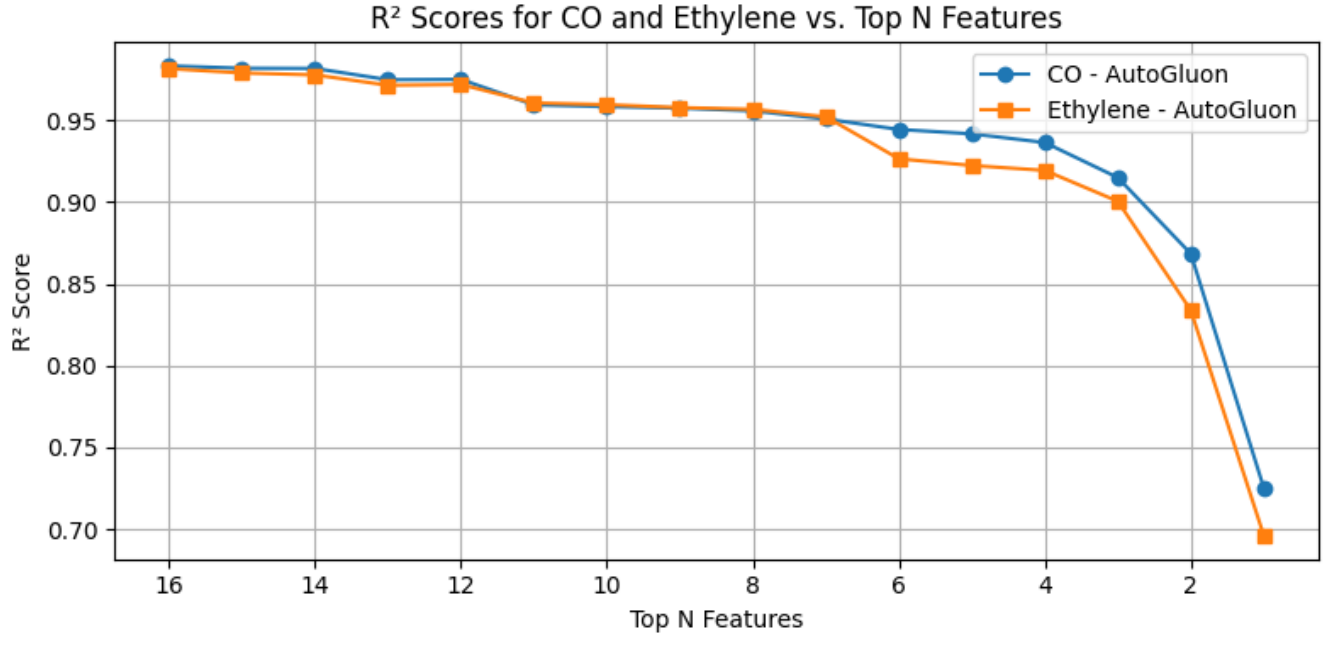


Fig. 3. Coefficient of determination scores for carbon monoxide and ethylene prediction using AutoGluon across varying numbers of top-ranked features.

4.3.3. Method 3: AutoGluon

AutoGluon, an automated machine learning framework, was utilized to construct high-performance ensembles through automated hyperparameter optimization, model selection, and ensemble construction. The framework combines predictions from multiple base models including Random Forest, Extra Trees, XGBoost, and Neural Network Torch into a single weighted ensemble, leveraging the complementary strengths of diverse algorithms [12]. AutoGluon was evaluated across the same correlation-ranked feature subsets as the other methods, ensuring a fair

 © Ariston Publications 2026. All rights reserved.

comparison. The automated optimization process identified optimal hyperparameters for each base model and determined the ensemble weights that minimized prediction error, yielding a robust and computationally efficient solution for gas concentration prediction with minimal manual intervention. Figure 3 displays the coefficient of determination scores obtained by AutoGluon for carbon monoxide and ethylene. The framework achieved near-perfect scores exceeding 0.97 for both gases when using sixteen features, and maintained values above 0.95 for feature counts as low as six. The decline in performance below six features was more gradual than observed for the other models, with scores remaining above 0.85 for four features. These results demonstrate AutoGluon's superior capability to exploit available features and its relative resilience to moderate feature reduction.

4.4. Performance Metrics

Model performance was evaluated using three primary metrics. The coefficient of determination quantifies the proportion of variance in the target variable that is explained by the model, with values approaching unity indicating superior explanatory power [24]. Mean absolute error provides the average magnitude of prediction errors without regard to their direction, offering an intuitive measure of practical accuracy. Root mean squared error penalizes larger deviations more heavily than smaller ones, thereby emphasizing the importance of minimizing outlier predictions [25]. These metrics were computed for carbon monoxide and ethylene predictions across all three modeling frameworks and all feature subsets, enabling comprehensive and comparative evaluation of model performance.

4.5. Comparative Analysis

A detailed comparative analysis was conducted to elucidate the relative strengths and limitations of the three modeling approaches. Performance trends were visualized through coefficient of determination, mean absolute error, and root mean squared error plots against the number of top features for each method. Heatmaps representing coefficient of determination scores across feature subsets highlighted the contribution of individual sensors to concentration prediction accuracy. Cross-method comparisons facilitated direct assessment of the three modeling frameworks under identical feature availability conditions, providing actionable guidance for model selection in practical gas sensing applications.

5. RESULTS AND DISCUSSION

The outcomes are illustrated with visualizations and comprehensive talks on the performance of each model. R^2, MAE, and RMSE were the main metrics analyzed for prediction of CO and ethylene. The use of comparative analyses and heatmaps gave a clearer view of the importance of features as well as the advantages and disadvantages of each modeling technique.

5.1. Method 1: Random Forest

5.1.1. R^2 vs. Top N Features

In the case of CO and ethylene predictions, the Random Forest model delivered R^2 values that were consistently high, thus confirming its strong and best predictive power even through the different feature subsets. Its stability in performance was seen even when there was a cut down in the number of features, especially going as low as 6 to 8 features, making it clear that these subsets are ideal for the mixture of accuracy and complexity. However, when less than 4 features were used, the R^2 values dropped sharply, pointing out the need to keep a good number of predictive features at all times.

5.1.2. MAE and RMSE Trends

The Random Forest algorithm was the one that always offered the lowest MAE and RMSE in all feature subsets for CO predictions and therefore, it was the most accurate and robust prediction method. The errors for Random Forest, however, went up quite significantly when less than 6 features were included in the model, which calls for the use of a sufficiently large feature subset in order to get accurate predictions. A similar pattern was noted for ethylene predictions where Random Forest not only won but also showed the best performance in both MAE and RMSE. The increase in error with less features were especially pronounced for ethylene, which means that it is more sensitive to feature reduction in that context. These results support Random Forest as a very efficient model for both gases and, at the same time, underline the importance of feature selection.

5.1.3. Heatmap Analysis

The heatmaps showing R^2 scores for CO and ethylene predictions additionally depicted Random Forest's even performance. It continued to have very good predictive accuracy despite the reduc- tion of features, showing high R^2 scores in a large number of the subsets. A significant decrease in R^2 was noted for both gases when using less than six features, which further emphasizes the necessity of sufficient feature subsets for consistently accurate predictions.

© Ariston Publications 2026. All rights reserved.

5.2. Method 2: TabNet

5.2.1. R^2 Performance

The TabNet model's performance for carbon monoxide and ethylene forecasting measured in terms of R^2, showed that the model had a fairly steady R^2 score when the complete feature set was used. The performance for CO predictions indicated that the model could take full advantage of the wide feature set. However, there was a significant drop in R^2 with just 6 features used and an even sharper decline with fewer than 3. It shows TabNet's dependency on adequate feature variety to obtain accurate predictions. Ethylene forecasting resembled this pattern; the larger feature subsets provided stable R^2 values while the smaller ones (less than 3 features) caused a marked decline in performance. TabNet in general showed good performance with a larger number of features but a sudden drop when the feature subsets were minimal.

5.2.2. Error Metrics

The plots of MAE and RMSE for TabNet have shown its good prediction performance over different feature subsets. Looking at the CO, MAE was low and stable for feature subsets of 16 to 6, which meant the model was very consistent in that area. On the other hand, MAE increased sharply when the number of features dropped below 6, with 2 features peaking very high indeed. RMSE for CO showed the same course of movement. For ethylene the same trends were derived, as there was a steady state of MAE and RMSE for the largest feature subsets (16 to 6) and a quick rise for the smallest ones. These plots stress that although TabNet is very good at working with the complete features, it does not perform well with a small number of the features, thus indicating that it is very important to keep a diverse set of features.

5.2.3. Heatmap Insights

The heat maps of TabNet's R^2 scores showed that the R^2 values remained very stable (more than 0.8) for CO when the feature subsets were reduced from 16 to 6. There was a decrease below 6 features, and R^2 dropped drastically for 3 features and got cut off very fast for just one feature. Ethylene behaved similarly, the R^2 values were above 0.76 for feature subsets from 16 to 6, and there was a considerable decline below 6 features. All these heat maps prove that TabNet is powerful at working with large dimensional feature spaces but also point to its vulnerability to such small feature subsets.

5.3. Method 3: AutoGluon

5.3.1. R^2 Trends

The R^2 plots of AutoGluon made it clear that the model was excellent for predicting the concentrations of CO and ethylene even when different subsets of features were used. For carbon monoxide, AutoGluon was able to get quite high R^2 scores for feature sets containing between 16 and 6 components, which means that its performance was quite strong indeed. A slight drop in performance was noted with the feature size reducing below 6, with R^2 getting very low for only 4 features and losing its reliability for less than 2 features. The same trend was witnessed in ethylene predictions which kept on showing high R^2 values when the number of features was from 16 to 6, while the downfall in accuracy started getting noticeable with feature sets of 4 or less. These outcomes demonstrate that AutoGluon is quite a trustworthy performer when larger feature sets are involved but at the same time it draws heavily upon the features being there in sufficient numbers for it to be able to maintain high prediction accuracy over time.

5.3.2. Error Metrics

The MAE and RMSE plots for AutoGluon clearly indicated the robust predictive power of the method for both CO and ethylene, and error values were always very low for feature subsets from 16 down to 6. That meant AutoGluon was able to use the different features for the predictions very effectively. But, when the number of features went below 6, then there was a steady increase in errors, and a sharp spike appeared for the subsets containing less than 3 features. This pattern tells us that AutoGluon needs a variety of features to be able to see the underlying data patterns. Errors were nearly nonexistent for larger feature sets but jumped rapidly for smaller subsets, which means that the model's ability to generalize was weakened then. The findings suggest that it is critical to keep at least 6 features for predictions to be trustworthy, as there was a drastic drop in performance with the smaller feature subsets.

5.3.3. Heatmap Findings

AutoGluon's R^2 scores heatmaps provided a clear picture of its excellent performance. Auto- Gluon marked R^2 scores of almost perfect with CO for feature subsets of 16 down to 13, indicating its superior data pattern capturing ability. The R^2 values were strong and only slightly decreased as the number of features dropped, with a large number of 5 features still providing an acceptable performance. There was a significant drop in the accuracy when the number of features went below 5. AutoGluon, on the other hand, gave up to 6 features very strong R^2 scores for ethylene, with a gradual decrease in scores corresponding to reduction in features. The R^2 kept

 © Ariston Publications 2026. All rights reserved.

being high for 6-feature subsets, then quickly dropping for smaller ones. The patterns here disclose not only the high prediction potential of AutoGluon given good features but also its reliance on feature variety for the accuracy to last longer.

5.4. Comparative Analysis

The combined coefficient of determination plot for Random Forest, TabNet, and AutoGluon provides a comprehensive view of their respective performances in predicting carbon monoxide and ethylene concentrations using different feature subsets. Figure 4 presents the comparative coefficient of determination scores across all three modeling frameworks. The comparative analysis reveals distinct performance hierarchies among the three frameworks. AutoGluon consistently achieved the highest coefficient of determination values across most feature subsets, demonstrating the best performance with diverse feature configurations. Random Forest followed closely, exhibiting strong performance throughout, particularly with larger subsets, and proving robust against feature reduction until the subset size decreased to six. TabNet performed comparably but generally lagged behind AutoGluon and Random Forest, with coefficient of determination values declining in correspondence with feature count reduction. For both carbon monoxide and ethylene predictions, all models experienced a sharp coefficient of determination drop when four or fewer features were employed, indicating that maintaining feature diversity at an adequate level is vital for accurate predictions.

5.4.1. R^2 Across Methods

The combined R^2 plots for Random Forest, TabNet, and AutoGluon give a nice view of their performances in predicting CO and ethylene concentrations using different feature subsets. Au- toGluon always got the highest R^2 values in most of the feature subsets showing the best with different features. Random Forest came right behind and had a strong performance throughout, especially with the larger subsets, and proved to be robust concerning feature reduction until the subset size reduced to 6. TabNet was in a similar boat performance-wise but actually, looked to be lagging behind AutoGluon and Random Forest most of the times, and that was the case with R^2 reduction corresponding to the reduction in number of features as well. In the case of predictions for both CO and ethylene, all models experienced a very sharp R^2 drop when they had 4 or fewer features which indicates that it is vital to keep feature diversity at a decent level for the predictions to be accurate. Thus, AutoGluon and Random Forest were the two most reliable models with TabNet being a solid alternative.

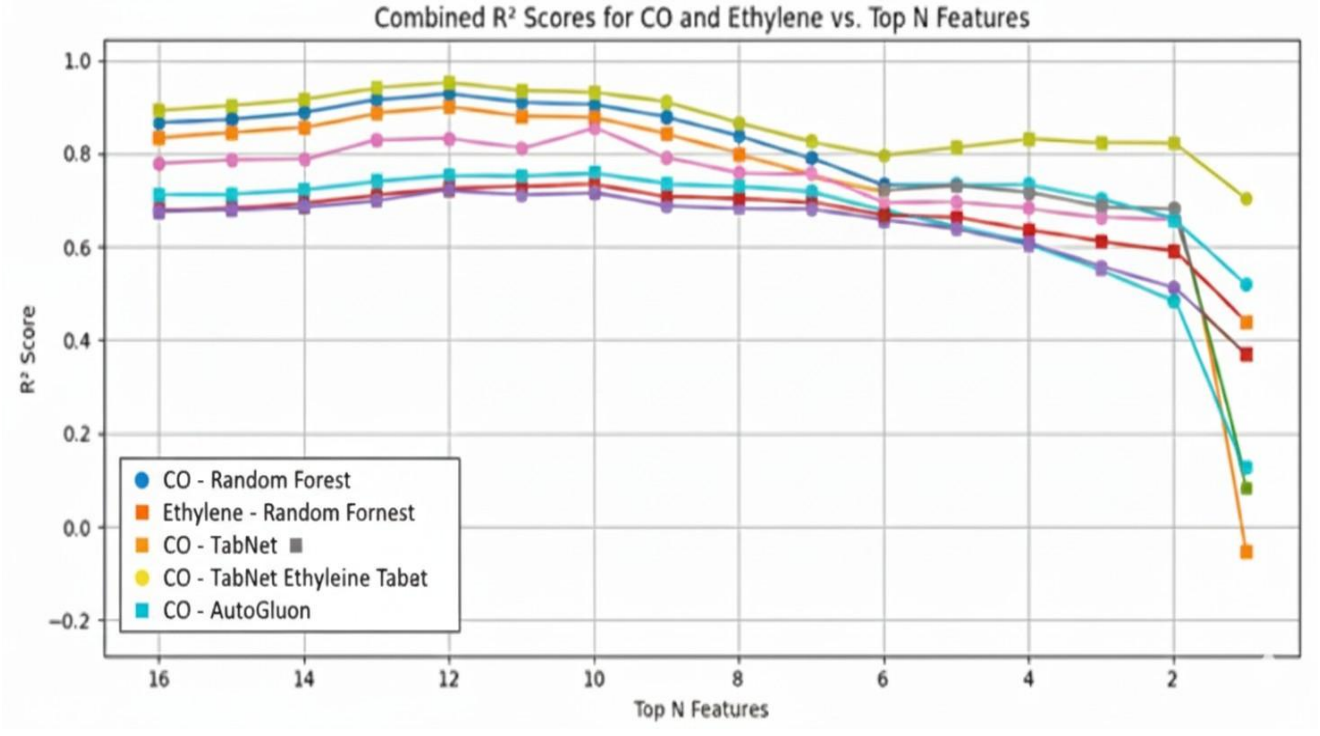


Fig. 4. Combined coefficient of determination scores for carbon monoxide and ethylene prediction across Random Forest, TabNet, and AutoGluon as a function of the number of top-ranked features.

© Ariston Publications 2026. All rights reserved.

5.4.2. MAE and RMSE Comparisons and Heatmap

The comparison through MAE and RMSE plots, together with R^2 heatmaps, gave more support to the observed trends. AutoGluon had lower errors all the time on feature subsets which means that it had the best prediction accuracy. Random Forest was able to keep errors low for larger subsets, but it experienced an increase in MAE and RMSE when features were cut down to the bare minimum. TabNet was in line with Random Forest as far as the trend is concerned, but it was more affected by the reduction of features. The combined analysis points to AutoGluon as the most powerful and reliable model, with Random Forest closely following, while TabNet is still in the race with good performance as long as there are enough features.

6. CONCLUSION

The present research has thoroughly investigated the various machine learning methods that can be applied to gas concentration prediction, giving a particular emphasis on using raw sensor resistance values for input. The exclusion of additional metrics like sensitivity percentage and recovery time not only simplifies the modeling but also makes the method more suitable for real- world applications. The study comprised an evaluation of three different approaches: the classic Random Forest model, the TabNet deep learning architecture, and the AutoGluon automated ensemble framework. The findings were crystal clear and showed the benefits of ensemble learning beyond any doubt. Auto-Gluon was always the best model concerning the prediction accuracy and error reduction, the differences being striking. Its natural power to merge several algorithms and to carry out hyperparame- ter tuning automatically makes it a great tool for difficult sensor- based applications. The others like TabNet, however, had a different approach and showed their strength in selecting dynamically the essential features which even resulted in getting cheaper but not lower quality inputs, thanks to their attention mechanisms. The Random Forest model was stable with its strong performance and at the same time reliable as a benchmark, always bringing along consistent results. Mainly, the study brought out along the way the great need for feature selection, by marking the top 10–16 features as the must-have ones in getting predictive accuracy that is very high. The merging of correlation-based feature ranking with machine learning models was a good way for optimization of the performance and at the same time the computational complexity was kept low. Heatmaps and comparative plots did more than just show; they made clear the complex connection existing between feature importance and model efficiency. To sum up, the experiment that has been carried out has successfully closed the loophole between the traditional way of modeling with sensor data and the application of sophisticated machine learning methodologies. Thus, a simple but very powerful method for predicting gas concentrations is introduced. The research findings draw attention to the great potential of raw sensor resistance data being used as a reliable input for different machine learning frameworks, so that the development of gas sensing solutions that are efficient, accurate, and scalable will be allowed. As the next step, the project could investigate more about the combination of different sensor types, data streaming in real-time, and hybrid modeling that would, in turn, improve the predictive systems' accuracy and applicability a lot.

DECLARATIONS

Ethical Approval

We affirm that this manuscript is an original work, has not been previously published, and is not currently under consideration for publication in any other journal or conference proceedings. All authors have reviewed and approved the manuscript, and the order of authorship has been mutually agreed upon.

Funding

No funding.

Availability of data and material

All of the data obtained or analyzed during this study is included in the report that was submitted.

Conflicts of Interest

The authors declare that they have no financial or personal interests that could have influenced the research and findings presented in this paper. The authors alone are responsible for the content and writing of this article.

Authors' contributions

All authors contributed equally to this work,

REFERENCE

[1] Korotcenkov, G., **2007**. Metal oxides for solid-state gas sensors: What determines our choice?. *Materials Science and Engineering: B*, *139*(1), pp.1-23. https://doi.org/10.1016/j.mseb.2007.01.044

[2] Wang, C., Yin, L., Zhang, L., Xiang, D. and Gao, R., **2010**. Metal oxide gas sensors: sensitivity and influencing factors. *sensors*, *10*(3), pp.2088-2106. https://doi.org/10.3390/s100302088

 © Ariston Publications 2026. All rights reserved.

[3] Benrekia, F., Attari, M. and Bouhedda, M., **2013**. Gas sensors characterization and multilayer perceptron (MLP) hardware implementation for gas identification using a field programmable gate array (FPGA). *Sensors*, *13*(3), pp.2967-2985. https://doi.org/10.3390/s130302967

[4] Sabilla, S.I., Sarno, R. and Siswantoro, J., **2017**. Estimating gas concentration using artificial neural network for electronic nose. *Procedia Computer Science*, *124*, pp.181-188. https://doi.org/10.1016/j.procs.2017.12.145

[5] Peng, P., Zhao, X., Pan, X. and Ye, W., **2018**. Gas classification using deep convolutional neural networks. *Sensors*, *18*(1), p.157. https://doi.org/10.3390/s18010157

[6] Wei, G., Li, G., Zhao, J. and He, A., **2019**. Development of a LeNet-5 gas identification CNN structure for electronic noses. *Sensors*, *19*(1), p.217. https://doi.org/10.3390/s19010217

[7] Chu, J., Li, W., Yang, X., Wu, Y., Wang, D., Yang, A., Yuan, H., Wang, X., Li, Y. and Rong, M., **2021**. Identification of gas mixtures via sensor array combining with neural networks. *Sensors and Actuators B: Chemical*, *329*, p.129090. https://doi.org/10.1016/j.snb.2020.129090

[8] Lai, W.I., Chen, Y.Y. and Sun, J.H., **2022**. Ensemble machine learning model for accurate air pollution detection using commercial gas sensors. *Sensors*, *22*(12), p.4393. https://doi.org/10.3390/s22124393

[9] Pareek, V. and Chaudhury, S., **2021**. Deep learning-based gas identification and quantification with auto-tuning of hyper-parameters. *Soft Computing*, *25*(22), pp.14155-14170. https://doi.org/10.1007/s00500-021-06222-1

[10] Breiman, L., **2001**. Random forests. *Machine learning*, *45*(1), pp.5-32. https://doi.org/10.1023/A:1010933404324

[11] Arik, S.Ö. and Pfister, T., **2021**, May. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 8, pp. 6679-6687). https://doi.org/10.1609/aaai.v35i8.16826

[12] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M. and Smola, A., **2020**. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*. https://doi.org/10.48550/arXiv.2003.06505

[13] Fonollosa, J., Fernandez, L., Gutiérrez-Gálvez, A., Huerta, R. and Marco, S., **2016**. Calibration transfer and drift counteraction in chemical sensor arrays using Direct Standardization. *Sensors and Actuators B: Chemical*, *236*, pp.1044-1053. https://doi.org/10.1016/j.snb.2016.05.089

[14] Yang, S., Lei, G., Xu, H., Lan, Z., Wang, Z. and Gu, H., **2023**. A review of the high-performance gas sensors using machine learning. *Machine Learning for Advanced Functional Materials*, pp.163-198. https://doi.org/10.1007/978-981-99-0393-1_8

[15] Fernandez, L., Guney, S., Gutierrez-Galvez, A. and Marco, S., **2016**. Calibration transfer in temperature modulated gas sensor arrays. *Sensors and Actuators B: Chemical*, *231*, pp.276-284. https://doi.org/10.1016/j.snb.2016.02.131

[16] Huang, S., Croy, A., Ibarlucea, B. and Cuniberti, G., **2023**. Machine learning-driven gas identification in gas sensors. In *Machine learning for advanced functional materials* (pp. 21-41). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-0393-1_2

[17] Wu, J., Zhao, W., Wu, F., Yan, J., Feng, P., Cui, H., Duan, S. and Peng, X., **2024**. A mixed gas concentration regression prediction method based on RESHA-ALW. *Sensors and Actuators B: Chemical*, *418*, p.136222. https://doi.org/10.1016/j.snb.2024.136222

[18] Jaleel, M., Kucukler, O.F., Alsalemi, A., Amira, A., Malekmohamadi, H. and Diao, K., **2023**. Analyzing gas data using deep learning and 2-D Gramian angular fields. *IEEE Sensors Journal*, *23*(6), pp.6109-6116. https://doi.org/10.1109/JSEN.2023.3243149

[19] Zhao, X., Wen, Z., Pan, X., Ye, W. and Bermak, A., **2019**. Mixture gases classification based on multi-label one-dimensional deep convolutional neural network. *IEEE access*, *7*, pp.12630-12637. https://doi.org/10.1109/ACCESS.2019.2892754

[20] Dhariwal, N., Bisht, S., Yadav, P., Ueda, T., Hyodo, T. and Kumar, V., **2026**. Machine Learning in Gas Sensing: A Paradigm Shift for Enhanced Detection and Selectivity–A Review. *IEEE Sensors Journal*. https://doi.org/10.1109/JSEN.2026.3655422

[21] Albanese, D., Visintainer, R., Merler, S., Riccadonna, S., Jurman, G. and Furlanello, C., **2012**. mlpy: Machine learning python. *arXiv preprint arXiv:1202.6548*. https://doi.org/10.48550/arXiv.1202.6548

[22] Barbieri, M.C., Grisci, B.I. and Dorn, M., **2024**. Analysis and comparison of feature selection methods towards performance and stability. *Expert Systems with*

© Ariston Publications 2026. All rights reserved.

Applications, *249*, p.123667. https://doi.org/10.1016/j.eswa.2024.123667

[23] Latinne, P., Debeir, O. and Decaestecker, C., **2001**, June. Limiting the number of trees in random forests. In *International workshop on multiple classifier systems* (pp. 178-187). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-48219-9_18

[24] Chicco, D., Warrens, M.J. and Jurman, G., **2021**. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, *7*, p.e623. http://dx.doi.org/10.7717/peerj-cs.623

[25] Chai, T. and Draxler, R.R., **2014**. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development*, *7*(3), pp.1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

 © Ariston Publications 2026. All rights reserved.