Applied Physics B (2022) 128: 161
https://doi.org/10.1007/s00340-022-07879-8

Applied Physics B
Lasers and Optics

A support vector machines framework for identification of infrared spectra

M. Arshad Zahangir Chowdhury[1] · Timothy E. Rice[1] · Matthew A. Oehlschlaeger[1]

Received: 3 July 2021 / Accepted: 20 July 2022 / Published online: 5 August 2022

Abstract

In prior work (Chowdhury, M.A.Z., Rice, T.E. & Oehlschlaeger, M.A., Appl. Phys. B 127, 34 (2021)), we found support vector machines (SVM) to be adept at learning patterns from spectral data within a THz frequency range (7.33–11 cm^{-1}) for the purposes of gas-phase speciation. Here, we implement SVM, in a one-versus-rest framework, for the classification of infrared spectra in a broad frequency range (400–4000 cm^{-1} or 2.5–25 μm) for 34 gas-phase compounds at pressures ranging from 0.1 to 1 atm and for absorber mole fractions from 1 ppm to 1 (pure gases). Within the SVM framework, hyperparameters for the classifier were optimized to choose an optimum kernel for the SVM and acceptable soft margin constant to minimize misclassifications. The framework is tested using cross-validation strategies to determine the dependence of performance on variation in pressure and absorber concentration. Validation was carried out by considering experimental absorption spectra, from the literature, in three random trials, where the combined experimental classification accuracy was greater than 90%. A simulated spectral dataset containing artificial noise was used to further evaluate the SVM classifier in studies where the frequency range and resolution were varied, to better interrogate the capabilities of the SVM framework.

1 Introduction

A basic problem in spectroscopy is the identification of an unknown substance(s) from a measured spectrum. The problem poses several complexities but at its fundamental level requires the matching of features or descriptors, that characterize the unknown spectrum, with a previously measured and identified spectrum. These descriptors could be frequency range, maximum absorbance peak, number of peaks, spacing between successive peaks, frequency location of peaks, and others. Manual matching such descriptors can be tedious, dificult, and unreliable. Therefore, an automated framework to consistently and accurately classify unknown spectra is desired. A successful framework for the automated classification of spectra could be much faster than human manual matching and is expected to result in fewer misclassifications. Additionally, such a framework should be extendable to a variety of different spectroscopy types and frequency ranges.

Supervised machine learning (ML) algorithms, are appropriate tools for automated spectral classification, as they learn patterns from training observations belonging to different categories or classes and can be used to determine the category or class of a new observation, based on the learned pattern. Many ML algorithms have been reported in the literature for pattern matching and in our prior work, we have evaluated the capabilities of several popular classifiers for the identification of absorption spectra in the THz region, that result from rotational transitions [1]. The classifiers evaluated in that study were decision trees [2, 3], random forests [4, 5], boosted decision tress and random forests [6, 7], k-nearest neighbors [8, 9], fully connected feed forward neural networks (i.e., multilayer perceptrons [10–13]), and support vector machines (SVMs) [14, 15]. While these methods are well established, their prior application in spectroscopy and/or gas sensing is relatively limited, particularly in regards to absorption spectroscopy.

SVM is an established method for pattern recognition, classification, and regression. Developed using statistical learning theory, SVMs recognize patterns by constructing a data-separating decision line or surface, mathematically known as a n-dimensional hyperplane, to establish a decision rule. The SVM hyperplane achieves optimal separation between training observations belonging to two classes, in

✉ M. Arshad Zahangir Chowdhury
chowdm@rpi.edu

[1] Department of Mechanical, Aerospace, and Nuclear Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA

our cases molecules, establishing the widest possible distances between those classes [15–20]. Since SVMs are by design a binary classifier, capable of separating two classes, they are extended to multi-class classification problems using the one-versus-rest (OVR) strategy [21, 22]. In our prior work, we found that OVR-SVM classifier provided the highest classification accuracy for identification of THz absorption spectra, relative to the other ML methods described above [1].

SVMs offer computational advantages, particularly for high-dimensional and large datasets. Instead of using all training observation, SVMs only use the borderline or marginal observations (support vectors) to construct decision rules. Therefore, with the introduction of new training observations, the decision boundary remains unchanged, unless those new training data are at the margins and modify the support vectors. Hence, once the decision boundary is determined from training observations, storing the training observations is not necessary and retraining is only required if new training observation alter the support vectors.

In the infrared (IR) frequency region, vibrationally active molecules have unique absorbance spectra owing to stretching, bending, or twisting oscillations [23, 24]. IR absorption spectra can be quite complex for polyatomic molecules, due to numerous vibrational modes and coupled rotation and vibration. The presence of specific functional groups within a molecule also uniquely contributes to the shape or pattern present in an infrared spectrum. Often functional groups (e.g., hydroxyl, carboxyl, amine) can be identified from a specific spectral signature within a sub-region of the IR, called the functional group region. The IR spectra of a molecule are often divided between a functional group region (1500–4000 cm^{-1}), indicating the presence of these specific functional groups, and a fingerprinting region (400–1500 cm^{-1}), containing a large number of peaks unique to a molecule. Manually matching both of these regions to reference spectra from curated IR databases help provide an understanding of the structure of a molecule. Although IR spectroscopy alone is rarely used to identify molecular structure, it is often used in conjunction with other methods such as nuclear magnetic resonance (NMR) spectroscopy. Such manual effort in matching spectra is prone to errors and motivates an automated ML pattern recognition framework to reduce misclassifications and allow rapid processing of vast numbers of IR spectra.

ML methods, particularly, neural networks have been implemented for classification of gas-phase species. Both fully connected feedforward neural networks, also known as multi-layer perceptrons and deep convolutional neural networks are capable of recognizing observations indicating the presence of gas-phase species [25, 26], however, at high computational training costs required for the optimization of thousands or millions of parameter. K-nearest neighbors, decision trees, and ensemble methods, with and without boosting, have also been applied to recognize patterns in gas sensor data for speciation [27–30]. Shortcomings of k-nearest neighbors include significant memory requirements and decision trees and ensemble methods are more apt to overfit training data, than other ML methods [11, 31].

SVMs have some advantages over the aforementioned methods. First, SVMs find the optimum or best decision boundary (hyperplane in n-dimensional space) to distinguish between classes [2, 14, 15, 20]. The hyperplane is constructed only using the support vectors, thus reducing memory requirements and allowing for fast retraining, compared to deep neural networks. Moreover, the number of parameters to be optimized within SVMs are fewer compared to other machine learning methods. SVMs can also use kernel functions to transform features in a non-linear feature space to a higher dimension for better classification performance. A regularization parameter of a SVM, the soft margin constant, also allows for a degree of user control over the number of misclassifications allowed by a SVM classifier. Furthermore, SVMs have been reported to identify gases from data generated by different sensor types, such as electronic noses [10, 11, 32, 33] and microelectronics-based THz spectrometers [34, 35].

The flexibility and advantages offered by SVM, coupled with the unique fingerprinting opportunities in the IR, motivates the present construction of a SVM-based framework for classification of gas-phase IR absorption spectra. The fast training times afforded by SVM also allow for the training of our framework over different frequency ranges and resolutions, depending on the availability of spectral patterns present within the IR and experimental data.

In this study, we constructed a SVM classifier to investigate its ability to recognize IR spectra in the frequency region from 400 to 4000 cm^{-1}. The classifier is based on a one-versus-rest (OVR) implementation of the SVM method and can be trained overs arbitrary frequency range, resolution, and with arbitrary number of compounds in the training set, where data are available. The OVR strategy allows for multiclass classification for identification of multiple compounds. The hyperparameters of the constructed SVM classifier were determined by a grid search cross-validation using a simulated dataset consisting of 1428 absorption spectra in a 70–30% training–testing split. A number of other datasets were prepared from simulations and experiments to evaluate the constructed SVM classifier and assess its capabilities. In total, absorption spectra of 34 compounds were simulated from fundamental spectroscopic parameters found in the HITRAN database [36]. K-fold cross-validations were performed for $k = 10$ and 7, to reveal the influence of pressure and absorber gas concentration on classifier performance, respectively.

2 Methodology

2.1 Problem overview and ML framework

The fundamental problem posed is the identification, or classification, of a molecule from two-dimensional data, presented in the form of absorption as a function of frequency, known as an absorption spectrum. In present classification problem, the spectra associated with each molecule, irrespective of their point of origin (i.e., from simulation or experiment), can be used as training data to construct a ML black box classifier algorithm, as schematically shown in Fig. 1. Data are preprocessed and then used to train, test, and validate a ML model that can identify associated labels. From Fig. 1, it is apparent that once the ML black box is trained, it alone sufices for making predictions directly from data without the need of intermittent preprocessing and classifier building steps. However, the stand-alone ML black box algorithm not only contains a trained classifier but it also contains the additional preprocessing steps, such as loading the data, feature extraction and check, resampling, etc.

The process flow of the ML approach used in the present study is depicted in Fig. 2. Absorption spectra are simulated and supplied as a dataset. A frequency region of interest is chosen which allows for feature extraction, where the number of features is defined. In the present study, features are used to construct a SVM-based classification model, as represented by a generic equation given below,

$$y = f(\boldsymbol{x}) \tag{1}$$

$$y \approx g(\boldsymbol{x}) \tag{2}$$

$$\boldsymbol{x} = x_1, x_2, \dots \dots \dots, x_N, \tag{3}$$

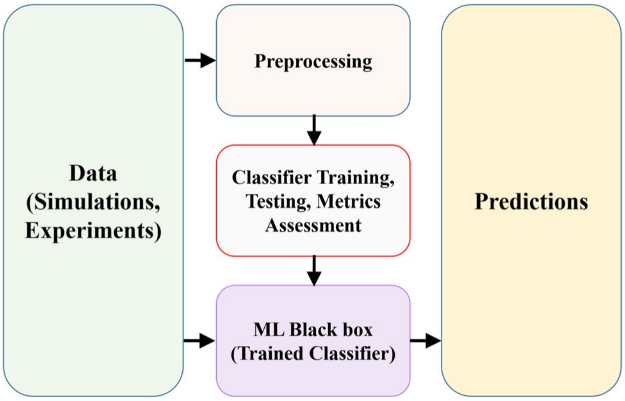

Data (Simulations, Experiments)
Preprocessing
Classifier Training, Testing, Metrics Assessment
ML Black box (Trained Classifier)
Predictions

Fig. 1 Schematic representation of a generic ML black box approach to make predictions from data

where $\boldsymbol{x}$ is the feature vector consisting of a number of features and N is the total number of features. The function f, the target function, represents the true relationship between the features and the molecule integer label index y. The function g is the hypothesis function and is an approximation of the target function f. Hence, a ML algorithm develops the function g by learning patterns in the data to predict label index y.

For the absorption spectra considered in the present study, the number of features is equal to the number of sampled points at frequency locations. For example, in our 400–4000 cm^{-1} frequency range, at a resolution of 1 cm^{-1} there are 3601 features. The features themselves are defined as the absorbance values at each frequency, for the conditions of the spectral simulation or experiment.

While not all features are equally important, and some could perhaps be excluded from consideration, due to the computational power available with modern consumer-grade processors, excluding such features does not provide any significant computational- or time-saving benefits. Moreover, while exclusion of features may be beneficial for identification of one molecule, it can often be disadvantageous for identification of another and sometimes that disadvantage is not easily detected. Hence, we have considered all available features with the exception of negligibly small absorbance values and compounds with negligible absorption within a frequency range under consideration are excluded from the training set.

Once the features for a simulated spectral dataset are established, the data are split into a training and a testing set, whose proportion may vary. In most cases, we used a 70–30% train–test split; however, other proportions have also been used to assess the performance of the model as specified in Table 1. Using the training and the testing dataset, the SVM classifier is constructed, where the hyperparameters (parameters that control how the model learns) are optimized using an empirical grid search cross-validation algorithm (GridSearchCV). Once the hyperparameters are optimized, model parameters (weights, w and bias, b) are determined from the training spectra. The GridSearchCV optimization process, adjusts the hyperparameters until a specified threshold for testing accuracy is met. For spectra over the broad frequency range of 400–4000 cm^{-1}, we found that 95% accuracy to be an acceptable threshold.

Once the hyperparameters were established and the SVM model trained, we treat the trained classifier as a machine learning black box, where the black box contains all the preprocessing steps, including resampling, necessary for the classifier to identify a given spectrum. In the present study, both simulated spectra, with and without artificial superimposed noise, and experimental spectra were used for testing and validation. In testing and validation studies, the classifier performance was determined using the common metrics of

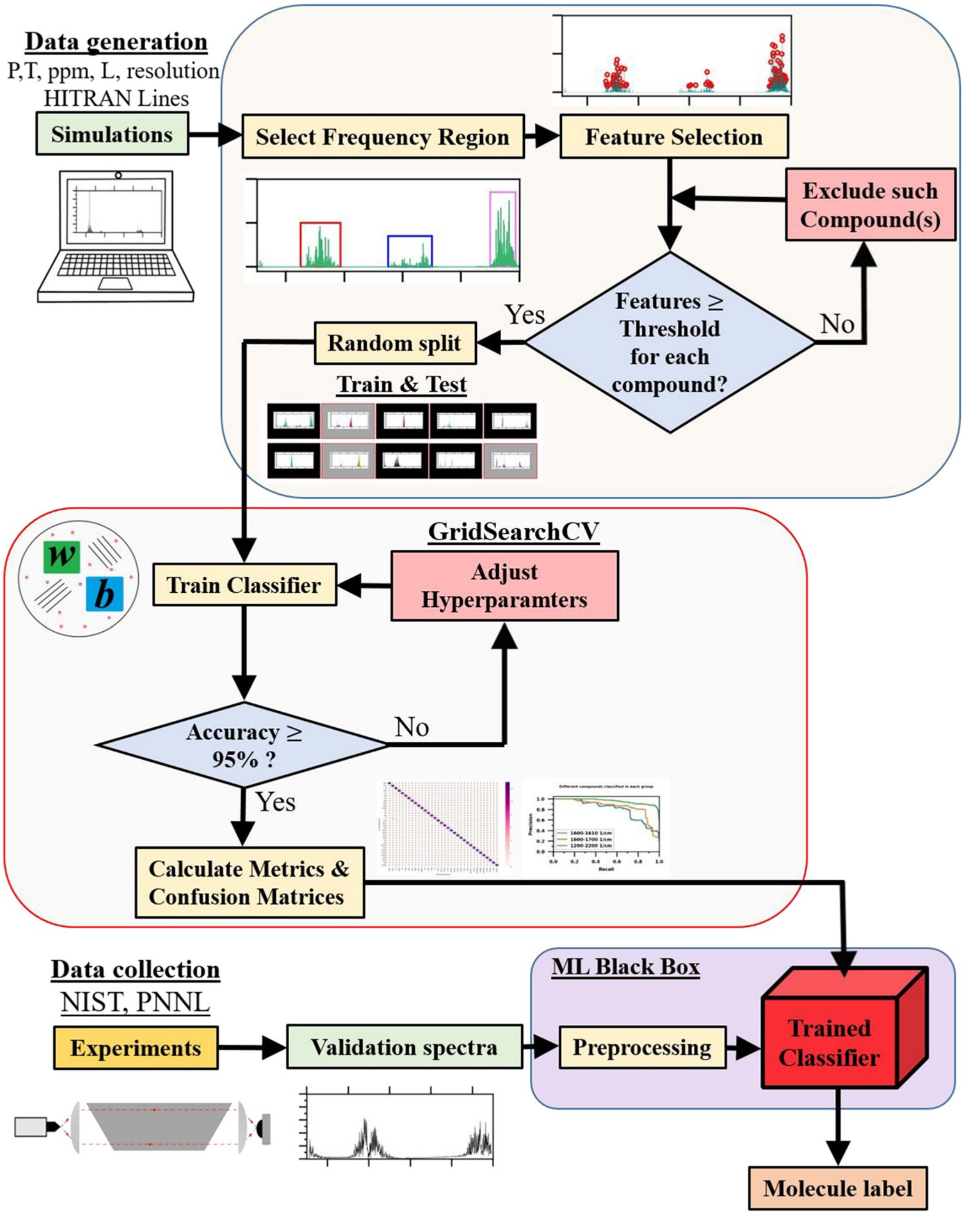


Fig. 2 Flowchart depicting ML implementation for prediction of molecule label

accuracy, precision, recall, and F1 score. Additionally, confusion matrices provide visual representation of classifier performance and precision-recall curves are also helpful for understanding classifier performance.

2.2 Datasets

A number of datasets were used in this study for training, testing, validation, and determination of hyperparameters for the classifier training, see Table 1. Datasets A-D consist of simulated spectra for all 34 compounds (Table 2), where all simulations were carried out at 297 K and 100 cm pathlength. Dataset A consists of noise-free simulated spectra for 0.1, 0.3, 0.5, 0.7, 0.9, and 1.0 atm and at 1 ppm, 10 ppm, 100 ppm, 1000 ppm, 1%, 10%, and 100% of each absorber dilute in N_2. The dataset A simulations were carried out at a range of frequency of 400–4000 cm^{-1} and at three resolutions (0.01, 0.1, and 1 cm^{-1}). Dataset B is the same as dataset A but with each spectra containing random artificial sinusoidal noise superimposed on those. In dataset A and B, each compound is equally represented with 42 spectra, with no imbalance; hence, each dataset consists of 1428 spectra per resolution. Dataset A was used for training and testing, in 70–30% training–testing split, hyperparameter optimization, and performance assessment. Dataset B was used for performance assessment under noisy spectral conditions.

Dataset C was designed to examine the influence of pressure variation on the performance of the classifier. It contains 340 simulated noise-free spectra, with ten spectra per compound (no imbalance). The total pressure is varied from

Table 1 Descriptions of the datasets used for training, testing, and validation of the SVM classifier

Dataset	Description	Purpose
A	- Simulated, noise-free - 34 compounds - 42 spectra/compound at each resolution - frequency range: 400–4000 cm^{-1} - resolution: 0.01, 0.1, and 1 cm^{-1} - conditions: 297 K; 0.1, 0.3, 0.5, 0.7, 0.9, 1.0 atm; 1 ppm, 10 ppm, 100 ppm, 1000 ppm, 1%, 10%, 100% absorber dilute in N_2; pathlength 100 cm - 4284 total spectra	- Classifier training and performance assessment (70%-30% train-test) - Hyperparameter optimization
B	- Dataset A with added random sinusoidal noise superimposed - 4284 total spectra	Performance assessment (validation) under noisy conditions
C	- Simulated, noise-free - 34 compounds - 10 spectra/compound at each resolution - frequency range: 400–4000 cm^{-1} - resolution: 1 cm^{-1} - conditions: 297 K; 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 atm; 100% absorber; pathlength 100 cm - 340 total spectra	Tenfold cross-validation to determine influence of pressure
D	- Simulated, noise-free - 34 compounds - 7 spectra/compound at each resolution - frequency range: 400–4000 cm^{-1} - resolution: 1 cm^{-1} - conditions: 297 K; 1.0 atm; 1 ppm, 10 ppm, 100 ppm, 1000 ppm, 1%, 10%, 100% absorber dilute in N_2; pathlength 100 cm - 238 total spectra	Sevenfold cross-validation to determine influence of absorber concentration
E	- 20 room temperature experimental spectra from NIST [38, 39] and PNNL [40]	Experimental validation

0.1 to 1.0 atm, in steps of 0.1 atm. The absorber concentration is fixed at 100% (undiluted) and the frequency resolution fixed at 1 cm^{-1} for the 400–4000 cm^{-1} frequency range. Dataset C was used for ten-fold cross-validation to investigate the classifier performance due to pressure broadening.

Dataset D was designed to examine the influence of absorber concentration on the performance of the classifier. It contains 238 simulated noise-free spectra, with 7 spectra per compound (no imbalance). The absorber concentration was varied from 1 ppm, 10 ppm, 100 ppm, 1000 ppm, 1%, 10%, to 100%. The total pressure was fixed at 1.0 atm and the frequency resolution fixed at 1 cm^{-1} for the 400–4000 cm^{-1} frequency range. Dataset D was used for seven-fold cross-validation to investigate the classifier performance with variation in absorber concentration or absorption signal intensity.

Dataset E consists of 20 experimental spectra, 1 for each molecule for which data was available. These compounds are listed in Table 2. Data were not available for all 34 simulated compounds across such a large frequency range.

All training–testing, hyperparameter optimization, cross-validations in pressure and concentration space, and noise investigations were performed with balanced datasets, with no bias towards any particular compound. For assessing classifier performance on experimental spectra, this was not possible, because only 20 experimental spectra were available. The counts of spectra, mean-normalized maximum absorbance and mean-normalized absorbance plots in the supplementary material for data analytics demonstrate that the different datasets are suficiently unique and there is good variability between the training and testing sets.

2.2.1 Simulated spectra

Simulated spectra were generated using the HITRAN spectroscopy database [36] and associated HITRAN Application Programming Interface (HAPI) code for spectral simulations [37]. Kochanov et. al. describes the absorbance spectra generation using HAPI in detail [37]. The HITRAN database contains data for 51 molecules. Here, we have considered a subset of 34 molecules out of the 51 molecules listed in HITRAN and presented their class labels, molecular formula and availability of experimental spectra in Table 2. Representative absorption spectra for these molecules are presented in Fig. 3, where it is observed that all molecules have a distinct spectral shape, i.e., “chemical fingerprint”, in the 400–4000 cm^{-1} frequency region. The label indices given in Table 2 are also referred to as global label indices

Table 2 List of compounds, labels, and sources of validation spectral data

Label Index	HITRAN ID	Name	Formula	Validation Experiment
0	1	Water	H_2O	NIST
1	2	Carbon dioxide	CO_2	NIST
2	3	Ozone	O_3	NIST
3	4	Nitrous oxide	N_2O	NIST
4	5	Carbon monoxide	CO	NIST
5	6	Methane	CH_4	NIST
6	8	Nitric oxide	NO	NIST
7	9	Sulfur dioxide	SO_2	NIST
8	10	Nitrogen dioxide	NO_2	
9	11	Ammonia	NH_3	NIST
10	12	Nitric acid	HNO_3	
11	14	Hydroflouric acid	HF	
12	15	Hydrochloric acid	HCl	PNNL
13	16	Hydrobromic acid	HBr	NIST
14	17	Hydroiodic acid	HI	
15	19	Carbonyl Sulfide	OCS	NIST
16	20	Formaldehyde	H_2CO	NIST
17	21	Hypochlorous acid	HOCl	
18	23	Hydrogen cyanide	HCN	
19	24	Chloromethane	CH_3Cl	NIST
20	25	Hydrogen peroxide	H_2O_2	
21	26	Acetylene	C_2H_2	NIST
22	27	Ethane	C_2H_6	NIST
23	28	Phosphine	PH_3	
24	31	Hydrogen sulfide	H_2S	NIST
25	32	Formic acid	HCOOH	
26	38	Ethylene	C_2H_4	NIST
27	39	Methanol	CH_3OH	
28	40	Methyl Bromide	CH_3Br	NIST
29	41	Acetonitrile	CH_3CN	
30	43	Diacetylene	C_4H_2	
31	44	Cyanoacetylene	HC_3N	NIST
32	47	Sulfur trioxide	SO_3	
33	49	Phosgene	$COCl_2$	

which is required for the one-vs-rest implementation of the SVM classifier, as described below.

Data generation, the first step in Fig. 2, involves the simulation of spectra at a range of pressure and absorber concentration for fixed temperature (297 K), pathlength (100 cm), and multiple frequency resolutions (0.01, 0.1, and 1 cm^{-1}), where the HITRAN database provides the absorber gas name, line strengths, ground state energies, and pressure broadening coeficients needed to carry out the spectral simulations.

2.2.2 Experimental spectra for validation

The experimental data used in the study are mostly from the National Institute of Standards and Technology (NIST) database [38, 39], with the exception of the spectra for HCl which comes from the Pacific Northwest National Laboratory (PNNL) database [40]. See Table 2 for compounds, labels, and sources of experimental data used for validation. Table 2 also lists the integer index (class labels in code) for all 34 molecules and specifically distinguishes molecules for which experimental data are available.

2.3 Support vector machine method

The support vector machine (SVM) method has been developed by Vapnik and colleagues based on statistical learning theory [14, 16, 41–43]. SVMs are well known for classifying texts and images [44, 45] and has also found applications in remote sensing [46] and various other fields [47–52].

Once the features are established, SVMs separate data by establishing a decision boundary, that distinguishes two different classes (in our case molecules). SVMs find an optimum decision boundary (mathematically an optimum hyperplane in n-dimensional space), that separates two classes or categories or molecules with the largest possible margin. If the optimum hyperplane is constructed such that there is no misclassification for training data, it is known as a hard margin classifier. Otherwise, it is known as a soft margin classifier. As an example, in Fig. 4, a hard margin decision boundary (line in two-dimensional feature space) is shown separating the water and hydrogen sulfide molecules. Vectors in the training spectra that are closest to the decision boundary are known as support vectors and identified in the Fig. 4 by individual numerals.

Here, each spectrum is represented by a number of features in vector $\vec{u}$ of length N, where the length is based on the frequency range and resolution of the spectrum. The SVM method is implemented in such a way that it can operate within any frequency range and resolution. The decision boundary separating the support vectors is established with the form

$$\vec{w} \cdot \vec{u} \geq c, \tag{4}$$

where $\vec{u}$ is an unlabeled feature vector, $\vec{w}$ represents a weight vector containing weights and c is an arbitrary scaler, neces-sary to establish the hyperplane. The values of weights and the scaler are determined from training. Equation 4 defines a decision rule for the identification of the "positive" class (in

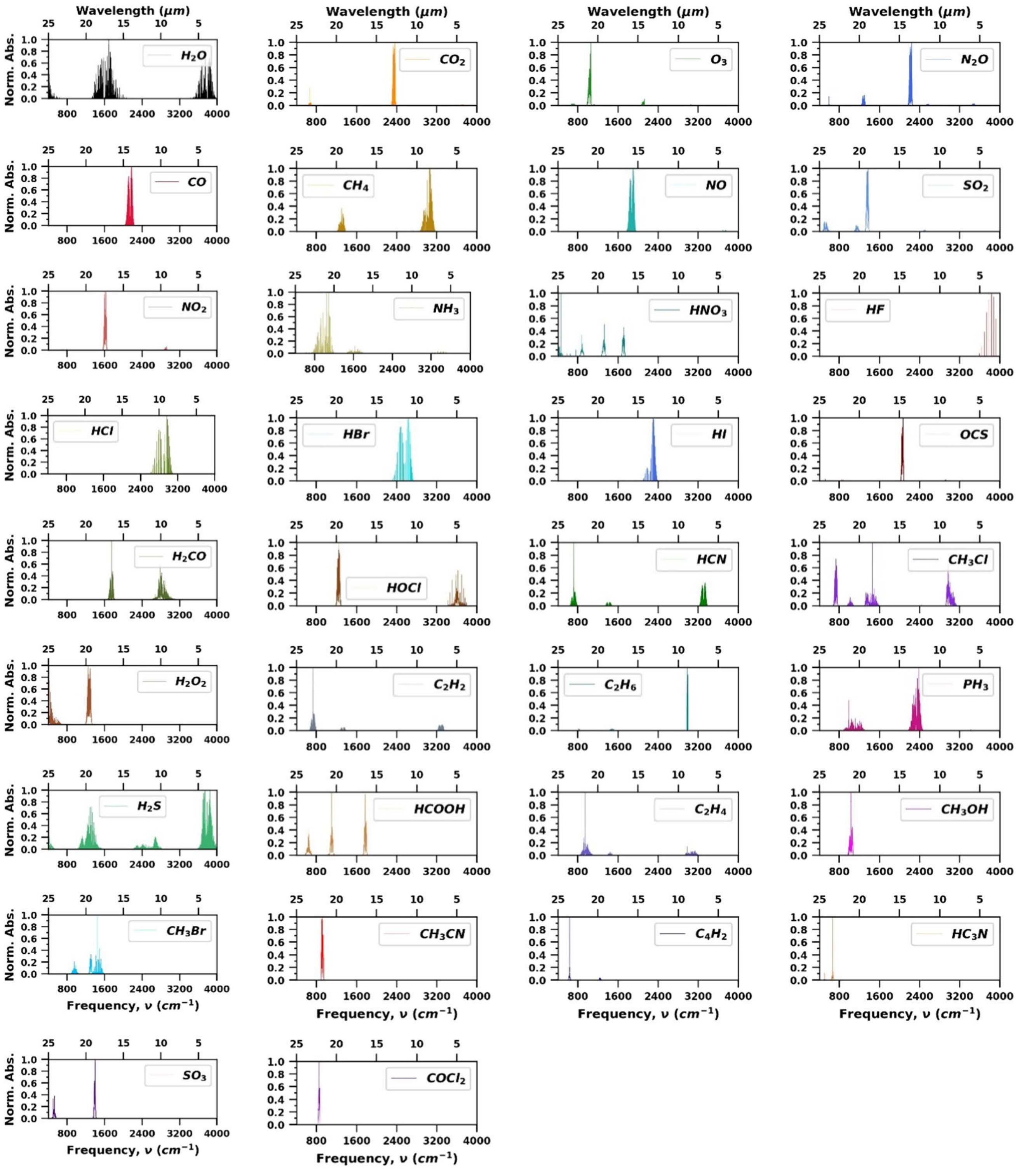


Fig. 3 Simulated absorption spectra for the molecules considered, plotted as normalized absorbance. Conditions: 1.0 atm, 297 K, 100 cm pathlength, 1000 ppm absorber in N_2, and frequency resolution of 0.01 cm^{-1}

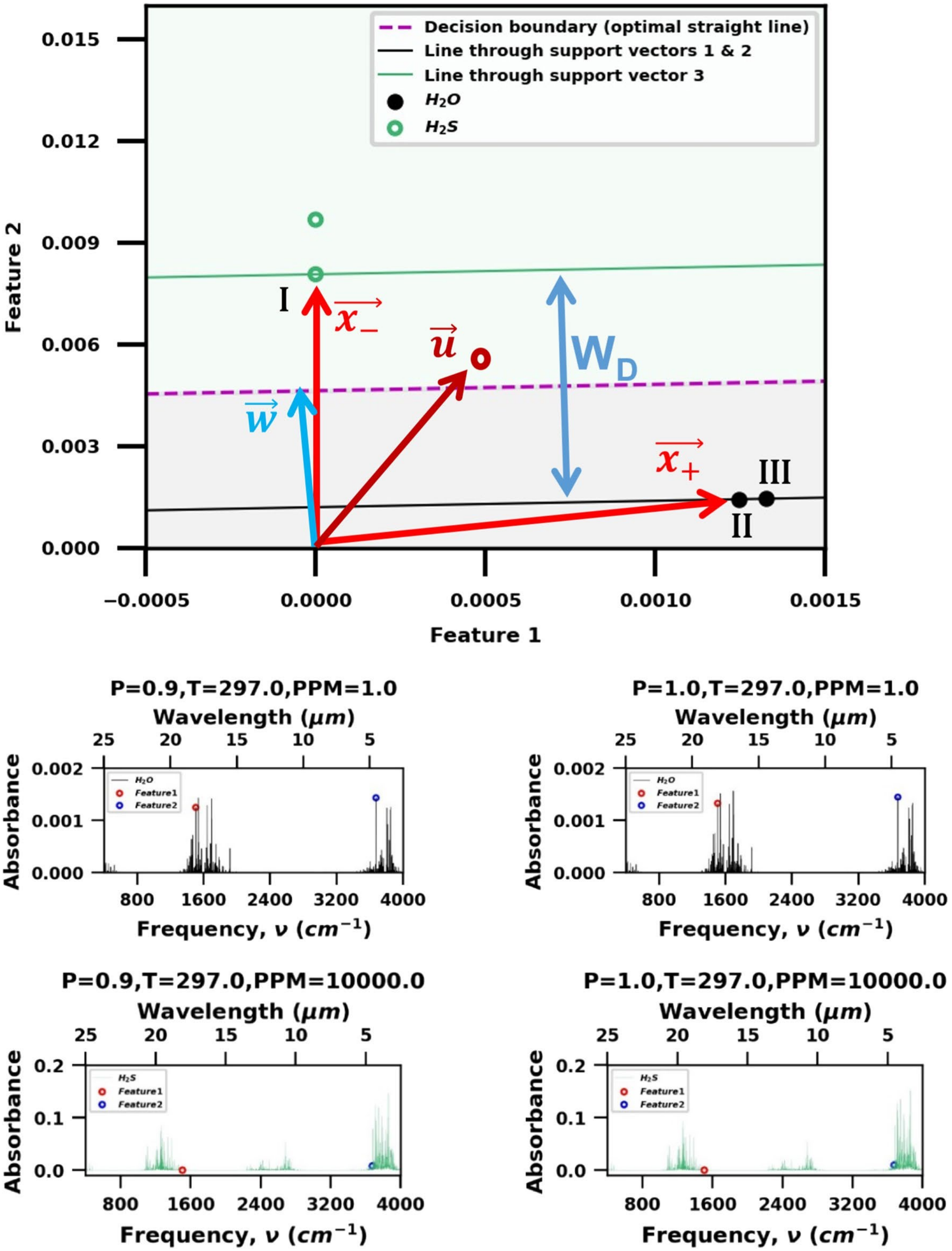


Fig. 4 A SVM decision boundary for distinguishing water and hydrogen sulfide based on two features defined as the absorbance at frequencies of 1507 and 3676 cm^{-1}. For this example, containing only four data points in 2-D feature space, the resulting boundary is a hard-margin decision boundary. The units of pressure and temperature are atm and K, respectively

the case shown in Fig. 4, H2O is the positive class and H_2S is the "negative" class). The decision rule can be simplified as follows,

$$w □□ u + b \geq 0, \quad (5)$$

□ □
where $b = -c$.

Since b and c are arbitrary, constraints can be imposed as follows, for a two-component vector as shown in the Fig. 4,

$$w □ x_+ □ □ □ + b \geq 1 \quad (6)$$

$$w □ x_- □ □ □ + b \leq -1, \quad (7)$$

where x_+ □ □ and x_- □ □ are the support vectors belonging in the positive and the negative classes, respectively.

For a supervised problem, each vector has an associated label index, y_i. For a binary problem, a simple choice for label indices are,

$$y_i = +1 \text{for} + \text{samples} \quad (8)$$

$$y_i = -1 \text{for} - \text{samples} \quad (9)$$

The width, W_D, between support vectors is given by,

$$\text{WD} = □ □ □ x_+ □ x_- □ □ \frac{w □ □ □}{□ w □}, \quad (10)$$

where Eqs. 6, 7, and 10 allow the width to be simplified as follows,

$$\text{WD} = \frac{2}{□ w □} \quad (11)$$

Since SVMs find the optimum straight line separating classes, the width, W_D, between support vectors should be maximized. Hence, $\frac{1}{2}□ w □^2$ should be minimized and there should be no intermediate data points between support vectors. Subject to the constraints in Eqs. 6 and 7 maximizing the width is equivalent to minimizing $\frac{1}{2}□ w □^2$

$$max \frac{2}{□ w □} \rightarrow \min □ w □ \rightarrow \min \frac{1}{2} □ w □^2 \quad (12)$$

Therefore, the following quadratic problem is posed subject to the constraint of Eqs. 6 and 7,

$$L = \frac{1}{2}□ w □^2 - \sum_{i=1}^{m} □_i [y_i □ w □ x_i + b - 1]. \quad (13)$$

Equation 13 is called the primal form of support vector machines, where $_i$ are Lagrange multipliers and m is the number of training spectra. It can be solved using training spectra via the Lagrange multipliers, resulting in:

$$\frac{□ L}{□ w} = w - \sum_{i=1}^{m} □_i y_i □ x_i = 0 \Rightarrow w = \sum_{i=1}^{m} □_i y_i □ x_i \quad (14)$$

$$\frac{L}{b} = \sum_{i=1}^{m} {}_i y_i = 0. \quad (15)$$

Next, w and b are substituted into the primal form to obtain the dual form which allows the use kernel functions. The dual form is given by:

$$L = \sum_{i=1}^{m} {}_i - \sum_{i=1}^{m}\sum_{j=1}^{m} {}_{ij} y_i y_j x_i, x_j, \quad (16)$$

where x_i, x_j is known as the kernel function and maps the features to a higher dimensional space which helps to classify linearly inseparable features. The advantage of using kernels is that SVM only calculates the relationship between pairs of spectra, as if they were transformed to a higher dimensional space, instead of performing the actual transformation. This process is known as the kernel trick [2, 15, 20, 53].

The dual form in Eq. 16 is solved under the constraint of Eq. 15 and the Lagrange multipliers obey the following two criteria:

$$_i \geq 0 \quad and \quad _i \leq C, \quad (17)$$

where C is a regularization parameter that controls the number of misclassifications allowed. In the context of SVMs, C is called soft margin constant. Larger values of C reduce misclassification.

The original kernel function in the dual form is the linear kernel given by:

$$x_i, x_j = x_i □ x_j. \quad (18)$$

Other choices of kernel functions, include, the radial basis function kernel (RBF) and the d-ordered polynomial kernel given, respectively, by:

$$x_i, x_j = e^{-\frac{x_i - x_j^2}{2}} = e^{- x_i - x_j^2} \quad (19)$$

$$x_i, x_j = (x_i □ x_j + k)^d, \quad (20)$$

where is a hyperparameter that relates sensitivity to variance in the feature vectors of the training spectra, $□ = \frac{1}{2□^2} > 0$, and k is any arbitrary scaler. In Scikit-learn implementation of SVM, the value for gamma can be specified and was optimized here.

The dual form is solved via convex optimization to yield non-zero solutions for the Lagrange multipliers. These are known as support vectors, S. From the number of support vectors, N_S, unlabeled features (unknown spectra) can be identified by modifying the constraints as follows,

$$\vec{w} = \sum_{i=1}^{m} \alpha_i y_i \vec{u} \tag{21}$$

$$b = \frac{1}{N_s} \sum_{i \in S} \left(y_i - \sum_{j \in S} \alpha_j y_j x_i \cdot x_j \right). \tag{22}$$

In this work, we assessed the performance of three different kernels, and ultimately chose to implement the linear kernel, since it offered the highest score in testing.

2.4 Implementation of SVM

In Fig. 4, we illustrate how a SVM classifier can be used to distinguish the spectra for water (negative class) and hydrogen sulfide (positive class). In two dimensions (Fig. 4), the hyperplane is simply a line. In the example, feature 1 is the absorbance at a frequency of 1507 cm^{-1} and feature 2 is the absorbance at a frequency of 3676 cm^{-1}, where these two features were extracted from four training spectra for water and hydrogen sulfide.

By examining Fig. 4, it is apparent that an infinite number of straight lines will separate the two water and two hydrogen sulfide data points. The support vectors are the points nearest to the decision boundary. It is clear that any of the lines through either support vectors 1 and 2 or support vector 3 will separate the data points and classify the molecules. However, this scenario would leave the support vector itself misclassified, since the decision boundary passes through it. Moreover, it is evident that, for a decision boundary passing through a support vector, any data points near that decision would have higher likelihood of misclassification.

The optimal straight line divides the region equally between the lines through the support vectors, thus minimizing misclassification. Using the primal form, described in the previous section, with a linear kernel, we can find this optimal line separating the four data points at equal distances from the support vectors. This particular decision boundary clearly separates the two compounds/classes/labels and thus is called a hard margin boundary, because none of the four data points are misclassified. The unlabeled compound represented by the vector, $\vec{u}$, can then be classified using the hard margin boundary as hydrogen sulfide.

However, for a complex problem, with many compounds and many features, a single straight line may not separate the data points and will result in some misclassifications. Such a decision boundary is called a soft decision boundary and the associated classifier is known as a soft margin classifier. The soft margin constant, C, defined in the previous section, acts as a regularization parameter and controls how much misclassification is allowed. Larger values of C reduce the number of misclassifications; however, for a soft margin classification problem, misclassification cannot be fully eliminated.

The SVM-based framework to classify infrared spectra was implemented in Python [54] using NumPy [55], SciPy [56], Pandas [57], Matplotlib [58] and most notably using the Scikit-learn machine learning library [59]. The SVM implementation in Scikit-learn is based on LIBSVM library [60]. The Python program was run on a Dell 5820 Precision Tower Workstation computer with 64 GB RAM, Intel Xeon 3.6 GHz processor with NVIDIA QUADRO RTX4000 GPU.

Since the SVM classifier is constructed as a binary classifier to distinguish between two molecules/labels/classes, a strategy known as one-versus-rest (OVR), is implemented. The OVR strategy is described in Fig. 5. The OVR classifier contains binary SVM classifiers for each molecule. Each of these binary SVM classifiers treats one molecule as the positive class. For example, the ozone (O_3) SVM classifier shown inside the OVR classifier trains on O_3 as the positive class ($y = +1$) and treats the rest of the compounds as a single negative class (y=-1). Therefore, when an unlabeled spectrum, in this example O_3, is considered by the OVR classifier, the O_3 SVM classifier outputs $y = +1$ and all SVM classifiers for the other molecules output $y = -1$. The outputs of all SVM classifiers are then taken into account by the OVR classifier and matched to the global OVR label index ($y = 2$ in this case), providing identification of the O_3 molecule.

2.4.1 Hyperparameter optimization and classifier construction

The SVM classifier has two key hyperparameters, the kernel function and the soft margin constant, C, that were determined by a grid search cross-validation using the Scikit-learn GridSearchCV function. Figure 6 illustrates classification performance (F1 score, see definition in Sect. 3 below) variation for the linear, radial basis function (RBF), and a third-degree polynomial kernel functions and for C values from 0.0001 to 1000. The RBF and a third-degree polynomial kernel functions requires an additional hyperparameter, in Eq. 19, which was optimized to maximize F1 score. The default Scikit-learn value was used as a starting value for optimization in both kernels, defined as $= \frac{1}{N_f \, {}_f}$, where N_f is the number of features and ${}_f$ the variance in the features.

During the hyperparameter search, a total of 144 cases were run using a 70–30% training–testing split and the corresponding scores were calculated. See Sect. 3 in supplementary materials for the associated table. Dataset A was used and two different frequency ranges, 400–4000 cm^{-1} and 1600–1610 cm^{-1}, were considered. After the feature

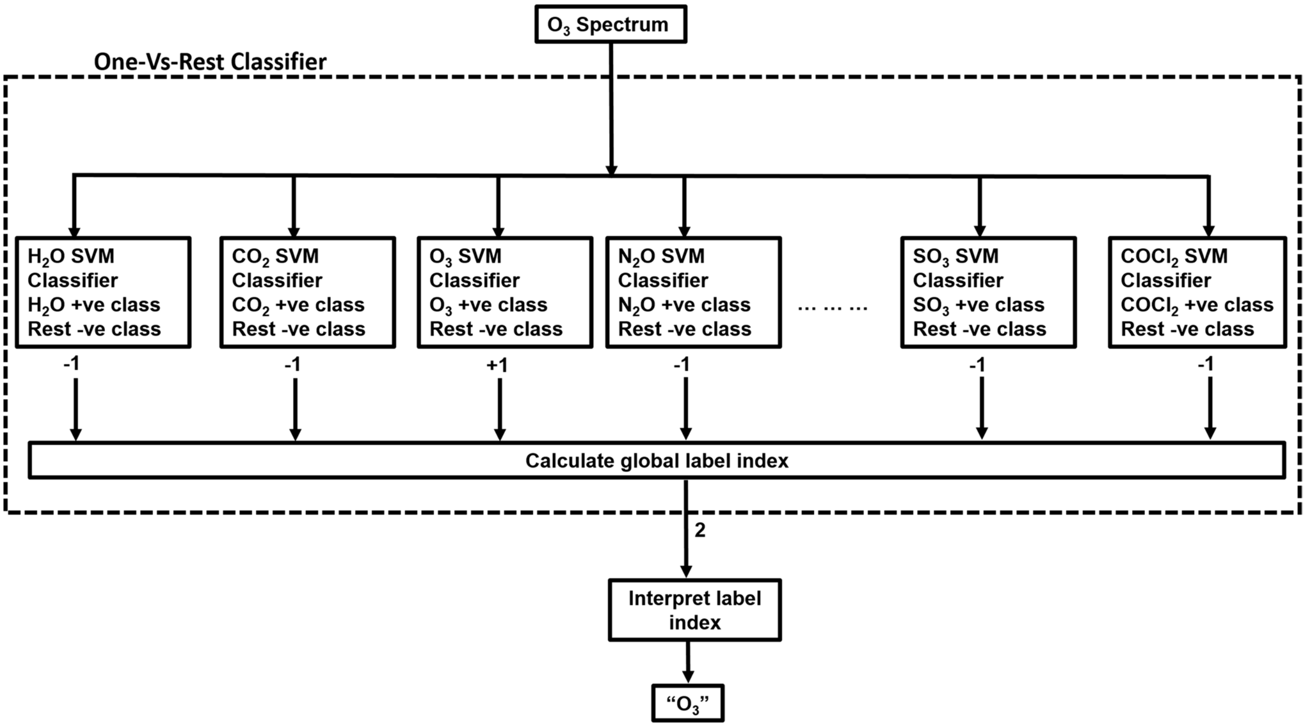

O_3 Spectrum
One-Vs-Rest Classifier
H_2O SVM Classifier H_2O +ve class Rest -ve class
CO_2 SVM Classifier CO_2 +ve class Rest -ve class
O_3 SVM Classifier O_3 +ve class Rest -ve class
N_2O SVM Classifier N_2O +ve class Rest -ve class
… … …
SO_3 SVM Classifier SO_3 +ve class Rest -ve class
$COCl_2$ SVM Classifier $COCl_2$ +ve class Rest -ve class
-1
-1
+1
-1
-1
-1
Calculate global label index
2
Interpret label index
"O_3"

Fig. 5 Flowchart depicting OVR implementation of SVM classifiers. In the example for O_3 spectrum identification, the internal SVM classifiers for each molecule first attempt to identify the spectrum, with the O_3 classifier yielding a positive result and all other yielding negative. This combination of internal results is then considered at the global OVR level, producing an identification as O_3

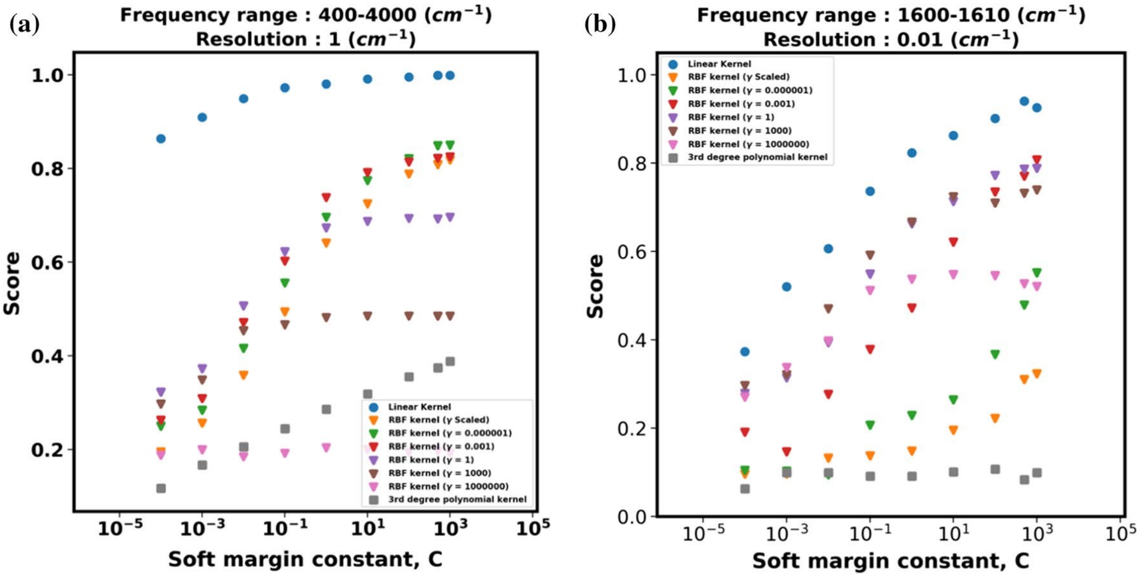


Fig. 6 F1 score dependence on soft margin constant and kernel function (linear, radial basis, and third-degree polynomial kernels) in two frequency ranges: a) 400–4000 cm^{-1} at 1 cm^{-1} resolution with 34 compounds and b) 1600–1610 cm^{-1} at 0.01 cm^{-1} resolution with 9 compounds. See Sect. 3 in supplementary materials for the associated data table

extraction step shown in Fig. 2, compounds were excluded if the absorbance for those features was smaller than 10^{-8} for the smaller frequency range and 10^{-19} for the larger frequency range. In the smaller frequency range (1600–1610 cm^{-1}), 25 compounds were excluded, as they have weak or no appreciable absorption features in that frequency range. For the larger frequency range (400–4000 cm^{-1}), all 34 compounds had appreciable absorption and remained in the hyperparameter search data set.

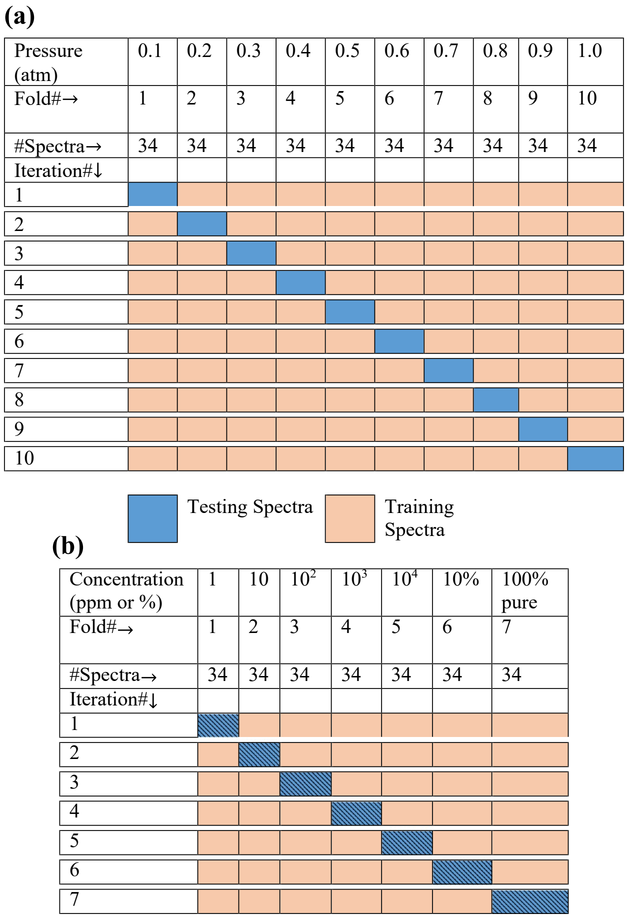
The influence on performance, in terms of F1 score, for variations in the soft margin constant *C* and the kernel function are shown in Fig. 6. For the 400–4000 cm^{-1}, the highest F1 score (0.999) occurs for the linear kernel with $C = 500$. For the 1600–1610 cm^{-1} range, the highest F1 score (0.940) occurs for the linear kernel with a $C = 500$. Hence, the linear kernel was chosen for construction of the SVM classifier. A generally increasing F1 score is also observed with increasing soft margin constant *C*, for most of the cases, as expected result since increasing *C* regularizes the classification model and minimizes misclassification through the soft decision boundary.

Among the three kernels considered, the third-degree polynomial kernel performed the worst with a scaled gamma value. Varying the gamma values for the polynomial kernel did not improve its performance. The RBF kernel's performance is improved for smaller gamma values combined with relatively larger C values. From Fig. 6, it is also observed that for a smaller number of compounds in a smaller frequency range, where the SVM classifier has relatively less information, a larger *C* value is required compared to a SVM classifier trained in the larger frequency range with more compounds. The performance improvement of the linear kernel relative to the RBF kernel is due to the spectral data containing a large number of features. Generally, RBF kernels are advantageous since they map the input features nonlinearly to an infinite-dimensional space. Thus, SVM can draw non-linear decision boundaries. However, the linear kernel is a special case of the RBF kernel and due to the abundance of features, the RBF kernel is unable to improve the classification performance relative to the linear kernel [61, 62]. Hence, the optimization of hyperparameters is crucial within the SVM framework to ensure high performance classification for any range of frequency selected.

2.4.2 Stratified k-fold cross-validation

Two stratified k-fold cross-validation studies were employed to investigate the influence of pressure and concentration on classifier performance. In all cases, the hyperparameters were fixed and the number of compounds were equally represented. The influence of pressure was investigated in a tenfold cross-validation process using dataset C. Whereas, the influence of concentration was investigated in a sevenfold cross-validation using dataset D. Datasets C and D, described in Table 1, were constructed such that only pressure and concentration were varied within those respective datasets.

In a stratified k-fold cross-validation, a dataset is split into k folds and there are a total of k iterations. In each iteration, (k-1) folds are used for training the classifier and the remaining fold is used for testing, as schematically shown in Fig. 7. Here, the folds in both cross-validations are arranged from the lowest pressure/concentration to the highest pressure/concentration. The results of different classification metrics calculated from the cross-validation are given in Table 3 and 4. The confusion matrices of both tenfold and sevenfold stratified cross-validations can be found in the supplementary material (Section S1).



(a)

Pressure (atm)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Fold#→	1	2	3	4	5	6	7	8	9	10
#Spectra→	34	34	34	34	34	34	34	34	34	34
Iteration#↓										
1	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
2	🟧	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
3	🟧	🟧	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟧
4	🟧	🟧	🟧	🟦	🟧	🟧	🟧	🟧	🟧	🟧
5	🟧	🟧	🟧	🟧	🟦	🟧	🟧	🟧	🟧	🟧
6	🟧	🟧	🟧	🟧	🟧	🟦	🟧	🟧	🟧	🟧
7	🟧	🟧	🟧	🟧	🟧	🟧	🟦	🟧	🟧	🟧
8	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟦	🟧	🟧
9	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟦	🟧
10	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟦

🟦 Testing Spectra 🟧 Training Spectra

(b)

Concentration (ppm or %)	1	10	10^2	10^3	10^4	10%	100% pure
Fold#→	1	2	3	4	5	6	7
#Spectra→	34	34	34	34	34	34	34
Iteration#↓							
1	▨	🟧	🟧	🟧	🟧	🟧	🟧
2	🟧	▨	🟧	🟧	🟧	🟧	🟧
3	🟧	🟧	▨	🟧	🟧	🟧	🟧
4	🟧	🟧	🟧	▨	🟧	🟧	🟧
5	🟧	🟧	🟧	🟧	▨	🟧	🟧
6	🟧	🟧	🟧	🟧	🟧	▨	🟧
7	🟧	🟧	🟧	🟧	🟧	🟧	▨

Fig. 7 Schematic of (**a**) tenfold cross-validation to observe the influence of pressure on classification and (**b**) sevenfold cross-validation to observe the influence of concentration on classification. In each testing fold, we see the progressive increase in pressure or concentration

Table 3 tenfold cross-validation results

Iteration number	Training time [ms]	Accuracy	Average Precision	Average Recall	Average F1 score
1	1,352	88.23%	0.846	0.882	0.855
2	1,253	97.06%	0.956	0.971	0.961
3	1,262	100%	1	1	1
4	1,293	100%	1	1	1
5	1,416	100%	1	1	1
6	1,161	100%	1	1	1
7	1,408	100%	1	1	1
8	1,210	100%	1	1	1
9	1,447	100%	1	1	1
10	1,180	97.06%	0.956	0.971	0.961

Table 4 sevenfold cross-validation results

Fold number	Time [ms]	Accuracy	Average Precision	Average Recall	Average F1 score
1	33,806	94.12%	0.9216	0.9411	0.9265
2	20,812	100%	1	1	1
3	68,470	100%	1	1	1
4	104,678	100%	1	1	1
5	39,684	100%	1	1	1
6	127,434	100%	1	1	1
7	78,021	100%	1	1	1

Table 5 SVM classifier (linear kernel, $C = 500$) performance for identification of simulated testing spectra, in 70–30% training–testing study

Trial	Accuracy (%)	Precision	Recall	F1 Score	Training time [ms]
1	99.77	0.9979	0.9977	0.9977	66,345
2	100.00	1.00	1.00	1.00	197,385
3	100.00	1.00	1.00	1.00	120,544
Total	99.92	0.9993	0.9992	0.9992	128,091

Python random number generator was used to seed different trials. Frequency range: 400–4000 cm^{-1}, dataset A. Metrics are calculated on testing spectra

3 Results and discussion

Performance of the SVM classifier framework for prediction of molecule labels for infrared spectra is characterized via several metrics as defined below.

a) Confusion matrix: A table or matrix that illustrates of the number of correct and incorrect classifications of spectra and the types of correct/incorrect classifications. From the confusion matrix, the following classification metrics can be calculated. All confusion matrices for the present study can be found in the supplementary material.

b) Classification accuracy (CA): The percentage of correctly classified spectra. Accuracy is calculated for the training, testing, and validation spectra and is called training, testing, and validation accuracy, respectively.

c) Precision (P): Precision is a measure of a classifier's ability to not misidentify a given spectrum. Precision represents the fraction of spectra identified as positives that were truly positive:

$$P = \frac{T_P}{T_P + F_P}, \tag{23}$$

where T_P are the true positives (correctly predicted spectra) and F_P are the false positives (incorrectly predictions of the positive class).

d) Recall (R): Recall indicates the fraction of actual spectra for a particular compound that are correctly predicted and is defined as:

$$R = \frac{T_P}{T_P + F_N}, \tag{24}$$

where F_N are the false negatives (incorrect predictions of the negative class).

e) F1 score: F1 score is the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{25}$$

f) Precision-recall curve: A plot of precision, the fraction of relevant results produced by the classifier, versus recall, the fraction of total relevant results correctly classified, provides a measure of the classifier's performance, where the area under the precision–recall curve is proportional to the classifier accuracy. It is not possible to continually increase both the precision and recall of a classifier. There is always a tradeoff between the two. Hence, comparison of the areas under precision–recall curves indicates relative classification performance.

3.1 Cross-validation studies

The performance of the OVR SVM classifier in the tenfold cross-validation study, in which the pressure was varied, are shown in Table 3, where performance metrics and the computer clock time for training and testing of the classifier are given. Except for iterations 1, 2, and 10, all the simulated spectra were correctly identified during testing. The lower accuracy, precision, recall, and F1 scores for iterations 1 and 2 occurs, because in these cases, the classifier has been trained on strong higher pressure spectra (0.3–1.0 atm) and then tested on lower pressure spectra (0.1–0.2 atm), where absorption is relatively weak. This is particularly prominent for iteration 1, where we have four misclassifications (see confusion matrix in supplementary material Sect. 2). For iteration 2, there is a single misclassification, which indicates that the increased absorption strength at 0.2 atm has resulted in significant improvements in classifier performance compared to 0.1 atm. Similarly, there is one misclassification in iteration 10, the highest pressure case. Since the SVM classifier uses a soft margin decision boundary, occasional misclassifications are expected at the limits of the training data or when the classifier is asked to extrapolate outside of its training data range.

The results of the sevenfold cross-validation study, for variation in absorber concentration, are given in .

Table 4. It is observed that the classifier performance is reduced when concentration is of the order of 1 ppm. However, the accuracy never decreases below 90% and, hence, the classification sensitivity to low concentrations appears to be not as strong as it is to low pressures.

3.2 70–30% training–testing

For the 70–30% training–testing study on simulated dataset A, confusion matrices were generated and classification performance was determined, as shown in Table 5. In Fig. 8, two example confusion matrices for the SVM classifier with two different kernels, the linear and the RBF kernels, are shown illustrating the better performance of the linear kernel. Interestingly, the vast number of misclassifications for the RBF kernel involve the confusion of the molecules with SO_3. The reduction in classification accuracy is similar to the reduction in F1 score observed in Fig. 6.

3.3 Experimental validation

With optimized hyperparameters, a linear kernel with a soft margin constant value of 500, yielding the best performance score against simulated spectra. Therefore, this classifier was trained with a 70–30% training–testing strategy and was tested against experimental spectra from the NIST and PNNL databases (dataset E). Classification was carried out in three trials or iterations and, because the experimental spectra are all reported over different frequency ranges, classification was performed over the experimental frequency ranges as given in Table 6. For each experimental spectrum, the SVM classifier is trained within the specific experimental frequency range. As listed in Table 6, there are misclassifications for HBr (1 of 3 trials), C_2H_2 (1 of 3 trials), and HCl (2 of 3 trials). All 17 other compounds are successfully identified in all 3 trials.

3.4 Influence of number of features and frequency resolution

The influence of the number of features (frequency range) and frequency resolution on the performance of the SVM classifier was examined through the consideration of spectra in a series of successively smaller frequency ranges within the infrared. The frequency ranges considered were 1200–2200 cm^{-1}, 1600–1700 cm^{-1}, and 1600–1610 cm^{-1}. Only a subset of 9 of the 34 molecules considered in the present study have absorption features in all 3 frequency ranges. Hence, the influence of number of features (frequency range) and resolution on classification performance was investigated for those nine molecules (H_2O, N_2O, CH_4, NO_2, NH_3, CH_3Cl, H_2O_2, C_2H_6, and CH_3Br for a total of 378 total spectra).

The SVM classifier, with unchanged hyperparameters, was trained using dataset A, in a 70–30% training–testing split. The classifier was then validated using dataset B (simulated spectra with artificial superimposed noise). For investigation of the influence of the number of features (frequency range), the coarsest spectra resolution of 1 cm^{-1} was used for all three frequency ranges: 1200–2200 cm^{-1}, 1600–1700 cm^{-1}, and 1600–1610 cm^{-1}. For the investigation of the influence of spectral resolution, the three frequency ranges of 1200–2200 cm^{-1}, 1600–1700 cm^{-1}, and

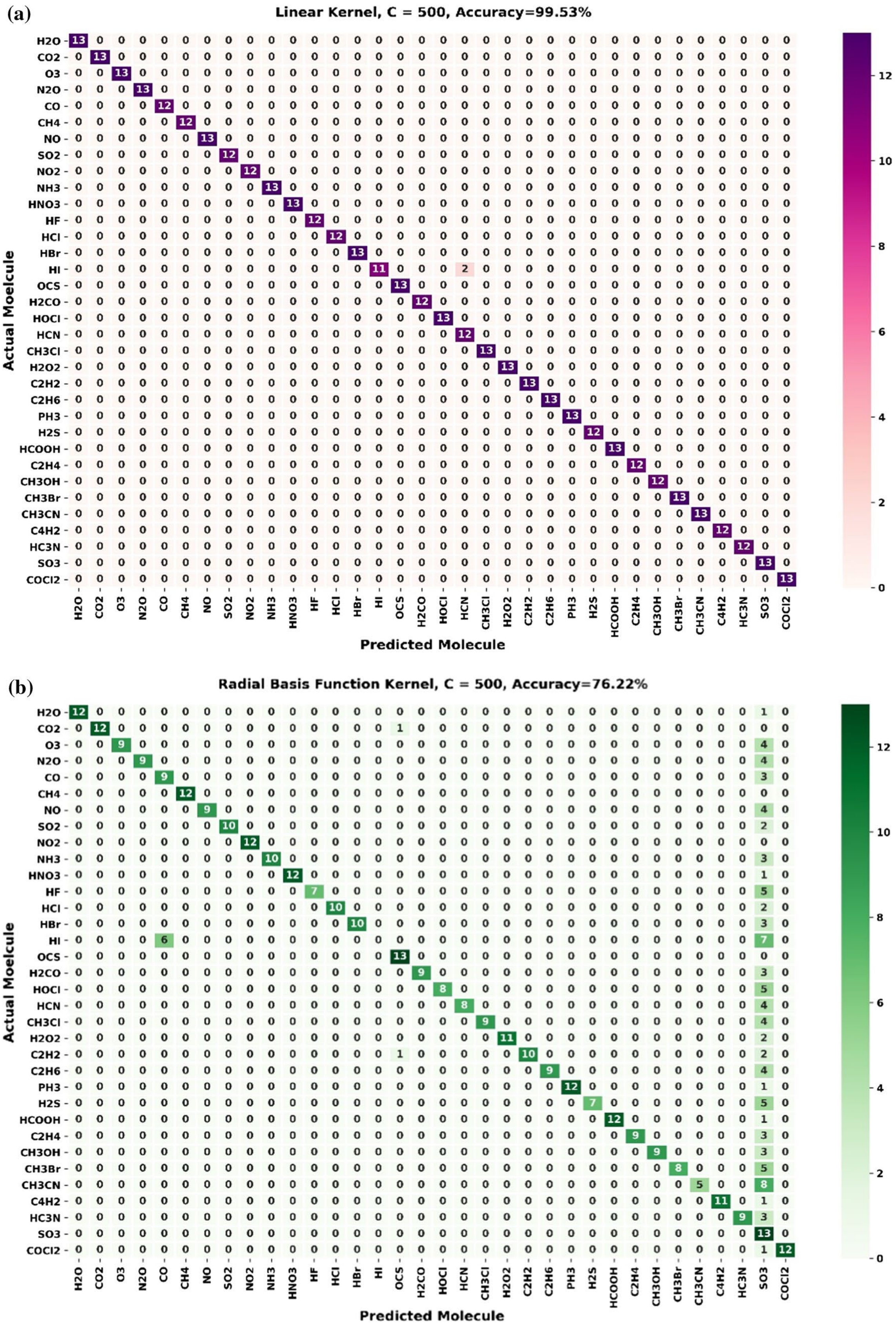


Fig. 8 Confusion matrices for SVM classifier (**a**) with a linear kernel, and (**b**) with a RBF kernel for prediction of dataset A testing spectra (frequency range 400–4000 cm^{-1})

Table 6 SVM classifier performance for predicting experimental spectra

Molecule	Experimental frequency range [cm^{-1}]	Accuracy on simulated spectra (70–30% studies)			Result on experimental spectra		
		Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3
H_2O	450–3966	99.77	99.77	99.07	✓	✓	✓
CO_2	459–3797	99.77	99.77	99.07	✓	✓	✓
CO	459–3803	99.77	100.00	99.53	✓	✓	✓
N_2O	478–3786	99.53	100.00	99.53	✓	✓	✓
CH_4	451–3801	99.77	100.00	99.53	✓	✓	✓
NO	626–3993	100.00	99.77	99.07	✓	✓	✓
NH_3	455–3798	100.00	99.77	99.30	✓	✓	✓
H_2CO	589–3902	99.77	100.00	99.53	✓	✓	✓
CH_3Cl	452–3803	99.77	100.00	99.53	✓	✓	✓
HBr	456–3741	99.77	99.77	99.07	H_2CO	✓	✓
OCS	462–3799	99.77	99.77	99.30	✓	✓	✓
C_2H_2	455–3788	99.77	100.00	99.07	✓	✓	CH_3Cl
C_2H_4	455–3795	99.77	100.00	99.53	✓	✓	✓ 99.30
C_2H_6	452–2400	98.37	98.13	✓	✓	✓	
SO_2	576–3975	100.0	99.77	99.07	✓	✓	✓ 99.77
O_3	404–3795		99.77	99.07	✓	✓	✓ 99.77
HCl	457–3765		99.77	99.07	H_2CO	✓	H_2CO
H_2S	459–3797	99.77	99.77	99.07	✓	✓	✓ 96.27
CH_3Br	576–2500		96.04	95.10	✓	✓	✓
HC_3N	554–3846	100.00	99.77	99.07	✓	✓	✓

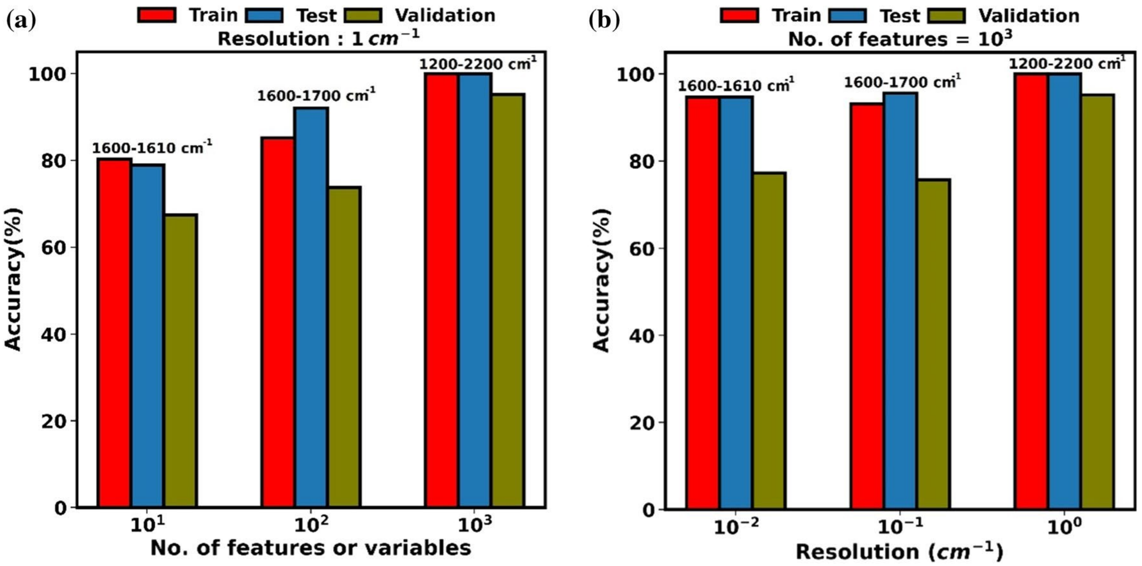


Fig. 9 Comparison of classification accuracy in training and testing (dataset A) and validation (dataset B), for variation in **a**) number of features (frequency range) and **b**) frequency resolution

1600–1610 cm^{-1} were considered at resolutions of 1, 0.1, and 0.01 cm^{-1}, respectively, resulting in three frequency ranges with the same number of features (1001 data points per spectra).

Figures 9 and 10 illustrate the performance of the classifier in these studies and confusion matrices can be found in the supplementary materials. From Fig. 9a, it is evident that by increasing the number of features the performance

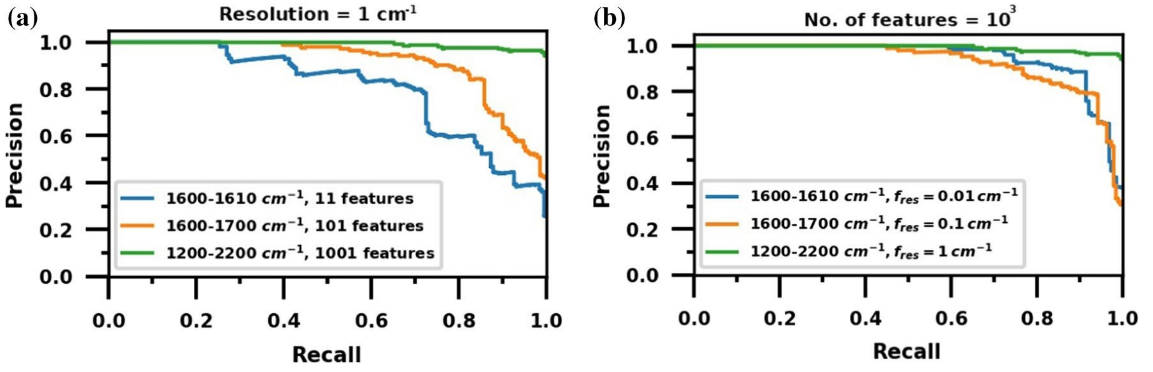


Fig. 10 Precision–recall curves for the cases considered in Fig. 9

of the classifier improves. We can conclude that providing more information, in this study a greater number of spectral features, is generally beneficial for construction a SVM classifier. Particularly, for cases, when an IR spectrometer is operating in a narrow frequency range. Additionally, not all features are equally discriminating and, therefore, providing a greater number of features increases the probability of the dataset containing discriminating features needed to develop a high-performance classifier.

The influence of frequency resolution, as shown in Fig. 9b, is less strong than the influence on the number of features, with an apparent reduction of performance with increasing resolution. Because the number of features (frequency points) is held constant in each case, the higher resolution spectra have smaller frequency ranges and hence contain less useful or discrimination information for spectral classifications. These results, show that the inclusion of greater numbers of spectral features is critical, rather than better resolving those spectral features.

4 Conclusions

A support vector machine (SVM)-based framework has been developed for the classification of infrared absorption spectra, taking advantage of the unique rotational-vibrational fingerprints for infrared-active molecules and the highly discriminating and automated capabilities of SVMs. The SVM classifier was trained using simulated spectra for 34 molecules in the 400–4000 cm^{-1} frequency range. Spectral simulations were carried out using fundamental spectroscopic parameters from the HITRAN database [36]. The performance of the SVM framework has been evaluated against simulated training spectra, both with and without artificial superimposed noise, and experimental spectra from NIST and PNNL databases. Hyperparameter optimization was performed and it was found that that the SVM classifier implemented in a one-vs-rest approach with a linear kernel and soft margin constant of 500 provided optimal performance. The resulting SVM classifier predicted simulated spectra at accuracies above 99% and correctly identified experimental spectra for 19 of 20 molecules (experimental classification accuracy of 93% across 3 random trials, 56 correct classifications in 60 attempts). The demonstrated performance indicates that the SVM classifier achieved accuracy suitable for the identification and monitoring of gas-phase species in real time. The framework proposed in this work is not specific to infrared absorption spectroscopy but can be extrapolated to other frequency ranges, spectroscopy types, or conditions. We intend spectroscopists in any frequency range to use this or a similar SVM framework for fast automated spectral detection, where the SVM classifier is trained on simulated spectra, from available databases (e.g., [36]).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00340-022-07879-8.

Acknowledgements This work was supported by the National Science Foundation under Grant CBET-1851291.

Data and code availability The code and experimental data have been made available on GitHub. GitHub repository: https://github.com/arshadzahangirchowdhury/IR_SVM.

References

1. M.A.Z. Chowdhury, T.E. Rice, M.A. Oehlschlaeger, Appl. Phys. B: Lasers Opt. **127**, 34 (2021)
2. C.M. Bishop, *Machine Learning and Pattern Recoginiton* (Springer, New York, NY, 2006)
3. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees* (Chapman & Hall/CRC, Boca Raton, Florida, 1984).
4. L. Breiman, Mach. Learn. **45**, 5 (2001)
5. P. Geurts, D. Ernst, L. Wehenkel, Mach. Learn. **63**, 3 (2006)
6. Y. Freund, Inf. Comput. **121**, 256 (1995)
7. Y. Freund, R. Schapire, Journal of Japanese Society for Artificial Intelligence **14**, 771 (1999)

8. L. Peterson, DOI: https://doi.org/10.4249/Scholarpedia.1883 (2009).
9. T.M. Cover, P.E. Hart, IEEE Trans. Inf. Theory **13**, 21 (1967)
10. A.E. Maxwell, T.A. Warner, F. Fang, Int. J. Remote Sens. **39**, 2784 (2018)
11. M. Pardo and G. Sberveglieri, Sensors and Actuators, B: Chemical **107**, 730 (2005).
12. S. Haykin, Soft Computing and Intelligent Systems 71 (2000).
13. J. Leonard, M.A. Kramer, Comput. Chem. Eng. **14**, 337 (1990)
14. C. Cortes, V. Vapnik, Mach. Learn. **20**, 273 (1995)
15. N. Chen, W. Lu, J. Yang, and G. Li, *Support Vector Machine in Chemistry* (WORLD SCIENTIFIC, 2004).
16. V.N. Vapnik, *Statistical Learning Theory* (Wiley, New York, NY, 1998)
17. J. Shawe-Taylor and S. Sun, Academic Press Library in Signal Processing: Volume 1 Signal Processing Theory and Machine Learning **1**, 857 (2014).
18. B. Scholkopf, A.J. Smola, *Learning with Kernels* (MIT Press, Cambridge, MA, 2001)
19. N. Cristianini, J. Shawe-Taylor, *An Introduction to support vector machines and other kernel-based learning methods* (Cambridge University Press, Cambridge, UK, 2013)
20. L. Wang, *Support vector machines: theory and applications* (Springer-Verlag, Berlin Heidelberg, 2005)
21. J. Weston and C. Watkins, Citeseer: Technical Report 23 (1998).
22. G. Anthony, H. Gregg, and M. Tshilidzi, 28th Asian Conference on Remote Sensing 2007, ACRS 2007 **2**, 801 (2007).
23. C. N. Banwell and E. M. McCash, *Fundamentals of Molecular Spectroscopy*, 4th ed. (McGraw-Hill Education, 2016).
24. R. K. Hanson, R. M. Spearrin, and C. S. Goldenstein, *Spectroscopy and Optical Diagnostics for Gases* (2016).
25. X. Zhai, A.A.S. Ali, A. Amira, F. Bensaali, IEEE Access **4**, 8138 (2016)
26. P. Peng, X. Zhao, X. Pan, W. Ye, Sensors (Switzerland) **18**, 1 (2018)
27. S. Güney and A. Atasoy, Sensors and Actuators, B: Chemical **166–167**, 721 (2012).
28. J. H. Cho and P. U. Kurup, Sensors and Actuators, B: Chemical **160**, 542 (2011).
29. H. Tian, H. Liu, Y. He, B. Chen, L. Xiao, Y. Fei, G. Wang, H. Yu, C. Chen, J. Food Measurement Characterization **14**, 573 (2020)
30. Y. Luo, W. Ye, X. Zhao, X. Pan, Y. Cao, Sensors (Switzerland) **17**, 1 (2017)
31. J. Mingers, Mach. Learn. **4**, 227 (1989)
32. K. Song, Q. Wang, Q. Liu, H. Zhang, Y. Cheng, Sensors **11**, 485 (2011)
33. L. Zhang, F. Tian, H. Nie, L. Dang, G. Li, Q. Ye, and C. Kadri, Sensors and Actuators, B: Chemical **174**, 114 (2012).
34. A. Tekawade, T.E. Rice, M.A. Oehlschlaeger, M.W. Mansha, K. Wu, M.M. Hella, I. Wilke, Appl. Phys. B: Lasers Opt. **124**, 105 (2018)
35. T.E. Rice, M.A.Z. Chowdhury, M.W. Mansha, M.M. Hella, I. Wilke, M.A. Oehlschlaeger, Appl. Phys. B: Lasers Opt. **126**, 152 (2020)
36. I. E. Gordon, L. S. Rothman, C. Hill, R. V. Kochanov, Y. Tan, P. F. Bernath, M. Birk, V. Boudon, A. Campargue, K. V. Chance, B. J. Drouin, J. M. Flaud, R. R. Gamache, J. T. Hodges, D. Jacquemart, V. I. Perevalov, A. Perrin, K. P. Shine, M. A. H. Smith, J. Tennyson, G. C. Toon, H. Tran, V. G. Tyuterev, A. Barbe, A. G. Császár, V. M. Devi, T. Furtenbacher, J. J. Harrison, J. M. Hartmann, A. Jolly, T. J. Johnson, T. Karman, I. Kleiner, A. A. Kyuberis, J. Loos, O. M. Lyulin, S. T. Massie, S. N. Mikhailenko, N. Moazzen-Ahmadi, H. S. P. Müller, O. V. Naumenko, A. V. Nikitin, O. L. Polyansky, M. Rey, M. Rotger, S. W. Sharpe, K. Sung, E. Starikova, S. A. Tashkun, J. Vander Auwera, G. Wagner, J. Wilzewski, P. Wcisło, S. Yu, and E. J. Zak, Journal of Quantitative Spectroscopy and Radiative Transfer **203**, 3 (2017).
37. R.V. Kochanov, I.E. Gordon, L.S. Rothman, P. Wcisło, C. Hill, J.S. Wilzewski, J. Quant. Spectrosc. Radiat. Transfer **177**, 15 (2016)
38. P. M. Chu, F. R. Guenther, G. C. Rhoderick, and W. J. Lafferty, *The NIST Quantitative Infrared Database* (n.d.).
39. A.L. Smith, *The Coblentz Society Desk Book of Infrared Spectra*, 2nd edn. (Coblentz Society, Kirkwood, Missouri, 1982)
40. S. W. Sharpe, T. J. Johnson, R. L. Sams, P. M. Chu, G. C. Rhoderick, and P. A. Johnson, *Gas-Phase Databases for Quantitative Infrared Spectroscopy* (2004).
41. C. Cortes and V. Vapnik, Patent no. US5640492A (1997).
42. I. Guyon, B. Boser, and V. Vapnik, Advances in Neural Information Processing Systems 147 (1993).
43. B. E. Boser, I. M. Guyon, and V. N. Vapnik, in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory* (Publ by ACM, 1992), pp. 144–152.
44. T. Joachims, in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '01* (ACM Press, New York, New York, USA, 2001), pp. 128–136.
45. O. Chapelle, P. Haffner, V.N. Vapnik, IEEE Trans. Neural Netw. **10**, 1055 (1999)
46. G. Mountrakis, J. Im, C. Ogole, ISPRS J. Photogramm. Remote. Sens. **66**, 247 (2011)
47. E. Gani, C. Manzie, Proceedings of the institution of mechanical engineers. Part D **221**, 1183 (2007)
48. A. Çevik, A. E. KURTOĞLU, M. Bilgehan, M. E. Gülşan, and H. M. Albegmprli, Journal of Civil Engineering and Management **21**, 261 (2015).
49. S. Raghavendra, P.C. Deka, Appl. Soft Comput. J. **19**, 372 (2014)
50. F. Gao, X. Shao, Environ. Sci. Pollut. Res. **28**, 21411 (2021)
51. X. Ma, R. Ge, L. Zhang, Kybernetes **43**, 1224 (2014)
52. F. Elmaz, B. Büyükçakır, Ö. Yücel, A.Y. Mutlu, Fuel **266**, 117066 (2020)
53. T. Kavzoglu, I. Colkesen, Int. J. Appl. Earth Obs. Geoinf. **11**, 352 (2009)
54. G. Van Rossum, *Python Reference Manual* (Amsterdam, 1995).
55. C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Nature **585**, 357 (2020)
56. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. Pietro Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G. L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P.

A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, and Y. Vázquez-Baeza, Nature Methods **17**, 261 (2020).

57. W. Mckinney, in (Proc. of The 9th Python in Science Conf. (SCIPY 2010), 2010).
58. J.D. Hunter, Comput. Sci. Eng. **9**, 90 (2007)
59. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011)
60. C.C. Chang, C.J. Lin, ACM Trans. Intell. Syst. Technol. **2**, 27 (2011)
61. S.S. Keerthi, C.J. Lin, Neural Comput. **15**, 1667 (2003)
62. C. Hsu, C. Chang, and C. Lin, National Taiwan University 1396 (2003).

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