Development and application of an optimised Bayesian shrinkage prior for spectroscopic biomedical diagnostics

Chu, Martin; Buchan, Emma; Smith, David; Goldberg Oppenheimer, Pola

DOI:
10.1016/j.cmpb.2024.108014

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Chu, M, Buchan, E, Smith, D & Goldberg Oppenheimer, P 2024, 'Development and application of an optimised Bayesian shrinkage prior for spectroscopic biomedical diagnostics', *Computer Methods and Programs in Biomedicine*, vol. 245, 108014. https://doi.org/10.1016/j.cmpb.2024.108014

Computer Methods and Programs in Biomedicine 245 (2024) 108014

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Development and application of an optimised Bayesian shrinkage prior for spectroscopic biomedical diagnostics

Hin On Chu [a], Emma Buchan [a], David Smith [b], Pola Goldberg Oppenheimer [a,c,*]

[a] School of Chemical Engineering, University of Birmingham, Birmingham B15 2TT, UK
[b] School of Mathematics, Watson Building, University of Birmingham, Birmingham B15 2TT, UK
[c] Healthcare Technologies Institute, Institute of Translational Medicine, Mindelsohn Way, Birmingham B15 2TH, UK

ARTICLE INFO

Keywords:
Vibrational spectroscopy
Bayesian analysis
Analytical platforms
Biomedical diagnostics

ABSTRACT

Background and objective: Classification of vibrational spectra is often challenging for biological substances containing similar molecular bonds, interfering with spectral outputs. To address this, various approaches are widely studied. However, whilst providing powerful estimations, these techniques are computationally extensive and frequently overfit the data. Shrinkage priors, which favour models with relatively few predictor variables, are often applied in Bayesian penalisation techniques to avoid overfitting.
Methods: Using the logit-normal continuous analogue of the spike-and-slab (LN–CASS) as the shrinkage *prior* and modelling, we have established classification for accurate analysis, with the established system found to be faster than conventional least absolute shrinkage and selection operator, horseshoe or spike-and-slab. These were examined *versus* coefficient data based on a linear regression model and vibrational spectra produced *via* density functional theory calculations. Then applied to Raman spectra from saliva to classify the sample sex.
Results: Subsequently applied to the acquired spectra from saliva, the evaluated models exhibited high accuracy (AUC>90 %) even when number of parameters was higher than the number of observations. Analyses of spectra for all Bayesian models yielded high-classification accuracy upon cross-validation. Further, for saliva sensing, LN–CASS was found to be the only classifier with 100 %-accuracy in predicting the output based on a leave-one-out cross validation.
Conclusions: With potential applications in aiding diagnosis from small spectroscopic datasets and are compatible with a range of spectroscopic data formats. As seen with the classification of IR and Raman spectra. These results are highly promising for emerging developments of spectroscopic platforms for biomedical diagnostic sensing systems.

1. Introduction

Raman spectroscopy (RS) and infrared spectroscopy (IR) provide powerful, non-invasive vibrational spectroscopic methods for analytical applications and for diagnostics with insight into molecular environments and relative concentrations. RS has been continuously exploited for many emerging healthcare-related applications in diagnostics including for instance traumatic brain injury [1–4], cancer [5–10], tuberculosis [11,12], and screening of keratitis [13–15]. While promising, the majority of clinical studies including healthy volunteers and patients tend to only have small, limited populations in early-stage studies to acquire samples from, especially when compared with the number of potential parameters involved in spectroscopic techniques.

The sample size required for an 80 % power, an accepted level or higher for determining whether the research study shows an actual effect [16], would not be a viable for most initial case studies for various diseases. For example, a disease that affects 4 % of a population and is only expected in 6 % of a study group would require over 850 samples for an 80 % power with 0.05 type-I error rate. Availability of patients and healthy controls capable to provide samples for early-stage spectroscopic diagnostic studies are typically significantly smaller, on the order of. This issue leads to small groups and therefore, small sample numbers in initial studies prior to progressing to larger trials where greater resources would be required. Therefore, the number of observations (n) is expected to be smaller than the number of possible parameters (p) that can be extracted from Raman or Infrared spectra.

* Corresponding author at: School of Chemical Engineering, University of Birmingham, Birmingham B15 2TT, UK.
E-mail address: GoldberP@bham.ac.uk (P. Goldberg Oppenheimer).

https://doi.org/10.1016/j.cmpb.2024.108014
Received 8 November 2023; Received in revised form 6 January 2024; Accepted 8 January 2024
Available online 9 January 2024

Several rules of thumb have been proposed regarding sample sizes, for PCA, to be either minimum 100 or, 5–10 times the number of variables [17–19]. This renders the commonly used principal component analysis (PCA) (for dimensional reduction analysis to extract spectral features) and linear discriminant analysis (LDA) as less reliable given the small n observations for each class during early-stage studies. Alternatively, other methods could be employed including the self-organising maps based on artificial neural networks (ANN), which tend to perform better than PCA-based approaches. However, even machine learning approaches, such as ANN, require sample sizes that are significantly larger than the weightings. For the ANN, Alwosheel *et* al. had proposed that the sample size should be 50 times greater than the weightings [20]. By contrast to ANN, Bayesian methods (an approach that observes new information to adapt the model to offer better accuracies) are known to provide better insight into incomplete or small datasets. Additionally, recent Bayesian approaches enable the parameter estimation in the oral minimal model of glucose dynamics [21], as well as predicting cardiovascular risk [22]. Thus, Bayesian methods offer a potentially powerful classification avenue for analysing vibrational spectroscopies, where the number of possible relevant peaks may well exceed the possible number of observations given the initial study group sizes being generally small. Unlike frequentist approaches, which are purely data-driven, Bayesian methods incorporate prior information into the analysis [23]. Where the prior updated with new information that is then used to obtain the posterior probability distribution. For shrinkage priors, there is a weighting at zero for smaller models that favour for sparse solutions, *i.e.,* those in which many coefficients are set to zero and so the corresponding features do not contribute to classification. Although Bayesian statistics is a well-established branch of statistics [24], it was not until relatively recently when Bayesian models have started to be applied to Raman spectroscopy analysis [25], and its application for disease diagnostics was not reported for a further four years [26]. More specifically, Raman spectroscopy towards the diagnostics of cancer was not reported in conjunction with a Bayesian method until 2013 [10]. In this study, the authors had applied a Naïve Bayes classifier to Resonance Raman spectra of breast tissue to determine if a sample was healthy or had cancer. From this, they were able to diagnose malignant invasive ductal carcinoma grade II with 99.9 % sensitivity and 100 % specificity. Furthermore, whilst many of the Bayesian methods provide a powerful method for parameter and measurement estimation, the gold-standard spike-and-slab regression methods can often become computationally intractable for large sets of parameters, which is expected when there are large number of potentially relevant peaks in the spectra.

To address the computational expense issues with the spike and slab approach, a Bayesian prior distribution [27], which uses a logit-normal continuous analogue of the spike-and-slab (LN–CASS) is presented as a Bayesian classifier approach for IR and Raman spectra. First, LN–CASS has been assessed for its performance against other Bayesian methods such as the Horseshoe prior, as well as the frequentist techniques of least absolute shrinkage, and selection operator (LASSO), ordinary least squares (OLS) and sparse group lasso (SGL). These approaches have then been examined with a parameter estimation based on a linear regression model. Subsequently, the LN–CASS prior has been assessed in a classifier model for simulated Raman and IR spectra of sucrose and glucose. Random forest (RF) and LASSO were also modelled on the same dataset for comparison. Finally, these three models and a self-organizing map (SOM) model which uses an artificial neural network were compared for their ability to classify biological sex based on Raman spectra of human saliva. The overall results lay the groundwork for the development of a classification model for vibrational spectroscopy studies, particularly for when sample numbers are small compared to the number of variables.

2. Materials and methods

2.1. Simulated vibrational spectra via DFT calculations

Density functional theory (DFT) calculations were carried out using ORCA 5.0.2 using the ORCA quantum chemistry package [28]. A modified version of Avogadro molecular visualization software (that enables ORCA-related extensions) was used to create input data files and process software outputs [29]. For DFT, B3LYP was chosen for its reasonable computational cost to accuracy-ratio paired with a Pople-style basis set (6–31G**). Spectra were then exported using the ORCA built-in utility ensuring the range between 300 and 2000 cm^{-1} wavenumbers was exported over 1024 data points, with a 15 cm^{-1} peak width. To produce a dataset, each spectra had jitter noise (factor set to 150 and random seed to 55) applied to produce unique spectra.

2.2. Raman characterisation

To compare with the simulated Raman spectra of D-(+)-glucose and sucrose, both materials were individually dissolved in minimum water (15 MΩ-cm) then deposited onto aluminum substrate and dried in a vacuum desiccator. For Raman spectroscopy measurements, Renishaw InVia Qontor Raman microscope system was employed using the 785 nm, excitation laser at 10 % laser power (100 % laser power at the sample, with no objective, was measured to be 120.2 mW using Thorlabs laser power meter (PM100D digital console equipped with an S121C standard Si photodiode sensor) with the 1200 lines/mm grating. The measurement settings selected were x50 objective with 5-second exposure over three accumulations.

2.3. Modelling with LN–CASS

The LN–CASS and other modelling was conducted with R code, R-Stan was used for the implementation of Bayesian methods [30], where the sampling of the posterior was done with a No-U-Turn sampler; this approach is more efficient than conventional random walk approaches and moreover provides adaptive step [31]. For the LN–CASS model, this was adapted from Thomson et al. code available at [32]. The LN–CASS prior requires the choice of hyperparameters, τ, $\mu_\lambda and \sigma_\lambda.\mu_\lambda$, these are fixed throughout the study as 5, logit(a).σ_λ and 10 respectively. τ is the standard deviation of the 'slab', the other two are the parameters of the logit-normal distribution that are the median of the logit-normal distribution and is based on the belief that priori that each coefficient has a probability a of being non-zero. In essence, to run the code, the spectra data should be prepared as text files (csv or txt), which are then imported into the IDE (RStudio was used as the IDE for this work) with the R code. Once all the data has been imported into the IDE, the modelling could begin and once complete, data plots would be generated. A more accessible guide to running this analysis on Raman spectroscopy can be seen in supporting information (S2.3).

3. Results

3.1. Recapitulation of Bayesian approaches for a linear regression model

This test expands to parameters values that are closer to Raman spectra data where typically the number of parameters $p >>$ observations n for typical studies. (Full range of values of p can be found in supporting information Fig. S1). The results of these data fits are shown in Figs. 1 and S1, where at low p, all methods were observed to work well as their accuracies are within 20 % for 10 parameters with 100 observations except the SGL, which was observed to consistently perform the worst among the models tested.

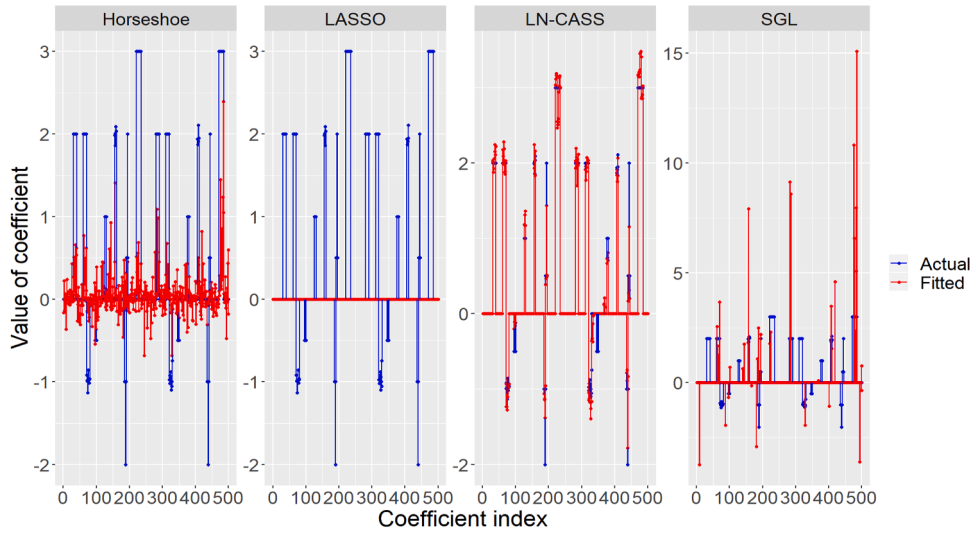

Fig. 1. Coefficient predictions with $n = 100$ observations and $p = 500$. Horseshoe, LASSO, LN–CASS, and OLS. The SGL method was omitted from these as that method breaks down for when $p > n$. The actual coefficient values are shown in blue and the fittings in red. **Fig. S1**, in the supplementary information has a larger range of different p values tested.

3.2. Classification of IR and Raman data

Fig. 2 presents the classification of Raman data for glucose and sucrose, comparing the different methods (LN–CASS, RF and LASSO) as well as showing the difference between experimentally and theoretically derived Raman spectra.

Table S1 summarising the performance of the models presented in Figs. 1, S1 and S2 across $p = 10$, 50, 100, 150, 250 and 500. Table 1 summarises the leave one out cross validation (LOOCV) for each method and spectra type.

Fig. 3 presents the classification of sucrose and glucose via the simulated IR spectra data, comparing LN–CASS, RF and LASSO.

3.3. Performance of the LN–CASS method on human saliva Raman data for biological sex classification

Fig. 4 presents the classification abilities of determining biological sex based off the Raman spectra of saliva using LN–CASS, RF, LASSO and the SOM approaches and Table 2 presents the Raman peak assignments for human saliva.

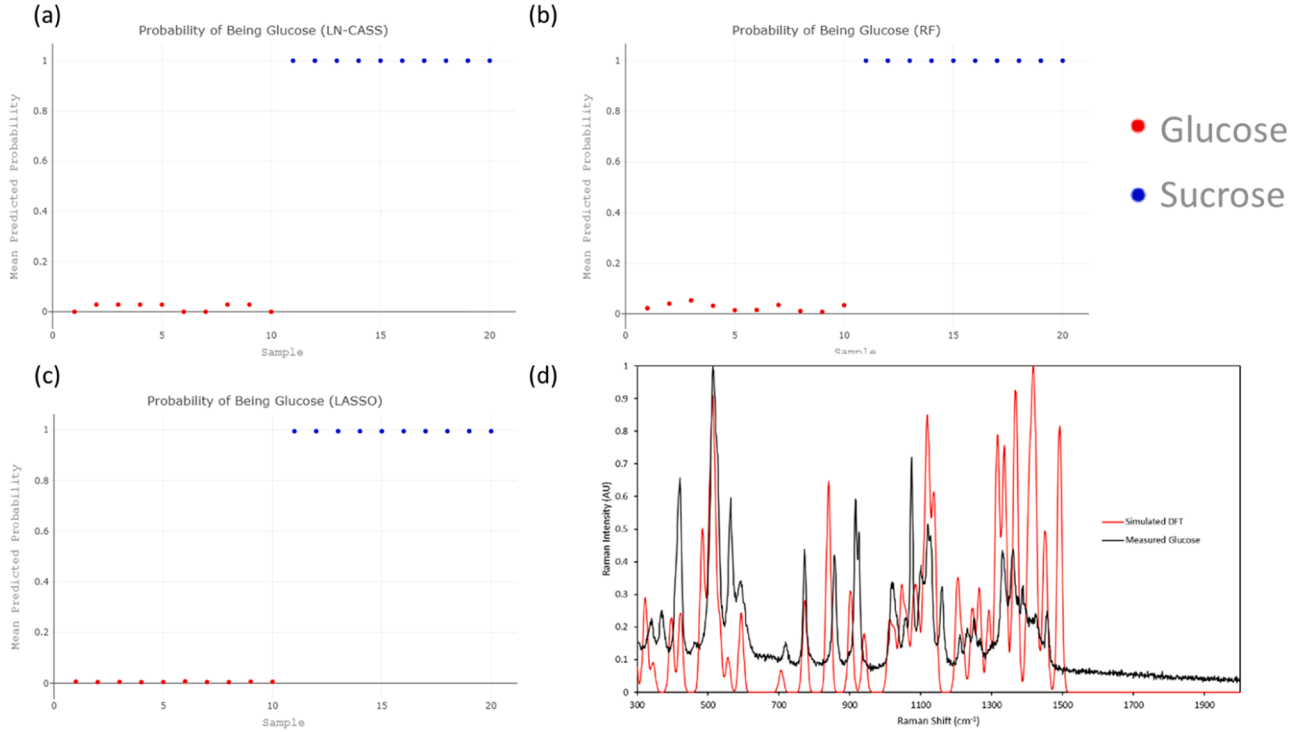

Fig. 2. *Classification of Raman spectra, comparing sucrose with glucose. Classification of glucose and sucrose assessing Raman spectra with* ***(a)*** LN–CASS ***(b)*** *Random Forest and* ***(c)*** *LASSO. For all models, these are based on a Wald test, and 35 Raman shifts are selected based on the magnitude of their z-scores.* ***(d)*** *Comparison of the simulated Raman spectra for glucose (red) to the experimentally observed spectra (black). Whilst most of the peaks align well with the measured observations, there are several peak intensities that are significantly different.*

Table 1
LOOCV derived AUC and MAE scores for LN–CASS, LASSO and RF classifying between D-glucose and Sucrose using simulated Raman spectra for $n = 20$ and $p = 35$.

Spectra Type	Method	AUC	MAE
Raman	LN–CASS	1.00	0.00857
	LASSO	1.00	0.00525
	RF	1.00	0.01320
IR	LN–CASS	1.00	0.00143
	LASSO	1.00	0.00524
	RF	1.00	0.01130

4. Discussion

Firstly, the LN–CASS prior was evaluated with randomised parameters and compared with other techniques such as the Horseshoe prior, in addition to the frequentist techniques LASSO, OLS, and SGL. Thomson et al. previously had conducted this test but with values of p tested only up to 120 [27]. Regardless of the model for $p = 10$ and $n = 100$, all were able to provide area under the receiver operator curve (AUC) values of 1. Where an AUC value of <0.5 means the method is worse than random guessing, >0.5 it is better than the aforementioned and at AUC=1 it means the method is perfectly able to make predictions. This result is not surprising considering the ratio of observations and parameters are suitable for techniques as data-hungry as PCA, to cluster samples based on their first 2–3 principal components [19]. The predictions observed in Fig. 1 for each model are consistent with the same test performed by Thomson *et* al. with different ranges of p and n [27], such that all methods perform well for $p \langle < n$, but for $p \sim n$, it was observed that OLS falls apart since it is ill-defined for cases when $p \rangle n$. From Table S1 summarising the performance of the models presented in Figs. 1 and S1 across $p = 10$, 50, 100, 150, 250 and 500, the LN–CASS and LASSO approaches achieved some of the best AUC values. In particular, LN–CASS resulted in the lowest mean absolute error (MAE) values. The Horseshoe method was seen to struggle in cases when the number of parameters is sufficiently higher than the number of observations. In the cases of $p = 150$ or higher, implementation of the Horseshoe prior resulted in difficulties with the Markov Chain Monte Carlo sampler not converging after 4 chains ran in parallel, suggesting that the sample was not close enough to the set posterior distribution. The SGL approach was observed to be the worst performer of the methods tested, often providing the greatest MAE values. From these results, potential candidates for a classifier applied to vibrational spectra should be either the LASSO or LN–CASS options.

Glucose and sucrose, common small molecules, were chosen since their vibrational spectra are well-defined and can be classified easily with visual observation. Density functional theory was used to simulate the IR and Raman spectrum of each molecule and a dataset was produced with a copy of each spectra having noise applied to it. This produced unique spectra between 300 and 2000 cm^{-1}. The datasets were preprocessed, first by applying a log-transformation, followed by subtraction of the mean and dividing by the standard deviation of the Raman intensity. Subsequently, the Raman intensities were passed through a Wald test and selected; 35 intensities with the largest Z-scores with absolute value. This would have represented 35 Raman peaks and for simple molecules such as sucrose this would represent approximately half of all possible Raman modes. Raising this selection number beyond that would mean selecting a parameter value that is greater than possible modes for a certain class. LN–CASS, LASSO and RF were modelled on the preprocessed Raman spectra with 10 observations from each class. The obtained output results are shown in Fig. 2 along with the comparison of the simulated and the experimentally observed

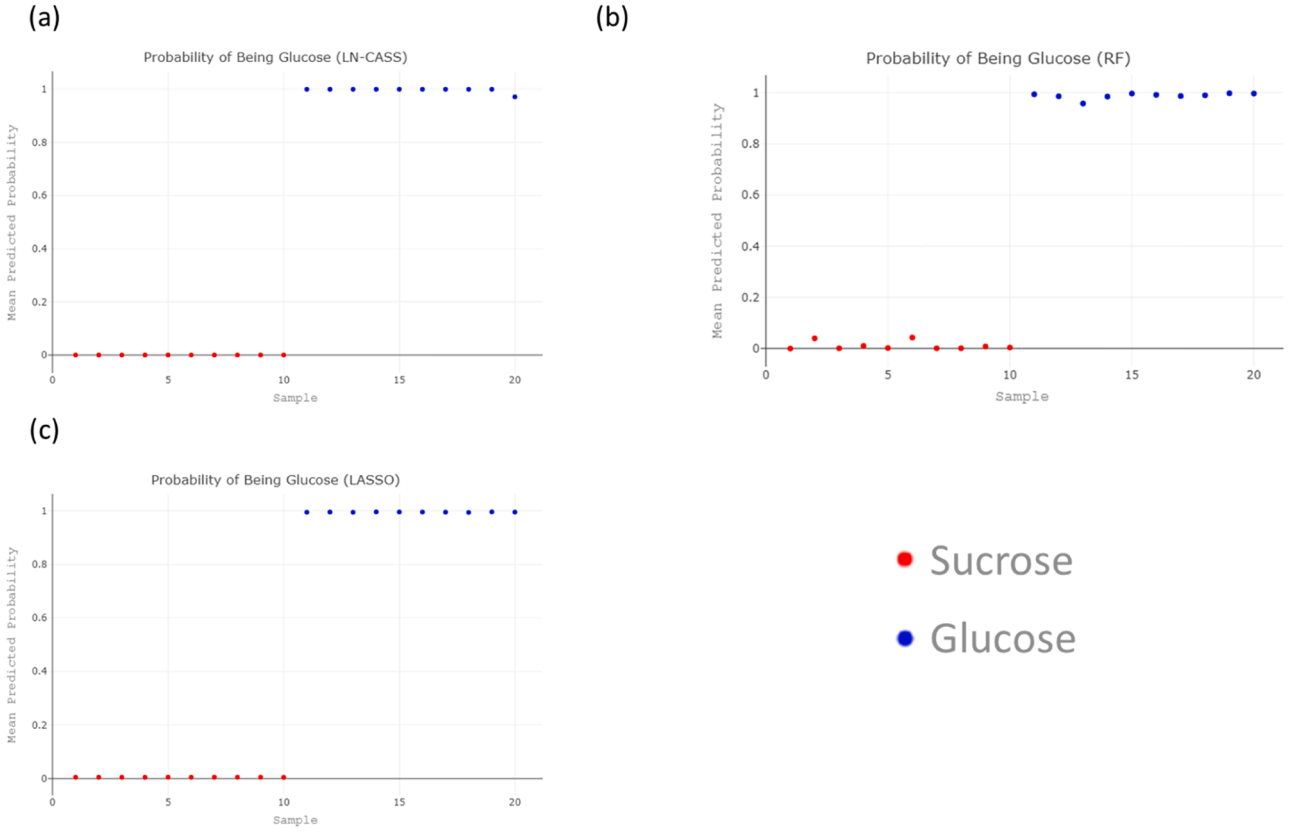

Fig. 3. *Classification of sucrose and glucose via the simulated IR spectra data using* ***(a)*** LN–*CASS* ***(b)*** *RF and* ***(c)*** *LASSO. While all approaches show excellent classification performance with IR, the amount of distance between the predicted values to the true value is observed to be reduced for* LN–*CASS by a factor of 6.99 when compared with the Raman dataset.*

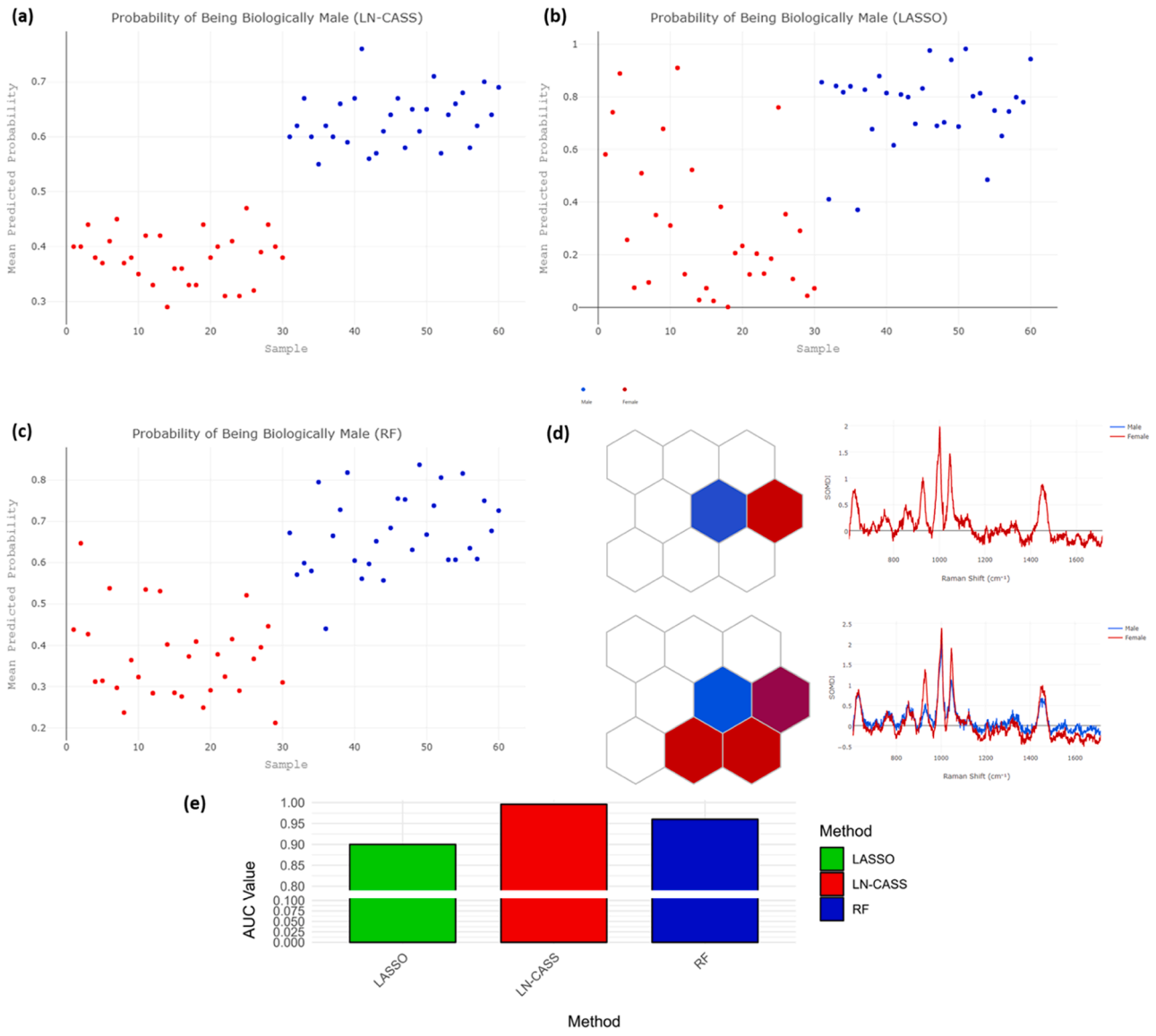

Fig. 4. Binary classification of human saliva samples predicting whether they are originating from male or female donors. Red dots represent samples from female donors and blue from male donors. Methods used for these classifiers are **(a)** LN–CASS, **(b)** LASSO **(c)** RF and **(d)** SOM. In **(d)**, the top row is for classification without the LVQ option applied to the trained map and the lower row is with LVQ applied. The SOMDI values are also presented, which presents which peaks contribute most to the classification. Dominant Raman peaks include the C–S stretch and C–C twist of proteins and tyrosine (628 cm^{-1}), C–C stretch of the amino acids (proline, hydroxyproline and valine), proteins (930 cm^{-1}), symmetric ring breathing mode (phenylalanine, tryptophan) at 1003–1005 cm^{-1}, C–O and C–N stretch at 1050 cm^{-1}, amide-III band at 1300 cm^{-1} and CH_2 and CH_3 deformation vibrations (proteins and lipids) at 1450 cm^{-1}. **(e)** Comparison of the AUC values for the LN–CASS, LASSO and RF methods. Since the values are close in value, the y-axis has been truncated from 0.1 to 0.8 for clarity.

representative spectra. The relative Raman shift positions of the calculated spectra were found to be consistent with the observed ones (Fig. 2d). Differences in experimental data presented small disparities in relative peak intensities and peak widths. The peak width disparity is due to the potential measurement of a sample with limited long-range order. In the 1300–1500 cm^{-1} range, the relative intensities are significantly lower than predicted in contrast to the 1000–1200 cm^{-1} range. As expected, the statistical methods worked well for classification of sugars based on their Raman spectra, all having achieved an AUC value of 1.00, revealing high classification capabilities. Therefore, the only critique from here would be to compare the MAE values, not as a comment on the method's ability for classification but rather an observation on how the models predict values. The machine learning model (RF) also showed promising results for the sugar classification, displaying similar MAE to the LN–CASS approach (0.013 for RF compared with 0.009 for LN–CASS). The results for classification based on the IR spectra (Fig. 3) showed a similar trend, except for the mean absolute error for LN–CASS, which was found to be significantly smaller than observed for the Raman data. While all methods showed 100 % accuracy in classification between glucose and sucrose based on the IR spectra, the random forest approach consistently showed the greatest amount of error in the predictions. For IR spectral analysis, LN–CASS revealed the lowest error values and LASSO showed impressive classification performance. The leave-one-out cross validation is summarized in Table 1, with results indicating that using any of these methods for classification of vibrational spectra with clear peak features will yield high accuracy and low errors. The lowest MAE obtained for Raman analysis was found to be 0.005 with the LASSO method, whereas for IR, the LN–CASS exhibited considerably reduced MAE of only 0.001, which is an improvement over LASSO approach by a factor of 6.99. This level of performance similarity between Raman and IR analysis was expected, as both datasets present intensity *versus* wavenumber. For such simple

Table 2
Peak assignments for the Raman spectra of human saliva [3].

Peak Wavenumber (cm^{-1})	Assignment
628	C-S stretch and C–C twist Protein; tyrosine
760	Ring breathing mode tryptophan; proteins
855	C–C; ring breathing mode tyrosine
960	Calcium-phosphate stretching band (cholesterol), α-helix. Proline, Valine (n (C–C))
1003	Symmetric ring breathing mode (phenylalanine, tryptophan)
1076	C-C (Lipids); symmetric stretch of phosphates in hydroxyapatite
1125	C–C skeletal stretch (lipids); C–N stretch (proteins)
1205	Amide III; CH_2 wagging and vibrations (glycine, proline, tyrosine and phenylalanine)
1337–1339	CH_2/CH_3 wagging and twisting (proteins, nucleic acid, lipids), nucleic acid bases (n (C–H))
1456	CH_2 and CH_3 deformation vibrations (proteins and lipids)
1655	Amide I region; C═C stretch (lipids); C═O stretch (proteins)

molecules, the 100 % classification accuracy with only 10 observations per class indicates that any of the Bayesian models or RF would be suitable for simple cases. In biological studies however, there may be many overlapping peaks with similar relative peak intensities which would be difficult to classify. We have thus applied a test case of any single Raman spectrum of complex biological dataset comprised of human saliva to determine biological sex.

Biological samples typically exhibit many overlapping peaks which render visual classification of Raman spectra difficult without additional post-processing. For comparison, in addition to the LASSO and RF models, the self-organising map algorithm that uses an artificial neural network developed by Banbury *et* al. was applied [33]. Our dataset contained 60 Raman spectra from 30 healthy males and 30 healthy female participants. Fig. 4d presents the Raman peaks in measured human saliva which contribute most to the classification. The identified peaks of importance are assigned to phenylalanine, tryptophan, CH modes from proteins and lipids, amide I and III. Table 2 summarises the overall peak assignments for Raman spectra of human saliva. The self-organising map is comprised of many empty neurons, which is due to the low sample numbers as neural networks by nature require significantly large sample numbers, which is not the case in this dataset. Despite this, the SOM was able to analyse the highly noisy spectral data and obtain good results when learning vector quantisation (LVQ) is applied (50 % accuracy on the training data without the learning vector quantisation applied to the trained map and 100 % accuracy when the LVQ is applied to the model) achieving a 10-fold cross-validation score of 0.75. Further details on how the SOM operates can be found in [33].

Compared with the vibrational spectra of single molecules, the saliva data is significantly noisier. Fig. 4 shows a significantly different performance for the Bayesian and RF classifier models on the relatively noisy data. In Fig. 4(a) LN–CASS was shown as the only classifier to not have any of the predicted samples proceed past the default decision boundary. In the other approaches, results indicate significant inaccuracy in determining the biological donor sex. Fig. 4(b) and (c) reveal at least 9 incorrect classifications for LASSO and 6 for RF. Despite these misclassifications, LN–CASS has predicted many of the samples close to the decision boundary therefore, obtaining the worst MAE score among the three methods (0.375). LASSO and RF methods acquired MAE values of 0.274 and 0.351, respectively. Nevertheless, it must be reinforced that the MAE value is not a useful metric for comparing the classifier performance but rather a comment on the computational modelling accuracy of predicted values. AUC is by far the more relevant statistic for the comparison between the classifiers. AUC values for LN–CASS, LASSO and RF were 0.996, 0.90 and 0.98, respectively. As presented in Fig. 4 (e), such comparison is omitted from the methods comparison using the simulated datasets as all tests showed AUC value of 1.00. Whilst these approaches may appear relatively high compared to the presented SOM, it must be stressed that the dataset examined in this study is considerably smaller than expected for applications of neural networks and hence, the lower cross-validation accuracy scores. This could be explained by the fact that ANN methods are significantly more data hungry than Bayesian modelling approaches, the latter of which is best suited to small or incomplete datasets [34]. One issue which remains for Bayesian modelling approaches is the sampling of the posterior distribution as one of the major rates determining steps in analysis. For the LN–CASS sampler, a Markov Chain Monte Carlo algorithm was employed and therefore, took longer among the tested classifiers when higher parallel chains are implemented. Despite this issue, on the timescale for clinical diagnostics, this extra time requirement of a few minutes would be negligible in practice. In future applications of the Bayesian classifiers, additional clustering approach could potentially benefit the LN–CASS approach to improve the MAE scores [35].

5. Conclusions

This communication presents the application of a Bayesian shrinkage prior in modelling the classification of vibrational spectra datasets. Parameter inference based on grouped LN–CASS prior around the regression coefficients in the simulated coefficients study, show that even at high parameters-to-observations ratios of 5:1, this approach was able to achieve an AUC of 0.98 with consistently the lowest mean absolute error values. Given the level of noise that was introduced to the simulated dataset to produce unique Raman spectra is at the same expected levels when measuring pure chemicals in vibrational spectroscopy, this AUC would be representative for those types of data. Simulated Raman and IR spectra modelling of small molecules of glucose and sucrose for low sample numbers (10 observations per class), *via* the Bayesian and Random Forest approaches can be accurately classified. However, when a biologically complex human saliva Raman dataset is analyzed, the LN–CASS prior was most capable of classifying the donor sex with 100 % accuracy and an AUC of 1.00 despite obtaining the greatest MAE value for that dataset among the tested machine learning approaches. Nevertheless, it should not detract from the achievements as a classifier since the MAE is not a statistic used for determining the performance of classifiers but rather the computational performance in predicted expected values. Since the dataset was sufficiently small to cause issues (lots of empty neurons in the self-organizing maps) for the ANN approach, a future application of the LN–CASS prior would potentially be to study different disease states based on the vibrational spectra, for the early-stage exploratory biomedical studies that involve low sample numbers.

Turning to the development and evaluation of diagnostic and prognostic models for clinical application, we note that penalization is not a 'magic bullet' to remove any difficulties associated with small sample sizes. For example, a recent systematic review of machine learning models in oncology found that there rarely is an explanation for the sample sizes used [36]. While penalization reduces overfitting, its use can incur increased variability in out-of-sample predictive performance, thus, it will be critical to assess this variability including the variability with tuning parameters carefully [37]. Thus, LN–CASS does not remove the need for adequate sample sizes for later stage model development and assessment in the clinical setting. Nevertheless, it is a valuable method for early-stage research and identification of predictive features for subsequent assessment and clinical translation.

Data accessibility

All code and data used in this study is available on GitHub. https://github.com/hinonchu/LN–CASS_for_vib-spectrsocopy. Alternatively, the code and data can also be accessed via the DOI (https://doi.org/10.5281/zenodo.8028655).

CRediT authorship contribution statement

Hin On Chu: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Emma Buchan:** Data curation, Investigation, Writing – original draft. **David Smith:** Data curation, Formal analysis, Methodology, Writing – review & editing. **Pola Goldberg Oppenheimer:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge funding from the Wellcome Trust (174ISSFPP) and the EPSRC (EP/W004593/1 and EP/V029983/1).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2024.108014.

References

[1] A.R. Stevens, C.A. Stickland, G. Harris, Z. Ahmed, P. Goldberg Oppenheimer, A. Belli, W. Huang, S.A. An, B.C. Shyu, M.S. Lin, et al., Raman spectroscopy as a neuromonitoring tool in traumatic brain injury: a systematic review and clinical perspectives, Cells 11 (2022) 1227, https://doi.org/10.3390/CELLS11071227, 2022Page.

[2] J.J.S. Rickard, V. Di-Pietro, D.J. Smith, D.J. Davies, A. Belli, P. Goldberg Oppenheimer, P.G. Oppenheimer, Rapid optofluidic detection of biomarkers for traumatic brain injury via surface-enhanced Raman spectroscopy, Nat. Biomed. Eng. 4 (2020) 610–623, https://doi.org/10.1038/s41551-019-0510-4.

[3] E. Buchan, L. Kelleher, M. Clancy, J.J. Stanley Rickard, P.G. Oppenheimer, Spectroscopic molecular-fingerprint profiling of saliva, Anal. Chim. Acta 1185 (2021) 339074, https://doi.org/10.1016/J.ACA.2021.339074.

[4] A. Logan, A. Belli, C. Banbury, E.R. Zanier, G. Vegliante, I. Styles, N. Eisenstein, P. G. Oppenheimer, Spectroscopic detection of traumatic brain injury severity and biochemistry from the retina, Biomed. Opt. Express 11 (11) (2020) 6249–6261, https://doi.org/10.1364/BOE.399473.

[5] L. Sun, J. Irudayaraj, Quantitative surface-enhanced raman for gene expression estimation, Biophys. J. 96 (2009) 4709–4716, https://doi.org/10.1016/j.bpj.2009.03.021.

[6] S. Uskoković-Marković, V. Kuntić, D. Bajuk-Bogdanović, I. Holclajtner-Antunović, Surface-enhanced Raman scattering (SERS) biochemical applications. Encyclopedia of Spectroscopy and Spectrometry, Third Edition, 2017, pp. 383–388.

[7] E. Lemoine, F. Dallaire, R. Yadav, R. Agarwal, S. Kadoury, D. Trudel, M.C. Guiot, K. Petrecca, F. Leblond, Feature engineering applied to intraoperative *in vivo* Raman spectroscopy sheds light on molecular processes in brain cancer: a retrospective study of 65 patients, Analyst 144 (2019) 6517–6532, https://doi.org/10.1039/c9an01144g.

[8] R. Kothari, V. Jones, D. Mena, V. Bermudez Reyes, Y. Shon, J.P. Smith, D. Schmolze, P.D. Cha, L. Lai, Y. Fong, et al., Raman spectroscopy and artificial intelligence to predict the bayesian probability of breast cancer, Sci. Rep. (2021) 11, https://doi.org/10.1038/s41598-021-85758-6.

[9] L. Guerrini, E. Garcia-Rico, A. O'loghlen, V. Giannini, R.A. Alvarez-Puebla, Surface-enhanced Raman scattering (SERS) spectroscopy for sensing and characterization of exosomes in cancer diagnosis, Cancers 13 (2021) 2179, https://doi.org/10.3390/CANCERS13092179, 2021Page.

[10] Y. Zhou, C. Liu, J. Li, L. Zhou, J. He, Y. Sun, Y. Pu, K. Zhu, Y. Liu, Q. Li, R.R. Alfano, S.G. Demos, et al., Resonance Raman spectroscopy for human cancer detection of key molecules with clinical diagnosis, in: Proceedings of the Optical Biopsy XI 8577, 2013.

[11] K. Shahzad, H. Nawaz, M.I. Majeed, R. Nazish, N. Rashid, A. Tariq, S. Shakeel, A. Shahzadi, S. Yousaf, N. Yaqoob, et al., Classification of tuberculosis by surface-enhanced Raman spectroscopy (SERS) with principal component analysis (pca) and partial least squares–discriminant analysis (PLS-DA), Anal. Lett. 55 (2022) 1731–1744, https://doi.org/10.1080/00032719.2021.2024218/SUPPL_FILE/LANL_A_2024218_SM2597.DOCX.

[12] R. Ullah, S. Khan, I.I. Chaudhary, S. Shahzad, H. Ali, M. Bilal, Cost effective and efficient screening of tuberculosis disease with Raman spectroscopy and machine learning algorithms, Photodiagn. Photodyn. Ther. 32 (2020) 101963, https://doi.org/10.1016/J.PDPDT.2020.101963.

[13] W. Wu, S.S. Huang, X.D. Xie, C. Chen, Z.W. Yan, X.Y. Lv, Y.Y. Fan, C. Chen, F. L. Yue, B. Yang, Raman spectroscopy may allow rapid noninvasive screening of keratitis and conjunctivitis, Photodiagn. Photodyn. Ther. 37 (2022), https://doi.org/10.1016/j.pdpdt.2021.102689.

[14] X. Xie, C. Chen, T. Sun, G. Mamati, X. Wan, W. Zhang, R. Gao, F. Chen, W. Wu, Y. Fan, et al., Rapid, non-invasive screening of keratitis based on Raman spectroscopy combined with multivariate statistical analysis, Photodiagn. Photodyn. Ther 31 (2020) 101932, https://doi.org/10.1016/j.pdpdt.2020.101932.

[15] H.Y. Liang, X.L. Cheng, S.X. Dong, H.Y. Wang, E.T. Liu, Y.X. Ru, Y.H. Li, X.D. Kong, Y.D. Gao, Rapid and non-invasive discrimination of acute leukemia bone marrow supernatants by Raman spectroscopy and multivariate statistical analysis, J. Pharm. Biomed. Anal. (2022) 210, https://doi.org/10.1016/j.jpba.2021.114560.

[16] S. Bezeau, R. Graves, Statistical power and effect sizes of clinical neuropsychology research, J. Clin. Exp. Neuropsychol 23 (2001) 399–406, https://doi.org/10.1076/JCEN.23.3.399.1181.

[17] J.W. Osborne, A.B. Costello, Sample size and subject to item ratio in principal components analysis, Pract. Assess. Res. Eval. 9 (2019) 11, https://doi.org/10.7275/ktzq-jq66.

[18] L. Hatcher, N. O'Rourke, A Step-By-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling, SAS Institute, 2013. ISBN 1612903878.

[19] S.S. Shaukat, T.A. Rao, M.A. Khan, Impact of sample size on principal component analysis ordination of an environmental data set: effects on eigenstructure, Ekol. Bratisl. 35 (2016) 173–190, https://doi.org/10.1515/EKO-2016-0014.

[20] A. Alwosheel, S. van Cranenburgh, C.G. Chorus, Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis, J. Choice Model. 28 (2018) 167–182, https://doi.org/10.1016/J.JOCM.2018.07.002.

[21] M.M. Eichenlaub, J.G. Hattersley, M.C. Gannon, F.Q. Nuttall, N.A. Khovanova, Bayesian parameter estimation in the oral minimal model of glucose dynamics from non-fasting conditions using a new function of glucose appearance, Comput. Methods Programs Biomed. 200 (2021) 105911, https://doi.org/10.1016/J.CMPB.2020.105911.

[22] J.M. Ordovas, D. Rios-Insua, A. Santos-Lozano, A. Lucia, A. Torres, A. Kosgodagan, J.M. Camacho, A Bayesian network model for predicting cardiovascular risk, Comput. Methods Programs Biomed. 231 (2023) 107405, https://doi.org/10.1016/J.CMPB.2023.107405.

[23] I. Fornacon-Wood, H. Mistry, C. Johnson-Hart, C. Faivre-Finn, J.P.B. O'Connor, G. J. Price, Understanding the differences between bayesian and frequentist statistics, Int. J. Radiat. Oncol. 112 (2022) 1076–1082, https://doi.org/10.1016/J.IJROBP.2021.12.011.

[24] T. Bayes, LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S, Philos. Trans. R. Soc. Lond. 53 (1763) 370–418, https://doi.org/10.1098/rstl.1763.0053.

[25] K. Routray, G. Deo, Kinetic parameter estimation for a multiresponse nonlinear reaction model, AIChE J. 51 (2005) 1733–1746, https://doi.org/10.1002/AIC.10446.

[26] C. Krafft, G. Steiner, C. Beleites, R. Salzer, Disease recognition by infrared and Raman spectroscopy, J. Biophotonics 2 (2009) 13–28, https://doi.org/10.1002/jbio.200810024.

[27] W. Thomson, S. Jabbari, A.E. Taylor, W. Arlt, D.J. Smith, Simultaneous parameter estimation and variable selection via the logit-normal continuous analogue of the spike-and-slab prior, J. R. Soc. Interface 16 (2019) 20180572, https://doi.org/10.1098/rsif.2018.0572.

[28] F. Neese, F. Wennmohs, U. Becker, C. Riplinger, The ORCA quantum chemistry program package, J. Chem. Phys. 152 (2020) 224108, https://doi.org/10.1063/5.0004608.

[29] M.D. Hanwell, D.E. Curtis, D.C. Lonie, T. Vandermeerschd, E. Zurek, G. R. Hutchison, Avogadro: an advanced semantic chemical editor, visualization, and analysis platform, J. Cheminform. 4 (2012) 1–17, https://doi.org/10.1186/1758-2946-4-17/FIGURES/14.

[30] Stan Development Team {RStan}: The {R} Interface to {Stan} 2020.

[31] M.D. Hoffman, A. Gelman, The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, J. Mach. Learn. Res. 15 (2014) 1593–1623.

[32] GitHub - Willthomson1/RS-Interface-Code: Code for J Roy Soc Interface Paper Available online: https://github.com/willthomson1/RS-Interface-code (accessed on 15 May 2020).

[33] C. Banbury, R. Mason, I. Styles, N. Eisenstein, M. Clancy, A. Belli, A. Logan, P. Goldberg Oppenheimer, Development of the self optimising kohonen index network (SKiNET) for Raman spectroscopy based detection of anatomical eye tissue, Sci. Rep. 9 (2019) 10812, https://doi.org/10.1038/s41598-019-47205-5.

[34] P. Kokol, M. Kokol, S. Zagoranski, Machine learning on small size samples: a synthetic knowledge synthesis, Sci. Prog. 105 (2022) 1–16, https://doi.org/10.1177/00368504211029777/ASSET/IMAGES/LARGE/10.1177_00368504211029777-FIG2.JPEG.

[35] F. Sağlam, E. Yıldırım, M.A. Cengiz, Clustered Bayesian classification for within-class separation, Expert Syst. Appl. 208 (2022) 118152, https://doi.org/10.1016/J.ESWA.2022.118152.

[36] P. Dhiman, J. Ma, C.L. Andaur Navarro, B. Speich, G. Bullock, J.A.A. Damen, L. Hooft, S. Kirtley, R.D. Riley, B. Van Calster, et al., Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review, BMC Med. Res. Methodol. 22 (2022) 1–16, https://doi.org/10.1186/S12874-022-01577-X, 2022 221.

[37] G.P. Martin, R.D. Riley, G.S. Collins, M. Sperrin, Developing clinical prediction models when adhering to minimum sample size recommendations: the importance of quantifying bootstrap variability in tuning parameters and predictive performance, Stat. Methods Med. Res. 30 (2021) 2545–2561, https://doi.org/10.1177/09622802211046388/ASSET/IMAGES/LARGE/10.1177_09622802211046388-FIG2.JPEG.