JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc

Activity

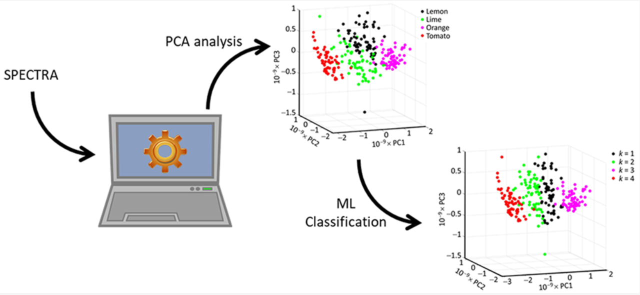
Exploring Machine Learning in Chemistry through the Classification of Spectra: An Undergraduate Project

Alanah Grant St James, Luke Hand, Thomas Mills, Liwen Song, Annabel S. J. Brunt, Patrick E. Bergstrom Mann, Andrew F. Worrall, Malcolm I. Stewart,* and Claire Vallance

Cite This: *J. Chem. Educ.* 2023, 100, 1343−1350

Read Online

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



ABSTRACT: Applications of machine learning in chemistry are many and varied, from prediction of structure−property relationships, to modeling of potential energy surfaces for large scale atomistic simulations. We describe a generalized approach for the application of machine learning to the classification of spectra which can be used as the basis for a wide variety of undergraduate projects. While our examples use FTIR and mass spectra, the approach could equally well be used with UV−visible, Raman, NMR, or indeed any other type of spectra. We summarize a number of different unsupervised and supervised machine learning algorithms that can be used to classify spectra into groups, and illustrate their application using data from three different projects carried out by fourth year chemistry undergraduates. The three projects investigated the ability of the various machine learning approaches to correctly classify spectra of a variety of fruits, whiskies, and teas, respectively. In all cases the algorithms were able to differentiate between the various samples used in each study, and the trained machine learning models could then be used to classify unknown samples with a high degree of accuracy (>98% in many cases). Depending on the extent to which students are expected to write their own code to perform the data analysis, the general model adopted in this work can be adapted for a variety of purposes, from short (one to two day) practical exercises and workshops, to much longer independent student projects.

KEYWORDS: *Upper-Division Undergraduate, Laboratory Instruction, Chemoinformatics, Interdisciplinary/Multidisciplinary, Computer-Based Learning, Chemometrics, Mass Spectrometry, Spectroscopy, Computational Chemistry*

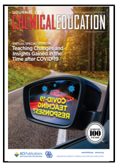
INTRODUCTION

Every chemist is familiar with the challenge of identifying an unknown compound or mixture of compounds. With a wide range of spectroscopic and mass spectrometric techniques at our fingertips, we are now able to perform sophisticated measurements on virtually any type of sample. In some cases, the sample and its spectra are simple enough to make identification straightforward. However, in many cases the spectra are sufficiently complex that our best approach is comparison with spectra from a reference library.[1−3] Spectral matching with library spectra is now almost exclusively performed by computer algorithms, making it easier than ever to identify individual chemical compounds and to characterize complex mixtures.

Machine learning (ML), a branch of Artificial Intelligence (AI), offers additional tools for the classification and identification of spectra. Machine learning algorithms use data to train a model, which can then be used to make predictions when presented with previously unseen data.[4] ML algorithms have already found a host of applications in chemistry, including the prediction of structure−property relationships,[5−7] the modeling of potential energy surfaces for large scale atomistic simulations,[8] the prediction of the electron densities of molecules,[9] the prediction of molecular structures from NMR spectra,[10] and new synthetic routes to complex chemicals.[11] However, despite its widespread use in research, there are not many examples of the use of ML in the chemical education literature.

One such example described an introductory ML exercise for undergraduate chemists, using Python notebooks to explore the physicochemical properties of a pre-existing data set of 6,497 wines.[12] The exercise assumes no prior coding knowledge and covers basic Python syntax and its usage prior to moving on to the ML activities. A supervised ML method known as *k*-nearest neighbors (*k*-NN) is implemented in the notebooks in order to classify wines as either red or white. Overall, the exercise is highly scaffolded to ensure that students with a mixture of prior coding experiences can complete all of the tasks. Another introductory

Received: July 15, 2022
Revised: January 11, 2023
Published: February 13, 2023



https://doi.org/10.1021/acs.jchemed.2c00682

ML task involves the use of binary classification algorithms alongside ML techniques to classify infrared spectra as "carbonyl" or "non-carbonyl".[13]

A number of authors have approached the subject of teaching machine learning via artificial neural networks.[14,15] However, the data sets that are required for training these networks need to be very reproducible to avoid the ANNs finding patterns in the data that do not exist. We therefore avoided the use of ANNs in this study, given the complex nature of the substrates that we planned to sample.

Prior to employing ML methods, it is usually necessary to carry out dimensionality reduction to reduce the large amount of data down to a data set of manageable size for computation. The most widely used dimensionality reduction approach, and the one used in our work, is Principal Component Analysis (PCA).[16−18] A detailed demonstration of PCA designed for chemistry undergraduates is demonstrated by Sidou and Borges, illustrating its use both as a data visualization technique and as a method of data reduction prior to ML classification.[19] Using the R programming language, the exercise focuses primarily on the use of PCA to extract correlations in the properties of chemical elements. Again, the exercise is structured to guide students through set examples of pre-existing data, with predetermined outcomes. Other such examples of PCA in the chemical education literature include the classification of vegetable oils by FTIR spectroscopy,[20] and the identification of a range of edible oils by NMR spectroscopy.[21]

In the following, we describe an approach suitable for use as a mini-research project in which students choose a set of samples (in our examples these were a variety of fruits, whiskies, and teas), record spectra for a large number of these samples, use the data to train an ML classifier, and then evaluate the ability of the resulting model to classify "unknown" samples based on their spectra. Our examples employ atmospheric-solids analysis probe mass spectrometry (ASAP-MS) and Fourier transform infrared (FTIR) spectroscopy to characterize the samples, but the approach would work equally well with other types of spectroscopy. We have previously applied a similar approach to the classification of spectra recorded using Raman[22,23] and vis−NIR reflectance[24] spectroscopies, for example. We have also demonstrated the use of ASAP-MS as an approachable technique in the undergraduate chemistry laboratory.[25] We also suggest that a cut-down version of this project may be suitable for the introduction of PCA/ML techniques in chemistry, into the undergraduate curriculum (see Supporting Information for practical details).

For the purposes of the present work, a "data point", sometimes known as a "predictor variable", is an FTIR or mass spectrum of one of the samples included in the study, and the goal is to use a classifier to assign each spectrum to the correct group. Taking our first student project as an example, given a large number of spectra recorded for oranges, lemons, limes, and tomatoes, we would like the ML algorithm to be able to assign a given spectrum to one of these four groups, e.g. to correctly recognize the spectrum recorded for a lemon as belong to the "lemon" group. Classifiers can be split into two different types, namely unsupervised and supervised methods. An unsupervised classifier looks for patterns in the data themselves to find the groups. In the present application such a classifier would be given only the set of spectra to work with, and would not be provided with any additional information about the samples. Given the task of sorting a large number of spectra into four groups, the classifier would aim to assign spectra to groups such that the variation between spectra within groups was minimized and between groups was maximized. The *k*-means clustering algorithm is an example of an unsupervised classifier.[26,27]

While unsupervised classifiers can be extremely helpful in finding patterns in the data, often we are able to train an ML classifier using a set of training data for which the identities of the individual data points are known. In this case we can use a supervised classifier, such as *k* nearest neighbors (*k*-NN), support vector machines (SVM), or linear discriminant analysis (LDA). The labels for each data point (in our case the identities of the corresponding samples) are known as "response variables", and correspond to the variable we would like to determine when presenting the trained algorithm with a new data point. To train the model, a set of predictor variables (mass or FTIR spectra) and corresponding response variables ("lemon", "lime", "orange", or "tomato"), collectively known as the "training data set", are presented to the algorithm, which then determines which features of the predictor variables correlate with the known response variables, i.e. which features of the spectra correlate with a given spectrum being that of an orange, lemon, lime, or tomato. The resulting model can then be tested by seeing how accurately a set of previously unseen "test data" is classified by the trained algorithm.[28,29] A key feature of ML is that more data can be added as it becomes available in order to improve the model. The more good quality data the model has to train on, the more robust it should become.[29]

The data analysis for our student projects was performed within the MATLAB[30] programming environment using the Statistics and Machine Learning toolbox,[31] which contains built-in functions for the various ML algorithms employed. Similar function libraries are available for other programming platforms, including Python, C++, and R, and the projects could be adapted for these platforms. Depending on how much coding the students are expected to do themselves, how many samples are studied, how much experimental characterization is performed, and what types of questions are posed, the approach described in the following can be developed into anything from a two-day practical exercise or workshop to an extended research project. Of course, the amount of prior experience that students have with MatLab will affect the length of time that students need to complete the tasks set; this is very much up to the individual institution to plan, but some suggestions of timings are given in the Supporting Information.

■ LEARNING OBJECTIVES

The exercise outlined here can be tailored to suit the needs of individual courses. The learning objectives of the exercise are likely to vary depending on the course, but possible objectives might include:

- To safely and appropriately prepare and run samples on analytical instruments.
- To introduce a programming language or build upon existing programming skills.
- To visualize data using principal component analysis.
- To classify data and make predictions using machine learning methods.
- To design a short project by identifying interesting yet achievable research questions.

■ METHODS

Mass spectra were acquired using an Advion expression^L Compact Mass Spectrometer—a single quadrupole instrument

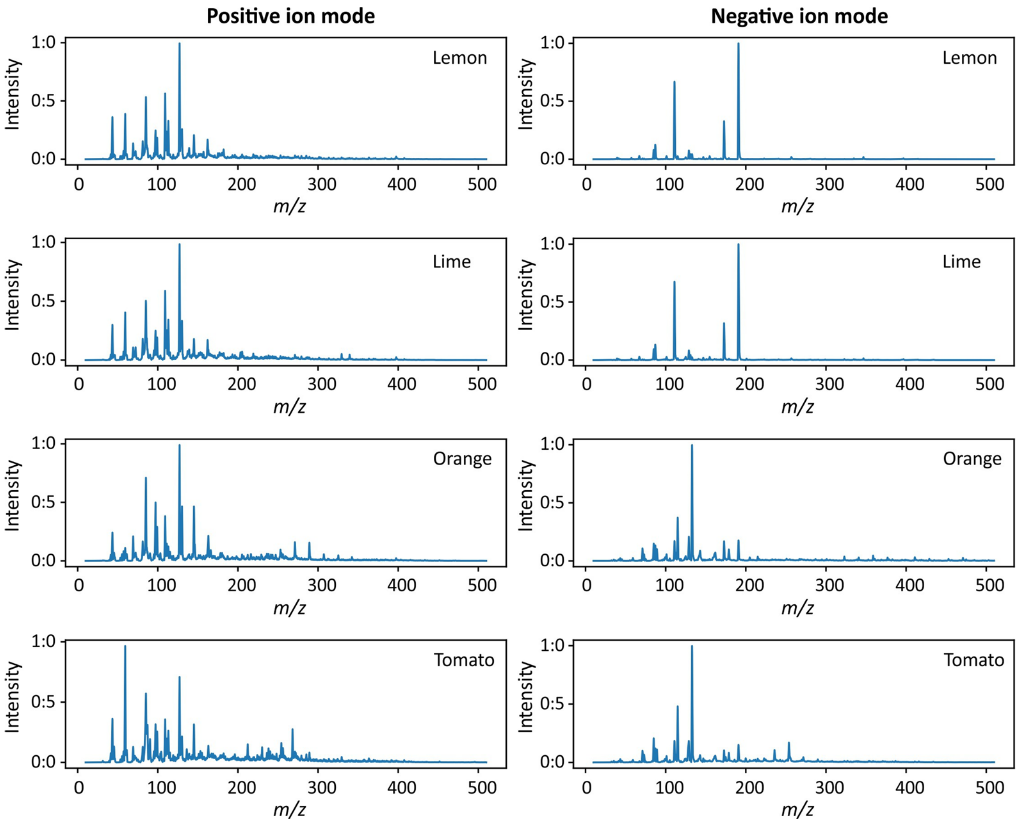


Figure 1. Mean positive ion and negative ion mass spectra recorded for each of the fruit species, plotted over the m/z range from 10 to 500.

with a mass range of 10−2000 m/z and resolution of 0.05 m/z over the entire mass range. The spectrometer is equipped with an Atmospheric Pressure Chemical Ionization (APCI) ion source, which was used in all experiments in combination with an atmospheric solids analysis probe (ASAP). The tip of the ASAP comprises a disposable sealed glass capillary. To record a mass spectrum, the capillary tip is placed in contact with the sample and the probe is then inserted into the ion source of the mass spectrometer. Spectra are recorded at 0.5 s intervals, and the spectra are typically averaged over the first 20 s after insertion of the probe.

FTIR spectra were recorded on a Shimadzu IRSpirit FTIR Spectrometer fitted with a QATR-S attenuated total reflectance accessory. Spectra were recorded between 400 and 4000 cm^{-1}, with a resolution of 1 cm^{-1}, and averaged over 10 scans.

The number of spectra that needs to be recorded for each sample for a successful outcome was left up to the student to investigate. It was found that this varied between 20 and 60 per sample to get sufficient resolution between the types of fruit/whisky/tea (see below). Students were able to obtain a data set from an individual sample (e.g. a type of fruit or tea) in 0.5−1.0 h, depending on the instrument used.

Fruit Study

Mass spectra were recorded for four different fruits: easy-peeler orange, lemon, lime, and salad tomato. To obtain a mass spectrum the fruit was cut in half, the closed end of the probe's glass capillary was dipped a few millimeters deep into the flesh of the fruit, and the probe was inserted into the mass spectrometer for analysis. Significant residue tends to build up over multiple acquisitions, so to preclude this the glass capillary tube was cleaned after each acquisition with a sponge soaked in detergent and water, before being rinsed in deionized water. The glass capillary tube on the probe was changed between each experiment with a new fruit and/or ionization mode, which was approximately every 10 acquisitions. Preliminary experiments showed that the "high-temperature, low-fragmentation" setting for the ion source yielded the best spectra, and this ionization mode was used for all subsequent measurements. Sixty mass spectra were recorded in both positive and negative ion mode for each fruit species, yielding a data set of 240 positive and negative ion mass spectra in total, examples of which are shown in Figure 1.

FTIR spectra were recorded for the same fruit types by placing a sample of skin/peel into the ATR attachment of the FTIR spectrometer. Sixty spectra were recorded for each fruit species, yielding a data set of 240 FTIR spectra in total. A different part of the skin/peel was sampled for each spectrum, the ATR prism was cleaned with isopropyl alcohol between measurements, and a new background spectrum was taken every 10 measurements.

Whisky Study

The mass spectra of four different whiskies were recorded: an American Bourbon, a Scottish blended, a Scottish single malt, and an English single malt. The whiskies were decanted into smaller sample bottles, which were stored in the dark until sampling to mitigate any potential chemical changes caused by prolonged light exposure over time. During preliminary testing it was determined that negative ion acquisitions displayed poor signal-to-noise ratio with this sample type, and thus the decision was taken to collect positive ion spectra only. Twenty positive-ion mass spectra were acquired for each of the four whiskies, for each of the nine possible preset ion source combinations of temperature (low, medium, high) and fragmentation conditions (low, medium, high).

Tea Study

Mass spectra were recorded for five different types of tea, namely Chai, Breakfast, Assam, Ceylon, and decaffeinated Earl Grey. To prepare samples for measurement, the 2.5 g contents of a teabag was brewed in 100 mL of water at 100 °C for 3 min. The ion source was run in "high-temperature, low-fragmentation" mode. Ten mass spectra were recorded for each brew, with the process repeated twice for each tea to give a total of 20 spectra per tea. The capillary was cooled and cleaned with methanol in between each acquisition.

Data Preprocessing and Analysis

A full data set for the Tea study is provided in the Supporting Information, together with the Python code used to analyze this. These data are supplied as .txt files, but by changing the file extension to.csv, the data may be loaded into a spreadsheet for manipulation. Data for the other studies is available from the authors on request from *bona fide* instructors.

Initial data preprocessing of the mass spectral data was carried out within the Advion Data Express software environment. For each measurement, the mass spectrum was averaged over the first 20 s after the ASAP was inserted into the mass spectrometer, and the resulting spectrum was saved as a comma-separated variables (.csv) file containing a list of m/z values together with the associated ion counts. For the tea study, a MATLAB program was written to perform this step automatically, generating individual data files for each measurement given the raw acquisition data file as input. For each study, the spectra were read into MATLAB, stored in a data store, and each spectrum was normalized either to a maximum peak height of unity (fruit study mass spectra) or for each spectrum, MS or FTIR, the total area between the spectral line and the baseline was calculated numerically by the code written, and this was then normalized to unity, to ensure that any variations in spectral intensity due to amount of substance were accounted for (whisky and tea study mass spectra; fruit study FTIR spectra). The latter is preferable, as normalization to the largest peak risks distortion of the data if there is a large, variable impurity peak. In these cases, either normalization protocol gave similar results in the hands of our students. The FTIR spectra were also interpolated to give data for the 400 to 4000 cm^{-1} range in regular 1 cm^{-1} intervals.

The data analysis for this project was performed within the MATLAB coding environment. Much of the analysis employed the Statistics and Machine Learning toolbox, which contains a variety of ready-to-use machine learning algorithms. Initially, PCA was carried out and the first two and three principal components plotted as a method of data visualization. This data reduction technique also prepared the data set for subsequent ML classification. The PCA and ML methods used are described in more detail in the Supporting Information.

■ EXERCISE OUTLINE

The projects described above were carried out by fourth year undergraduates during the first few weeks of the final year of an M.Chem. integrated master's degree. These students were about to embark upon a more substantial research project in an established research group, under the supervision of a Principal Investigator, and these preliminary projects, carried out within the research group, were designed to introduce students to the area of research that became the subject of their master's thesis. Since all students in our institution are placed individually in research groups, the projects described in this paper have, so far, only been undertaken by a small number. However, we suggest that the flexibility of approach (see below) would allow for the earlier introduction of this area in the undergraduate course—in our course this would be in the third year, as part of a suite of optional laboratories. It is likely, especially if mass spectrometry was the chosen technique, to continue to be limited to a small number of undergraduates, due to instrument availability. The projects described here served to teach the students how to use the instruments and how to analyze their data using a variety of ML classifiers. The students were encouraged to design their own short projects, selecting their own samples and identifying the key questions they would attempt to answer. This project-based approach can encourage students to take responsibility for their work and helps them to develop skills critical for any future career.[32] Depending on the analytical technique employed and samples chosen, the list of possible questions is almost endless, but potential areas to explore include:

- Comparison between different spectroscopic techniques.
- Effects of sample preparation methods and/or sample degradation.
- Effects of data set size.
- Variations between sample brands, batches, etc., and how these affect the analysis.
- Choice of machine learning methods for different problems/questions.
- Sensitivity of classifiers to small variations between samples.

Given the relatively large amount of data that needs to be acquired to implement ML classifiers, a short data acquisition time for each measurement is essential. ASAP-MS and FTIR are very well suited for such applications, particularly given the ease of use and applicability to a broad range of sample types. As noted previously, our approach could equally well be used with data acquired using other spectroscopic or spectrometric techniques. Students could be given a selection of instruments to use and prompted to consider the practicalities of data collection as part of the project.

The amount of coding the students are expected to do themselves can be varied depending on the anticipated length of the project or practical and the previous coding experience of the students. For the studies described above, the students had no prior programming experience and completed several MATLAB Self-Paced courses before beginning the project, namely MATLAB Onramp (~2 h),[33] MATLAB Fundamentals (~21 h),[34] and Machine Learning with MATLAB (~14 h).[35] Exact implementation of the exercise could be varied depending on the coding experience required, or desired, by the student. For example, a two-day exercise with minimal coding knowledge could be carried out with data collection and preprocessing on day one, followed by PCA and ML analysis using prefilled coding notebooks on day two. In this implementation students should focus on using one or two ML algorithms, such as *k*-means clustering (unsupervised) and *k*-NN (supervised), and understanding them well (see Supporting Information for details on the ML algorithms used). Alternatively, a longer (1+ week) project could be imagined in which students choose the samples they would like to analyze, for example different food stuffs or different batches of the same type of sample. They would then need to implement their own code and optimize their data collection for the best results. Students could be presented with a range of unsupervised and supervised ML

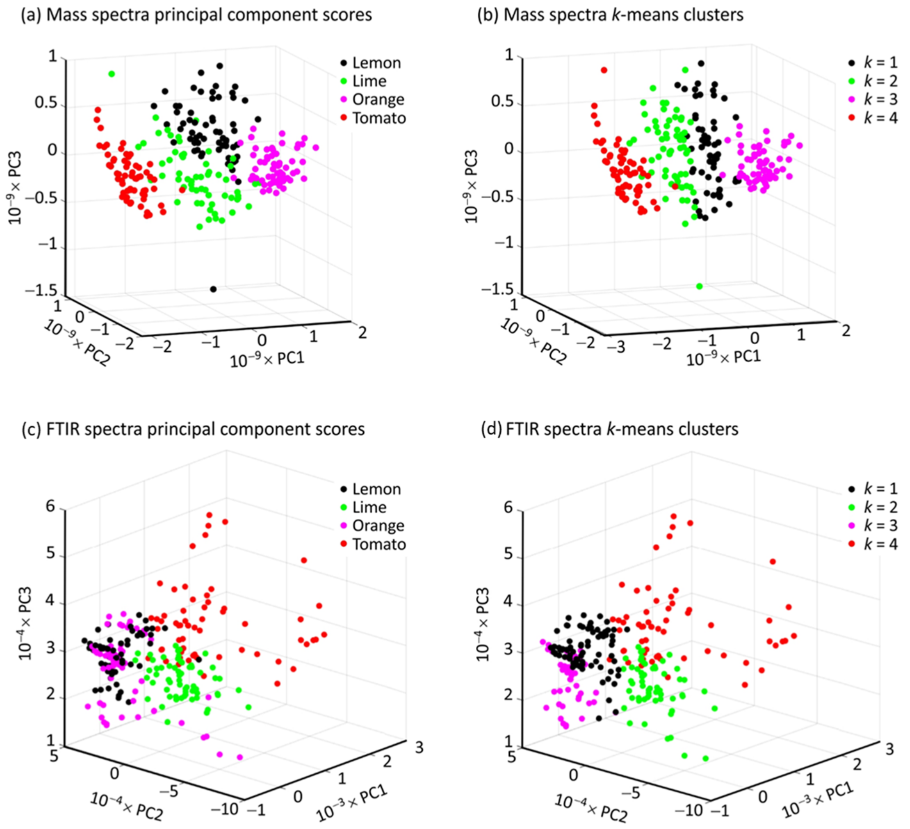


Figure 2. (a) 3D plot of the first three principal component scores for each data point (mass spectrum) recorded in the fruit study. Data points are colored according to the identity of the sample. (b) Result of a *k*-means clustering analysis in which the mass spectral data was grouped into four clusters. Data points are colored according to the group to which they are assigned. (c) 3D plot of the first three principal component scores for each FTIR data point recorded in the fruit study. (d) Result of a *k*-means clustering analysis in which the FTIR data set was grouped into four clusters.

methods and asked to determine which are most suitable for their data.

As mentioned, the data preprocessing and analysis can be carried out in Python, C++, R, or similar in place of MATLAB, with all platforms offering dedicated libraries for PCA and ML methods. As such, the exercise can be tailored to follow on from any pre-existing coding experience as appropriate. Since the projects described here have been completed, introductory exercises in Python and MATLAB have been implemented within the second year of our own undergraduate course. Students carrying out such projects in future will already be familiar with the syntax of these two languages, and will be able to make an immediate start on the required data processing and ML analysis. This exemplifies our spiral curriculum approach to undergraduate teaching, with key skills, in this case writing and using code, being revisited at later stages of the course and with a higher degree of complexity.[36]

■ EXAMPLE RESULTS

The fruit study was the first student project performed, with the goal of exploring how well a variety of unsupervised and supervised ML algorithms were able to classify spectra recorded for the student's chosen fruit samples. Visual inspection of the mass spectra reveals that the lemon and lime spectra are very similar in appearance in both positive and negative ion mode, while the spectra recorded for oranges and tomatoes are markedly different. Figure 2(a) shows a plot of the first three principal component scores for each mass spectrum, with the data points for each type of fruit represented in different colors. Not unexpectedly given the visual appearance of the spectra, we see some overlap between the data points for lemons and limes in principal component space, while the data points for oranges and tomatoes are reasonably well separated both from each other and from the lemon and lime data points.

The data were first analyzed using an unsupervised *k*-means clustering algorithm, which was instructed to group the data into four clusters. The algorithm uses one of several available distance metrics to measure the distances between data points, and uses these to cluster nearby data points into groups. The first 20 principal component scores were used to calculate the distances. The results of this analysis are plotted in Figure 2(b) in the same principal component space as above, but this time the data points are colored according to the group they were assigned by the clustering algorithm rather than by their true identity. Comparison of Figure 2(a) and (b) reveals that the *k*-means clustering algorithm is able to assign oranges and tomatoes to their own groups with a high degree of accuracy: nearly all tomato data points are assigned correctly to group 4, and nearly all orange spectra are assigned to group 3. However, lemon and lime spectra are often confused by the algorithm, with the result that groups 1 and 2 each contain a mixture of lemon and lime spectra.

Similar PCA and *k*-means clustering analysis of the FTIR data was carried out and complements the results obtained for the mass spectra. Figure 2(c) shows the first three principal component scores for each normalized FTIR spectrum, color-coded according to sample type. Tomatoes, limes, and oranges/lemons are seen to occupy separate regions of principal

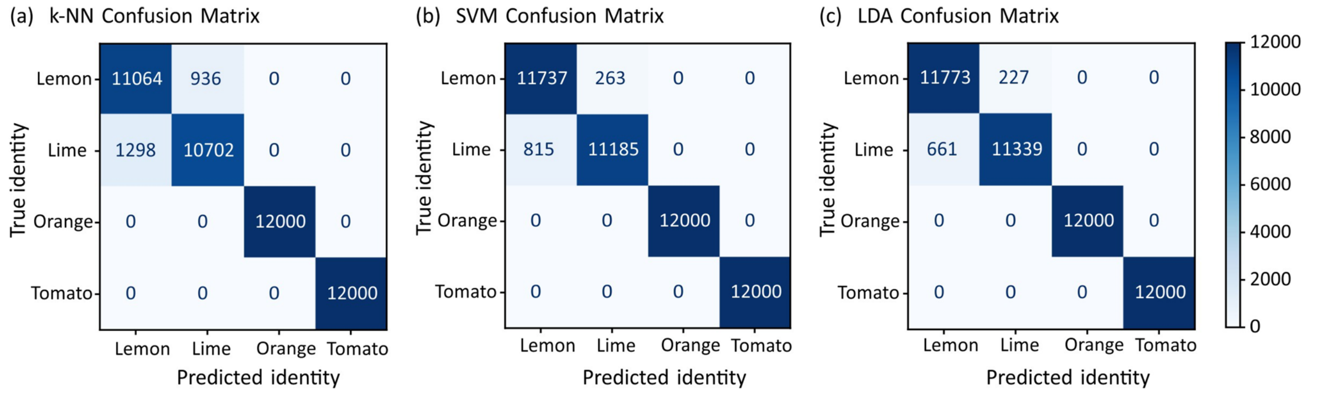


Figure 3. Confusion matrices showing the results of analyzing the fruit mass spectra with three supervised machine learning classification algorithms: (a) k-NN, (b) SVM, (c) LDA.

component space, but there is significant overlap between the principal component scores of oranges and lemons. Meanwhile, Figure 2(d) shows that *k*-means clustering analysis assigns tomato spectra exclusively to group 3, and limes mostly to group 2, with erroneous assignment in ∼17% of cases to group 4. Lemons and oranges are often confused, with both assigned mostly either to groups 3 or 4, and occasionally to group 2. Combining this analysis with that for the mass spectra, we can see that while some fruits are misidentified for both techniques, different pairs of fruits are confused in each case, and so a more accurate assignment could be elucidated through a combination of the results.

The mass spectra were also analyzed with a number of different *supervised* ML classifiers, namely *k*-NN, SVM, and LDA. Each algorithm was run 1000 times with a different 80:20 split of the spectra into training and test data on each run to minimize any biases arising from the choice of specific training data points. The results for each classifier, summed over all 1000 runs, are shown in the form of a confusion matrix in Figure 3. The sensitivities and specificities achieved by each classifier are shown in Table 1. With 20% of the 60 spectra (i.e., 12 spectra) for each fruit being used as test data for each run, perfect performance would correspond to 12,000 data points being assigned to each fruit and appearing along the diagonal of the confusion matrix, with all off diagonal elements (corresponding to mis-assignments) being zero. We see that in reality, while the proportion of correct classifications is very high, there are some nonzero off-diagonal elements, the most significant corresponding to misclassifications of lemon spectra as belonging to limes, and vice versa.

Table 1. Percentage Sensitivity and Specificity Achieved by the Supervised ML Classifiers for Each Species of Fruit

	SVM		*LDA*		*k-NN*	
	Sens. (%)	Spec. (%)	*Sens. (%)*	*Spec. (%)*	*Sens. (%)*	*Spec. (%)*
Lemon	96.4	99.8	98.0	98.5	96.1	95.9
Lime	99.6	98.8	97.6	99.2	91.3	98.7
Orange	99.9	100	100	100	100	100
Tomato	100	100	100	100	100	100

Unsurprisingly, the performance of all of the supervised ML algorithms is much better than that of the unsupervised *k*-means clustering algorithm. Even the simplest supervised algorithm, *k*-NN, is able to assign the (principal component scores of the) spectra to the correct fruit well over 90% of the time, with perfect assignment of oranges and tomatoes to the correct groups. The best performing algorithm, LDA, assigns oranges and tomatoes correctly 100% of the time and lemons and limes correctly around 98% of the time. The success of this first project was very encouraging, and opened the way for students to pose a variety of new questions in subsequent projects.

While slightly different questions were posed, similarly encouraging results were obtained for the whisky and tea studies, with further information and results from these studies included in the Supporting Information.

■ STUDENT FEEDBACK

Students were asked to complete a questionnaire, designed to monitor changes in skills and perceptions as a result of working through this project. A copy of the questionnaire is provided (Supporting Information), and Figure 4 shows how the students' responses were affected by the experience they had been through

In addition, students were asked to rate two statements: *I enjoyed doing the project on ML and PCA in Chemistry*, and *I would recommend this project to other students at my university*. These statements scored 4.75 and 5.00 respectively, on the same scale.

The results show that the project was very popular with the students, but Q7 shows that the students who undertook the work were already interested in this area. The opportunity to respond in free text produced comments such as "I think integrating this project into Chemistry undergraduate degrees would advance valuable computational skills in students, better equipping them to thrive in a career in Chemistry—and countless other fields—in the modern world." and "It is very easy for students to see how the project is applicable to current scientific research. Much of the code written for such a mini-project can be easily adapted for research settings."

■ SUMMARY

Overall, these short ML projects served as useful introductions to PCA and ML analysis, but also to self-directed learning in preparation for students' final year projects. After completion of the short projects, all of the students were able to use their newfound skills successfully in their main year-long research projects. These longer projects employed the same methods but applied them to the analysis of human tissue and plasma samples in the context of exploring new approaches to clinical diagnostics and patient risk stratification. The students found

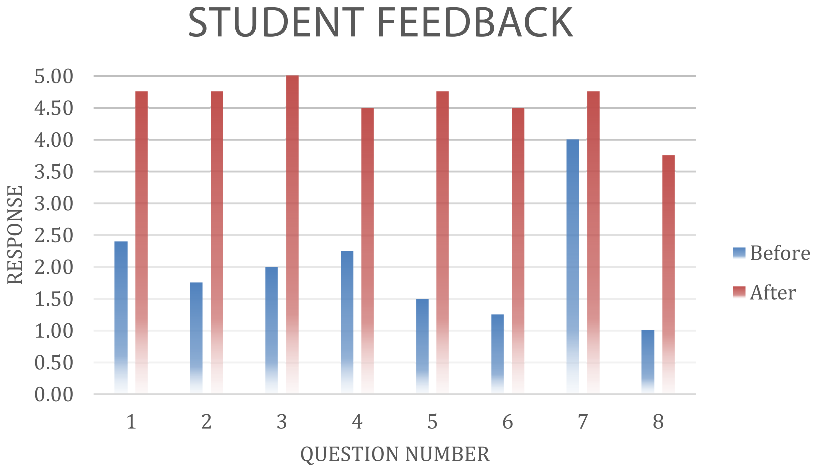


Figure 4. Analysis of students' responses to postproject questionnaire. Responses were collated and scores averaged ($n = 4$). Responses to questions were on a 5-point scale (1 is "I strongly disagree" through 5, "I strongly agree"). Q1. I understood the term "Machine Learning (ML)". Q2. I understood the term "Principal Component Analysis (PCA)". Q3. I was aware of the application on ML and PCA to areas of Chemistry. Q4. I would rate myself as confident in writing computer code using at least one language. Q5. I was able to read MatLab code. Q6. I was able to write MatLab code. Q7. I was interested in the application of computers to the processing of large data sets. Q8. I would be confident in applying my MatLab skills outside of Chemistry.

the MATLAB courses to be very useful in introducing them to the basics of coding, such as reading in data files, creating and editing variables, visualizing data, for loops, and the implementation of the ML algorithms. However, with a sufficiently informative lab manual they felt that the project would have been manageable even if they had only completed the two-hour MATLAB Onramp course. All students were very proficient at coding in MATLAB by the end of their projects, and felt that their newly acquired coding skills would serve them well during future careers either in scientific research or in other professions such as consulting or finance. They also greatly enjoyed the opportunity to gain insight into the vast capabilities of modern computational methods.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.2c00682.

Notes for Instructors (PDF, DOCX)

MatLab code and sample data for Tea Study (ZIP)

Tea_analysis_simple_code_and_output (PDF)

Evaulation survey (PDF, DOCX)

AUTHOR INFORMATION

Corresponding Author

Malcolm I. Stewart − *Department of Chemistry, University of Oxford, Chemistry Teaching Laboratory, Oxford OX1 3PS, United Kingdom;* orcid.org/0000-0002-5724-9160; Email: malcolm.stewart@chem.ox.ac.uk

Authors

Alanah Grant St James − *Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Oxford OX1 3TA, United Kingdom*

Luke Hand − *Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Oxford OX1 3TA, United Kingdom*

Thomas Mills − *Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Oxford OX1 3TA, United Kingdom*

Liwen Song − *Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Oxford OX1 3TA, United Kingdom*

Annabel S. J. Brunt − *Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Oxford OX1 3TA, United Kingdom;* orcid.org/0000-0002-8815-6768

Patrick E. Bergstrom Mann − *Department of Chemistry, University of Oxford, Chemistry Teaching Laboratory, Oxford OX1 3PS, United Kingdom;* orcid.org/0000-0003-1428-1440

Andrew F. Worrall − *Department of Chemistry, University of Oxford, Chemistry Teaching Laboratory, Oxford OX1 3PS, United Kingdom;* orcid.org/0000-0002-2875-6905

Claire Vallance − *Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Oxford OX1 3TA, United Kingdom;* orcid.org/0000-0003-3880-8614

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jchemed.2c00682

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank laboratory technician Thomas Birkett for his assistance in the acquisition of FTIR spectra, and to Mark Allen and Mark Baumert from Advion, Inc. for facilitating access to an ASAP mass spectrometer during parts of the study. One of us (C.V.) gratefully acknowledges financial support from the John Fell Fund.

REFERENCES

(1) Stein, S. Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification. *Anal. Chem.* **2012**, *84* (17), 7274−7282.

(2) Spectra and Spectral Data. University of Texas Libraries. https://guides.lib.utexas.edu/chemistry/spectra (accessed 2022-03-08).

(3) Mass Spectrometry Data Center. National Institute of Standards and Technology. https://chemdata.nist.gov/ (accessed 2022-03-08).

(4) Mjolsness, E.; DeCoste, D. Machine Learning for Science: State of the Art and Future Prospects. *Science* **2001**, *293* (5537), 2051−2055.

(5) Li, R.; Herreros, J. M.; Tsolakis, A.; Yang, W. Machine Learning-Quantitative Structure Property Relationship (ML-QSPR) Method for Fuel Physicochemical Properties Prediction of Multiple Fuel Types. *Fuel* **2021**, *304* (July), 121437.

(6) Deng, Q.; Lin, B. Automated Machine Learning Structure-Composition-Property Relationships of Perovskite Materials for Energy Conversion and Storage. *Energy Mater.* **2022**, *1*, 100006.

(7) Mao, J.; Akhtar, J.; Zhang, X.; Sun, L.; Guan, S.; Li, X.; Chen, G.; Liu, J.; Jeon, H. N.; Kim, M. S.; No, K. T.; Wang, G. Comprehensive Strategies of Machine-Learning-Based Quantitative Structure-Activity Relationship Models. *iScience* **2021**, *24* (9), 103052.

(8) Ceriotti, M.; Clementi, C.; Anatole von Lilienfeld, O. Machine Learning Meets Chemical Physics. *J. Chem. Phys.* **2021**, *154* (16), 160401.

(9) Fabrizio, A.; Grisafi, A.; Meyer, B.; Ceriotti, M.; Corminboeuf, C. Electron Density Learning of Non-Covalent Systems. *Chem. Sci.* **2019**, *10* (41), 9424−9432.

(10) Howarth, A.; Ermanis, K.; Goodman, J. M. DP4-AI Automated NMR Data Analysis: Straight from Spectrometer to Structure. *Chem. Sci.* **2020**, *11* (17), 4351−4359.

(11) Gajewska, E. P.; Szymkuć, S.; Dittwald, P.; Startek, M.; Popik, O.; Mlynarski, J.; Grzybowski, B. A. Algorithmic Discovery of Tactical Combinations for Advanced Organic Syntheses. *Chem.* **2020**, *6* (1), 280−293.

(12) Lafuente, D.; Cohen, B.; Fiorini, G.; García, A. A.; Bringas, M.; Morzan, E.; Onna, D. A Gentle Introduction to Machine Learning for Chemists: An Undergraduate Workshop Using Python Notebooks for Visualization, Data Processing, Analysis, and Modeling. *J. Chem. Educ.* **2021**, *98* (9), 2892−2898.

(13) Thrall, E. S.; Lee, S. E.; Schrier, J.; Zhao, Y. Machine Learning for Functional Group Identification in Vibrational Spectroscopy: A Pedagogical Lab for Undergraduate Chemistry Students. *J. Chem. Educ.* **2021**, *98* (10), 3269−3276.

(14) Joss, L.; Müller, E. A. Machine Learning for Fluid Property Correlations: Classroom Examples with MATLAB. *J. Chem. Educ.* **2019**, *96* (4), 697−703.

(15) Revignas, D.; Amendola, V. Artificial Neural Networks Applied to Colorimetric Nanosensors: An Undergraduate Experience Tailorable from Gold Nanoparticles Synthesis to Optical Spectroscopy and Machine Learning. *J. Chem. Educ.* **2022**, *99* (5), 2112−2120.

(16) Hotelling, H. Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.* **1933**, *24* (6), 417−441.

(17) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *London Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2* (11), 559−572.

(18) Jolliffe, I. T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **2016**, *374* (2065), 20150202.

(19) Sidou, L. F.; Borges, E. M. Teaching Principal Component Analysis Using a Free and Open Source Software Program and Exercises Applying PCA to Real-World Examples. *J. Chem. Educ.* **2020**, *97* (6), 1666−1676.

(20) Rusak, D. A.; Brown, L. M.; Martin, S. D. Classification of Vegetable Oils by Principal Component Analysis of FTIR Spectra. *J. Chem. Educ.* **2003**, *80* (5), 541−543.

(21) Anderson, S. L.; Rovnyak, D.; Strein, T. G. Identification of Edible Oils by Principal Component Analysis of 1H NMR Spectra. *J. Chem. Educ.* **2017**, *94* (9), 1377−1382.

(22) Livermore, L. J.; Isabelle, M.; Bell, I. M.; Scott, C.; Walsby-Tickle, J.; Gannon, J.; Plaha, P.; Vallance, C.; Ansorge, O. Rapid Intraoperative Molecular Genetic Classification of Gliomas Using Raman Spectroscopy. *Neuro-Oncol. Adv.* **2019**, *1* (1), vdz008.

(23) Livermore, L. J.; Isabelle, M.; Bell, I. M.; Edgar, O.; Voets, N. L.; Stacey, R.; Ansorge, O.; Vallance, C.; Plaha, P. Raman Spectroscopy to Differentiate between Fresh Tissue Samples of Glioma and Normal Brain: A Comparison with 5-ALA-Induced Fluorescence-Guided Surgery. *J. Neurosurg.* **2020**, *135* (2), 469−479.

(24) De Maria, G. L.; Lee, R.; Alkhalil, M.; Borlotti, A.; Kotronias, R.; Langrish, J.; Lucking, A.; Dawkins, S.; Choudhury, R. P.; Kharbanda, R.; Banning, A. P.; Vallance, C.; Channon, K. M. Reflectance Spectral Analysis for Novel Characterization and Clinical Assessment of Aspirated Coronary Thrombi in Patients with ST Elevation Myocardial Infarction. *Physiol. Meas.* **2020**, *41* (4), 045001.

(25) Moloney, J. G.; Campbell, C. D.; Worrall, A. F.; Stewart, M. I. Hands-on Inquiry-Based Qualitative Identification of Metals in Coins Utilizing Atmospheric Pressure Chemical Ionization Mass Spectrometry. *J. Chem. Educ.* **2022**, *99*, 2697.

(26) Forgy, E. W. Cluster Analysis of Multivariate Data : Efficiency versus Interpretability of Classifications. *Biometrics* **1965**, *21*, 768−769.

(27) Lloyd, S. P. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28* (2), 129−137.

(28) Love, B. C. Comparing Supervised and Unsupervised Category Learning. *Psychon. Bull. Rev.* **2002**, *9* (4), 829−835.

(29) Howard, W. R. Pattern Recognition and Machine Learning. *Kybernetes* **2007**, *36* (2), 275.

(30) MATLAB. *9.6.0.1150989 (R2019a)*; The MathWorks Inc.: Natick, MA, 2019.

(31) MATLAB. *Statistical and Machine Learning Toolbox*; The MathWorks Inc.: Natick, MA, 2019.

(32) Burnham, J. A. J. Skills for Success: Student-Focused, Chemistry-Based, Skills-Developing, Open-Ended Project Work. *J. Chem. Educ.* **2020**, *97* (2), 344−350.

(33) MATLAB. MATLAB Onramp. https://matlabacademy.mathworks.com/details/matlab-onramp/gettingstarted?s_tid=course_mlor_start1 (accessed 2022-04-14).

(34) MATLAB. MATLAB Fundamentals. https://matlabacademy.mathworks.com/details/matlab-fundamentals/mlbe (accessed 2022-04-14).

(35) MATLAB. Machine Learning with MATLAB. https://matlabacademy.mathworks.com/details/machine-learning-onramp/machinelearning?s_tid=course_mlor_start (accessed 2022-04-14).

(36) Campbell, C. D.; Midson, M. O.; Bergstrom Mann, P. E.; Cahill, S. T.; Green, N. J. B.; Harris, M. T.; Hibble, S. J.; O'Sullivan, S. K. E.; To, T.; Rowlands, L. J.; Smallwood, Z. M.; Vallance, C.; Worrall, A. F.; Stewart, M. I. Developing a Skills-Based Practical Chemistry Programme: An Integrated, Spiral Curriculum Approach. *Chem. Teach. Int.* **2022**, *4* (3), 243.