Machine Learning and Convolutional Neural Network based Trace Identification of Pesticides and Dye Mixtures from Raman Spectra

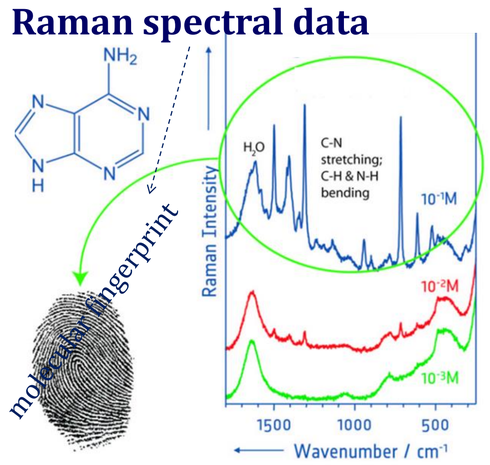
Venkat Suprabath Bitra
International Institute of Information Technology Bangalore, India
venkat.suprabath@iiitb.ac.in

Shweta Verma
Raja Ramanna Centre for Advanced Technology, Indore, India
shwetaverma@rrcat.gov.in

B. Tirumala Rao
Raja Ramanna Centre for Advanced Technology, Indore, India
trao@rrcat.gov.in

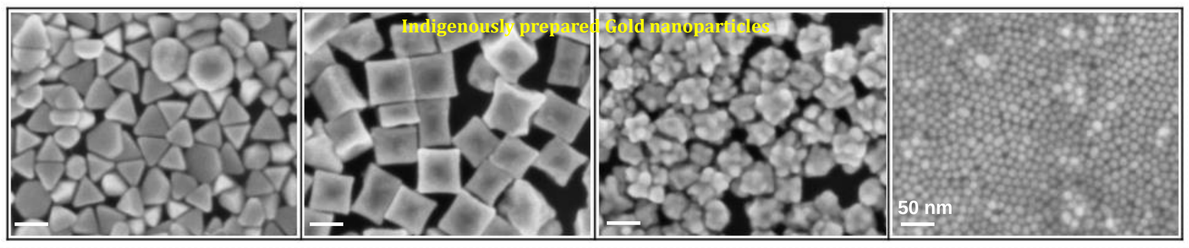
Raman spectral data

Background knowledge



- Non-destructive detection of compounds for
  - Food safety
  - Pharmaceuticals
  - Narcotics
  - Explosives
  - Life sciences
- **Limitation**: Low detection limit
- **Solution**: **S**urface **E**nhanced **R**aman **S**cattering (**SERS**)
  - Raman spectra in presence Gold (or Silver) nanoparticles

Signal enhancement 10^4 - 10^6 times, allows trace detection up to ppm or ppb



Current status:

- Difficult to identify compounds from standard library.

(spectral features: line-profiles, intensity and noise alters with SERS substrate, detector, laser wavelength & power, integration time, analyte concentration etc.)

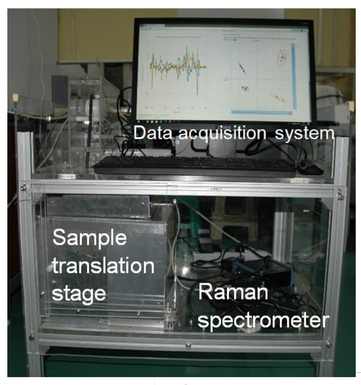
- No public Raman spectra datasets for reliable ML/DL studies.
- Difficult in prediction of multiple compound mixtures.

Objectives:

- Aim to identify various compositions of binary and ternary mixtures of two pesticides and one dye
- Image-based analysis of SERS spectra of varied dimensions for accurate predictions of mixtures (to overcome limitations of spectrometer hardware)

Methodology

Indigenously developed novel system for high throughput data acquisition:



Within 5 hrs. it can acquire 1000 labelled data (i.e. spectra, of 5-10 s integration time) with subsequent pre-processing.

Spectrometer: Portable fibre optic
make: Ocean Optics, model: QEpro

Indigenously developed software for batch analysis

Experimental

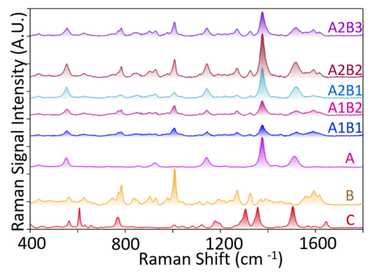
Compound labels:
A -Thiram
B - Thiabendazole
C - Rhodamine 6G

Compound concentrations:
A1: 0.25 µM & A2: 0.5 µM
B1: 6 µM, B2:12.5 µM & B3:25 µM
C1: 0.5 µM, C2: 2.5 µM & C3: 5 µM

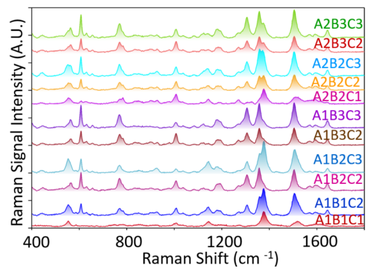
Data set: 300 measurements for each mixture
5 samples with 60 different combinations of: integration time (1,2,5,8 & 10 sec), laser power (25, 50, 75 &100 mW) and mixture concentration (33%, 50% & 100%).

Results and Analysis

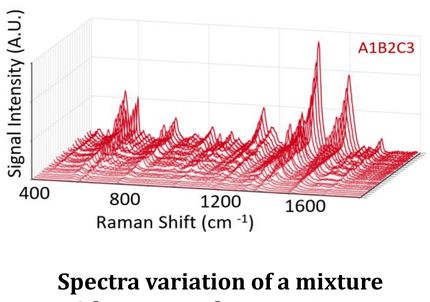
Indigenously developed low cost SERS technique to produce the dataset.



Spectra of individual compounds and binary mixture



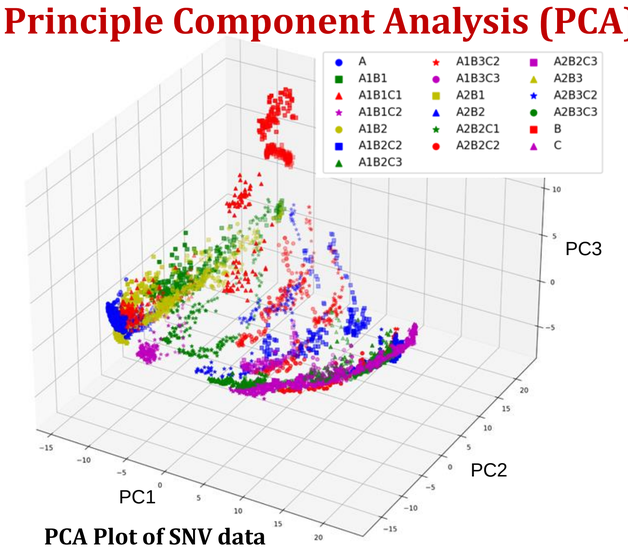
Spectra of tertiary mixture



Spectra variation of a mixture with measured parameters

- Total 5980 samples and 20% stratified split test set.

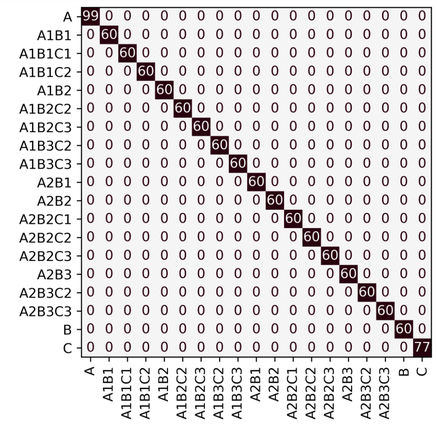
Principle Component Analysis (PCA)



PCA Plot of SNV data

Data normalization with SNV is important in PCA for accurate composition prediction.

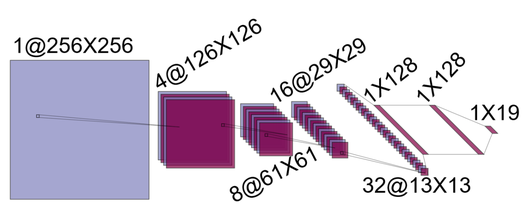
Model	Parameters	Optimal Parameter Value	Accuracy
SVM	Kernel Type	RBF	100%
	C	8	
	γ	0.003615	
KNN	Query Type	Ball Tree	99.91%
	Leaf Size	10	
	No. of Neighbors	4	
Logistic Regression	Penalty Type	L2 Norm	100%
	Solver	lbfgs	
	C	128	



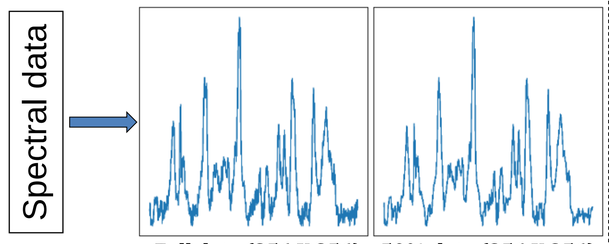
Confusion matrix: SNV + PCA-SVM

- Presence of certain narrow fingerprint feature of each compound resulted high accuracy.

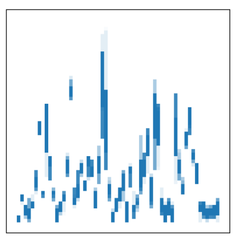
Spectral images processing with CNN



CNN architecture diagram



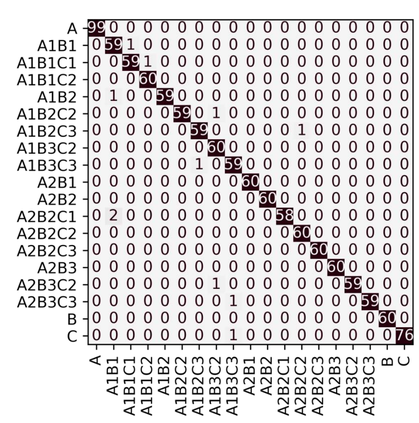
Full data (256 X 256) **50% data (256 X 256)**



Full data (64 X 64)

Image Size	Training Accuracy	Validation Accuracy	Testing Accuracy
64X64	90.59%	86.73%	86.95%
128X128	99.08%	97.38%	96.90%
256X256	99.97%	98.53%	99.08%
256X256 (50% data removal)	98.98%	97.86%	97.90%

- Data without normalization where PCA failed to predict, CNN provided 97.24% accuracy.
- Image based technique for accurate predictions from different dimensions of train and test data



Confusion matrix: image CNN

Conclusion

- Demonstration of ML based accurate prediction of composition of compound mixtures by integrating indigenously developed technologies
- Data analysis with SNV normalized PCA-SVM is useful when large data sets are available for a given hardware
- Spectra image analysis with CNN has advantage for data with and without normalization and different dimensions, independent of hardware
- The study is useful for development of ML driven Raman systems for Pharmaceuticals, Narcotics, Military, Food Safety, Bio-medical etc.

References

1. Lussier F., Thibault V. *et al* 2020. Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering *TrAC Trends in Analytical Chemistry* 124 (2020) 115796.
2. Cooman T., Trejos T., Romero A.H., Arroyo L.E., 2022. Implementing machine learning for the identification and classification of compound and mixtures in portable Raman instruments, *Chem Phys. Lett.* 787 (2021), 139283.
3. Pedregosa F.*et al* 2011. Scikit-learn: Machine Learning in Python *J. Machine Learning Research* 12 (Nov. 11) 2825-2830.