Defense against adversarial spectroscopic attacks

Casey J. Smith, Youlin Liu, & Garth J. Simpson*
Department of Chemistry, Purdue University, West Lafayette, IN, 47906
*gsimpson@purdue.edu

ABSTRACT:
As the volume of data used for chemical decision-making increases, so too grows its susceptibility to nefarious manipulation to alter outcomes. The vulnerability of conventional spectral classification to digital adversarial attack is explored using Raman spectroscopy as a representative example. Following supervised training by linear discriminant analysis (LDA), addition of patterned "noise" to an initial spectrum enabled reclassification to an alternative target spectrum. The "attacked" spectra visually retained key spectral features of the true, initial spectra, but were misclassified with high statistical confidence as the target. The digital attack demonstrated herein is intended to help serve as a testbed for evaluation of provenance strategies for ensuring the integrity of data-intensive decision-making.

Keywords: Raman Spectroscopy, Genetic Algorithm, Spectroscopic Attack, Big Data, Dimension Reduction, Linear Discriminant Analysis (LDA), Adversarial Spectroscopy,

1. INTRODUCTION:

Modern instrumentation has dramatically increased the amount of data available for decision making[1,2]. Those tasked with making decisions based on this wealth of knowledge routinely turn to methods and techniques that can handle the vast quantity of information, such as artificial neural networks, principle component analysis, linear discriminant analysis, support vector machines, and other dimension reduction methods[2]. As instrumentation continues to improve, the greater becomes our collective reliance on algorithmic data analysis approaches. These data analysis approaches are being used; and will be relied on in the future to make decisions in very important fields such as, drug testing, DNA matching, regulation of pharmaceutical manufacturing, voice/facial recognition, among many other fields[3,4,5].

Given the societal importance of these fields, it is imperative that researchers have reliable and robust classification methods. However, intentional non-probabilistic digital perturbations have the ability to be used to misrepresent the actual statistical confidence of a decision or result in intentional misclassification. Researchers in the field of computer science and electrical engineering have demonstrated various avenues of attacking an artificial neural network designed to classify images, with most methods being imperceptible to the human eye[6,7,8,9].

To date, adversarial attacks on chemical measurements have not been demonstrated. Such a perturbation could be applied to the background subtraction file and defy routine detection but have a profound impact on decisions based on that measurement. This type of attack would have profound consequences especially in the pharmaceutical industry where getting a drug to market holds a significant financial importance. Opportunities for subtlety increase as the dimensionality of the spectral data.

In the present work, we demonstrate a method for a digital adversarial spectroscopy attack. We hope that this work provides a test-bed for i) developing methods for detecting spectroscopic malfeasance and ii) enable novel nondeterministic classification strategies less vulnerable to adversarial attacks. In designing the adversarial attack, we determine the optimal perturbation through a genetic algorithm, and then apply the perturbation to induce misclassification in a reduced-dimensional space defined by linear discriminant analysis (LDA).

2. MATHEMATICAL FRAMEWORK:

2.1 Cost function in the reduced dimensional analysis.

The main objective of the adversarial attack here is to produce a perturbation, $\boldsymbol{\delta}$, that is not detectable by visual inspection, yet produces a substantial deviation in the reduced-dimensionality space. For the initial sample spectrum $\boldsymbol{x}_s$, the perturbed spectrum is given by $\boldsymbol{x}'$ according to Eq. 1.

$$\boldsymbol{x}' = \boldsymbol{x}_s + \boldsymbol{\delta} \tag{1}$$

LDA in optimizes the resolution between the defined classes. Each wavelength channel in the original spectral space results in a vector that contributes to the position of the spectrum in the reduced dimensional space. While the main spectral features combine to determine the general position within the LDA space, normally distributed noise produces a spread about that mean position.

The method of attack could be thought of as patterning the noise levels in the original spectra to relocate the position of the spectra in the LDA space. Effectively, it is the addition of many low-amplitude perturbations that both induces misclassification and makes it visually difficult to detect. The deviations $\boldsymbol{d}$ from the initial sample spectrum $\boldsymbol{x}_s$ to the "target" $\boldsymbol{\mu}_t$ is given by the following expression.

$$\boldsymbol{d} = \left(\boldsymbol{x}_s + \boldsymbol{\delta}\right) - \boldsymbol{\mu}_t \tag{2}$$

Applying the Euclidean distance formula to the deviations, one can compute the distance of the perturbed spectra to the target. The squared deviations are given by projection of $\boldsymbol{d}$ onto the reduced dimensional space through the matrix D, given by $\left\|D\boldsymbol{d}\right\|^2$. Where D is a matrix of eigenvectors corresponding to a reduced-dimensional space (e.g., LDA or PCA). In the present study, two dimensions in LDA-space were considered in the analysis of spectra with 1340 elements, such that D is a 2×1340 matrix in the present study.

To apply the requirement that the attack be difficult to visually detect, another requirement must be built into the cost function. An additional cost function term was added such that original spectrum remained as unchanged as possible. Mathematically, this is done by minimizing the sum of squared deviations in the spectra space. The total cost function for the reduced dimensional analysis was given by the sum of the two terms, which collectively minimized the squared deviation to the target while minimizing the overall magnitude of the perturbation.

$$\hat{\delta} = argmin\left[||D(x_s + \delta - \mu_t)||^2 + \beta \sum\nolimits_i \left(x_i - (x_i + \delta)\right)^2\right] \tag{3}$$

The scaling parameter β in Eq. (3) allows for empirical adjustment of the relative importance given to proximity to target relative to minimizing perturbation to the major spectral features. In the present study, a value of β=1 was used with unit weighting on each value of $\boldsymbol{\delta}$.

2.2 Statistical assessment of classification.

Statistical assessment provides a means for evaluating the confidence with which an initial or modified spectrum can be assigned as belonging to each of the finite set of possible classes.

For compatibility with the limited training data (252 spectra, 84 per class) used in the Raman analysis, the data were assumed to be normally distributed about the mean at each wavelength. To test this assumption, the skewness and kurtosis at each wavelength were evaluated, the mean values of which were 0.30 and 3.28, in agreement with the expected values of 0 and 3 for normal distributions. The assumption of a normal distribution is also qualitatively consistent with the symmetrically distributed observed projections in LDA-space. To employ common statistical methods, the data were converted to the z parameter, which is given by $z_{ni} = \left(x_{si} - \mu_{ni}\right)/\sigma_{ni}$. Where $x_{s,}$ is the sample spectrum for the class n at each wavelength, i.

It is helpful to first consider the confidence with which a single scalar value can be assigned to each of two classes in a two-class system, followed by extension to full spectral analysis and an arbitrary number of classes.

Considering just a single measurement at the wavelength *i*, the ratio *r* of probabilities that $\left(x_s\right)_i$ belongs to either class 1 or class 2 (P_1 or P_2, respectively) is given by the following expression.

$$\left(r_{12}\right)_i \equiv \frac{P_1}{P_2} = \frac{f_1\left(z_{1i}\right)dz_{1i}}{f_2\left(z_{2i}\right)dz_{2i}} = \frac{f_1\left(z_{1i}\right)}{f_2\left(z_{2i}\right)} = \exp\left[-\tfrac{1}{2}\left(z_{1i}^2 - z_{2i}^2\right)\right] \tag{4}$$

Evaluation of the probability ratio in Eq. (4) is simplified considerably by recognizing that $dz_{1i} = dz_{2i}$ because of the normalization inherent in the definition of the *z*-statistic. In Eq. (4), *f* is the normalized probability density function for the *z* statistic, given by $f\left(z\right) = \exp\left(-\tfrac{1}{2}z^2\right)$. In the absence of covariance, the total probability including all wavelengths is recovered from the product of probabilities for each wavelength.

$$r_{12} = \frac{\prod_i P_{1i}}{\prod_i P_{2i}} = \exp\left[-\tfrac{1}{2}\left(\left\|\boldsymbol{z}_1\right\|^2 - \left\|\boldsymbol{z}_2\right\|^2\right)\right] \tag{5}$$

Since $P_1 + P_2 = 1$ for a spectrum that must fall into one of the two classes, substitution of $r_{12} = \left(1 - P_2\right)/P_2$ yields the expression $P_2 = \left(1 + r_{12}\right)^{-1}$. Bearing in mind that $r_{22} = 1$, the expression for P_2 can be equivalently written as $P_2 = \left(r_{22} + r_{12}\right)^{-1}$. Extension from two to an arbitrary number of *N*-classes is straightforward. By analogy with the 2-class system, the following expression describes the probability of a spectrum being assigned to a class *n* from a set of *N* possibilities.

$$P_n = \left(\sum_{j=1}^{N} r_{jn}\right)^{-1} \tag{6}$$

In the case of analyses performed in a reduced dimensional space defined by multiplication by a dimension reduction matrix *D*, the expression for $\boldsymbol{z}$ is modified to $\boldsymbol{z}_n = \left[\left(D\boldsymbol{x}_s\right) - \boldsymbol{\mu}_n\right]/\boldsymbol{\sigma}_n$, in which $\boldsymbol{z}$, $\boldsymbol{\mu}_n$, and $\boldsymbol{\sigma}_n$ are evaluated in the reduced-dimensional space.

2.3 Genetic algorithm for "blind" optimization of the perturbation δ.

Optimization of the perturbation, δ, was performed using a genetic algorithm. Use of a genetic algorithm was chosen because the optimal perturbation was non-trivial to derive. The genetic algorithm employed uses two primary functions to minimize the cost function, Equation 3. Namely, these functions are mutation and splicing or cross-over. Mutation is performed by randomly selecting a position (in this case, 1 through 1340) and multiplying that positional value by a random number selected from a uniform distribution from -2 to 2 excluding zero (to prevent the perturbation from getting trapped in a local minimum) to generate a new spectrum from the parent spectrum. Splicing was performed by selecting two parent spectra and selecting a random position, all wavelengths beyond or before that position were exchanged with the other parent spectrum to generate a new spectrum. After each generation, the best two perturbations were selected to be the parents for a new generation of mutations and cross-overs. The genetic algorithm ran for 400 generations with 2000 generated spectra per generation.

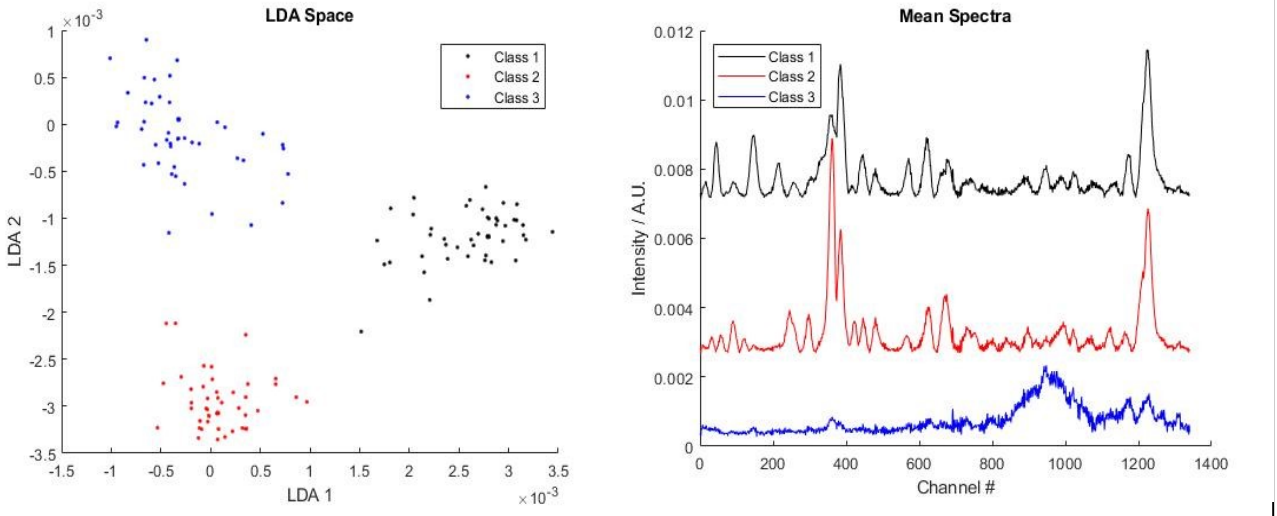

Figure 2. Projection of experimental Raman spectra in LDA-space, with the corresponding mean spectra for each class. Mean spectra is an average of 84 measurements.

3. EXPERIMENTAL METHODS:

Raman spectra were acquired using a custom Raman microscope, built in-house and described in detail previously.[10] In brief, a continuous wave diode laser (Toptica, 785nm wavelength) coupled into a Raman probe (InPhotonics, RPS785/24) was collimated by a 12.5 mm fused silica lens, and directed through an X-Y scan head composed of two galvanometer scanning mirrors. Two additional 25 mm diameter fused silica lenses formed a 4*f* configuration to deliver a collimated beam on the back of a 10x objective (Nikon). The Raman signal from the sample was collected through the same objective and descanned back through the same beam path into the Raman probe. A notch filter was built in the Raman probe to reject the laser signal. Raman spectra were acquired using an Acton SP-300i spectrometer with a 100x1340 CCD array, and controlled by a computer running WinSpec32 software.

Pure clopidogrel bisulfate Form I and Form II were produced in-house at Dr. Reddy's Laboratories. Both the Form I and Form II particles were spherical with similar particle size distributions (diameter: ~25 μm). The laser power measured at the sample place was ~30 mW. The exposure time was 0.5 s per spectral frame. To achieve higher signal to noise ratio for high quality training data for classification, 30 consecutive frames were averaged for each spectrum acquired over a spot size of ~2-3 μm diameter within the field of view. A Savitzky-Golay filter was applied to smooth the spectra[11], and a rolling ball filter was used to remove the fluorescence background[12]. Finally, the spectra were normalized to their integrated intensities, i.e., the area under the curves. The integrated intensity information of every spectrum was recorded so it can be retrieved when intensity information within each spectrum was needed for subsequent analysis.

4. RESULTS AND DISCUSSION:

The mean spectra, average of 84 measurements, for three classes are shown in **Figure 2.** The spectra corresponding to the background is classified as class 3 (bottom spectra, blue). The spectra belonging to the two polymorphs of clopidogrel bisulfate are classified as class 1 and class 2 (black top and red middle respectively). The recorded spectra for class 3 (background) signal has one major feature of note, that being a large rolling peak around channel number 950. Class 1 and class 2 while similar, show notable differences and can be distinguished by relative peak intensities of the major feature at 363 & 385, several peaks in class 1 that do not occur in class 2, and peak shifts between the first two classes. Linear discriminant analysis provided adequate separation and reduced the dimensionality of the data from 1340 channels to just two channels, which were plotted.

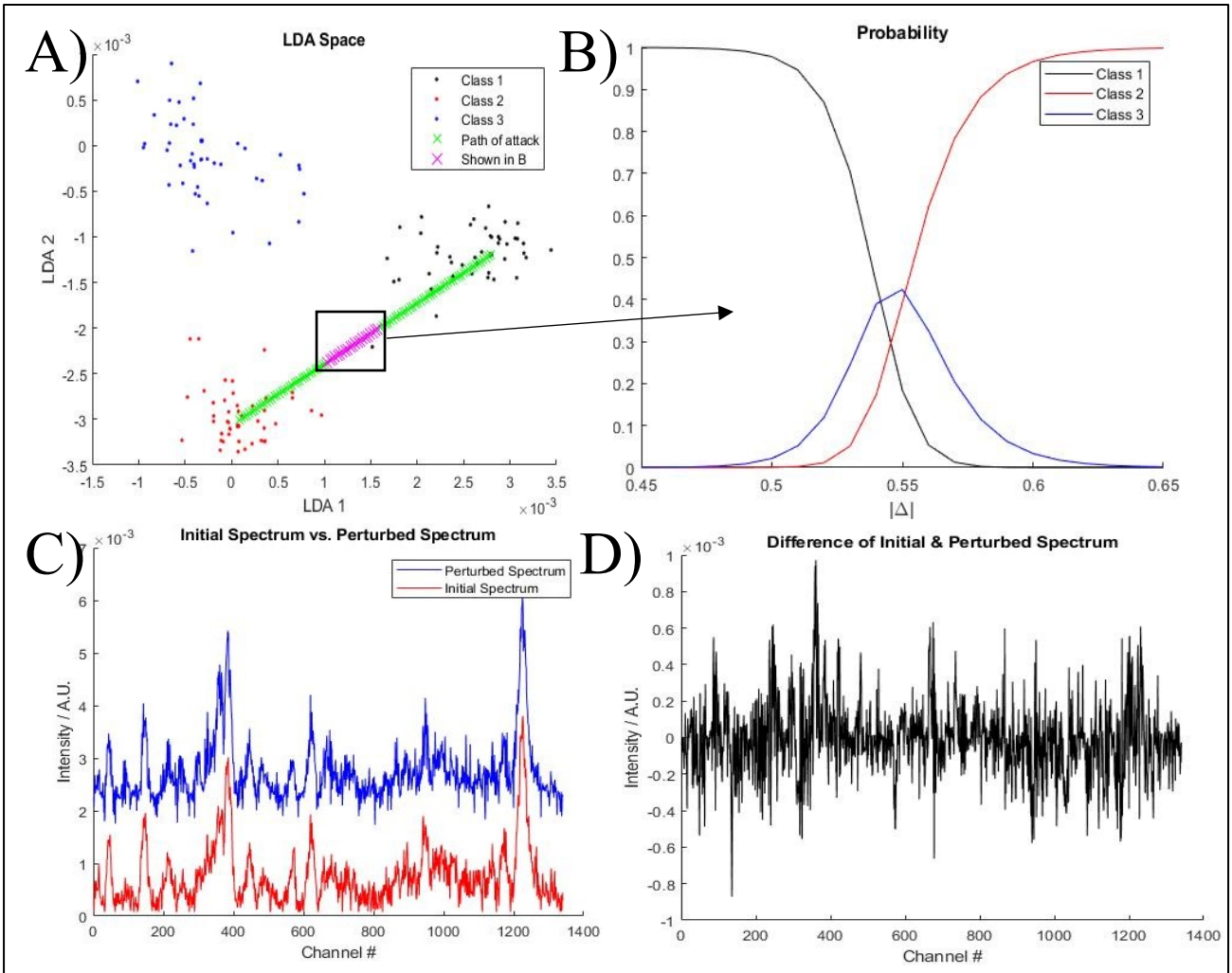

Figure 3. The optimal perturbation for moving class 1 to class 2 was scaled from 0 to 1 as shown in A. The green x's in A are the scaled perturbations. B) shows the zoom in of the region where the classification based on probability changed from class 1 to class 2. C) Shows the initial spectra vs. the perturbed spectra. The two spectra are offset for clarity. D) The perturbation used to induce misclassification.

Using this data set of 252 spectra, a demonstration of an attack in the reduced-dimensional LDA space was performed and shown in **Figure 3**. Figure 3A shows the reduced dimensional space with the 3 classes well separated. The perturbation in figure 3 was designed to misclassify a spectra belonging to class 1 as class 2. The optimal perturbation from the genetic algorithm moved the selected spectra from class 1 to the mean of the target class (class 2). The perturbation was multiplied by a scaling factor from 0 to 1 to produce the green x's on Figure 3A. The Equation 8 was used to determine the probability of the perturbed spectra belonging to a particular class which is shown in Figure 3B and denoted by the purple x's in 3A. When the optimal perturbation was rescaled to ~60% the probability of the perturbed spectra belonging to the target class was greater than 95%. Figure 4C highlights how difficult this attack is to detect upon visual inspection. The two spectra, although they classify differently in LDA space, show high visual similarity. Figure 4D shows the actual perturbation used to misclassify the selected spectra as the target class. These results show that the approach resulted in an attack pattern that is visually difficult to detect in spectral space, while squarely moving the spectra to the target class in the reduced dimensional space.

These trends are repeated in considerations of perturbations to induce misclassification from class 1 to 3 (not shown) and in class 3 to 2 in **Figure 4.**

This attack approach could be particularly successful if integrated into background-subtracted spectral analyses. Where the attack perturbation file is to the digital file corresponding to the background correction. It has the potential to dramatically alter outcomes based on chemical spectroscopic analyses in a manner that is challenging to detect forensically as shown in the Figure 3 and 4. Comparison of the magnitudes of the perturbations relative to the initial spectrum suggest that perturbations on the order of 12% are sufficient to unequivocally alter the spectral classification. Such relatively subtle changes spread over the entire spectrum would generally be challenging to discriminate from random noise in a background subtraction architecture.

These results highlight the growing challenges in ensuring statistical validity in regulatory, business, and legal decisions derived from data-intensive measurements. As demonstrated herein, subtle adversarial attacks on spectral information can completely change decision outcomes. Subtle digital alteration of files used in the routine operation of background subtraction can result in an adversarial attack in which the attack successfully misclassifies the outcome based on spectral analysis. As the volume of data integrated for decision-making increases, decisions based on chemical analysis is poised to be increasingly vulnerable to manipulation through adversarial perturbations.

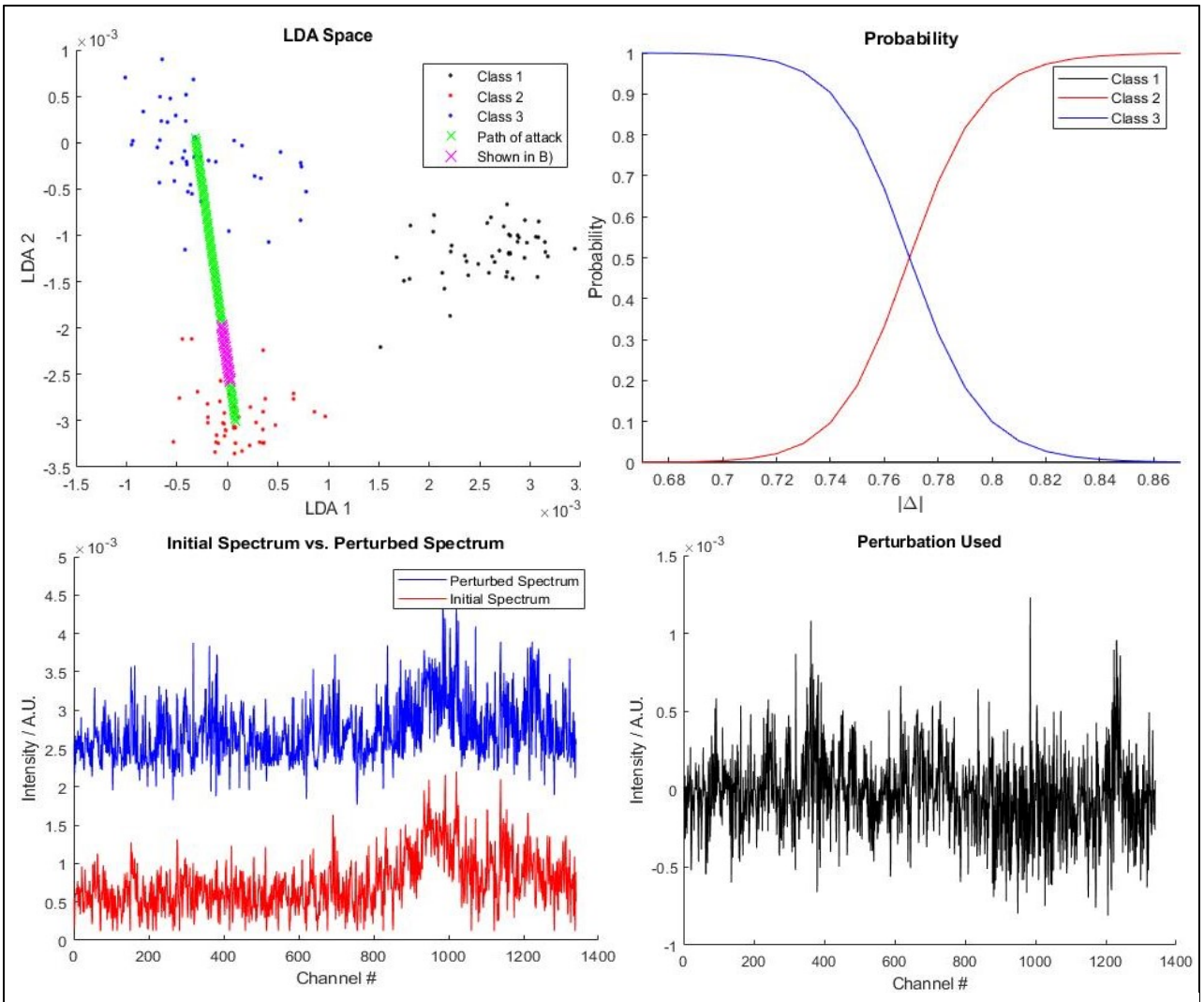

Figure 4. The optimal perturbation for moving class 3 to class 2 was scaled from 0 to 1 as shown in A. The green x's in A are the scaled perturbations. B) shows the zoom in of the region where the classification based on probability changed from class 1 to class 2. C) Shows the initial spectra vs. the perturbed spectra. The two spectra are offset for clarity. D) The perturbation used to induce misclassification.

5. CONCLUSIONS:

We present a method for performing digital attacks on chemical spectra, in this case Raman spectra of clopidogrel bisulfate, in which addition of a digital perturbation can result in high-confidence mis-classification. Because the additive perturbations were low amplitude (~12% of the initial amplitude) and spread out over the entire spectra, retain visual similarity of the initial spectra, but classify with high confidence to the target spectral class in a reduced dimensional space, LDA-space for this example.

6. REFERENCES:

(1) Provost, F.; Fawcett, T. Data Science and Its Relationship to Big Data and Data-Driven Decision Making. *Big Data* **2013**, *1*, 51-59.

(2) Hinton, G. E.; Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504-507.

(3) Ojha, V. K.; Jackowski, K.; Snasel, V.; Abraham, A.: Dimensionality Reduction and Prediction of the Protein Macromolecule Dissolution Profile. In *Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications*; Kromer, P., Abraham, A., Snasel, V., Eds.; Advances in Intelligent Systems and Computing; Springer-Verlag Berlin: Berlin, 2014; Vol. 303; pp 301-310.

(4) Clarke, R.; Ressom, H. W.; Wang, A. T.; Xuan, J. H.; Liu, M. C.; Gehan, E. A.; Wang, Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* **2008**, *8*, 37-49.

(5) Liberati, C.; Mariani, P.: Big Data Meet Pharmaceutical Industry: An Application on Social Media Data. In *Classification*; Mola, F., Conversano, C., Vichi, M., Eds.; Studies in Classification Data Analysis and Knowledge Organization; Springer International Publishing Ag: Cham, 2018; pp 23-30.

(6) Su, J.; Vargas, D. V.; Kouichi, S. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864* **2017**.

(7) Kwon, H.; Kim, Y.; Park, K. W.; Yoon, H.; Choi, D. Advanced Ensemble Adversarial Example on Unknown Deep Neural Network Classifiers. *IEICE Trans. Inf. Syst.* **2018**, *E101D*, 2485-2500.

(8) Zhang, G.; Yan, C.; Ji, X.; Zhang, T.; Zhang, T.; Xu, W. In *Tilte*2017; ACM.

(9) Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* **2016**.

(10) Zhang, S. J.; Song, Z. T.; Godaliyadda, G.; Ye, D. H.; Chowdhury, A. U.; Sengupta, A.; Buzzard, G. T.; Bouman, C. A.; Simpson, G. J. Dynamic Sparse Sampling for Confocal Raman Microscopy. *Analytical Chemistry* **2018**, *90*, 4461-4469.

(11) Ehrentreich, F.; Summchen, L. Spike removal and denoising of Raman spectra by wavelet transform methods. *Anal Chem* **2001**, *73*, 4364-4373.

(12) Liland, K. H.; Almoy, T.; Mevik, B. H. Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra. *Appl Spectrosc* **2010**, *64*, 1007-1016.