Machine Learning Based Detection of Trace Organic Contaminants in Water

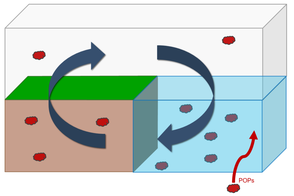
Authors: Vishnu Jayaprakash, Dr. Jae Bem You, Dr. Xuehua Zhang, Dr. Esma Khatan

In Concert with: MITACs Globalink, Kyungpook National University, University of Alberta

Background:

Chemistry has progressed immensely as a science and a business over the course of the twentieth century. However, our past rapid development was done without the knowledge for the perverse impact that human activities could have on the world ecosystem. As climate change and other human caused environmental effects came into light during the latter half of the 1900s it became clear that our progress without foresight came at a large cost. Many chemical additives that were discovered over the past 100 years have recently been found to be linked to many environmental, ecological, and health issues. It is the responsibility of modern scientists and engineers to manage and reverse the harm that humanity has so far caused to the biosphere.

Figure 1: Bioaccumulation and Cycling of persistent organic pollutants (POPs).



Amongst chemical additives that are negatively impacting our world, one class of compounds stands alone in the complex issues they creates. Persistent organic pollutants (POPs) are chemical compounds that are produced through human activity and have no natural remediation mechanisms. As such, these compounds can infiltrate many areas of the ecosystem and, through bioaccumulation, pose great risk to flora, fauna, and human health. As they lack natural remediation the removal of these POPs from the ecosystem must be a human endeavour.

Particularly difficult to manage POPs are those that are typically emitted in such low quantities that they can be close to undetectable until they build up and fully infiltrate the environment. These are known as trace organic compounds (TOCs) and pose a unique challenge in their detection and removal from the ecosystem. TOCs are produced through a variety of human activity including pharmaceuticals, cosmetics, industrial processing, and consumer products. Their effects are also very diverse from cancer risk to reducing human reproductive health to damage to flora and fauna.

A new advancement in the detection of TOCs is surface enhanced Raman spectroscopy (SERS), which has recently been shown to be able to detect TOCs at concentrations below 10^{-5} mol/L. Despite this considerable advancement, in situ detection at such low concentrations is limited by large variance in the Raman spectra of low level samples. With proper and accurate detection of these TOCs governments can refine their regulations and companies/municipalities can better control their wastewater.

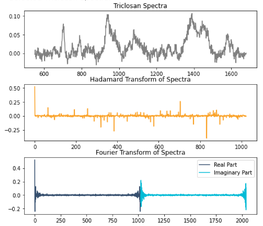
Objective:

This work uses nanodroplet formation techniques coupled with SERS to create a dataset of spectra for analysis. Machine learning and signal processing methods are explored to achieve accurate and generalizable predictions of concentration of three key TOCs—R6g, Triclosan and Chlorpyrifos. This investigation focuses on noisy, small datasets which are more representative of field applications.

Methodology and Comparisons to Literature:

The determination of concentration via machine learning from Raman spectral data is not a well studied problem. Standard applications of machine learning are well developed to differentiate between chemicals and between mixtures, but not between concentrations of a specific chemical. Determination of specific concentration is generally poorly conditioned as spectra from different concentrations are typically quite similar. Additionally, at the low concentrations that are expected from TOCs, spectral noise and interference from droplet media become a large issue. There is also a considerable difficulty in acquiring a large dataset for concentration determination as many different preparations of samples are required, unlike for differentiation between chemicals.

Figure 2: Application of Fourier and Hadamard Transform to a Selected Triclosan Spectra.



To manage some of these issues, supervised classification techniques are used over the typically regression that would be expected of numerical data. This is done by treating orders of magnitudes of concentration as classes to be assigned by the machine. This approach leads to a better conditioned problem even for smaller datasets. Secondarily, the approach of this work involves considering the spectral classification problem as a time series classification problem, which is well studied in machine learning science. This work uses frequency domain transforms, specifically the Fourier transform as well as the Hadamard transform. These are both techniques commonly in signal processing to improve the usability of time series data. These techniques have the added benefit of improving the models performance with data that is not baseline corrected. Other frequency domain techniques, such as power spectral density and z-transform were also considered.

Data Exploration:

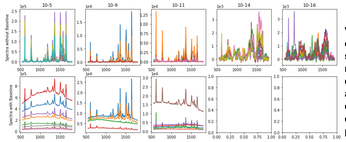
The collected spectra are split between five datasets—r6g via evaporating Ouzo, r6g via silver nanoparticles, triclosan via silver nanoparticles, triclosan via evaporating Ouzo, and chlorpyrifos via silver nano particles. The r6g ouzo data includes include spectra that are not baseline corrected, and both the r6g and triclosan ouzo data contain considerable noise from droplet generation.

Table 1: Spectral Count and Distribution for Chemicals

	R6g				Triclosan				Chlorpyrifos	
	Evaporating Ouzo		Silver Nanoparticles		Silver Nanoparticles		Evaporating Ouzo		Silver Nanoparticles	
	Conc. (mol/L)	Count	Conc. (mol/L)	Count	Conc. (mol/L)	Count	Conc. (mol/L)	Count	Conc. (mol/L)	Count
	10^{-4}	13[1]	10^{-4}	10	10^{-4}	5	10^{-4}	6[2]	10^{-4}	10
	10^{-5}	8[1]	10^{-5}	10	$5\cdot10^{-5}$	5	10^{-5}	6[2]	10^{-5}	10
	10^{-6}	14[1]	10^{-6}	10	10^{-5}	5	10^{-6}	6	10^{-6}	10
	10^{-7}	9[1]	10^{-7}	10	$5\cdot10^{-7}$	5	10^{-7}	6[2]	10^{-7}	10
	10^{-8}	16[3]	10^{-8}	10	10^{-7}	5	-	-	10^{-8}	10
Total:	60		50		25		24		50	

[1] Contains Baseline Uncorrected Data
[2] Contains Increased Noise

Figure 3: Evaporating Ouzo Raman Spectra for R6g, Baseline Corrected Spectra on Top Row and Uncorrected spectra on Bottom Row

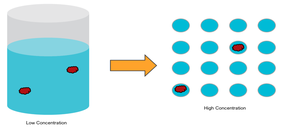


Variation across the same concentration involves the overall stretching or shrinking of the peaks in the spectra. Between concentrations some key peaks are modified. This variation is very slight and each concentration category has considerable overlap with adjacent categories. Correlation matrices were created for all three chemicals and the average correlation coefficient is generally low indicating the peaks are relatively independent in their variance.

Droplet Generation and Spectral Measurement:

e

Figure 4: Concentration of TOC infiltrated water via nanodroplet formation.

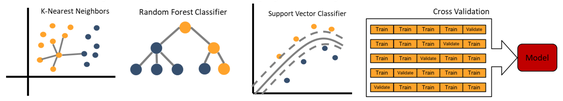


Classical Machine Learning:

The first machine learning techniques that are considered are standard classification models. Random Forest, k-Nearest Neighbors and Linear Support Vector classification methods were used. Random Forest Classification is a technique that consists of the machine using training data to create a series of decision trees which are then used to new place spectra into a appropriate class. K-Nearest Neighbors compares spectra to a set amount to nearest neighbors grouping the spectra into classes based on similarity and then makes predictions by placing new spectra into one of these groupings. Support Vector Classification works by creating a vector that best separates each class from each other and uses that as a boundary to predict the placement of new spectra.

Each of these methods was tested with default hyperparameters and then later refined with Bayesian hyperparameter tuning, which is a gradient descent method. This process was done with simple between spectra scaling as well as the Fourier and Hadamard transforms. Across spectra normalization was also investigated but had inconsistent results. Scaling is combined with the frequency domain transforms whenever considerable improvements are seen, with both pre and post-transform scaling being considered. For the evaluation of methods and comparison a five-old cross validation is done in the r6g case and four-fold in the triclosan and chlorpyrifos case. Cross validation involves splitting the set of training data into several subsets and then using one of the subsets to validate the training of the rest. This process is repeated with each subset being used for validation in order to improve the generalizability of the model. Average cross validation scores and standard deviations are complied as results for the training. Additionally, confusion matrices were generated for the test set prediction for each combination of transform and machine learning model.

Figure 5: Diagrams of the Machine Learning Algorithms Investigated and Cross Validation.

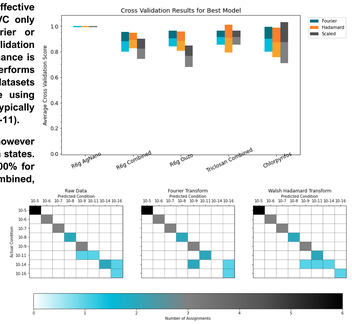


Classical Approach Results:

Models are compared based upon average cross validation results. Random Forest proved to be the most effective method out of the standard models. KNN and SVC only performed comparable when improved by Fourier or Hadamard. The two transforms improve cross validation results in all cases except triclosan, where performance is he same as the scaled. The Hadamard transform performs best in the combined r6g dataset. Across all datasets accuracies of greater than 85% were achievable using frequency transforms. Errors that are seen are typically incorrect by one or two classes (10-9 classified as 10-11).

Results are all reported on the random state 117, however consistent results are obtained from several random states. Actual prediction accuracy across datasets is - 100% for R6g AgNano, 83% for R6g Ouzo, 91% for R6g Combined, 88% for Triclosan Combined and 78% for Chlorpyrifos. The hyperparameters for RF are max_features, n_estimators and criterion with a log-normal sampling for numerical hyperparameters. Similarly, kernel and C coefficient were tuned for SVC and n_neighbors, weights and metric was tuned for k-NN.

Figure 6: Cross Validation Results for Classical Machine Learning Approach. Best ML Method Results Displayed (Top). Confusion Matrix for R6g Combined Random Forest (Bottom).

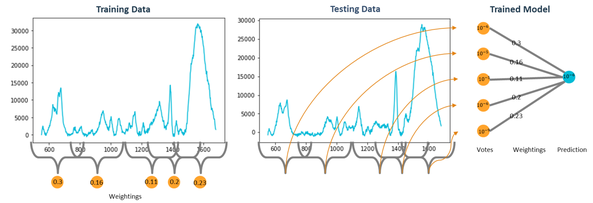


Convolutional Neural Network:

The literature standard approach for machine assisted spectral analysis is deep learning via Convolutional Neural Network (CNN). CNNs are a deep learning technique in which classification or recognition is achieved by tuning several decision making neurons that each handle one aspect of the data. In CNNs, generally, pictures are used as input data in the form of 2-D grids. The neurons are then trained to each recognize one aspect of the picture. For example, in a bird classification software one neuron would handle colouration, another would handle size, and a third would deal with wing shape. Each of these neurons then gives their input into the final classification decision. The typical CNN to handle Raman data involves reshaping the long spectral vector into a 2-D grid of numerical values. This approach is dependant on very large datasets with considerable variation, as would be expected of classification between different compounds. Additionally, this technique does not lend itself to make use of the frequency domain transformations that have been explored.

Therefore, in this work, a time series classification approach is taken via 1-D convolution. This is the approach taken by previous work to handle other time series data such as human activity recognition. In this method transformed or untransformed spectral data is fed directly to the CNN and each neuron is tuned to process a particular part of the overall spectra.

Figure 7: 1-D Convolutional Neural Network of Spectral Data.

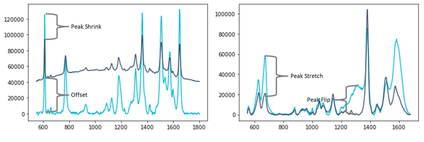


Data Augmentation:

One of the largest drawbacks of deep learning methods is the reliance on large datasets to tune the network of several neurons. For spectral data the collection of a sufficiently large dataset is near impossible. Therefore, modern techniques involve data augmentation which is the creation of additional data that can realistically supplement the original dataset leading, enabling deep learning. The construction of augmented data is something that needs to be taken with care, if the additional data is representative of the original data it can improve performance and generalizability. However, if the augmentation is not true to the original performance will be severely hindered. There is also a risk in overfitting by constructing an augmentation strategy that is too specific to the particular testing set.

The focus of all data augmentation is to create realistic modified version of the original data. In 2-D picture classification via CNN the traditional method of augmenting data is rotation, transposition and translation of the picture. While past research has implemented this strategy for spectral CNNs it is difficult to implement in a way that is representative of real spectra. As the approach in this project is based on time series classification the approach for data augmentation changes. The augmentation algorithm used takes the original spectral data as input. It then detects offsets dataset and checks the variance in peak distribution across the dataset. Next, a random spectra from the set is selected. If the spectra has an offset then the offset is increased or decreased by a random amount based on the standard deviation of all offsets in the set. If the variance in the peak distribution is larger than a key value it means peak flipping is present in the dataset. This is the semi-random increase in intensity of some of the smaller peaks due to the orientation of the compound or noise. If this is the case, the algorithm will randomly select 1-3 smaller peaks and increase their intensity to within 20% of the large peaks. For all spectra, stretching or shrinking of the peaks is implemented. This stretch is a fraction on the standard deviation of all peaks in the spectra.

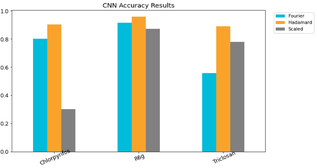
Figure 8: Examples of Variation in Dataset Considered for Augmentation of Data.



CNN Results:

The implemented CNN uses architecture based on the architecture typically used for time series classification. The final model consists of two 1-D convolutional layers with 64 filters and 'relu' activation, a dropout layer, a max pooling layer, a flatten layer and two dense layers with a final 'softmax' activation. Twenty epochs were used with early stopping procedure. Across all runs Hadamard transformed data performed best with accuracies of 96%, 89% and 90% for r6g, triclosan and chlorpyrifos respectively. As is typical with deep learning applications, there is variation across random states. As such, all reported values are from the random state 117 to maintain consistency.

Figure 9: Accuracy Results for CNN



Conclusions:

T