ACS SENSORS

pubs.acs.org/acssensors Article

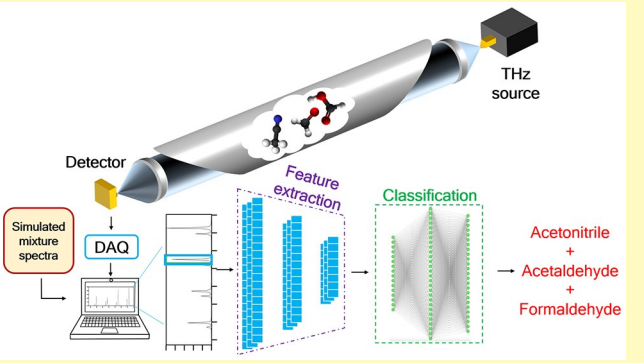
TSMC-Net: Deep-Learning Multigas Classification Using THz Absorption Spectra

M. Arshad Zahangir Chowdhury,* Timothy E. Rice, and Matthew A. Oehlschlaeger

Cite This: https://doi.org/10.1021/acssensors.2c02615 Read Online

ACCESS | Metrics & More | Article Recommendations | * Supporting Information

ABSTRACT: The identification of gas mixture speciation from a complex multicomponent absorption spectrum is a problem in gas sensing that can be addressed using machine-learning approaches. Here, we report on a deep convolutional neural network for multigas classification using terahertz (THz) absorption spectra, THz spectra mixture classifier network or TSMC-Net. TSMC-Net has been developed to identify eight volatile organic compounds in mixtures based on their fingerprint rotational absorption spectra in the 220−330 GHz frequency range. A data set consisting of simulated absorption spectra for randomly generated mixtures, with absorption greater than thresholds representing detectable limits and annotated with multiple labels, was prepared for model development. The supervised multilabel classification problem, i.e., the identification of individual gases in a mixture, is converted to a supervised multiclass classification problem via label powerset conversion. The trained model is validated and tested against simulated spectra for gas mixtures, with and without white Gaussian noise. The trained model exhibits high precision, recall, and accuracy for each pure compound. Class activation maps illustrate the complex decision-making process of the model and highlight relevant frequency regions that are needed to identify unique mixtures. Finally, the model was demonstrated against measured THz absorption spectra for pure species and mixtures, acquired using a microelectronics-based THz absorption spectrometer. The data set generation strategy and deep convolutional neural network approach are generalized and can be extrapolated to other spectroscopy types, frequency ranges, and sensors.



KEYWORDS: *THz spectroscopy, convolutional neural network, gas mixtures, deep learning, species identification, classification*

Gas sensing in industrial, scientific, and environmental applications, such as process control and remote monitoring, often requires the determination of the speciation of complex mixtures from spectra containing dozens to perhaps hundreds of features, obtained using absorption spectroscopy or other methods.[1−3] Such gas-sensing applications include the continuous sensing of targeted toxic chemical species as required in personnel safety scenarios,[4] occupational health,[5,6] pharmaceutical and medical,[7−11] automotive,[12,13] national security,[14,15] pollutant monitoring,[16−20] and other safety[21,22] applications. Complex spectra, arising from multicomponent mixtures, can be analyzed using multiple component analysis (MCA),[23] independent component analysis (ICA),[24] multivariate calibration (MVC),[25] self-modeling curve resolution (SMCR),[25,26] and multivariate curve resolution (MCR);[27] however, these methods typically require knowledge of the species present in the mixture prior to their application. Furthermore, to determine quantitative concentrations of mixture components, a regression problem needs to be solved, in which a signal or feature is fit against a calibration or model that defines the concentration. This is often done using principal component regression (PCR), non-negative matrix factorization,[28] partial least-squares (PLS), inverse least-squares (ILS), or ordinary least-squares (OLS) methods.[29−32]

Machine learning (ML) and deep learning (DL) have been used in several recent chemical sensing applications. For example, DL has been applied in solid-state gas sensors.[33−36] Two- and three-dimensional deep convolutional neural network (CNN) models trained on optical micrographs for liquid-crystal-based mixture speciation have been reported.[37−39] One-dimensional CNNs have been used to extract nonlinear features to classify pure gases in a solid-state electronic nose sensor.[40] Classical ML models, for instance, classification trees, random forests, multilayer perceptrons, and support vector machines, have been demonstrated to achieve high classification accuracy for the identification of pure gases from terahertz (THz)[41] and

Received: November 29, 2022
Accepted: February 14, 2023

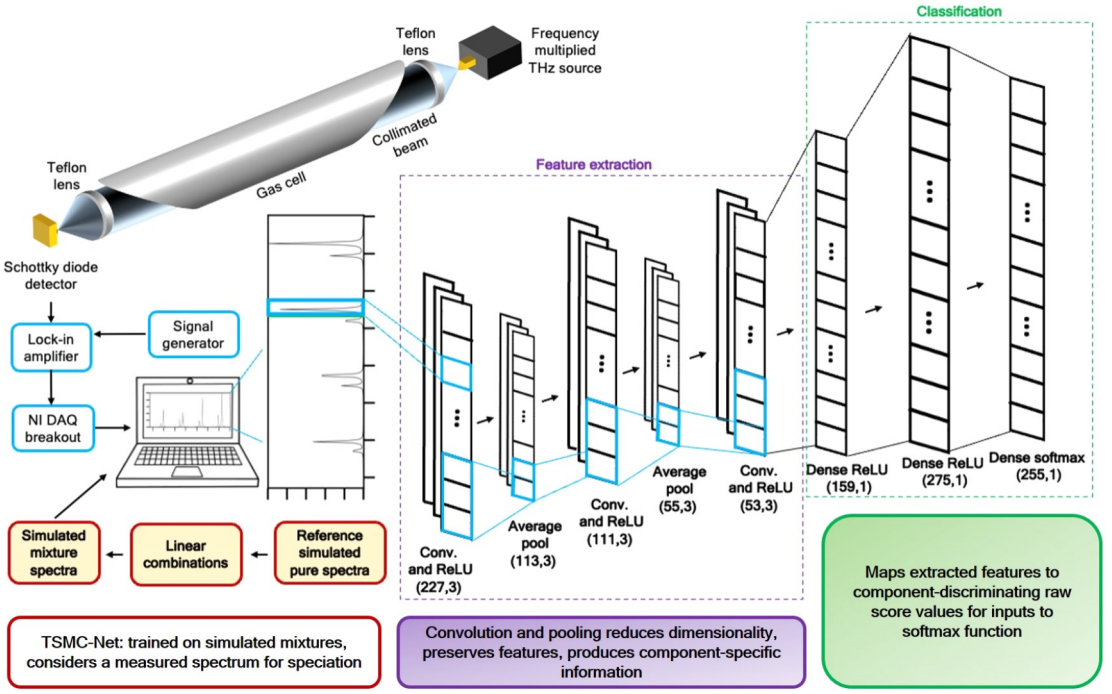


Figure 1. Schematic diagram of TSMC-Net and a THz spectrometer for classification of complex mixtures. TSMC-Net consists of a feature extraction block and a dense downstream classification block.

infrared spectra.[42] In recent work, we have reported a deep convolutional network (VOC-Net) for classification of pure rotational spectra in the 220−330 GHz frequency range and examined it with gradient-weighted class activation maps[43] to understand its inner workings.[44] Interpretable classification networks, such as VOC-Net, illustrated that DL models are capable of achieving high classification performance measures by generalizing spectral features. Deep convolutional neural network models offer several key advantages. First, deep CNNs can extract features (sets of derived values from the absorption spectrum). Second, the universal approximation theorem guarantees that a deep neural network will be able to represent the complex relationship that exists between a spectrum and its corresponding class-distinctive label.[45] Third, a deep convolutional neural network can be further examined using the GRAD-CAM method,[43] which, when implemented in 1-D for spectra, visually highlights the spectral features that contribute to a positive class distinction decision.

The THz absorption spectrum for a given polar gas-phase molecule, within in a suficient frequency band, is dependent on the unique rotational structure of the molecule and can retain a self-similar spectral fingerprint irrespective of thermodynamic conditions, concentration, bath gas composition, or frequency of the transitions.[46−48] The rotational fingerprint allows for the application of one-dimensional CNNs or other ML models to learn and extract relevant features for speciation. Furthermore, due to the recent advances in THz microelectronic sources, miniature and portable THz gas sensors are feasible.[49,50] Thus, DL-assisted THz multigas speciation can contribute toward advanced sensor design for portable, wearable,[7,22] and unmanned vehicle[51] applications.

The problem of determining mixture speciation from a spectrum is similar to the problem of blind source separation (BSS), which is the separation of an original source signal from a mixed signal without any information regarding the mixing process.[52,53] Here, we present a DL[54] approach to determine the speciation of a mixture from its THz absorption spectrum. We assume a mixture spectrum is a linear combination of component spectra and train a CNN followed by a dense neural network using high-dimensional simulated spectra. The present approach is only limited by the availability of pure reference spectra and can be adapted to any arbitrary and potentially large number of mixture components. This is advantageous, since the absence of a known number of mixture components makes the application of traditional methods (i.e., MVC, SMCR, MCR, PCR, PLS, and OLS) dificult.

The present DL model is developed to identify eight possible volatile organic compounds (VOCs) that are rotationally active in the frequency range of interest (220−330 GHz) and have line positions and strengths that are documented in the HITRAN[55] and JPL[56] spectroscopic databases. For many of the considered VOCs, we have previously reported spectral absorption measurements in the present frequency range.[49,57,58] The DL model is trained and validated using simulated spectra, is tested against noisy simulated spectra, and is demonstrated via the classification of experimental spectra. Lastly, the model is analyzed using class activation maps to examine how it learns from the spectral fingerprints containing class-discriminating spectral features.

METHODOLOGY

The data sets, their generation, the architecture of the THz spectra mixture classifier network (TSMC-Net), and network training and validation approaches are described in the following subsections. A schematic of TSMC-Net with the experimental THz spectrometer, illustrating the overall methodology, is shown in Figure 1. TSMC-Net is trained using simulated mixture spectra. A laboratory spectrometer measures the THz spectrum of a gas mixture of interest, and

https://doi.org/10.1021/acssensors.2c02615
ACS Sens. XXXX, XXX, XXX−XXX

TSMC-Net outputs estimated probabilities for the mixture speciation. The TSMC-Net codes and the analysis presented in this paper are hosted at https://github.com/arshadzahangirchowdhury/TSMC-Net.

Problem Statement and Solution Approach. Multicomponent speciation based on a single absorption spectrum is a multilabel supervised classification problem. Since a single spectrum can be associated with the presence of multiple gases, the spectrum will have multiple labels, presenting two challenges. First, labeling or annotating a multigas spectrum for model training is much more expensive compared to a pure gas spectrum. Second, not all machine-learning models can be directly applied to classify spectra with multiple labels, thus motivating conversion of the multilabel problem to a multiclass classification problem.

Here, the maximum number of mixture components considered is 8 and each spectrum considered consists of 229 absorbance values from 220 to 330 GHz (frequency resolution of 0.016 cm^{-1}). The label powerset method[59] is implemented to convert an 8-label classification problem to a $2^8 - 1 = 255$ class classification problem. Because each spectrum is comprised of 229 features, the input layer to the TSMC-Net has 229 neurons, and the output layer contains 255 neurons, corresponding to the number of unique mixtures (classes). The input and output layers of the network are trained using simulated spectral data sets and their corresponding multiclass integer label indices. Each neuron in the output layer provides a raw score for each mixture class. Thus, the model establishes a mapping between the input spectra and raw scores corresponding to its label. These scores are converted to estimated probabilities via softmax activation.

The softmax scores, σ, are given by

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (1)$$

where z is the input to the softmax layer and K is the number of molecule classes.

The feature extraction scheme highlighted in Figure 1 applies convolution and pooling operations to reduce the dimensionality of the input spectrum while preserving the mixture-distinguish information. Each convolution converts the input vector to a new vector containing learned features. The size of the convolution output is given by $\frac{1}{S}(W - K_s + 2P) + 1$, where, W, K_s, P, and S are the sizes of the input, kernel, padding, and the stride, respectively, for each filter. TSMC-Net uses three filters for every convolution. Consequently, the output of the first convolutional layer is given by $\frac{1}{1}(229 - 3 + 2 \times 0) + 1 = 227$, since the input size is 229, the kernel size is 3, the stride size is 1, and valid (zero) padding was used. Since we implemented three filters, the total output of the first convolution is given by (227,3). For the subsequent pooling layer, for each filter, the output size is $\frac{W - K_s + 1}{S}$. For a stride size of 2 and a pooling kernel size of 2, we obtain the total output shape of (113,3).

Each convolution passes a sliding window over its inputs, and each pooling layer downsamples its inputs; thus, collectively they produce a derived set of values from the input spectrum during the forward pass of the model. The weights of these convolutional and pooling layers are trained and regularized and ultimately output a small set of derived values after the last convolutional layer. The derived set of values contain the most relevant information contained in the input spectrum in a lower-dimensional vector. The process of reducing the input features of the spectrum to a new set of features in a lower-dimensional space is known as feature extraction. The features extracted after the last convolutional layer are fed to a flattened dense layer. The dense layer learns from the extracted features and during the forward pass transforms its inputs to the supervised class label of each mixture. Each dense layer applies learned weights and biases followed by nonlinear ReLU activation to obtain raw scores. Afterward, the Adam optimizer[60] updates weights and biases with sparse categorical cross-entropy loss with a batch size of 32. Once training is complete, the raw scores are converted with the softmax activation function and the estimated probabilities for all 255 classes are reported. The class with the highest estimated probability is the predicted speciation for the considered mixture spectrum. The classification scheme shown in Figure 1 simply maps the extracted features to raw scores. For multiclass classification, these raw scores are converted to estimated probabilites via the softmax function.

EXPERIMENTAL METHODS

Experimental spectra were acquired using an spectrometer that has been described in prior work.[49,50,57,58,61] THz radiation is generated with a microelectronics source (VDi SGX Model WR 3.4 220330 GHz), passed through a gas cell containing the absorbing gas mixture (path length of 21.6 cm), and detected using a Schottky diode detector (VDi Model QOD 315). A brief schematic of the setup is shown in Figure 1.

The absorbance is determined by first measuring a baseline reference intensity, I_0, while the gas cell is under vacuum. Then the cell is pressurized with the gas mixture of interest and the transmitted intensity, I, is measured. Gas mixtures were made *in situ* via partial pressures from pure chemicals (>99%, Sigma-Aldrich). The spectral absorbance of the gas mixture, $A_{mixture}$, is given by

$$A_{\text{mixture}} = -\ln\left(\frac{I}{I_0}\right) = \sum_i \varepsilon_i c_i l \quad (2)$$

where ε_i is the spectral absorption coeficient for the ith species, which is a function of the spectroscopic parameters (i.e., line positions, shapes, strengths) and temperature, pressure, and gas concentration, c_i is the concentration of the ith absorbing gas species, and l is the absorption path length.

An absorption spectrum is calculated from the measured reference and transmitted intensities and sent to the TSMC-Net model for classification ofline (after completion of the experiment). The measured experimental spectrum is resampled to match the 0.016 cm^{-1} resolution of the simulated spectra used to train TSMC-Net. In this work, ten measured spectra for gas mixtures and pure species are used to demonstrate TSMC-Net against experimental data. These experimental spectra are described in Table S1 of the Supporting Information.

COMPUTATIONAL METHODS

Training, Validation, and Test Spectra. DL models require significant amounts of data for training and validation. A key challenge in the training of DL models to recognize spectra is the lack of experimental spectra. In prior pure gas spectra identification problems in the THz and infrared frequency ranges,[41,42,44] simulated spectra proved to be excellent for training various ML models. Model training for mixtures necessitates the use of simulated spectra, given the lack of a suficient number of available experiments on multigas mixtures.

To generate simulated mixture spectra, we simulated the spectra for the eight pure components of interest (listed in Table 1) at a total pressure of 1 Torr, a temperature of 297 K, and a path length of 21.6

https://doi.org/10.1021/acssensors.2c02615
ACS Sens. XXXX, XXX, XXX−XXX

Table 1. Summary of Pure Compounds and Source of the Spectroscopic Parameters Used in Spectral Simulations

compound	formula	source of spectroscopic data
acetaldehyde	CH_3CHO	JPL[56]
acetonitrile	CH_3CN	JPL
chloromethane	CH_3Cl	HITRAN[55]
methanol	CH_3OH	HITRAN
ethanol	C_2H_5OH	JPL
formic acid	$HCOOH$	JPL
nitric acid	HNO_3	HITRAN
formaldehyde	H_2CO	HITRAN

cm without dilution. Simulations were carried out using the HAPI software[62] and the spectroscopic parameters found in the HITRAN[55] and JPL[56] databases. The spectra were simulated at a frequency resolution of 0.016 cm^{-1}). These spectra are illustrated in Figure 2. Mixture spectra were generated by the linear combination of these eight pure component reference spectra with randomly generated concentrations for each pure component. The simulated mixture spectra generated by a linear combination of the pure compound reference spectra assume collisional line broadening that is independent of mixture composition (i.e., all collisions result in broadening that is the same as a self-broadening collision). Unfortunately, information on broadening parameters is almost always restricted to self-broadening and broadening by air or nitrogen, owing to the lack of cross-collisional broadening parameters in the literature.

For simulated spectra, the composition of the mixture is generated randomly by the selection of component concentrations for the linear combination from a uniform distribution on the interval [0, 1], while obeying two constraints. First, the maximum absorbance for each species in the mixture must be above 0.01. For kHz-rate experiments, minimum detectivity typically occurs for an absorbance of around 0.001. Hence, the requirement of a maximum absorbance of 0.01 for each species corresponds to a maximum signal-to-noise ratio (SNR) of 10 for the detection of each species at its peak value of absorbance within the 220−330 GHz frequency range. As a second constraint, to ensure that each component absorbance suficiently stands out compared to the total absorbance, we require that the ratio of the maximum absorbance for each species to the maximum absorbance for the mixture be 0.01 or greater. These thresholds ensure that the DL model will not learn from a spectrum which (a) is very weak, (b) contains practically undetectable components, and/or (c) has weaker absorbers whose fingerprint is overwhelmed by strong absorbers.

The spectra generation process was as follows.

Step 1: A set of random concentration values for 8 components and the diluent are generated such that their sum equals 1.

Step 2: Using reference spectra at 1 Torr for the 8 pure compounds, a mixture spectrum is calculated via linear combination.

Step 3: The mixture spectrum is checked for the two constraints on absolute and relative absorbance.

Step 4: If the spectrum passes the two constraints, it is retained; otherwise, steps 1−3 are repeated.

Step 5: Once the desired number of mixture spectra are generated, 90 spectra from each of the 255 mixture classes are

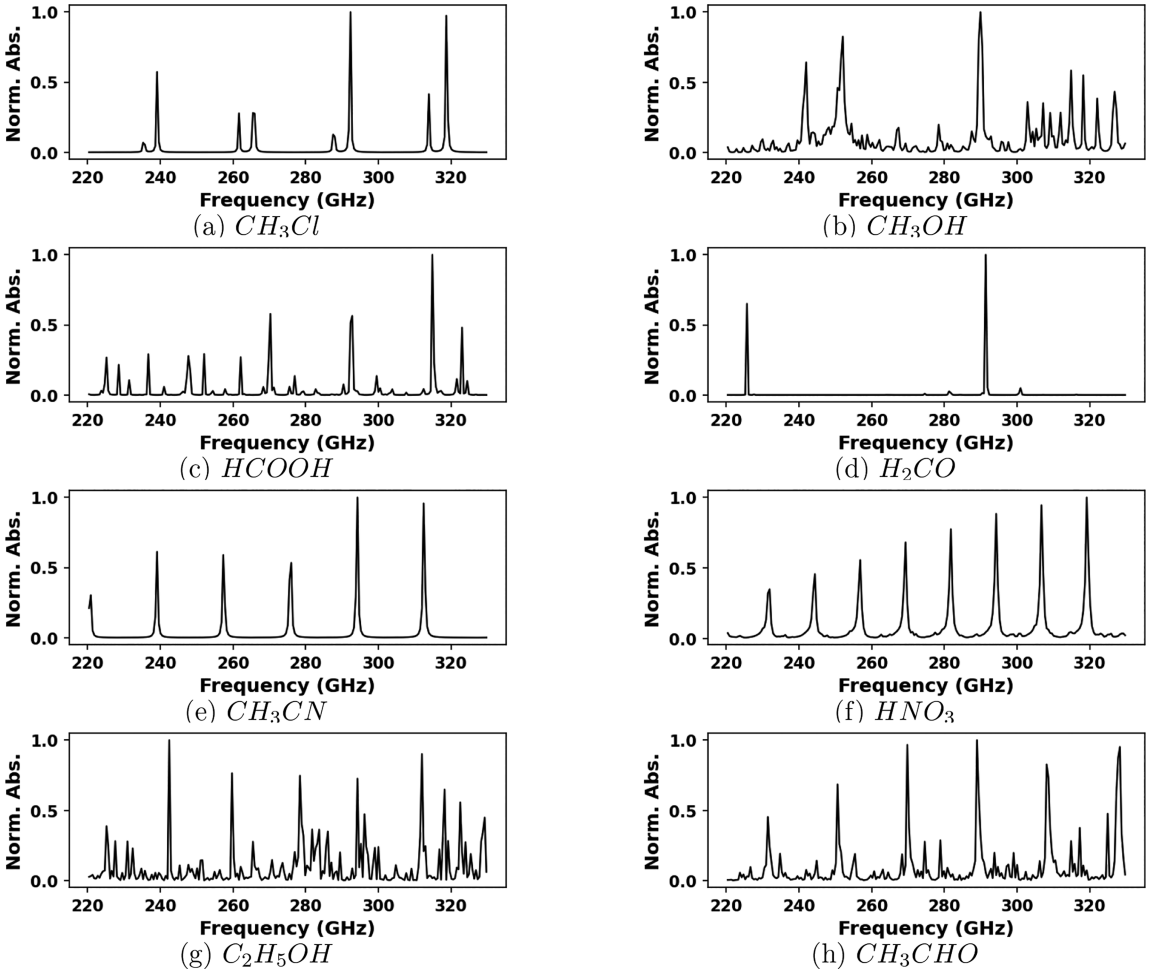


Figure 2. Simulated reference spectra. Conditions: 1 Torr, 297 K, 21.6 cm path length.

https://doi.org/10.1021/acssensors.2c02615
ACS Sens. XXXX, XXX, XXX−XXX

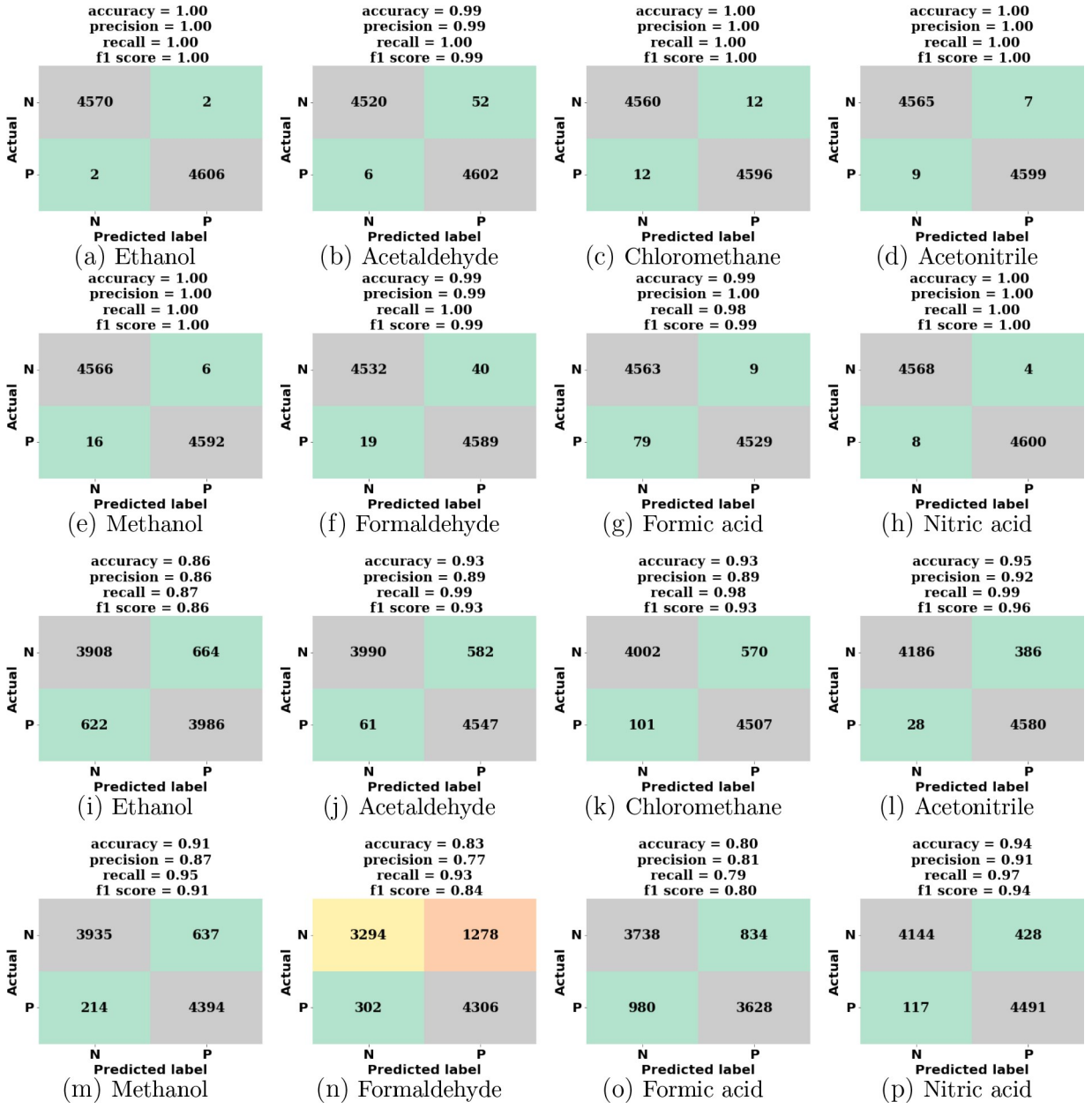


Figure 3. Accuracy, precision, recall, and f1 score for species identification: (a−h) simulated validation spectra without added noise. (i−p) test spectra with added Gaussian noise (SNR = 30).

randomly sampled, to generate an unbiased data set with each mixture type equally represented.

Step 6: These spectra are split into a training and a validation set.

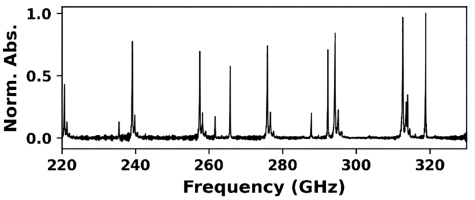
For the unique 255 mixture types, 22950 unique simulated spectra were generated for the development of TSMC-Net (90 spectra per mixture type). Approximately 97% of the simulated spectra were multicomponent, while the remaining 3% of the training spectra were for pure components diluted in air, such that the trained network would also accurately predict those single absorber mixtures. The spectra were split into 60%−40% training and validation sets, yielding 13770 training spectra (54 spectra per mixture type) and 9180 validation spectra (36 spectra per mixture type). Thus, the matrices containing the training and validation spectra have shapes of 13770 × 229 and 9180 × 229, respectively. The representation of spectra in the training and validation sets from each of the unique combinations of pure compounds is shown in Figure S1 and the description of each data set is provided in Table S2 of the Supporting Information.

To understand the influence of noise on TSMC-Net classification performance, a data set was prepared where Gaussian noise was added to the 9180 simulated validation spectra; we call this the test set in Table S2 of the Supporting Information. Noise was added to generate spectra with a SNR of 30, where SNR has been defined as the maximum absorbance value for the mixture spectra divided by the amplitude of absorbance white noise added.

The training time is on average 145 ± 1 s. The prediction time on the entire validation data set was measured to be less than 2 s. The timing measurements were carried out on an Intel(R) Xeon(R) W-2223 CPU with a clock speed of 3.60 GHz.

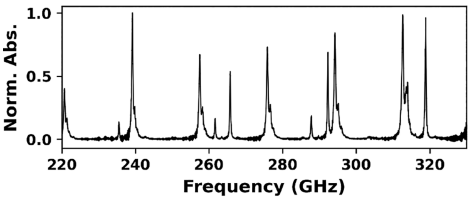
1D Convolutional Neural Network Architecture (TSMC-Net). For the classification of pure spectra in the 220−330 GHz range, we previously reported a convolutional deep neural network (VOC-Net).[44] TSMC-Net borrows the VOC-Net architecture and adds an additional convolutional layer for improved feature extraction and a denser penultimate flattened layer for classification of 255 classes of spectra (mixture types).

https://doi.org/10.1021/acssensors.2c02615
ACS Sens. XXXX, XXX, XXX−XXX



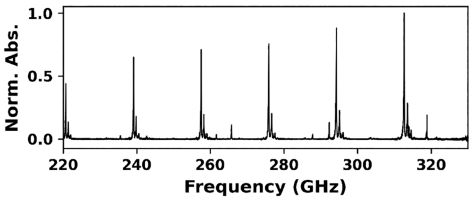
(a) 30% CH_3Cl and 70% CH_3CN.

Species present	Score
CH_3Cl+CH_3CN	0.999458
C_2H_5OH+CH_3Cl+CH_3CN	0.000540
CH_3Cl+CH_3CN+CH_3OH	0.000002
C_2H_5OH+CH_3Cl+CH_3CN+$HCOOH$	0.000000
CH_3Cl+CH_3CN+$HCOOH$	0.000000



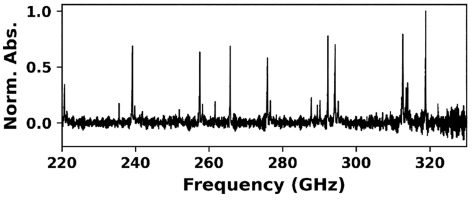
(b) 30% CH_3Cl and 70% CH_3CN.

Species present	Score
CH_3Cl+CH_3CN	0.646248
C_2H_5OH+CH_3Cl+CH_3CN	0.353748
CH_3Cl+CH_3CN+CH_3OH	0.000003
C_2H_5OH+CH_3Cl+CH_3CN+$HCOOH$	0.000000
C_2H_5OH+CH_3Cl+CH_3CN+CH_3OH	0.000000



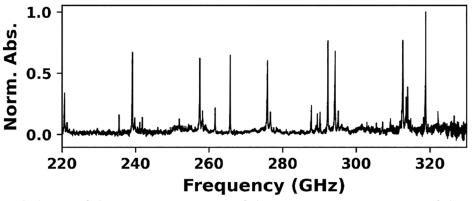
(c) 30% CH_3Cl and 70% CH_3CN.

Species present	Score
CH_3Cl+CH_3CN	0.909778
CH_3Cl+CH_3CN+CH_3OH	0.084952
C_2H_5OH+CH_3Cl+CH_3CN	0.005270
CH_3CN	0.000000
CH_3Cl+CH_3CN+H_2CO	0.000000



(d) 90% CH_3OH, 5% CH_3Cl, and 5% CH_3CN.

Species present	Score
CH_3Cl+CH_3CN+CH_3OH	0.999953
CH_3Cl+CH_3CN	0.000047
CH_3Cl+CH_3CN+CH_3OH+$HCOOH$	0.000000
CH_3Cl+CH_3CN+CH_3OH+HNO_3	0.000000
CH_3Cl+CH_3CN+H_2CO	0.000000



(e) 90% CH_3OH, 5% CH_3Cl, and 5% CH_3CN.

Species present	Score
CH_3Cl+CH_3CN+CH_3OH	0.999165
C_2H_5OH+CH_3Cl+CH_3CN+CH_3OH	0.000830
CH_3Cl+CH_3CN+CH_3OH+$HCOOH$	0.000004
CH_3Cl+CH_3CN+CH_3OH+HNO_3	0.000000
CH_3Cl+CH_3OH	0.000000

Figure 4. Measured mixture spectra under the conditions given in Table S1. The tabular data to the right of each spectrum give the top mixture probabilities predicted by TSMC-Net.

In Figure 1, the TSMC-Net architecture is illustrated. The network consists of three sections: the feature extraction block, a dense classification block, and a softmax probability model (not shown in Figure 1). Raw absorbance values from simulated spectra serve as features, and each spectrum serves as a training instance. The feature extraction block is tasked with examining regions of the spectra via convolutional filter kernels to progressively extract mixture-discriminating information from the raw absorbance. These extracted features are then sent to a dense model to calculate raw classification scores. The raw classification scores are next converted to probabilities in a softmax layer. The final layer of the dense model contains $2^8 - 1 = 255$ neurons, a neuron for every unique mixture, to convert from a multilabel to a multiclass classification problem, a step known as label powerset conversion of multilabel classification.

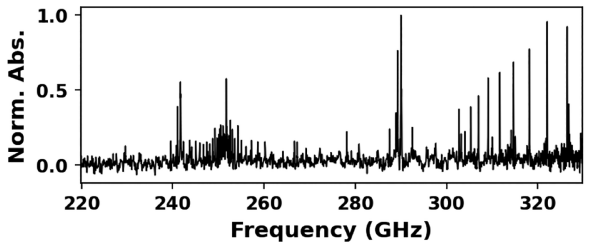
For the present complex classification task, TSMC-Net must learn the training examples closely and a relatively complex model is needed; hence, early stopping was implemented to regularize TSMC-Net instead of dropout. Training was stopped after 37 epochs with accuracies on training and validation at 97.1% for both.

RESULTS AND DISCUSSION

Validation and Testing Results. In Figure 3a–h, the confusion matrices for the identification of each compound in the validation data set are shown. Each confusion matrix lists the true positives, true negatives, false positives, and false negatives for the classification of a single compound for all instances where the compound is present in the validation spectra. The model classification performance against simulated noise-free validation spectra is very strong, with overall accuracy, precision, and recall for predicting individual compounds trending toward greater than 99%. As previously shown, the accuracy for predicting the correct mixture is greater than 97% for the validation data set.

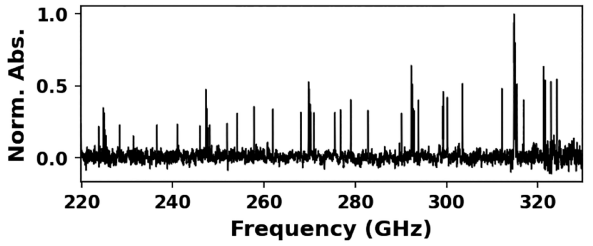
The addition of Gaussian noise to the data set (SNR = 30) results in a slightly reduced classification performance, as shown in the Figures 3i–p confusion matrices. For the noise-added test data set, the overall accuracies range from 80% for

https://doi.org/10.1021/acssensors.2c02615
ACS Sens. XXXX, XXX, XXX−XXX



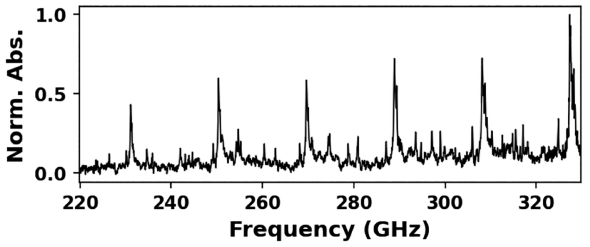
(a) CH_3OH.

Species present	Score
CH_3OH	0.999994
$CH_3CHO+CH_3CN+CH_3OH$	0.000006
CH_3CN+CH_3OH	0.000000
$C_2H_5OH+CH_3CN+CH_3OH$	0.000000
CH_3Cl+CH_3OH	0.000000



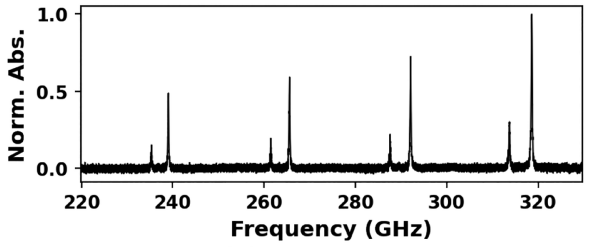
(b) $HCOOH$.

Species present	Score
$HCOOH$	0.999988
$C_2H_5OH+HCOOH$	0.000007
$C_2H_5OH+CH_3Cl+HCOOH$	0.000005
$CH_3Cl+HCOOH$	0.000000
$H_2CO+HCOOH$	0.000000



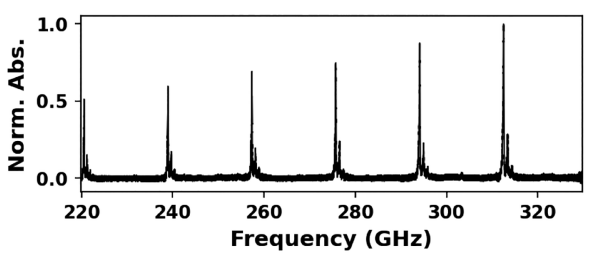
(c) CH_3CHO.

Species present	Score
$C_2H_5OH+CH_3CHO+CH_3CN$	0.843115
$C_2H_5OH+CH_3CHO+CH_3CN+HNO_3$	0.136387
$C_2H_5OH+CH_3CHO+HNO_3$	0.006015
$CH_3CHO+HNO_3$	0.005360
$CH_3CHO+CH_3CN+HNO_3$	0.004531



(d) CH_3Cl.

Species present	Score
CH_3Cl	0.999838
CH_3Cl+CH_3CN	0.000125
$CH_3CHO+CH_3Cl+CH_3CN$	0.000035
$CH_3CHO+CH_3Cl$	0.000001
$CH_3CHO+CH_3Cl+CH_3CN+HNO_3$	0.000000



(e) CH_3CN.

Species present	Score
CH_3CN	0.998151
CH_3CN+CH_3OH	0.001848
CH_3Cl+CH_3CN	0.000000
$CH_3Cl+CH_3CN+CH_3OH$	0.000000
$C_2H_5OH+CH_3CN$	0.000000

Figure 5. Measured pure spectra under the conditions given in Table S1. The tabular data to the right of each spectrum give the top mixture probabilities predicted by TSMC-Net.

formic acid (HCOOH), a weak absorber, to 95% for acetonitrile (CH_3CN), a strong absorber. The precision ranges from 0.77 to 0.99 and the recall from 0.79 to 0.99. The classification metrics are the poorest for the weakest absorbers (H_2CO, CH_3CHO, HCOOH) that are more easily overwhelmed in mixtures by competing absorbers and noise, when present.

TSMC-Net was compared to two statistical learning models, decision tree[63] and random forest.[64] TSMC-Net achieved higher precision, recall, accuracy, and f1 scores across all the compounds and mixtures. Both the decision tree and random forest methods also overfit the training spectra and achieved poor accuracies in our prior work on pure compound identification in the same frequency range.[41] Furthermore, the prediction time for these models were found to be relatively slower, since TSMC-Net can be parallelized.

Demonstration Results. In Figure 4, the performance of TSMC-Net for the classification of five experimental mixture spectra is demonstrated. The measured spectra and the corresponding softmax scores from TSMC-Net are shown. TSMC-Net outputs 255 softmax scores corresponding to the probabilities associated with the respective spectra belonging

https://doi.org/10.1021/acssensors.2c02615
ACS Sens. XXXX, XXX, XXX−XXX

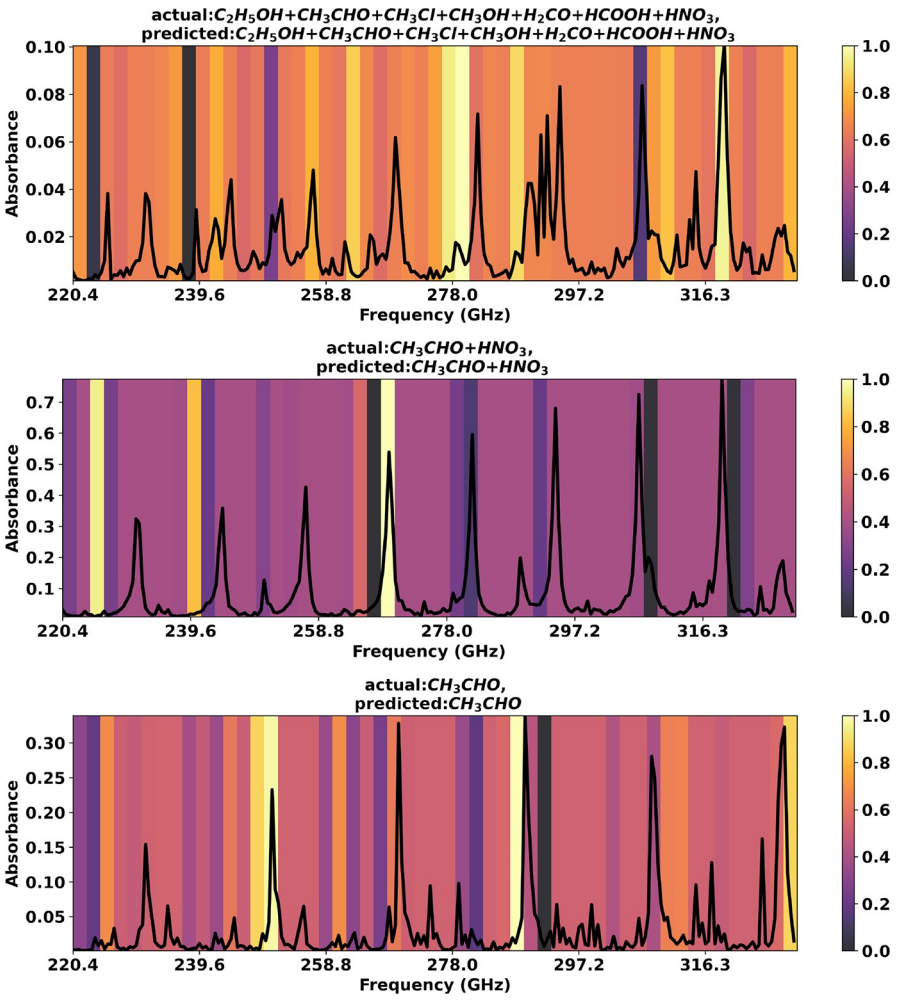


Figure 6. Class activation maps for the classification of three simulated validation spectra: (top) 38% C_2H_5OH−2% CH_3CHO−2% CH_3Cl−16% CH_3OH−10% H_2CO−20% HCOOH−8% HNO_3−4% air; (middle) 20.5% CH_3CHO−75.8% HNO_3−3.7% air; (bottom) 40% CH_3CHO−60% air.

to any one of the mixture classes. In Figure 4, the top 5 softmax scores are given. Three measurements of 30% CH_3Cl−70% CH_3CN are illustrated and all correctly classified. The first (Figure 4a) shows only high probability for the correct classification (CH_3Cl and CH_3CN), but the second two (Figure 4b,c) show marginal probabilities (35% and 8%) that the mixture could include C_2H_5OH in addition to CH_3Cl and CH_3CN. Figure 4d,e demonstrates two experimental spectra for a 3-component mixture (90% CH_3OH−30% CH_3Cl−70% CH_3CN) with different levels of noise. The model yields 99.9% probability for the correct classification in both cases and illustrates that the model can deal with variations in the noise floor.

In Figure 5 TSMC-Net is demonstrated for the identification of five single-component spectra. With the exception of acetaldehyde (CH_3CHO), the model produces the correct classification with over 99% probability. The acetaldehyde spectrum is misclassfied as a three-component mixture of acetaldehyde, ethanol, and acetonitrile, with a reported probability of approximately 84%. For this misclassified spectrum, the model predicts acetaldehyde in all of the top five classification probabilities but predicts false positives. In this case, the model is misconstruing some of the absorption features within the acetaldehyde spectrum as arising from other species, which have similar and overlapping absorption features in this frequency range (see Figure 2).

Interpretability. Gradient-weighted class activation mapping (Grad-CAM)[43] has been employed to interpret TSMC-Net. The last convolutional layer within the model produces a set of extracted features which originate from the input features. Grad-CAM gives a visualization of the importance of each extracted feature using a heat map that is superimposed on the original raw spectrum, where the intensity (color) of the output indicates the weight associated with each extracted feature. See several examples in Figure 6. The color scale indicates whether a feature has a positive or negative contribution to the model class prediction. Higher weights associated with extracted features indicate a positive contribution (bright colors) toward the identification of the positive class, and lower weights indicate a negative contribution (dark colors) to the final classification. The class activation maps contain useful information for sensor design, since critical localized spectral peaks within a spectrum for classification decision-making can be visualized, thus distinguishing mixture-discriminating regions in each spectrum.

Three class activation maps (CAM) are illustrated in Figure 6. The example 7-component CAM (top graph) demonstrates that a significant portion of the spectrum is used to identify the

https://doi.org/10.1021/acssensors.2c02615
ACS Sens. XXXX, XXX, XXX−XXX

mixture; i.e., much of the spectrum has colors that tend toward higher Grad-CAM weights. For mixtures containing fewer components, we see that smaller portions of the spectra contribute positively to the mixture classification. In the cases of 2- and 1-component mixtures a smaller number of important features are identified by the model as providing the important class-discriminating fingerprint.

CONCLUSIONS

A convolutional deep-learning neural network model, TSMC-Net, has been developed and demonstrated for the automated identification of gas mixtures comprised of up to eight industrially relevant VOCs based on their fingerprint rotational absorption spectra in the 220−330 GHz frequency range. TSMC-Net achieved greater than 97% accuracy for identifying mixtures on a balanced validation data set consisting of 9180 simulated spectra. The accuracy in the identification of individual compounds was greater than 99%. The model is applied to ten experimental spectra, yielding correct classification for nine of those spectra and for the other spectrum (pure acetaldehyde) identifying the true positive and two false positives. Class activation maps for TSMC-Net illustrate the decision-making process of the model. TSMC-Net can be applied in gas sensing and other situations where spectral identification is required and for sensitive and selective sensor design. TSMC-Net may be extended to different frequency regions, types of spectroscopy, or other spectral classification problems, provided suitable training data sets exist or can be generated. The method can also be extended in the future to provide quantitative concentrations, in addition to speciation. Additionally, extension of the sensor operating frequency range can allow for improved sensitivity and access to additional species.

ASSOCIATED CONTENT

Data Availability Statement

TSMC-Net is hosted at https://github.com/arshadzahangirchowdhury/TSMC-Net

* Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssensors.2c02615.

Further information regarding the data sets used for training, validation, testing, and demonstration (PDF)

AUTHOR INFORMATION

Corresponding Author

M. Arshad Zahangir Chowdhury − *Department of Mechanical, Aerospace and Nuclear Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180-3522, United States*; Email: chowdm@rpi.edu

Authors

Timothy E. Rice − *Department of Mechanical, Aerospace and Nuclear Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180-3522, United States*

Matthew A. Oehlschlaeger − *Department of Mechanical, Aerospace and Nuclear Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180-3522, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssensors.2c02615

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge the support of the National Science Foundation under Grant CBET-1851291.

REFERENCES

(1) Hodgkinson, J.; Tatam, R. P. Optical gas sensing: a review. *Measurement science and technology* 2013, *24*, 012004.

(2) Bogue, R. Detecting gases with light: A review of optical gas sensor technologies. *Sensor Review* 2015, *35*, 133.

(3) Chen, Z.; Chen, Z.; Song, Z.; Ye, W.; Fan, Z. Smart gas sensor arrays powered by artificial intelligence. *Journal of Semiconductors* 2019, *40*, 111601.

(4) Pandey, S. K.; Kim, K.-H.; Tang, K.-T. A review of sensor-based methods for monitoring hydrogen sulfide. *TrAC Trends in Analytical Chemistry* 2012, *32*, 87−99.

(5) Ashley, K. Developments in electrochemical sensors for occupational and environmental health applications. *Journal of hazardous materials* 2003, *102*, 1−12.

(6) Pitarma, R.; Marques, G.; Caetano, F. *New advances in information systems and technologies*; Springer: 2016; pp 13−21.

(7) Tricoli, A.; Nasiri, N.; De, S. Wearable and miniaturized sensor technologies for personalized and preventive medicine. *Adv. Funct. Mater.* 2017, *27*, 1605271.

(8) Kiani, S.; Minaei, S.; Ghasemi-Varnamkhasti, M. Application of electronic nose systems for assessing quality of medicinal and aromatic plant products: A review. *Journal of Applied Research on Medicinal and Aromatic Plants* 2016, *3*, 1−9.

(9) Nasiri, N.; Clarke, C. Nanostructured gas sensors for medical and health applications: Low to high dimensional materials. *Biosensors* 2019, *9*, 43.

(10) Wu, J.; Yue, G.; Chen, W.; Xing, Z.; Wang, J.; Wong, W. R.; Cheng, Z.; Set, S. Y.; Senthil Murugan, G.; Wang, X.; Liu, T. On-chip optical gas sensors based on group-IV materials. *ACS photonics* 2020, *7*, 2923−2940.

(11) Righettoni, M.; Tricoli, A.; Pratsinis, S. E. Si: WO3 sensors for highly selective detection of acetone for easy diagnosis of diabetes by breath analysis. *Analytical chemistry* 2010, *82*, 3581−3587.

(12) Tardy, P.; Coulon, J.-R.; Lucat, C.; Menil, F. Dynamic thermal conductivity sensor for gas detection. *Sens. Actuators, B* 2004, *98*, 63−68.

(13) Liu, X.; Cheng, S.; Liu, H.; Hu, S.; Zhang, D.; Ning, H. A survey on gas sensing technology. *Sensors* 2012, *12*, 9635−9665.

(14) Sekhar, P. K.; Wignes, F. Trace detection of research department explosive (RDX) using electrochemical gas sensor. *Sens. Actuators, B* 2016, *227*, 185−190.

(15) Chen, L.; Wu, D.; Yoon, J. Recent advances in the development of chromophore-based chemosensors for nerve agents and phosgene. *ACS sensors* 2018, *3*, 27−43.

(16) Hong, T.; Sung, S.; Kang, H.; Hong, J.; Kim, H.; Lee, D.-E. Advanced real-time pollutant monitoring systems for automatic environmental management of construction projects focusing on field applicability. *Journal of Management in Engineering* 2022, *38*, 04021075.

(17) Schauer, J. J.; Fraser, M. P.; Cass, G. R.; Simoneit, B. R. Source reconciliation of atmospheric gas-phase and particle-phase pollutants during a severe photochemical smog episode. *Environ. Sci. Technol.* 2002, *36*, 3806−3814.

(18) Wang, H.; Yuan, B.; Hao, R.; Zhao, Y.; Wang, X. A critical review on the method of simultaneous removal of multi-air-pollutant in flue gas. *Chemical Engineering Journal* 2019, *378*, 122155.

(19) Rai, R.; Rajput, M.; Agrawal, M.; Agrawal, S. Gaseous air pollutants: a review on current and future trends of emissions and impact on agriculture. *Journal of Scientific Research* 2011, *55*, 1.

(20) Fowler, D.; Cape, J. N.; Coyle, M.; Flechard, C.; Kuylenstierna, J.; Hicks, K.; Derwent, D.; Johnson, C.; Stevenson, D. The global

https://doi.org/10.1021/acssensors.2c02615
ACS Sens. XXXX, XXX, XXX−XXX

exposure of forests to air pollutants. *Water, Air, and Soil Pollution* 1999, *116*, 5−32.
(21) Essl, C.; Seifert, L.; Rabe, M.; Fuchs, A. Early detection of failing automotive batteries using gas sensors. *Batteries* 2021, *7*, 25.
(22) Bag, A.; Lee, N.-E. Recent advancements in development of wearable gas sensors. *Advanced Materials Technologies* 2021, *6*, 2000883.
(23) Bigourd, D.; Cuisset, A.; Hindle, F.; Matton, S.; Bocquet, R.; Mouret, G.; Cazier, F.; Dewaele, D.; Nouali, H. Multiple component analysis of cigarette smoke using THz spectroscopy, comparison with standard chemical analytical methods. *Appl. Phys. B: Laser Opt.* 2007, *86*, 579−586.
(24) Monakhova, Y. B.; Astakhov, S. A.; Kraskov, A.; Mushtakova, S. P. Independent components in spectroscopic analysis of complex mixtures. *Chemometrics and Intelligent Laboratory Systems* 2010, *103*, 108−115.
(25) Malinowski, E. R.; Howery, D. G. *Factor analysis in chemistry*; Wiley: 1980; Vol. *3*.
(26) Lawton, W. H.; Sylvestre, E. A. Self modeling curve resolution. *Technometrics* 1971, *13*, 617−633.
(27) De Juan, A.; Tauler, R. Multivariate curve resolution (MCR) from 2000: progress in concepts and applications. *Critical reviews in analytical chemistry* 2006, *36*, 163−176.
(28) Cichocki, A.; Zdunek, R.; Amari, S.-i. New algorithms for non-negative matrix factorization in applications to blind source separation. 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 2006; pp V−V.
(29) Gemperline, P. *Practical guide to chemometrics*; CRC Press: 2006.
(30) Issa, M. M.; Nejem, R. M.; Shanab, A. M. A.; Shaat, N. T. Resolution of five-component mixture using mean centering ratio and inverse least squares chemometrics. *Chemistry Central Journal* 2013, *7*, 1−11.
(31) Baikadi, A.; Bhatt, N.; Narasimhan, S. Extraction of Pure Species Spectra from Labeled Mixture Spectral Data. *Ind. Eng. Chem. Res.* 2019, *58*, 13437−13447.
(32) Oliveri, P.; Malegori, C.; Casale, M. *Chemical analysis of food*; Elsevier: 2020; pp 33−76.
(33) Hayasaka, T.; Lin, A.; Copa, V. C.; Lopez, L. P.; Loberternos, R. A.; Ballesteros, L. I. M.; Kubota, Y.; Liu, Y.; Salvador, A. A.; Lin, L. An electronic nose using a single graphene FET and machine learning for water, methanol, and ethanol. *Microsystems & nanoengineering* 2020, *6*, 1−13.
(34) Kang, M.; Cho, I.; Park, J.; Jeong, J.; Lee, K.; Lee, B.; Del Orbe Henriquez, D.; Yoon, K.; Park, I. High Accuracy Real-Time Multi-Gas Identification by a Batch-Uniform Gas Sensor Array and Deep Learning Algorithm. *ACS sensors* 2022, *7*, 430−440.
(35) Peng, P.; Zhao, X.; Pan, X.; Ye, W. Gas classification using deep convolutional neural networks. *Sensors* 2018, *18*, 157.
(36) Zhai, X.; Ali, A. A. S.; Amira, A.; Bensaali, F. MLP neural network based gas classification system on Zynq SoC. *IEEE Access* 2016, *4*, 8138−8146.
(37) Bao, N.; Jiang, S.; Smith, A.; Schauer, J. J.; Mavrikakis, M.; Van Lehn, R. C.; Zavala, V. M.; Abbott, N. L. Sensing Gas Mixtures by Analyzing the Spatiotemporal Optical Responses of Liquid Crystals Using 3D Convolutional Neural Networks. *ACS sensors* 2022, *7*, 2545−2555.
(38) Smith, A. D.; Abbott, N.; Zavala, V. M. Convolutional network analysis of optical micrographs for liquid crystal sensors. *J. Phys. Chem. C* 2020, *124*, 15152−15161.
(39) Cao, Y.; Yu, H.; Abbott, N. L.; Zavala, V. M. Machine learning algorithms for liquid crystal-based sensors. *ACS sensors* 2018, *3*, 2237−2245.
(40) Zhan, C.; He, J.; Pan, M.; Luo, D. Component Analysis of Gas Mixture Based on One-Dimensional Convolutional Neural Network. *Sensors* 2021, *21*, 347.
(41) Chowdhury, M. A. Z.; Rice, T. E.; Oehlschlaeger, M. A. Evaluation of machine learning methods for classification of rotational absorption spectra for gases in the 220−330 GHz range. *Appl. Phys. B: Laser Opt.* 2021, *127*, 1−20.
(42) Chowdhury, M. A. Z.; Rice, T. E.; Oehlschlaeger, M. A. A support vector machines framework for identification of infrared spectra. *Applied Physics B: Lasers and Optics* 2022, *128*, 161.
(43) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision, 2017; pp 618−626.
(44) Chowdhury, M. A. Z.; Rice, T. E.; Oehlschlaeger, M. A. VOC-Net: ADeep Learning Model for the Automated Classification of Rotational THz Spectra of Volatile Organic Compounds. *Applied Sciences* 2022, *12*, 8447.
(45) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep learning*; MIT Press: 2016.
(46) Fischer, B. M.; Helm, H.; Jepsen, P. U. Chemical recognition with broadband THZ spectroscopy. *Proceedings of the IEEE* 2007, *95*, 1592−1604.
(47) Neese, C. F.; Medvedev, I. R.; Plummer, G. M.; Frank, A. J.; Ball, C. D.; De Lucia, F. C. Compact submillimeter/terahertz gas sensor with efficient gas collection, preconcentration, and ppt sensitivity. *IEEE Sensors Journal* 2012, *12*, 2565−2574.
(48) Smith, R. M.; Arnold, M. A. Selectivity of Terahertz Gas-Phase Spectroscopy. *Anal. Chem.* 2015, *87*, 10679−10683.
(49) Rice, T. E.; Chowdhury, M. A. Z.; Mansha, M. W.; Hella, M. M.; Wilke, I.; Oehlschlaeger, M. A. VOC gas sensing via micro-electronics-based absorption spectroscopy at 220−330 GHz. *Appl. Phys. B: Laser Opt.* 2020, *126*, 1−12.
(50) Rice, T. E.; Mansha, M. W.; Chowdhury, A.; Hella, M. M.; Wilke, I.; Oehlschlaeger, M. A. All Electronic THz Wave Absorption Spectroscopy of Volatile Organic Compounds Between 220−330 GHz. 2020 45th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz), 2020; pp 01−02.
(51) Klemas, V. V. Coastal and environmental remote sensing from unmanned aerial vehicles: An overview. *Journal of coastal research* 2015, *315*, 1260−1267.
(52) Pal, M.; Roy, R.; Basu, J.; Bepari, M. S. Blind source separation: A review and analysis. 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013; pp 1−5.
(53) Kopriva, I.; Jeric, I. Multi-component analysis: blind extraction of pure components mass spectra using sparse component analysis. *Journal of mass spectrometry* 2009, *44*, 1378−1388.
(54) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* 2015, *521*, 436−444.
(55) Gordon, I. E.; Rothman, L. S.; Hill, C.; et al. The HITRAN2016 molecular spectroscopic database. *Journal of Quantitative Spectroscopy and Radiative Transfer* 2017, *203*, 3−69.
(56) Pickett, H.; Poynter, R.; Cohen, E.; Delitsky, M.; Pearson, J.; Muller, H. Submillimeter, millimeter, and microwave spectral line catalog. *Journal of Quantitative Spectroscopy and Radiative Transfer* 1998, *60*, 883−890.
(57) Rice, T. E.; Chowdhury, M. A. Z.; Powers, M. N.; Mansha, M. W.; Wilke, I.; Hella, M. M.; Oehlschlaeger, M. A. Gas Sensing for Industrial Relevant Nitrogen-Containing Compounds Using a Micro-electronics-Based Absorption Spectrometer in the 220 to 330 GHz Frequency Range. *Sens. Actuators, B* 2022, *367*, 132030.
(58) Rice, T. E.; Chowdhury, M. A. Z.; Mansha, M. W.; Hella, M. M.; Wilke, I.; Oehlschlaeger, M. A. Halogenated hydrocarbon gas sensing by rotational absorption spectroscopy in the 220−330 GHz frequency range. *Appl. Phys. B: Laser Opt.* 2021, *127*, 1−9.
(59) Bogatinovski, J.; Todorovski, L.; Dzeroski, S.; Kocev, D. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications* 2022, *203*, 117215.
(60) Kingma, D. P.; Ba, J. *Adam: A method for stochastic optimization*; arXiv preprint arXiv:1412.6980 2014. https://arxiv.org/pdf/1412.6980.pdf.

https://doi.org/10.1021/acssensors.2c02615
ACS Sens. XXXX, XXX, XXX−XXX

(61) Chowdhury, M. A. Z.; Rice, T. E.; Powers, M. N.; Mansha, M. W.; Wilke, I.; Hella, M. M.; Oehlschlaeger, M. A. Gas Sensing for Commercial Refrigerants R-134a and R-1234yf Using Rotational Absorption Spectroscopy in the 220−330 GHz Frequency Range. *Journal of Infrared, Millimeter, and Terahertz Waves* 2022, 1−12.
(62) Kochanov, R. V.; Gordon, I.; Rothman, L.; Wcisło, P.; Hill, C.; Wilzewski, J. HITRAN Application Programming Interface (HAPI): A comprehensive approach to working with spectroscopic data. *Journal of Quantitative Spectroscopy and Radiative Transfer* 2016, *177*, 15−30.
(63) Breiman, L. *Classification and regression trees*; Routledge: 2017.
(64) Breiman, L. Random Forests. *Machine learning* 2001, *45*, 5−32.

https://doi.org/10.1021/acssensors.2c02615
ACS Sens. XXXX, XXX, XXX−XXX