Detecting forged alcohol non-invasively through vibrational spectroscopy and machine learning

James Large[1], E. Kate Kemsley[2], Nikolaus Wellner[2], Ian Goodall[3], and Anthony Bagnall[1](✉)

[1] School of Computing Sciences, University of East Anglia, UK
{james.large, anthony.bagnall}@uea.ac.uk
[2] Quadram Institute, Norwich Research Park, UK
{kate.kemsley, nikolaus.wellner}@quadram.ac.uk
[3] Scotch Whisky Research Institute, Research Avenue North, Edinburgh, UK
ian.goodall@swri.co.uk

Abstract. Alcoholic spirits are a common target for counterfeiting and adulteration, with potential costs to public health, the taxpayer and brand integrity. Current methods to authenticate spirits include examinations of superficial appearance and consistency, or require the tester to open the bottle and remove a sample. The former is inexact, while the latter is not suitable for widespread screening or for high-value spirits, which lose value once opened. We study whether non-invasive near infrared spectroscopy, in combination with traditional and time series classification methods, can correctly classify the alcohol content (a key factor in determining authenticity) of synthesised spirits sealed in real bottles. Such an experimental setup could allow for a portable, cheap to operate, and fast authentication device. We find that ethanol content can be classified with high accuracy, however methanol content proved difficult with the algorithms evaluated.

Keywords: classification, spectroscopy, non-invasive, authentication

1 Introduction

Up to 25% of licensed premises in some parts of the UK have been found to have counterfeit alcohol for sale[4]. Brown-Forman, the company that makes Jack Daniels, estimates that around 30% of all alcohol in China is fake[5].

Counterfeit alcohol poses a health risk to the consumer, as illegally produced spirits may contain harmful contaminants such as methanol, an economic risk due to the avoidance of taxes, and a risk to brand integrity in cases where the fakes are being sold as named brands.

Forgeries can sometimes be detected through external appearance such as inconsistent labelling or bottling, but currently there is no way to conclusively tell

[4] http://www.bbc.co.uk/news/uk-12456360

[5] https://www.theguardian.com/sustainable-business/2015/sep/16/china-fake-alcohol-industry-counterfeit-bathtub-booze-whisky

whether spirits are forged without opening the bottle to analyse a sample directly. Breaking the seal and taking samples from a bottle is effectively a destructive process, because even if authenticity is confirmed the bottle cannot later be sold on store shelves or at auction, and collectors' whisky will be greatly devalued. Also, testing of samples can be an expensive and time consuming process that is not suitable for mass screening. No matter what process is used it will require one or more of: transport of the sample to a centralised lab; expert knowledge and handling; consumable materials used in the analysis; and time for methods such as chromatography. It is therefore desirable to develop a system that can non-invasively determine authenticity of a suspect bottle in a cheap, portable, simple and fast manner.

Vibrational spectroscopy in combination with modern machine learning techniques provides a promising potential solution to these problems. Ever improving computing power, spectroscopy equipment and algorithms mean that on-site classification using cost effective equipment is becoming evermore feasible.

The alcohol concentration of genuine spirits in the UK is tightly controlled. For example, Scotch whisky must contain the level stated on the bottle to within 0.3% (v/v). Forgeries typically do not have this level of quality control, with the alcohol content often being lower than reported. Alternatively, methanol and many higher alcohols and heavy metals have regulations prohibiting their presence in spirits to within certain maximal concentrations to ensure safe consumption, and are also tightly controlled. Both ethanol level and methanol level can in principle be characterised by vibrational spectroscopy, and ultimately determined with chemometric and machine learning techniques.

We wish to evaluate to what extent non-invasive determination of alcohol concentrations in arbitrary sealed bottles using vibrational spectroscopy is possible and worth pursuing. We describe experiments carried out on synthesised alcohol-water solutions, analysed through-bottle using near infrared spectroscopy (NIRS), and classified using a set of benchmark machine learning algorithms into 'genuine' and 'forged' categories based on their ethanol and methanol concentrations.

First, related work is reviewed in Section 2, and an overview of the data collection process and a high-level analysis of it is given in Section 3. The experimental and evaluation methods used are outlined in Section 4 and results presented in 5, before conclusions are drawn in Section 6.

2 Background

2.1 Spectroscopy

Vibrational spectroscopy (VS) is the term used to describe two complementary analytical techniques, infrared spectroscopy (IRS) and Raman spectroscopy (RS). These are non-destructive, non-invasive tools that provide information about the molecular composition of a sample by measuring the intensities of different vibrational interactions between a light source and sample. The spectra produced by VS acts like a fingerprint for the contents, and can be used qualitatively and quantitatively for identification, characterisation, and quality control.

VS methods, along with many competing techniques, are much researched within the food and drink sector [4, 9] due to VS's non-invasive and relatively low operating-cost nature as an analytical technique. However, VS suffers from lower discriminatory power when compared with more time consuming and destructive techniques such as gas or liquid chromatography, one of which is often the technique used to determine the ground truth of studied samples.

As early as 2005, [10] carried out a comparison of NIR and Raman spectrometries for their suitability in combination with regression techniques for the determination of alcohol content in whisky and vodka contained within clear and coloured glass bottles. The study was conducted to evaluate the techniques for possible use in non-invasive, in-situ quality assurance in bottling plants.

Univariate regression models for each type of drink were calibrated for the Raman data in the first derivative spectrum, while a multivariate Partial Least Squares (PLS) model was calibrated for the NIR data. The latter calibration procedure involved some optimisations on the test data, and therefore the results specifically should be treated with caution. However, the higher level conclusions in terms of the relative difficulty of different aspects of the experiments are still insightful: that differences between bottles accounted for the greatest variation and difficulty in the analysis, relative to differences in bottle positioning and time of measurement. The authors concluded that both NIR and Raman were not suited to the analysis of samples within coloured glass in particular, due to the effect of large amounts of fluorescence on the spectra. They also found that for the doubly-transmitted NIR method, a signal could not be collected from the widest part (70mm path length) of the largest bottles, whereas comparable signals to that of the smallest bottles could be found by measuring through the neck of the bottle (40mm path length).

More recent work in this area is described in [7]. The ability for Raman spectroscopy to analyse and discriminate between certain Scotch whisky production factors from within their original containers is tested. 44 whisky samples, three of which had samples transferred to glass vials due to their original bottles being made of green glass, were measured directly through the glass walls using an Avantes Raman instrument. Although not detailed in full, the authors note that the location of measurement (from the neck, base or centre) had no influence on the quality of the readings. Furthermore, the stability of the sampling suggested excellent reproducibility, with normalised spectra being 'virtually identical'.

In an initial Principal Component Analysis (PCA) visualisation, separation could already be found between the type of cask each whisky was matured in. However, factors such as the source distillery and use of artificial caramel colourings could not be defined by the first three principal components (PC).

PLS Regression (PLSR) and Principal Component Regression (PCR) were subsequently evaluated. However PLSR reportedly delivered far better results and was therefore the only method discussed. Leave-one-out cross validation was used as an evaluation procedure. A quantitative analysis of important factors related to authentication was described: age; ethanol concentration; and a second attempt at the presence of artificial colourings. Age could be estimated within

0.42 years (root mean squared error (RMSE)), from the samples in the range 3-22 years. On average ethanol concentration could be estimated to within 0.44% (RMSE), which is only just outside the regulatory limits of Scotch Whisky (0.3%). These are very strong results, suggesting not only feasibility but even an existing ability to quantitatively determine key factors to whisky authentication. Ethanol level determination is perhaps less surprising, as it does form 40-55.8% by vol of the samples, but determining unintuitive properties like age within such an accuracy is a positive result.

Continuing on from these works, our own investigation into this problem focuses on portability, simplicity, and speed in all aspects of the analysis of a sample. The final aim is to allow a non-expert to determine the authenticity of an arbitrary spirit on-site and within seconds.

2.2 Classification

Classically, machine learning and chemometric methods handling spectral data have been linear regressive models built on top of (automatically or manually) selected attributes or PCA-transformed spaces. The physical interactions giving rise to the spectra are understood to be linear in nature, and the resonances of molecules being looked for are known to occur at certain wavelengths, even if the particular wavelengths are not known a priori. Given this, more complex systems may not have much room to increase predictive accuracy, be prone to overfitting, and in some cases may lose interpretability of results. Linear systems on reduced attribute spaces work satisfactorily for clean spectra collected under professional and standardised conditions. However, they may be unable to handle structural changes in the data.

The nature of the problem suggests a regression model. However, through consultation with industry, the ultimate use case designed to aid field use is a traffic light classification scheme; green (genuine), yellow (suspect), and red (forged). The confidence thresholds that define the boundaries of each class can be defined by the user in response to factors such as the costs measurement, verification, and screening.

The classification of spectra can be phrased as a time series classification (TSC) problem [1]. A time series is a set of (typically numerous) ordered and numeric attributes. While different sets of TSC data will have different underlying properties, the typical higher-order structures informing classification are the shapes and patterns of series and/or subseries.

Recent large scale evaluations on entire dataset archives in both traditional classification [6] and TSC [1] give indications as to the classification methods that could be suitable for this particular problem space. [1] found that for spectral datasets (of which there were 7 of 85 datasets in total) classifiers that considered the full series similarity were consistently better than those considering subseries similarity, frequency or distribution. Throughout both evaluations, the effectiveness of ensembling was clearly evident. In the benchmark experiments presented in this work we use a range of classifiers classically used in chemometrics, in addition to state of the art and ensemble classification methods.

3 Data

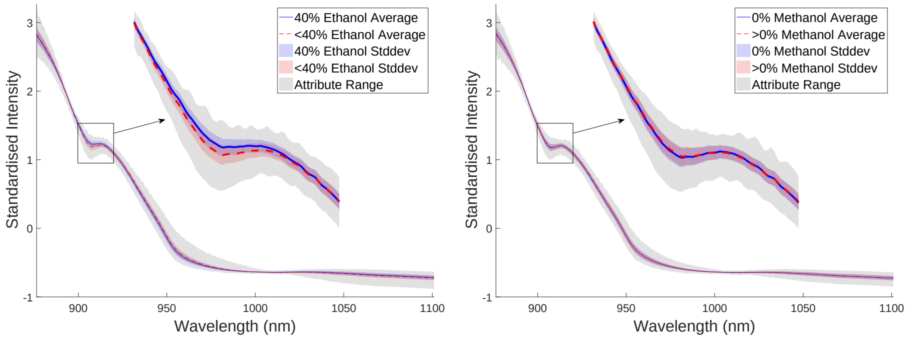
There are many experimental factors that could confound a non-invasive alcohol classification system in the field: ambient light and environmental conditions; variation in spectral hardware; and the measurement habits of different users may all cause variation in the resulting spectra. However, we believe one of the largest sources of variation which needs to be accounted for arises from the properties of the bottle a sample is contained in. Bottle shape and size, glass thickness and colour, and interfering labeling and embossing can all work to frustrate the collection of consistent, reliable spectra. Therefore, with these experiments, we primarily wish to determine the difficulty of measuring and classifying the alcohol content of samples in arbitrary bottles.

We have conducted experiments using 44 different examples of real, non-standardised bottles. While most of the bottles are transparent and cylindrical, some are coloured, rectangular or skewed. Using a single StellarNet BLACK-Comet-SR spectrometer, transmission near-infrared spectra over a one second integration time of ethanol, methanol and water solutions within each bottle were collected to form two datasets. For the ethanol concentration experiments, 40% ethanol (with the remainder being water) is taken to be the 'genuine' case, while concentrations of 35% and 38% ethanol are taken to be 'forgeries'. The second dataset is detecting the presence of methanol. With 40% total alcohol concentration being maintained, solutions with 1%, 2% and 5% methanol (v/v) form the forged class, while 0% methanol (i.e 40% ethanol) constitutes not forged. The two classification problems are therefore to determine from a spectra whether or not a solution within an arbitrary sealed bottle 1) has less than 40% alcohol or 2) contains dangerous levels of methanol. Information on the bottles and the raw data, including labels for bottle and concentration for each reading, can be downloaded at[6].

Three batches of each alcohol concentration were produced, and for each solution in each bottle the sample is placed, a spectra taken, and placed again for a total of three readings. A total of over 2000 readings were taken. Bottles were positioned such that the light travels through the widest part of the bottle while avoiding labelling, embossing and seals as much as possible. However, to mimic future conditions a precise recreation of the exact path on each placement was intentionally not attempted. For simplicity, and to mimic a possible portable sampling station, the geometry of the light source and receiver was fixed at 15cm; enough to accommodate the widest bottles tested. Dark readings were subtracted from each spectra. Data collection took place over the course of multiple weeks by a single tester. Batches of each concentration were spread out over that time-frame, to reduce the chance of any patterns based on time of measurement forming. Spectra are presented in the wavelength range 876.5nm - 1101nm, sampled every 0.5nm, and each spectrum is standardised.

To help give an intuition of the classification problem, Figure 1 shows the average series of each class to demonstrate their differences. The progressively

[6] http://research.cmp.uea.ac.uk/DetectingForgedAlcohol/Data/



(a) Ethanol

(b) Methanol

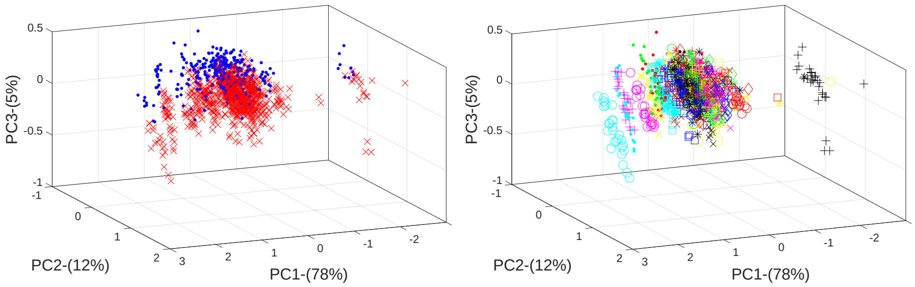
Fig. 1. Graphs showing the average series of each class, overall standard deviation and range for the ethanol and methanol concentration datasets. For each image, the main discriminatory region is zoomed.

shaded regions show the overall standard deviation and range of intensities at each wavelength. The fact that these are difficult to distinguish by eye is itself quite telling. The overall variance in the dataset is very low, and the inter-class variance a fraction of that.

The zoomed regions show the wavelength ranges where alcohols are known to have a strong resonance. A clear separation between classes can be seen within the ethanol problem. However, for methanol the classes appear to be indistinguishable. Ethanol and methanol have overlapping resonances, and therefore the fact that the overall concentration of alcohol remains at 40% means any difference between the class values in the resulting spectra is drastically reduced.

Relative to the apparent differences in the average class spectra, individual series are greatly affected by noise introduced by a variety of means through the nature of the experiment, further increasing classification difficulty. For example, an individual series may be skewed by the lensing effects of a uniquely shaped bottle. This is evidenced by Figure 2. It shows the first three PCs of the transformed ethanol dataset, which explain 95% of the total variance. In (a), the instances are categorised by their ethanol concentrations. While some separation is found between the two classes, this is observed mostly in the second and third PCs, which account for only 17% of the total variation. The first PC, as (b) shows, for the most part explains variance due to the bottles. This is in line with our expectations that bottle variation would be one of the larger obstacles to overcome for the final use case of an authentication system. While many bottles are clustered close together, there are some that form clear and separate clusters of their own. As might be expected, these are bottles that have some particularly non-standard bottle property, such as irregular shape or colour.

Promisingly, the PCA transform does suggest a good separation between ethanol concentrations within a particular type of bottle, as illustrated by the outlying bottle clusters when compared between figures. The equivalent figures



(a) Categorised by concentration

(b) Categorised by bottle

Fig. 2. Graphs of the top three PCs of the PCA-transformed ethanol forgery dataset, with samples categorised by (a) 'genuine' (blue dot) and 'forgery' (red cross), and (b) by bottle.

for the methanol dataset are not included in this paper for the sake of readability and space, however they (and the source ethanol images including keys) are available on-line[7]. What they show is analogous to figure 1(b); that the PCA is almost entirely unable to distinguish between the alcohol concentrations, however trends by bottle type are largely the same.

4 Experimental Setup

For this application, our long-term hypothesis is that TSC methods that consider overall shape may be able to correct for structural defects in the spectra brought about by the many sources of noise involved with non-invasive spectra collection. For example, a linear method built on a small number of selected attributes may not be able to account for high-level structural changes in a new test case caused by an abnormally shaped bottle. Using the datasets that have been formed, we perform benchmark and exploratory evaluations with a wide variety of classification schemes.

The classifiers evaluated are: Logistic Regression (LR); Partial Least Squares Regression (PLSR); Multilayer Perceptron (MLP); 1-Nearest-Neighbour with Euclidean Distance (1NN); C4.5 Decision Tree (C45); linear SVM (SVML); quadratic SVM (SVMQ); radial basis function SVM (SVMRBF); Rotation Forest [11] (RotF); Random Forest [3] (RandF); Heterogeneous Ensemble of Standard Classification Algorithms [8] (HESCA), and for the TSC-specific classifiers: Random Interval Spectral Ensemble [2] (RISE); Bag of SFA Symbols [12] (BOSS); and Time Series Forest [5] (TSF).

We evaluated each classifier on the datasets using a leave-one-bottle-out (LOBO) cross validation. In this scheme, all samples contained within a single bottle are reserved for the test set, with the remainder forming the training

[7] http://research.cmp.uea.ac.uk/DetectingForgedAlcohol/FiguresAndTables/

set. By evaluating in this manner, classifiers should not be able to leverage any discriminatory features caused by the bottle itself, focusing on alcohol level as the only commonly varying factor.

To avoid ambiguity, we stress that in all cases the training of a classifier, including any hyper-parameter tuning and model selection required, is performed independently on the train set of a given fold, and the trained classifier is evaluated exactly once on the corresponding test set. Our code[8] reproduces the splits used in this evaluation exactly, and results[9] are able to be recreated.

Our primary concern is generally accuracy (ACC) because of its ease of motivation and interpretability. However, in applications such as this the costs of measurement, verification, and misclassification externally influence the ways in which decisions need to be made. For example, if the costs of confirming the legitimacy of a suspect bottle are high, relative to the resources available to the tester, then the decision boundary may be skewed to favour the 'genuine' label. As a result, only samples that the device is more confident are fake will be seized or sent for further analysis. Accuracy cannot entirely capture these factors. Therefore balanced accuracy (BALACC) and measures that assess the quality of the classifiers' probabilistic outputs are also reported; the Log-Likelihood (LL) and the Area Under the Receiver Operating Characteristic (AUROC).

5 Results

Table 1 details the average accuracies achieved by each classifier on all datasets formed for the sake of space, however each subset of experiments is separately discussed in turn with the superscripts denoting the particular columns of results being discussed.

5.1 Leave-one-bottle-out Cross Validation^{1}

When considering the LOBO experiments on the original (time series form) data, two trends are immediately apparent: ethanol concentration, with the correct models, can be classified with high accuracy; determining methanol concentration in a constant overall alcohol level is much more difficult. Only some of the classifiers tested achieving higher than the minimum expected accuracy of 0.75, the proportion of the majority class.

To discover what classifiers are best for each evaluation statistics, we can perform statistical tests of difference over fold scores because all classifiers are evaluated on identical splits, which are reproducible with the published code. Figure 3 is a critical difference diagram over all folds of the LOBO-sampled ethanol and methanol datasets combined. Classifiers are ordered by average rank over fold scores, and those connected by a bar are pairwise not significantly different between each other, $p = 0.05$.

[8] http://research.cmp.uea.ac.uk/DetectingForgedAlcohol/Code/
[9] http://research.cmp.uea.ac.uk/DetectingForgedAlcohol/Results/

Table 1. Average accuracies (and standard deviations) over all folds of the alcohol datasets. The best classification accuracies on each dataset are bold.

Classifier	Ethanol LOBO[1]	Methanol LOBO[1]	PCA Ethanol LOBO[3]	PCA Methanol LOBO[3]	Bottle[2]
1NN	0.866(0.093)	0.672(0.103)	0.779(0.104)	0.627(0.101)	0.541(0.017)
BOSS	0.913(0.086)	0.786(0.050)	-	-	0.622(0.021)
C45	0.824(0.132)	0.658(0.098)	0.796(0.106)	**0.750**(0.001)	0.412(0.017)
HESCA	**0.965**(0.069)	0.843(0.079)	**0.818**(0.104)	**0.750**(0.001)	0.639(0.020)
LR	0.964(0.045)	0.809(0.100)	0.807(0.092)	0.744(0.030)	0.430(0.018)
MLP	0.960(0.068)	0.834(0.083)	0.813(0.108)	**0.750**(0.001)	0.617(0.027)
PLSR	**0.965**(0.053)	0.860(0.073)	0.801(0.089)	0.745(0.026)	0.061(0.010)
RandF	0.888(0.105)	0.758(0.047)	0.817(0.093)	0.714(0.060)	0.587(0.015)
RISE	0.776(0.115)	0.780(0.031)	-	-	0.622(0.016)
RotF	0.938(0.078)	0.839(0.049)	0.815(0.104)	**0.750**(0.001)	0.653(0.014)
SVML	0.945(0.075)	0.838(0.077)	0.801(0.094)	**0.750**(0.001)	0.517(0.017)
SVMQ	0.959(0.103)	**0.864**(0.102)	0.803(0.092)	**0.750**(0.001)	**0.656**(0.019)
SVMRBF	0.881(0.098)	0.841(0.092)	0.806(0.091)	**0.750**(0.001)	0.349(0.015)
TSF	0.868(0.112)	0.769(0.029)	-	-	0.635(0.018)

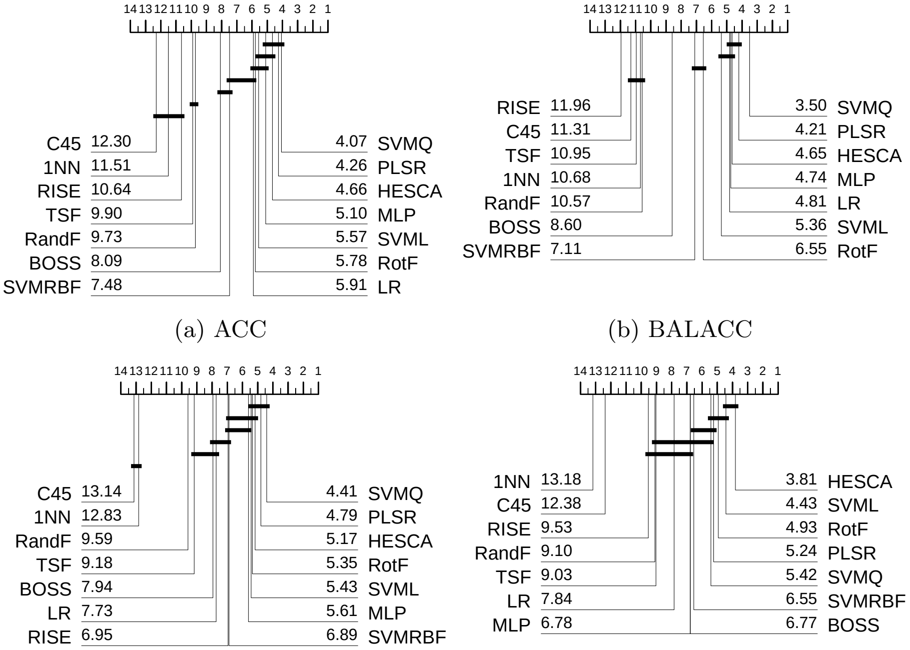
Because each test fold represents a single bottle, the accuracy on a fold gives an indication of the difficulty that a particular bottle adds to the classification problem. We took the top four classifiers on the ethanol problem (PLS, LR, MLP, and HESCA) which all achieved similarly strong performances, and looked at which bottles were preventing perfect classification. On 34 of the 44 folds, at least one of these top four classifiers achieved an accuracy of 1. Where only a subset of the four classifiers met this criteria, the rest only ever misclassified only one or two test cases.

The worst fold accuracy represents the Bernheim Original Kentucky Straight wheat whiskey bottle, where the average accuracy across the four classifiers is 0.76. Only five bottles had an average accuracy of less than 0.9, and all of them are irregular in some way. This does lend credence to the idea that the determination of alcohol concentration cannot be done *entirely* irrespective of bottle. However, the fact that there is clearly some transferability (evidenced by better-than-guessing accuracy in this LOBO format) is promising.

In [7] and [10], coloured glass posed challenges for the collection of Raman spectra, which particularly struggles to handle fluorescence, but also for NIRS in [10]. Our experiments included three green-glass bottles, however on these no significant drop in predictive accuracy was observed in the same analysis of the top four classifiers.These three bottles also showed no clear separation from the largest central cluster in the PCA transform presented in Figure 2b.

5.2 Classifying the bottle[2]

The PCA transform of the ethanol dataset, shown in Figure 2b, indicated that the majority of the variance corresponded with differences in the containing bottle's properties. Further, most of the first PC was caused by a small number of irregularly shaped bottles. The majority of bottles otherwise formed a dense cluster. To further investigate the extent to which features of the bottle are detectable in the spectra, we ran experiments with the same set of classifiers but



(a) ACC

(b) BALACC

(c) AUROC

(d) LL

Fig. 3. Critical difference diagrams for the four evaluation statistics on the alcohol datasets.

with the containing bottle as the class label, instead of alcohol concentration. The full dataset was split 30 times using random stratified sampling with a 70/30 train/test split. We would expect the outlying bottles on the PCA transform to be the easiest to classify, while most standard bottles can only be guessed at. The best accuracies achieved were up to 0.656 (SVMQ), which on a 44 class problem is quite high. It is worth noting also that the non-linear methods and especially TSC methods make a relative improvement on this problem, signifying the different nature of the discriminatory features.

In the interest of finding where the classifiers were making their errors, we grouped bottles by whether they could be described as being standard (clear glass and cylindrical, 28 bottles) or irregular (coloured glass and/or non-cylindrical, 16 bottles). Considering the SVMQ's predictions, we counted the incorrect classifications for cases with a standard bottle label classified as a standard bottle, standard classified as irregular, irregular classified as standard, and irregular classified as irregular. The first of these four cases accounts for 69% of the total error, while the remaining three account for a little over 10% each. When correcting for the number of possible ways to misclassify in each scenario, SVMQ was still twice as likely to make the first kind of error as the last.

These results have positive implications for the original goal of non-invasive alcohol level determination. The fact that it is relatively much easier to mistake one standard bottle for another suggests that a classifier could be reliably trained under the assumption that the test sample bottle has certain properties matching those in the train set. In terms of the practical use and production costs of a device, the worst case is that each individual type of bottle requires its own adequately populated training data for a model to learn on. While this or at least a two-stage classification procedure may still be needed for irregular bottles, a device that can effectively classify the contents of bottles within some particular range of properties is still a worthwhile improvement over the worst case.

5.3 PCA Transforms[3]

Lastly, we repeated the LOBO classification experiments again with PCA-transformed versions of the datasets (calculated and applied to each resample individually), maintaining components that explain 95% of the variance. Analysis of spectral data in the literature often involves a dimensionality-reducing transformation such as PCA, both to highlight discriminatory variance and reduce the computation time of analysis. However, in this case it appears to reduce accuracy relative to classification performed on the time series, in agreement with [7].

The methanol PCA transform seemingly cannot discriminate between concentrations at all, with all classifiers simply picking the majority class. For ethanol, all classifiers except 1NN achieve very similar accuracies. Referring to Figure 2a, it would seem that most of the classifiers are forming almost identical decision boundaries, the same that a human naively would by eye.

6 Discussion

We have demonstrated the feasibility of determining alcohol concentration non-invasively in arbitrary bottles using near infrared spectroscopy in combination with machine learning. While ethanol level could be classified with high accuracy, methanol concentration within a consistent overall alcohol level was much more difficult to detect. However, some classifiers demonstrated results significantly better than random guessing, suggesting that the discriminatory features are not entirely lost at the physical hardware level. There may still be room for improvement with different optical geometries and better-tailored data processing and model selection. Bottles with particularly unique properties introduced extra difficulty, but the contents of more standard bottles could be learned and determined with very good transferability.

Traditional methods within chemometrics such as Logistic and Partial Least Squares regression were strong, as expected. However, Principal Component Analysis led to significantly decreased performance. A quadratic support vector machine and simple neural network architecture also performed well. A larger computational investment for more thorough tuning would likely lead to improved results for these. In a subsequent evaluation incorporating the presented test results, ensembling these also led to insignificant improvements in accuracy and probabilistic outputs. Modern algorithms bespoke to time series classification did not provide an immediate increase in predictive power.

Acknowledgements

This work is supported by the Biotechnology and Biological Sciences Research Council [grant number BB/M011216/1], and the UK Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/M015087/1]. Work was performed in collaboration with the Quadram Institute, Norwich and the Scotch Whisky Research Institute, Edinburgh. The classification experiments were carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia.

References

1. A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advance. *Data Mining and Knowledge Discovery*, pages 1–55, 2016.
2. A. Bagnall, J. Lines, J. Hills, and A. Bostrom. Time-series classification with COTE: The collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 27:2522–2535, 2015.
3. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
4. GP. Danezis, AS. Tsagkaris, F. Camin, V. Brusic, and CA. Georgiou. Food authentication: Techniques, trends & emerging approaches. *TrAC - Trends in Analytical Chemistry*, 2016.
5. H. Deng, G. Runger, E. Tuv, and M. Vladimir. A time series forest for classification and feature extraction. *Information Sciences*, 239:142–153, 2013.
6. M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.
7. J. Kiefer and A. Lynda Cromwell. Analysis of single malt Scotch whisky using Raman spectroscopy. *Anal. Methods*, 91:790–794, 2017.
8. J. Large, J. Lines, and A. Bagnall. The Heterogeneous Ensembles of Standard Classification Algorithms (HESCA): the Whole is Greater than the Sum of its Parts. 2017.
9. S. Lohumi, S. Lee, H. Lee, and BK. Cho. A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration. *Trends in Food Science and Technology*, 46(1):85–98, 2015.
10. A. Nordon, A. Mills, R. Burn, F. Cusick, and D. Littlejohn. Comparison of non-invasive NIR and Raman spectrometries for determination of alcohol content of spirits. *Analytica Chimica Acta*, 548(1-2):148–158, 2005.
11. J. Rodriguez, L. Kuncheva, and C. Alonso. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, 2006.
12. P. Schäfer. The BOSS is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29(6):1505–1530, 2015.