Iowa Initiative for Artificial Intelligence
Final Report

Project title:	AI/ML supported analysis of multiple volatile organic compounds (VOCs) detection data with high spectral overlap
Principal Investigator:	Fatima Toor (Electrical and Computer Engineering)
Prepared by (IIAI):	Avinash Mudireddy
Other investigators:	-
Date:	06/10/2022
Were specific aims fulfilled:	Yes
Readiness for extramural proposal?	The work supported by this pilot project has certainly prepared us to move forward with our multi-gas sensor development in an efficient manner.
If yes … Planned submission date	Proposals submitted to DHA, NSF. Submission for DoD RFP due in September 2022 in final stages that will require the use of AI to deconvolve the sensors data.
Funding agency	DHA, NSF, DoD
Grant mechanism	SBIR/STTR
If no … Why not? What went wrong?	

Brief summary of accomplished results:

Research report:

Aims (provided by PI):

The investigators will provide gas absorption spectral data for VOCs available from the NIST database [1] with the following parameters:

1. Three VOCs without spectral overlap: butyric acid, acetone, dimethyl sulfide
   a. We will provide spectral data for 10 different concentrations of each gas
   b. We will then mix each of these 10 different concentrations of each gas with the other two to generate permutations of spectral data
2. Three VOCs with spectral overlap: dimethyl amine, ethanol, phenylacetic acid.
   a. We will provide spectral data for 10 different concentrations of each gas
   b. We will then mix each of these 10 different concentrations of each gas with the other two to generate permutations of spectral data
3. We will add interfering spectral data due to water vapor at different concentrations within the pure gas mixture data

Six VOCs were selected - each of them are a biomarker for a particular disease when released from human skin or sweat. This project data will help strengthen our NIH U01 application.

Modified aims after collaboration with IIAI:

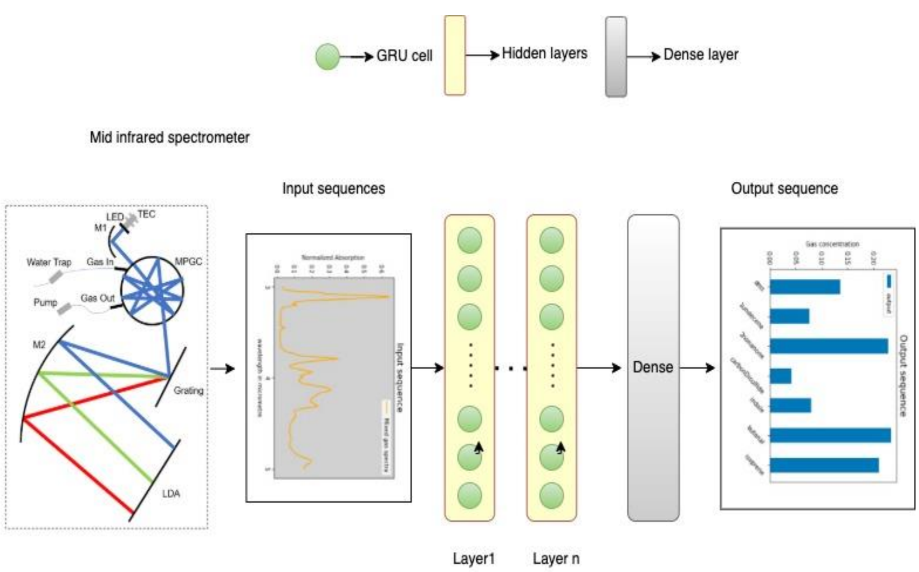
- Aim 1: Determination of percentage concentrations of elementary VOCs in synthetic VOC-mixture spectra. The number of gases increased from 6 to 7. The new gases for analysis are dimethyl sulfide, 1-undecene, 2-nonanone, isoprene, carbon disulfide, butanal, indole, and isoprene.
- Additional aim (Aim 2): Determination of bacteria emitting the synthetic VOC-mixture spectra taking the spectrum for the gas as input. The first 4 gases in the above aim are released by P. aeruginosa and the rest by E. coli

Data:

The dataset for the project is generated synthetically for the pilot. We evaluated the effect of using a spectrum between 3 to 10 μm wavelengths (dataset 1)and 3 to 5 μm wavelengths (dataset 2) divided into a 300-length array by interpolation method.
For Aim 1 the input dataset contains 1.5 million samples divided into 70/20/10 split of training/validation/testing containing the mixed spectra of the 7 gases. The target is to predict the proportions of individual gases in the mixed spectra.
For Aim 2 the input is the same as Aim 1. The target is to predict the probabilities of 3 bacteria classes namely mixed class, P. aeruginosa, and E. coli.

AI/ML Approach:

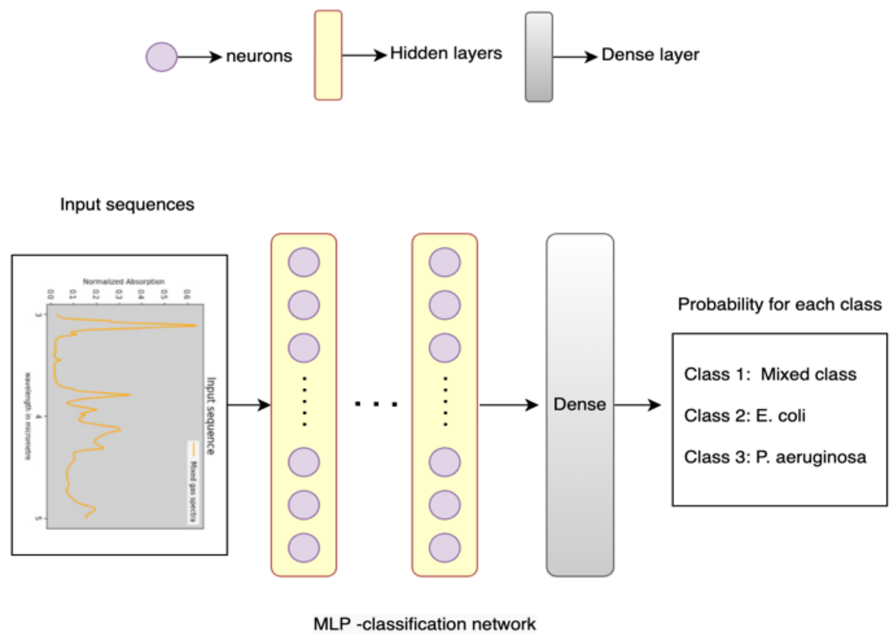
We proposed GRU-based Recurrent neural networks for AIM 1 and MLP-based ANN for Aim2. The model setup for both the Aims are illustrated in the below pictures.
Aim 1:



GRU - RNN regression network

The figure illustrates the input created by a spectrometer. However, in our case, the inputs are generated synthetically. The model constitutes 4 hidden layers with 256, 256, 128, and 128 GRU cells in each layer respectively followed by a dense layer with 7 neurons.

Aim 2:



The model constitutes 3 hidden layers with 64, 32, and 16 neurons in each layer respectively followed by a dense layer with 3 neurons.

Experimental methods, validation approach:

For Aim 1:
Both datasets are further feature-engineered using principal component analysis (PCA) with an explained variance of 0.99. After applying PCA to our synthetic datasets, the features per sample were reduced to 5 and 7 respectively for 3 to 10 µm and 3 to 5 µm wavelengths.

We used Adam optimizer with a learning rate of 0.0001. The model is trained for 50 epochs with 256 as batch size. The performance of the model is evaluated using the root mean square error (RMSE) and the coefficient of determination (R2 score) as metrics.

For Aim 2:
Both input datasets for the bacteria prediction model remain the same as that of Aim 1 with PCA.

We used Adam optimizer with learning rate of 0.001. The model is trained for 20 epochs with 256 as batch size. The model is trained to minimize the binary cross entropy loss (BCE) between the prediction and actual concentration.

Results:

The results for both the Aim 1 and Aim 2 are represented below in Table 1 and Table 2, respectively.

Table.1

Model	Root mean square error (RMSE)	Coefficient of determination (R^2)
3 to 10 µm dataset	0.0185	0.9820
3 to 5 µm dataset	0.0038	0.9992

Table.2

Model	class	precision	recall	f1-score	accuracy
3 to 10 µm dataset	Mixed class	1.0000	1.0000	1.0000	0.9998
	P. aeruginosa	1.0000	1.0000	1.0000	0.9998
	E. coli	1.0000	1.0000	1.0000	0.9998
3 to 5 µm dataset	Mixed class	1.0000	1.0000	1.0000	0.9997
	P. aeruginosa	1.0000	1.0000	1.0000	0.9997
	E. coli	1.0000	1.0000	1.0000	0.9997

Ideas/aims for future extramural project:

Environmental pollutants and VOC based biomarkers, both drive the need for mid-IR gas sensors, which in turn AI alogrithms to improve sensor detection process. PI Toor is targeting the EPA and DoD for funding opportunities to further build upon the work of the pilot grant.