OPEN ACCESS
IOP Publishing
Journal of Physics D: Applied Physics
J. Phys. D: Appl. Phys. **57** (2024) 345202 (11pp) https://doi.org/10.1088/1361-6463/ad4f97

Distinguishing methane from other hydrocarbons using machine learning and atmospheric pressure plasma optical emission spectroscopy

Tahereh Shah Mansouri[1,*], **Hui Wang**[2], **Davide Mariotti**[1] **and Paul Maguire**[1]

[1] NIBEC Engineering, Ulster University, Belfast, United Kingdom
[2] Queens University, Belfast, United Kingdom

E-mail: t.shah_mansouri@ulster.ac.uk

Received 20 December 2023, revised 11 April 2024
Accepted for publication 23 May 2024
Published 4 June 2024

Abstract

The ability to detect gas molecule and assign a concentration offers an inventive solution in the field of plasma integrated with machine learning. The most important finding of this work is firstly, to develop an algorithm for gas-molecule identification using three different hydrocarbons (CH_4, C_2H_2, C_2H_6) and secondly, organize a model for detecting gas concentration (classification). For this reason, initially eight different gases evaluated. The study confirms the present of the unique emission lines as a gas indicator, i.e., a wavelength peak related to hydrocarbons identified via increasing in C_xH_y concentration. By means of unique variable important in projection, hydrocarbons can be distinguished. Our proposed Chemometric analysis strategy examined on >1000 samples and results development of suitable techniques that are sufficiently rapid, accurate and innovative. This demonstrates the potential for real-time, portable, and continuous monitoring of trace gases with potential applications in medical, environmental, and industrial gas sensing.

Keywords: methane identification, optical emission spectroscopy (OES), variable importance in projection (VIP), unique VIP, partial least square discriminant analysis, hydrocarbons, classification

1. Introduction

The application of supervised learning models based on partial least squares discriminant analysis (PLS-DA) was investigated for the identification of trace gases using data obtained from plasma optical emission spectroscopy [1]. As a result, the parts-per-million (ppm) classification of a single type of hydrocarbon gas (methane) was achieved down to concentrations in the low ppm range. The next stage is to attempt the detection of various hydrocarbon gases and investigate the capability of the machine learning (ML) approach in more complex scenarios.

In this study, we aimed to distinguish different hydrocarbon gases using the PLS-DA method and its features, such as variable importance on projection (VIP) selection. Detecting hydrocarbons such as CH_4 and C_2H_6 can be useful in breath analysis (VOCs) or climate change monitoring applications. Carbon bonding determines the hydrocarbon type, with single-bond C forming an alkane, double-bond C forming an alkene, and triple-bond C forming an alkyne [2]. To determine the capability for general hydrocarbons (C_xH_y) detection, two alkanes (CH_4 and C_2H_6) and one alkyne (C_2H_2) were selected with the

[*] Author to whom any correspondence should be addressed.

aim of using them as a basis for detecting other gases that are very similar in their composition. Thus, this technique based on plasma optical emission spectra coupled with ML is fundamental for a much wider field of hydrocarbon or breath VOC detection [3] and identification applications.

The miniature plasma source for generating optical emission spectral data is aimed at low-cost portable or field deployment, capable of autonomous and continuous monitoring of environmental trace gases. Plasma contains a high concentration of energetic electrons with a wide distribution of energies [4]. Various collision processes occur between electrons and gas atoms or molecules leading to ionization, electronic excitation, molecular dissociation, attachment, and vibrational excitation with the probability of each process dependent on the electron and target species concentrations, electron energy distribution, and process energy cross-section. This is a complex non-equilibrium environment, which also contains other impurity molecules such as O_2, N_2 and H_2O, and is therefore not amenable to the direct and absolute measurement of its characteristics. Thus, a machine-learning approach is essential. However, additional challenges arise in identifying different hydrocarbons. In hot or energetic plasmas, it is expected that all hydrocarbons will be fully dissociated by electron bombardment into atomic and molecular fragments [5–7], examples of which are given below for methane.

$$\mathrm{e} + \mathrm{CH_4} \rightarrow e + \mathrm{CH_3} + \mathrm{H} \tag{1}$$

$$\mathrm{e} + \mathrm{CH_4} \rightarrow e + \mathrm{CH_2} + \mathrm{H_2} \tag{2}$$

$$\mathrm{e} + \mathrm{CH_4} \rightarrow e + \mathrm{CH} + \mathrm{H_2} + \mathrm{H} \tag{3}$$

$$\mathrm{e} + \mathrm{CH_4} \rightarrow e + \mathrm{C} + 2\mathrm{H_2}. \tag{4}$$

There is also the possibility of superideal quenching [8] of noble gas metastable states, leading to molecular dissociation.

$$\mathrm{H}e\left(2s^1S\right) + \mathrm{CH_4} \rightarrow He + \mathrm{CH} + \mathrm{H_2} + \mathrm{H}. \tag{5}$$

With significant dissociation, hydrocarbons would lose their identity, and discrimination is not possible. Therefore, we maintain the plasma at a low gas temperature [9], and while a degree of dissociation must occur because the dissociation energy is lower than that for ionization, the aim is to maintain a sufficiently low level to enable discrimination. However, there have been few investigations into low-temperature plasma interactions with hydrocarbon molecules at atmospheric pressure. In our recent paper [1], plasma modeling provides evidence for limited dissociation, whereas algorithm detection of methane is based on both indirect mechanisms through plasma interaction with impurities, and on low-intensity carbon species emission- namely CH and C_2—the former producing an identifiable emission line at 431 nm [10–13]. In recent years, few techniques proposed to detect hydrocarbons and gas impurities in the form of molecular using Glow discharge and kinetics of fast electrons [14, 15]. In this study, different gas mixtures of methane, acetylene, and ethane at various concentrations from 1 ppm to 100 ppm in helium were used to provide emission spectra in the range of 194–1122 nm as input to ML classification models based on PLS-DA with the aim of detecting hydrocarbon type and concentration.

2. Experimental methods

In this study, the main device used to generate optical emission samples was atmospheric pressure plasma. To produce plasma, helium flows through a quartz capillary with inner diameter of 2 mm in the middle of two aligned ring electrodes with 5 mm distance. Presence of helium is necessary as a carrier gas for sustain of plasma when measuring different gases. Mass flow controller (MFC) is employed to adjust flow rate of helium and other gases. Spectrometer works alongside with ocean view software (Ocean View Spectrometer Operating Software: Version 1.5.2) to measure wavelength scope. The overall view of the apparatus and their connection in this study is shown in figure 1. Plasma device is located at the center of a gas line network and is driven by RF power for gas ionization. The spectra were measured using an Ocean Optics HR4000CG-UV-NIR spectrometer (optical resolution < 1.0 nm FWHM, slit width 5 μm), with a total of 3648 wavelength points in a range of 194 nm–1122 nm (interval 0.25 nm). The spectrometer from one side is connected to optic fiber, and from other side connected to a computer where an Oceanview spectroscopy software process the data.

The target species (N_2, H_2, CH_4, C_2H_6 & C_2H_2, H_2O & helium) are diluted in helium to reach the desired concentration. Each gas was measured individually. Each species concentration varied from 0–100 ppm (0, 1, 2, 4, 6, 12, 23, 30, 47, 60, 77, 85, 100 ppm) where at least 100 samples recorded for each ppm. Other non-hydrocarbons species, such as N_2, He, H_2O, and H_2 were recorded at different concentrations for comparison and further investigation. In this study, the term ‘category’, ‘spectra’, ‘concentration’ and ‘ppm’ are interchangeable, and all may refer to the recorded sample. Meanwhile, in many places, the wavelength is referred to as ‘variable’.

Datasets from three gases (CH_4–He, C_2H_6–He and C_2H_2–He) were collected in a matrix of 3648 variables (wavelengths) for each, using two RF plasmas formed in a quartz capillary between two exterior ring electrodes. In this study, two plasmas were utilized to measure samples: plasma no. 1 with a special structure and covering chamber that had the least air diffusion, and plasma no. 2, which was portable, but may record samples with more impurities (figure 2).

There are four classes of species: helium, carbon-based (C I, C II, CH, etc), hydrogen-based (H I, H II, etc), and impurity (e.g., N, O, OH/H_2O). As an example, figure 3(a) compares the three spectra for methane at different concentrations. No peak, except possibly near 516 nm, could be assigned unambiguously to any species. In this study, the appearance and disappearance of a peak at 431 nm was triggered by the addition or removal of hydrocarbons to the helium mixture at the time of experimentation.

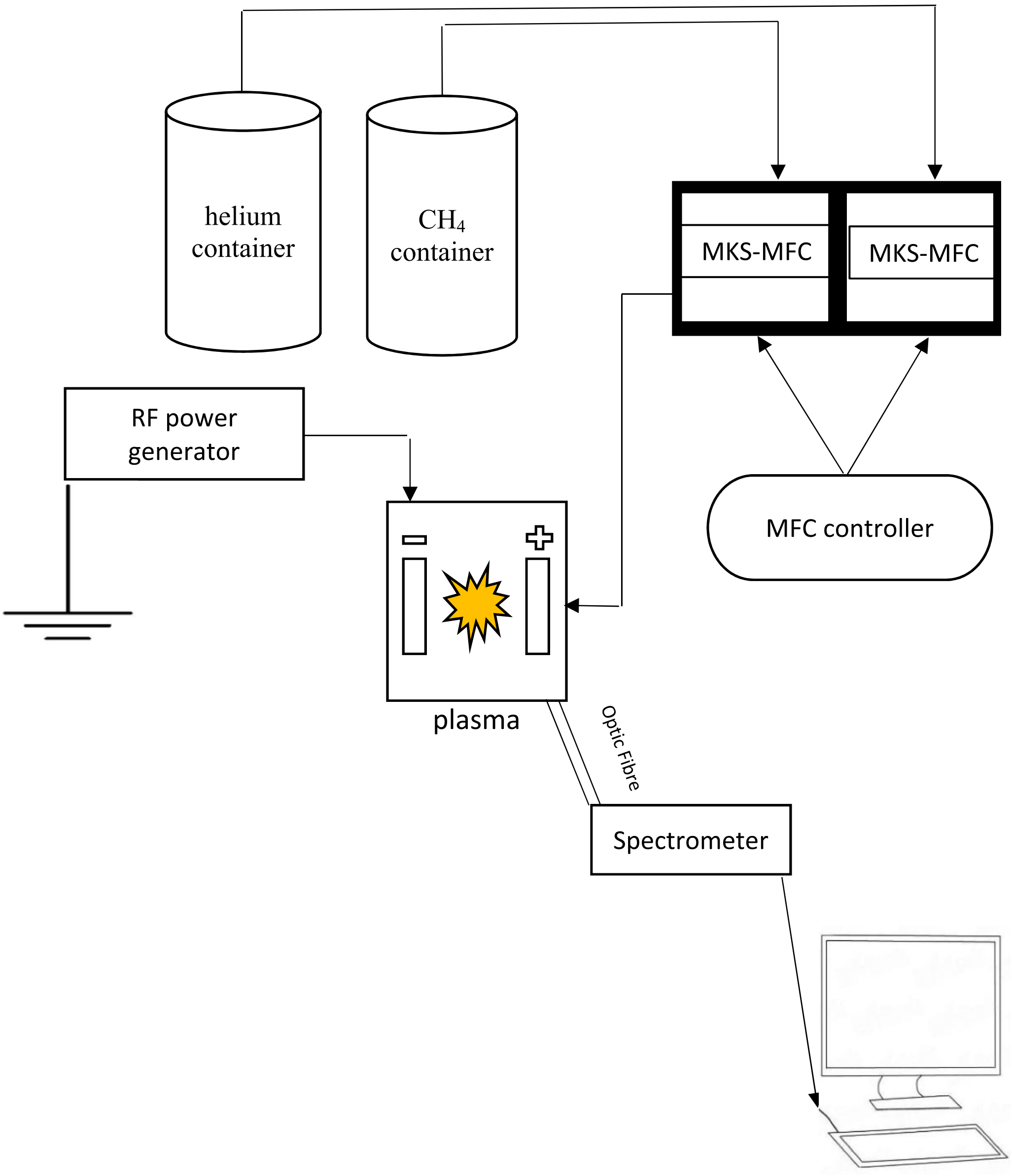

Figure 1. Apparatus and its set up for measuring CH_4 from He—CH_4 mixture. Emission spectra of He—CH_4 is obtained from a plasma operated at atmospheric pressure. The mixture can be replaced by other mixtures (e.g., C_2H_6, C_2H_2, N_2) for recording their value. In this work, RF power generator (Cesar- RS232) connected to the plasma system when forwarded power (W_f) for all samples remained in a constant value of 90 w and the Reflected power (W_r) also could not exceed to more than 2% of W_f and usually retained at zero. Ocean Optics (HR4000CG-UV-NIR) spectrometer is connected from one side to the plasma via an Optic Fiber and from other side to the computer. Mass Flow Controller (MFC) via two MKS-MFC is set to manage and quantify gas and flow rate.

Comparing other recorded gases (N, H_2O, He, and H_2 with available hydrocarbon [C_xH_y]) with the same setup at a wavelength of 431 nm shows that the highest intensity is obtained in the presence of hydrocarbons, and therefore could signify a C_xH_y species, most likely CH [11–13] (figure 4(a)). Two concentrations have been considered for H_2O, as the behavior of H_2O samples is quite different as compared to other gases, which is outside the scope of this study. Figure 4(b) shows the peaks at a wavelength of 431.382 nm, where an almost direct relationship was observed between C_2H_2 concentration and intensity up to 12 ppm.

Therefore, in the current study, we propose to use the CH peak at 431.382 nm as an indicator and the presence of hydrocarbon species (C_xH_y). This is especially obvious for the C_2H_2-He mixture when the C-H bonding is more robust.

3. Algorithmic techniques

PLS is a regression algorithm that reduces the number of observations to a smaller number of components,

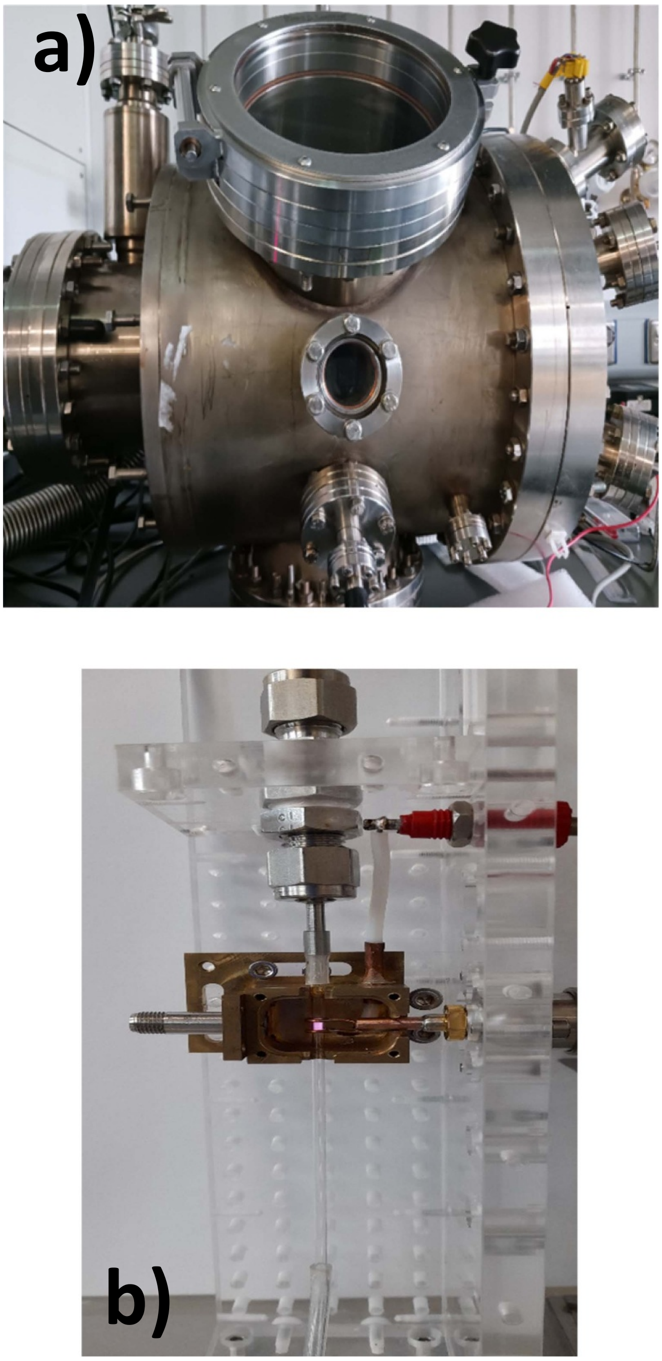

Figure 2. (a) Plasma no.1 device used in this study with covering chamber. (b). plasma no. 2 device (no- chamber), with connected plastic tube outlet to prevent air diffusion.
Note. RF capillary plasma system operated with helium carrier at atmospheric pressure. The electrode gap was 5 mm.

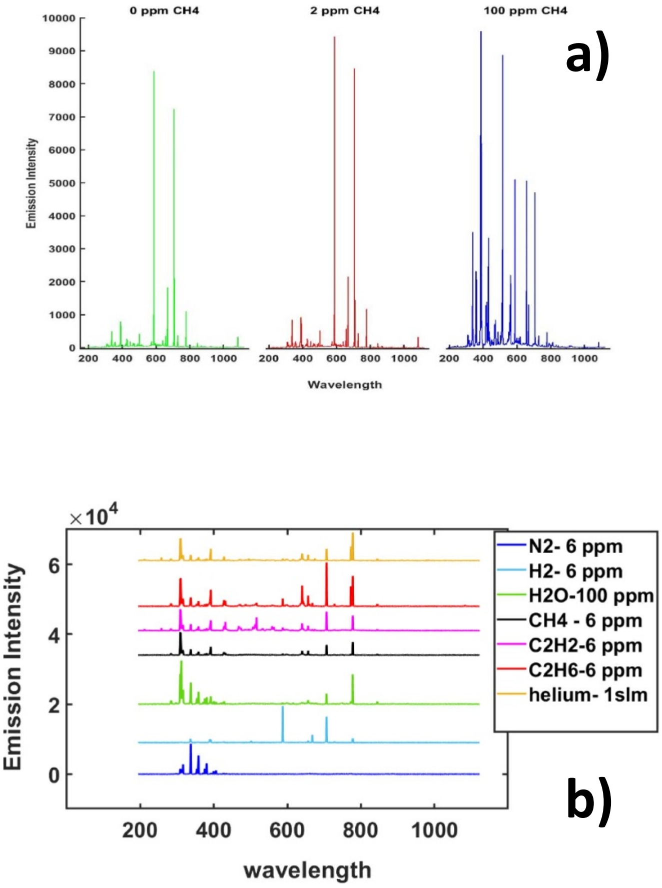

Figure 3. (a). Illustrative spectra from samples with 0 ppm, 2 ppm and 100 ppm CH_4, truncated to the wavelength range 194–1150 nm. Category increment will cause a boost in peaks number and lead these data to a categorical nature. The spectra are collected via plasma no.2. (b). Comparing spectra from two different plasma (plasma no.1 &2)- *note*: H2 and CH4 recorded via plasma no. 1 and remaining spectra are collected via plasma no.2 in this figure.

called Latent Variables (LV), and conducts Least Square regression on these component sets [16]. Classification with PLS is termed PLS-DA, where DA represents discriminant analysis.

This classification technique can deal with multivariate and high-dimensional data by reducing the number of independent variables X or samples to LV with a maximum covariance with the corresponding dependent variable Y or classes [17]. This algorithm has favorable properties compared to other algorithms when applied to spectral data. It can provide weight and model VIP, which can be utilized to identify variables that participate the most in the model [18]. In addition, comparison of loadings and score plots can help discover important variables in each class [19–21]. As described in a previous work [1], PLSDA represents the best model performance on our CH4–He spectral data. The algorithm was applied to separate and group the samples (ppm) in classification section. Figure 5 shows the application of leave-one-out cross-validation (LOOCV) on the raw CH4 data when $LV = 15$. In this approach, the cross-validation algorithm validates each sample against the others in each LV. Increasing the number of LV improved the model performance by up to 98% for $LV = 15$. The PLS-DA accuracy is the fraction of predictions that the model obtained accurately when classifying PPMs. The process is discussed in the following sections.

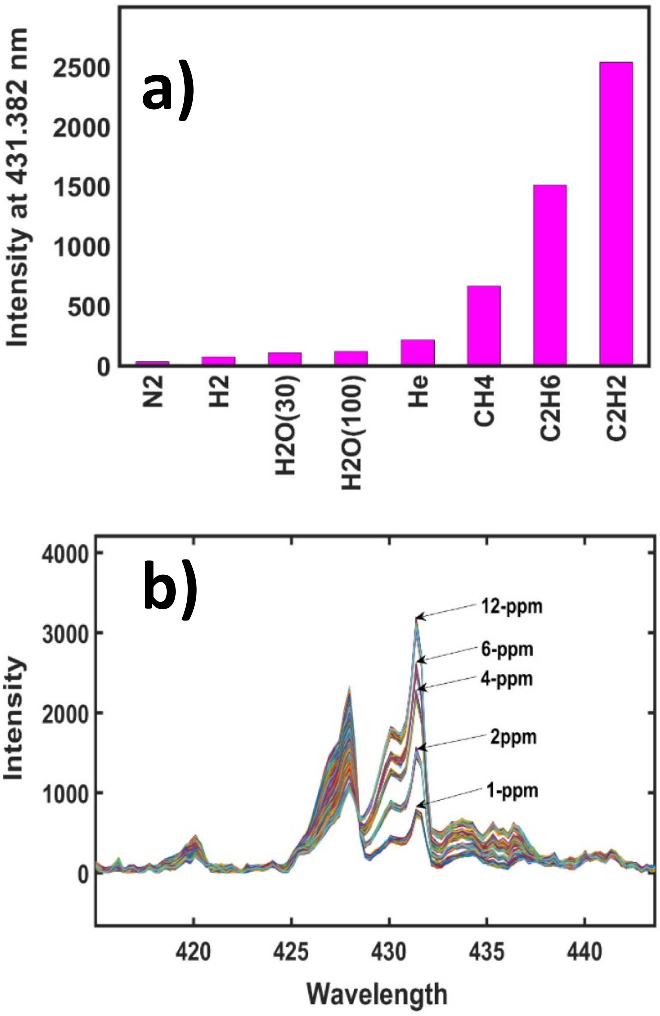

Figure 4. (a) Spectral intensity at wavelength 431.382 nm for different gases in helium. (b) Increasing intensity with C_2H_2 concentration's growth at wavelength 431.382 nm. Hundred samples are recorded for each ppm.

3.1. Gas identification via unique VIPs

PLSDA has been shown ability to detect a single trace gas in helium, but its accuracy is significantly reduced when multiple gases are included in the model [1]. In this section, the objective is to search for specific peaks that are unique to individual gas, and then with an unknown spectrum, such features are used to select gas-specific models. Figure 6 illustrates the decision-making process for the gas identification and concentration determination.

Therefore, to determine the gas-specific peaks, they must be unique to the gas or have a significantly larger intensity than the same peak as other gases. Such peaks are also likely to be considered important variables by the gas model i.e. they have a VIP score >1 [22]. Therefore, the VIP scores were analyzed to find the required features. In this study, the number of VIP score outputs for each complete dataset matched the total number of variables (i.e. 3648 scores).

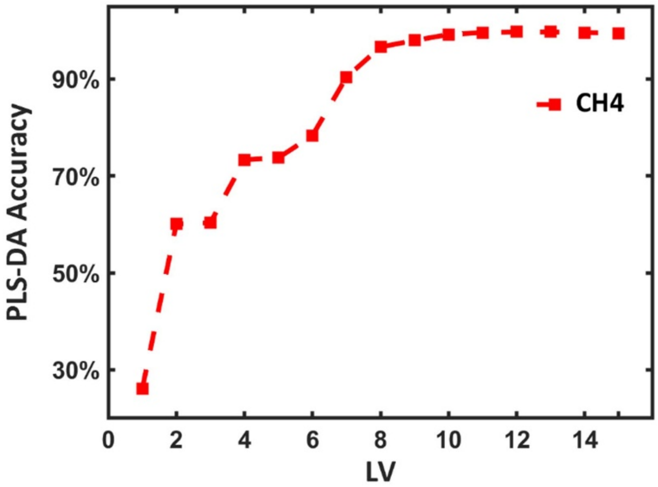

Figure 5. Applying Leave One Out Cross Validation on CH_4 -He dataset. PLSDA LOOCV's iteration is usually equal to the number of latent variables. Here, 15 latent variables are considered.

Figure 7 compares variables with a VIP score >1 for the CH_4 dataset.

The most important variables were located on the peaks, and there was a direct correlation between the VIP height and intensity. If a variable has a VIP score of < 1, it can be considered less important and possibly removed from the model. Variables with a VIP score > 1 are important in the model.

The VIP score plot is complicated. Variables can have different VIP scores in each category (concentration). Each gas may also have different VIPs scores. Finally, each peak region covers multiple wavelength variables, and may therefore contain multiple VIPs. To address the indicated challenges, a three-phase process was developed to filter the number of VIPs and select one VIP per peak, as shown in figure 8. Phase one: Select all VIP scores > 1(black line); phase two: within each peak region, select the highest value VIP score (green line); phase three: select the 10–20 highest value VIP scores from phase two (blue line) using the VIP threshold.

All three phases is applied on all gases and their concentrations separately. Figure 9 shows selected VIPs (phase-1) for 6ppm C_2H_6 at wavelength interval 406–442 nm. As the figure indicates, six VIPs are observed near the 431 nm line (range 429.28–431.90 nm, containing 8 wavelength variables). Adjacent to this peak, there is the helium line at 427 nm (with a peak width from 425.35 nm until 428.76 nm) that supplies another six VIPs. As shown in figure 4(a), the 431 nm line is only observed as a significant peak when hydrocarbons are present, and any single observed VIP in this interval (429.28–431.90 nm) indicates the presence of hydrocarbons in the mixture.

Although each mixture displays a high degree of similarity in the highest-value VIP wavelengths selected in phase three, it was found that each gas has a unique VIP score located at a specific wavelength interval that is not available at the same point for other gases (figure 10). Therefore, some wavelength

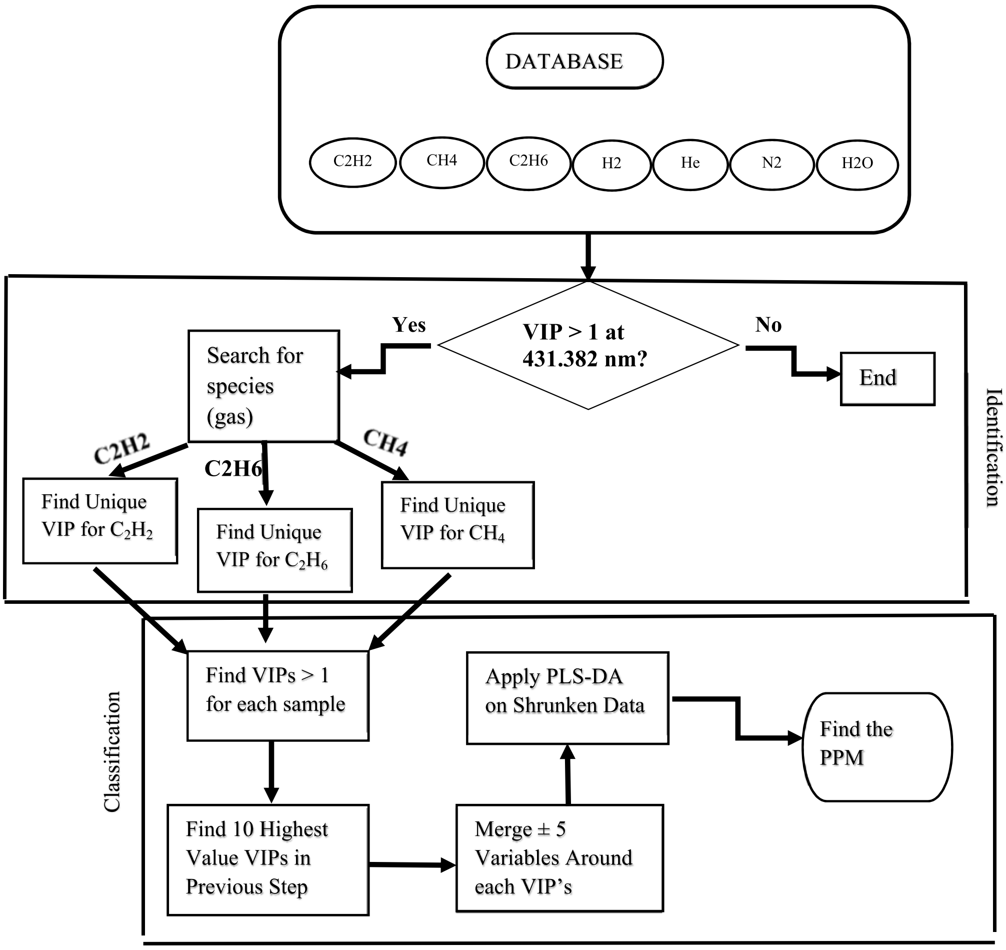
DATABASE
C2H2
CH4
C2H6
H2
He
N2
H2O
VIP > 1 at 431.382 nm?
Yes
No
End
Search for species (gas)
C2H2
C2H6
CH4
Find Unique VIP for C_2H_2
Find Unique VIP for C_2H_6
Find Unique VIP for CH_4
Identification
Classification
Find VIPs > 1 for each sample
Find 10 Highest Value VIPs in Previous Step
Merge ± 5 Variables Around each VIP's
Apply PLS-DA on Shrunken Data
Find the PPM

Figure 6. To identify hydrocarbons (CH_4–He, C_2H_6–He, and C_2H_2–He), eight different gases with concentrations (0–100 ppm) were recorded separately. Step 1 (Identification): searching for VIP > 1 at 431 nm, and if it exists, identifying gas via a unique VIP. Step 2 (classification): Merge ± 5 variables around each peak to cope with model overfitting and acknowledge each gas concentration.

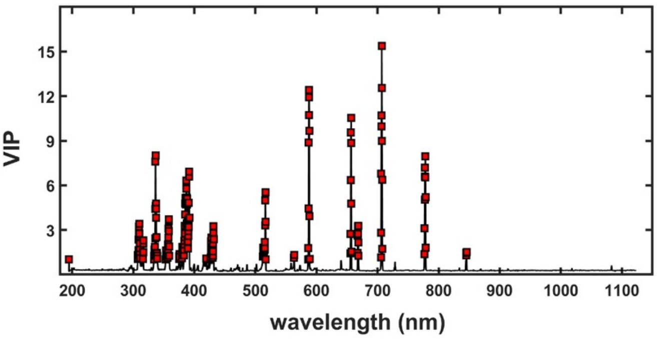

Figure 7. Selected VIPs for CH_4–He recorded with plasma 1. The red dot shows VIPs > 1.

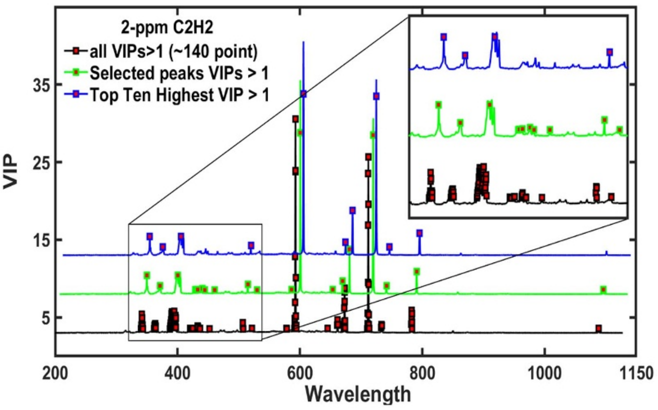

Figure 8. Finding highest value VIP in three stages, zooming window demonstrates wavelength interval ~ 300–534 nm. Red squares show selected VIPs in each step. Stage1: black line indicates selected all VIPs > 1 for 2-ppm C_2H_2. Stage2: green line selected highest VIP for each wavelength range for the same sample. As the zooming window shows the number of VIPs have decreased from black line to green line (select highest VIP for each peak). Stage3: blue line shows selected 10–20 highest VIPs that located on highest peak.

peaks could be utilized to distinguish one gas from the others.

A comparison of the three gases (figure 10(b)) shows that there are unique peaks for C_2H_2 in the wavelength interval 510 nm–610 nm, which provide unique VIP scores > 1. However, the peak intensity was not high enough for CH_4 and

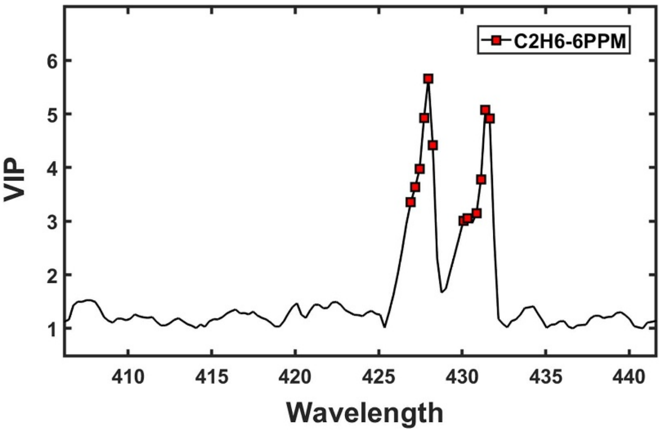

Figure 9. Selected VIPs (phase 1) for 6ppm C_2H_6 at wavelength 431 nm. (Hydrocarbon identifier spot).

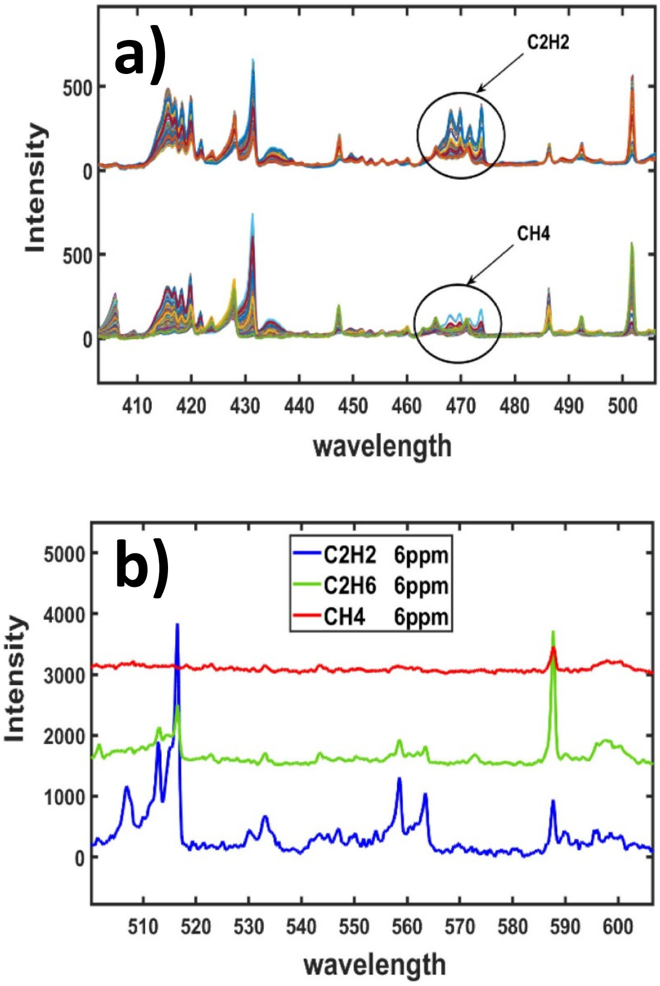

Figure 10. (a) Peak differences between CH_4 & C_2H_2 at wavelength interval 410 nm–500 nm.(b). Comparing peaks for CH_4, C_2H_6 & C_2H_2 at wavelength interval (∼510–610 nm). Two unique peaks (VIPs) belong to C_2H_2 are identified at 513 & 557 nm. The baseline for CH_4 & C_2H_6 is shifted to avoid line overlap.

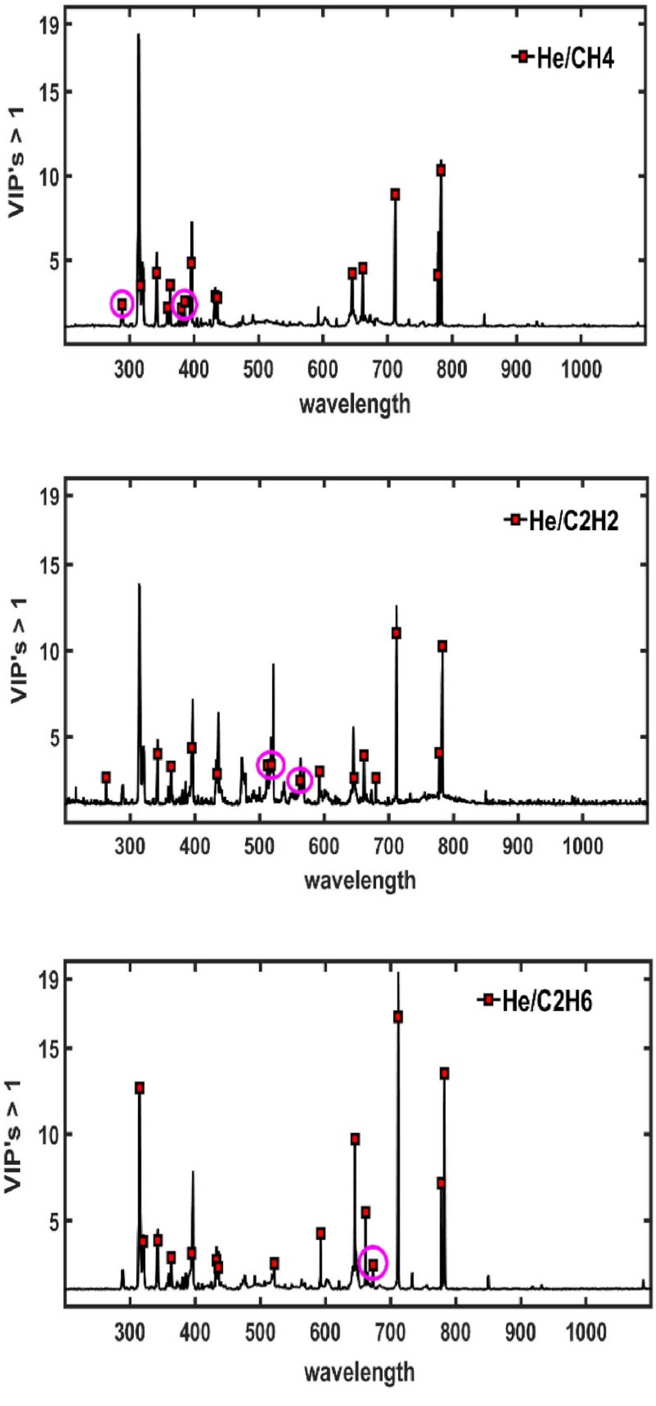

Figure 11. Black line: selected VIP scores >1 for CH_4, C_2H_6 & C_2H_2. Red circles on this line shows unique VIP for each. Unique VIP for CH_4 at wavelength 283 & 375 nm. Unique VIP for C_2H_6 at wavelength 668.07nm. Unique VIPs for C_2H_2 at wavelength 513.36 nm & 557.72 nm.

C_2H_6 to be captured by means of the VIP algorithm in this interval. The unique VIPs for each hydrocarbon gas are shown in figure 11.

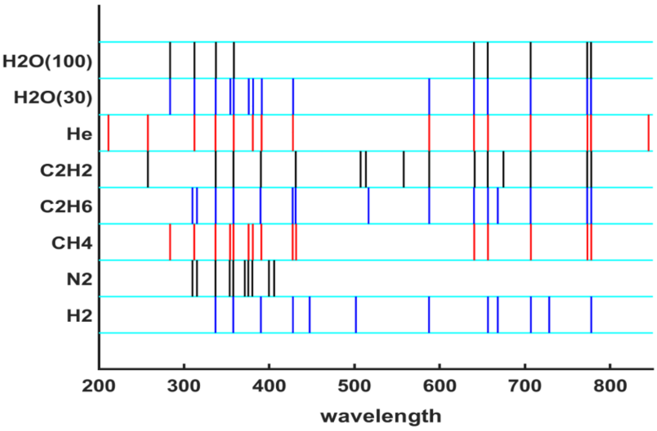

Figure 12. Selected highest value VIPs (phase 3) belong to eight recorded gases (N_2, H_2, CH_4, C_2H_6 & C_2H_2, H_2O & helium). Two concentrations have been considered for H_2O.

Table 1. Protocols for implementation of algorithm training & evaluation. No C_2H_6 available in our data repository from plasma no.1. As discusses in experimental set up, 100 samples were recorded for each concentration.

Protocol	Description	Plasma 1	Plasma 2
1	Model training and evaluated using C_2H_2 dataset		✓
2	Model training and evaluated using CH_4 dataset		✓
3	Model training and evaluated using C_2H_6 dataset		✓
4	Model trained and evaluated using two C_2H_2 datasets.	✓	✓
5	Model trained and evaluated using two CH_4 dataset.	✓	✓

A comparison of the highest value VIPs (phase 3) belonging to the eight recorded gases is shown in figure 12. H_2O was recorded twice for two concentrations, i.e., 30 ppm and 100 ppm, as the nature of its data is slightly dissimilar to other gases and VIP numbers usually decrease with increasing ppm.

It is important to point out that although the same threshold value for the VIP algorithm was used for each gas, three mixtures (N_2, H_2, and H_2O-100) provided >15 VIPs. VIPs corresponding to hydrocarbon species (approximately 431 nm) were only found in the C_xH_y mixtures, namely 431.382 nm for CH_4, 430.858 nm for C_2H_6, and 429.285 nm for C_2H_2. The line related to helium (i.e. 427) can be identified in most mixtures, except (N_2, C_2H_2 and H_2O). Apart from nitrogen, all samples indicated a VIP $>$ 1 score at 777 nm due to the presence of oxygen (OI).

3.2. Concentration classification via merging VIPs

To evaluate and improve the ppm prediction accuracy of the algorithm, the model performance was evaluated using five different protocols, as listed in table 1.

This approach was assessed on two different CH_4 datasets and has been discussed in a previous study [1]. As table 1 shows, protocols 1–3 represent individual or combined session evaluations, while protocols 4 and 5 use datasets from plasma one for training and testing on similar species from plasma two. As a LOO-CV approach was applied to acquire an estimate of the model sensitivity to the number of LV used to build the model, from which a plot of accuracy versus LV was obtained (figure 13). LOOCV is a special case of k-fold cross validation that test each sample against all other individuals [23]. Therefore, if a dataset contains **n** number of samples, LOOCV can be iterated **n** times for all dataset (i.e. **k** $=$ **n**), when **n–1** samples will be used for training and one sample for testing in each permutation.

In each protocol, the training and validation samples were swapped, and the results changed by approximately $>4\%$. For each individual gas model tested via protocols 1–3, the accuracy increases with LVs growth and saturates at >90% for LV values $>$ 8. However, for protocols 4 and 5, the models were unsatisfactory, and the accuracy was poor. The previously mentioned protocols represent an extreme test using entirely different plasma sources, and the low accuracy is likely due to overfitting of the unique response characteristics of each system.

The software can be locked to a gas as soon as identification is complete, and from there, the VIP summation can proceed to cope with model overfitting, as can be seen in protocols 4 and 5. The analyzing process will occur in real-time, and the identification-classification result is simultaneous. The identification accuracy for $C_2H_{2,}$ C_2H_4 and C_2H_6 varied between 95%–97% for each gas.

From [1, 24], it was found that peak merging is a valid solution to improve the accuracy and ppm classification. In this approach, the highest value peaks over a wavelength range $\Delta\lambda$, selected by a VIP threshold, and shrinks into a single intensity value. The VIP algorithm merges $\pm$ 5 intensities over λ; as a result, a single compressed intensity value within $\Delta\lambda$ remains, and the remaining variables are eliminated from the model. This technique reduces the correlated variables around a peak and minimizes the problem of model overfitting [25].

As mentioned in the previous section, the most significant variables (VIPs) are located at their peaks. The VIP algorithm will be repeated to identify peaks and perform ± 5 merging around each selected VIP (figure 14). This method can abolish the ambiguity and changeability of each category; however, as explained earlier, each mixture displays much similarity in the highest-value VIP wavelengths. Therefore, the summation procedure (peak merging) can be performed only after mixture identification.

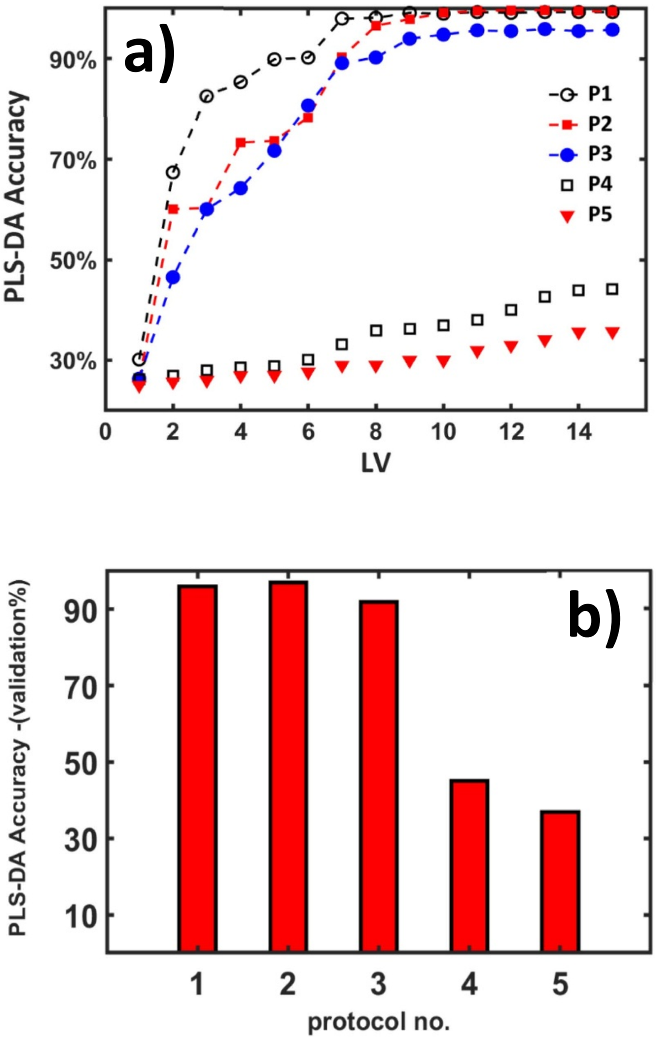

Figure 13. (a) Comparison of LOOCV PLS-DA accuracy versus the number of model Latent Variables for five protocols given in table 1. (b) Corresponding PLSDA validation result- Each bar shows multiclassification result for defined protocols.

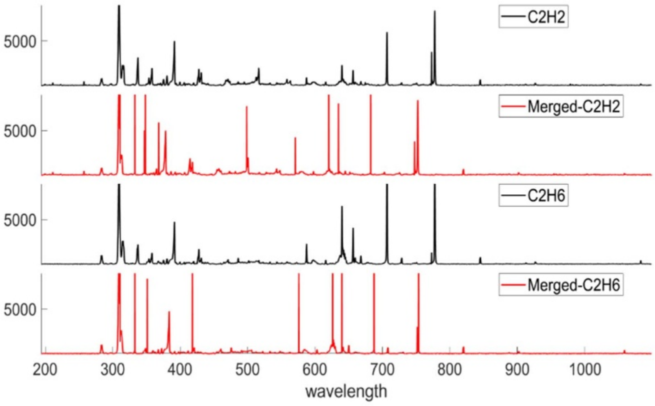

Figure 14. 2 ppm C_2H_2 & C_2H_6 after & before merging as the comparison between merged and unmerged samples show there is slightly shift in wavelength when merging algorithm reducing the number of wavelengths by 100 variables. The result of this concentration classification via merging is shown in figure 15.

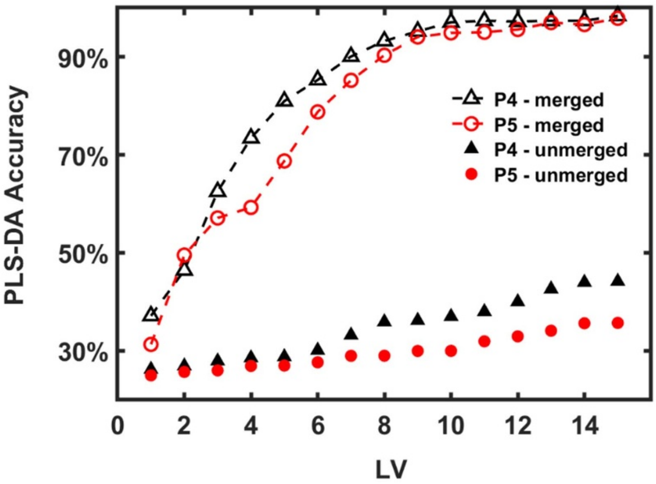

Figure 15. PLS-DA accuracy versus LV using peak merging for Protocol 4 (C_2H_2) and 5 (CH_4), in comparison with accuracy obtained from unmerged peaks.

The peak-merging algorithm was not necessary for protocols 1, 2, and 3, as the accuracy was >95% with only the PLSDA algorithm. However, for protocols 4 and 5, it was also necessary to include the merging algorithm, which it ensure an accuracy of ⩾95% for the nine LVs, as shown in figure 15.

4. Discussion

PLSDA algorithm in addition to being a dimensionality reduction technique, it is also equipped with properties such as VIP selection, which makes it an appropriate approach for gas identification. An analysis of the full variable count models using VIP indicated, as expected, that the primary contributors to the models were at the spectral peaks. Specific peaks unique to individual gases can be identified using the VIP threshold. Therefore, for each gas, a VIP threshold outputs the highest value peaks that are rarely shared with other gases. Each selected peak covers features that they call the VIP scores. Initially, the total number of VIP scores is equal to the total number of variables (i.e. 3648 scores), which will be filtered in three different phases, and only around 10–15 scores will enter the final stage for identification and classification purposes. Under realistic test scenarios, trace-gas mixtures are likely to contain air. In this study, all gases were mixed with helium, and the effect of air was minimized by adding a longer capillary followed by a 50 cm tube. The preliminary investigation of the effects of air inclusions on gas detection and plasma operation was beyond the scope of this study. Pure helium (zero C_xH_y) VIPs were precisely evaluated to investigate the effect of helium on each gas.

Figure 12 and highest value VIPs shows that the line at a wavelength of 337 nm, captured via the VIP algorithm, can be observed in pure helium. This wavelength was also used to identify OH (as shown by [26]) and N_2 (as shown by [27]). CH lines in the range 410–440 nm- transition: $A^2\Delta$—$X^2\Pi$ are not found in pure helium, other than the specific line 427 nm, which correlates with helium. It seems that line 427, which belongs to helium, can be seen in all recorded sets, apart from N2, C_2H_2 and heavy molecules of H_2O (i.e. 100 ppm). It is possible that the stronger energy bond between C=H in C_2H_2 and N=N diminishes the present of helium at 427 nm. A stronger C–H bond results in decreasing hydrocarbon VIPs prior to 431 nm, i.e. 9 wavelength lines are detected for CH_4 ahead of 431, six lines are discovered for C_2H_6 and four for C_2H_2. It appears that the VIP lines belonging to N_2 are virtually congested at interval of 300 nm–400 nm.

The VIP identified at 380 nm may belong to N_2 [28, 29] and is found in nitrogen sets, but not in helium mixture. The line at 375 was used by [29] to identify CO_2 and by [28] for N_2, it cannot be seen in pure helium mixture. Conversely, VIP captured at 656 nm was observed in the helium set. C II lines those found at 588.8 and 589.3 nm by [30], can be seen in H_2O, C_2H_6, C_2H_2 and helium. The 777 nm line that may belong to OI [31], can be seen in all sets apart from the N2 samples. The hydrocarbon detector line (431 nm) was captured via the VIPs algorithm for all C_xH_y mixtures, although at a completely different height for each mixture. Finally, some other line differences between sets have been identified, such as 257 nm, 283 nm, and 640 nm, but their appearance is inconsistent.

Regarding PPMs classification, the potential for overfitting of spectral data and a decrease in accuracy for a given LV is obvious from protocols 4 and 5, where the training and validation were from two different plasmas. In contrast, the peak uniting algorithm allowed 97% success with protocols 4 and 5, a similar outcome to that obtained from protocols 1–3. By shrinking the number of correlated wavelength variables around each highest-value VIP to a single value, the consequences of overfitting were minimized.

5. Conclusion

The ability to detect hydrocarbons and assign a concentration classification offers a fundamental solution for model overfitting and vapor identification, particularly when dealing with gas spectral data. In this study, we initially considered eight different gases. The first step in identifying a hydrocarbon is to confirm the presence of a wavelength line at 431 nm. This is the only wavelength where an increase in the peak height indicates an increase in the C_xH_y concentration. The unique VIP of each gas was used to distinguish the hydrocarbons. Although there was much similarity between the VIP scores of each gas, each hydrocarbon shows a few different VIPs lines that may not be present for other gases. While certain specific wavelength points were identified, which are indicating the presence of C_xH_y species, the general hypothesis is that a unique VIP at a wavelength interval of 200–400 nm indicates methane, 500–600 nm shows acetylene and 600–700 nm indicates ethane.

Diminishing algorithm execution is visible when moving from one plasma set to another; consequently, model overfit the training data and has reduced generality. The PLSDA algorithm adjoining variable merging could deal with model overfitting and improve the performance by up to 98%. Future research will involve assessing the potential role of the algorithm in identifying unique VIPs for other hydrocarbons and vapors. This would make the current work more robust and generalizable to other fields. Once this has been achieved, the improvement and further development of current algorithmic methods will be the next step. More work needs to be performed to assess how the algorithm responds to complex mixtures of hydrocarbons, such as a combination of CH_4 and C_2H_6 mixed with air. Finally, this identification applies to the plasmas (1 and 2) that we investigated and to those hydrocarbons. It remains to be seen how this applies to other plasmas and hydrocarbons.

Data availability statement

The data cannot be made publicly available upon publication because they are not available in a format that is sufficiently accessible or reusable by other researchers. The data that support the findings of this study are available upon reasonable request from the authors.

Conflict of interest

There is no conflict of interest statement.

ORCID iDs

Tahereh Shah Mansouri https://orcid.org/0000-0003-4710-0546
Davide Mariotti https://orcid.org/0000-0003-1504-4383
Paul Maguire https://orcid.org/0000-0002-2725-4647

References

[1] Mansouri S, Tahereh H W, Mariotti D and Maguire P 2022 Methane detection to 1ppm using machine learning analysis of atmospheric pressure plasma optical emission spectra *J. Appl. Phys.* **55** 225205

[2] Ferris S W 1955 Introduction *Handbook of Hydrocarbons* (Academic) pp 1–17

[3] Rothbart N, Schmalz K, Borngräber J, Kissinger D and Hübers H-W 2018 Detection of volatile organic compounds in exhaled human breath by millimeter- wave/terahertz spectroscopy *2018 43rd Int. Conf. on Infrared, Millimeter, and Terahertz Waves (Irmmw-thz)* (*Nagoya, Japan*) p 1

[4] Thompson W B 2013 *An Introduction to Plasma Physics* 2nd edn (Elsevier Science)

[5] Janev R K and Reiter D 2002 Collision processes of CH_y and $CH_y{}^+$ hydrocarbons with plasma electrons and protons *Phys. Plasmas* **9** 4071

[6] Sun W, Uddi M, Won S H, Ombrello T, Carter C and Ju Y 2012 Kinetic effects of non-equilibrium plasma-assisted methane oxidation on diffusion flame extinction limits *Combust. Flame* **159** 221–9

[7] Fantz U and Meir S 2005 Correlation of the intensity ratio of C_2/CH molecular bands with the flux ratio of C_2Hy/CH_4 particles *J. Nucl. Mater.* **337–339** 1087–91

[8] Vermeiren V and Bogaerts A 2020 Plasma-Based CO_2 Conversion: to quench or not to quench? *J. Phys. Chem.* C **124** 18401–15

[9] Hendawy N, McQuaid H, Mariotti D and Maguire P 2020 Continuous gas temperature measurement of cold plasma jets containing microdroplets, using a focussed spot IR sensor *Plasma Sources Sci. Technol.* **29** 085010

[10] Kojima J, Ikeda Y and Nakajima T 2005 Basic aspects of OH(A), CH(A), and C_2(d) chemiluminescence in the reaction zone of laminar methane–air premixed flames *Combust. Flame* **140** 34–45

[11] Shogun V, Tyablikov A, Schreiter S, Scharff W, Wallendorf T and Marke S 1998 Emission actinometric investigations of atomic hydrogen and CH radicals in plasma-enhanced chemical vapour deposition processes of hexamethyl disiloxane *Surf. Coat. Technol.* **98** 1382–6

[12] Ma J, Ashfold M and Mankelevich Y 2009 Validating optical emission spectroscopy as a diagnostic of microwave activated $CH+/Ar/H_2$ plasmas used for diamond chemical vapor deposition *J. Appl. Phys.* **105** 043302

[13] McCord W, Gragston M, Wu Y, Zhang Z, Hsu P, Rein K, Jiang N, Roy S and Gord J 2019 Quantitative fuel-to-air ratio determination for elevated-pressure methane/air flames using chemiluminescence emission *Appl. Opt.* **58** C61

[14] Zhou C *et al* 2021 Use of plasma electron spectroscopy method to detect hydrocarbons,alcohols, and ammonia in nonlocal plasma of short glow discharge *Plasma Sources Sci. Technol.* **30** 117001

[15] Saifutdinov A I and Sysoev S S 2023 Numerical simulation and experimental diagnostics offast electron kinetics and plasma parameters in a microhollow cathode discharges in helium *Plasma Sources Sci. Technol.* **32** 114001

[16] Abdi H 2010 Partial least squares regression and projection on latent structure regression (PLS Regression) *WIREs Comput. Stat.* **2** 97–106

[17] Song W, Wang H, Maguire P and Nibouche O 2017 Local partial least square classifier in high dimensionality classification *Neurocomputing* **234** 126–36 .

[18] Krishnan A, Williams L, McIntosh A and Abdi H 2011 Partial least squares (PLS) methods for neuroimaging: a tutorial and review *NeuroImage* **56** 455–75

[19] Wold S, Sjöström M and Eriksson L 2001 PLS-regression: a basic tool of chemometrics *Chemometr. Intell. Lab. Syst.* **58** 109–30

[20] Barker M and Rayens W 2003 Partial least squares for discrimination *J. Chemom.* **17** 166–73

[21] Hasegawa K and Funatsu K 2012 Evolution of PLS for modeling sar and omics data *Mol. Inform.* **31** 766–75

[22] Mehmood T, Liland K, Snipen L and Sæbø S 2012 A review of variable selection methods in partial least squares regression *Chemometr. Intell. Lab. Syst.* **118** 62–69

[23] Berrar D 2019 Cross-validation *Encyclopedia of Bioinformatics and Computational Biology* (Elsevier) pp 542–5

[24] Vincent J, Wang H, Nibouche O and Maguire P 2020 Detecting trace methane levels with plasma optical emission spectroscopy and supervised machine learning *Plasma Sources Sci. Technol.* **29** 85018

[25] Ying X 2019 An Overview of overfitting and its solutions *J. Phys.: Conf. Ser.* **1168** 022022

[26] Thiyagarajan M, Sarani A and Nicula C 2013 Optical emission spectroscopic diagnostics of a non-thermal atmospheric pressure helium-oxygen plasma jet for biomedical applications *J. Appl. Phys.* **113** 233302

[27] Sahu B B, Jin S B and Han J G 2017 Development and characterization of a multi-electrode cold atmospheric pressure DBD plasma jet aiming plasma application *J. Anal. At. Spectrom.* **32** 782–95

[28] Zaplotnik R, Primc G and Vesel A 2021 Optical Emission spectroscopy as a diagnostic tool for characterization of atmospheric plasma jets *Appl. Sci.* **11** 2275

[29] Khan M, Rehman N, Khan S, Ullah N, Masood A and Ullah A 2019 Spectroscopic study of CO_2 and CO_2–N_2 mixture plasma using dielectric barrier discharge *AIP Adv.* **9** 085015

[30] Fazekas P, Keszler A, Bódis E, Drotár E, Klébert S, Károly Z and Szépvölgyi J 2014 Optical emission spectra analysis of thermal plasma treatment of poly (vinyl chloride) *Open Chem.* **13** 549–56

[31] Milosavljević V, Donegan M, Cullen P and Dowling D 2014 Diagnostics of an O_2–He RF atmospheric plasma discharge by spectral emission *J. Phys. Soc. Japan* **83** 014501