Research Article

Journal of RAMAN SPECTROSCOPY

Received: 12 May 2010 Accepted: 6 August 2010 Published online in Wiley Online Library:

(wileyonlinelibrary.com) DOI 10.1002/jrs.2798

Assessment of robustness and transferability of classification models built for cancer diagnostics using Raman spectroscopy

Martina Sattlecker,[a] Nick Stone,[b] Jennifer Smith[b] and Conrad Bessant[a]*

Over recent years, Raman spectroscopy has been demonstrated as a prospective tool for application in cancer diagnostics. The use of Raman spectroscopy for this purpose relies on pattern recognition methods that have been developed to perform well on data achieved under laboratory conditions. However, the application of Raman spectroscopy as a routine clinical tool is likely to result in imperfect data due to instrument-to-instrument variation. Such corruption to the pure tissue spectral data is expected to negatively impact the classification performance of the diagnostic model. In this paper, we present a thorough assessment of the robustness of the Raman approach. This was achieved by perturbing a set of spectra in different ways, including various linear shifts, nonlinear shifts and random noise and using previously optimised classification models to predict the class membership of each spectrum in a testing set. The loss of predictive power with increased corruption was used to calculate a score, which allows an easy comparison of the model robustness. For this approach, three different types of classification models, including linear discriminant analysis (LDA), partial least square discriminant analysis (PLS-DA) and support vector machine (SVM), built for lymph node diagnostics were the subject of the robustness testing. The results showed that a linear perturbation had the highest impact on the performance of all classification models. Among all linear corruption methods, a gradient *y*-shift resulted in the highest performance loss. Thus, the factor most likely to affect the predictive outcome of models when using different systems is a gradient *y*-shift.

Supporting information may be found in the online version of this article.

Keywords: robustness; classification models; cancer diagnostics; Raman spectroscopy

Introduction

Cancer can be considered an epidemic since the number of incidences is rising rapidly. In the United Kingdom alone, one in four deaths is due to cancer.[1] The highest impact on the mortality rate is the stage of the cancer at the point of detection. Since an early detection improves the survival rate, it is desirable to detect the development of malignant abnormal cell growth as soon as possible, preferably when the first biochemical changes occur. Current methods, including white light spectroscopy, often can detect a tumour only when invasion has already taken place. Hence there is a need for diagnostic methods that can detect cancer development at a pre-malignant level.

Currently, histopathology is the gold standard technique for cancer diagnosis and staging. Typically, tissue samples are taken and examined by pathologists using various staining techniques. This procedure not only delays diagnostic results but it also relies upon a subjective method, which can result in inter-observer disagreement.[2,3] Furthermore, excisional biopsy of vulnerable organs, including the central nervous system and vascular system, can be of increased hazard.[4] In light of these limitations, an ideal diagnostic test would be rapid, non-invasive and high-throughput, and would not require any tissue processing before analysis.

Vibrational spectroscopic methods, such as Raman and infrared spectroscopy have shown promise as techniques to aid histopathologists in the procedure of cancer detection and staging. These methods are capable of measuring subtle biochemical changes in tissue within malignant disease development. This feature makes these techniques highly suitable for early cancer detection, especially because the detection can take place as soon as the first chemical changes occur, which would not be detectable by traditional methods. Additionally, these methods are fast, high throughput, objective and non-destructive.

The future application of vibrational spectroscopy as a routine technique for cancer diagnosis strongly depends on chemometric pattern recognition techniques. For this purpose, various methods, including linear discriminant analysis (LDA),[5–7] artificial neural networks (ANNs),[8,9] random forests[10] and support vector machines (SVM),[11,12] were investigated to build classification models for cancer diagnostics. For instance Teh *et al*.[7] applied LDA for diagnosing gastric cancer and achieved a sensitivity of 95.2% and a specificity of 90.9%. ANNs were applied for the diagnoses of melanoma and achieved 85% sensitivity and 99% specificity.[8] Similarly, SVMs were investigated for the classification of colonic tissue, where normal tissue, polyps and adenocarcinomatous colonic tissue were classified with a diagnostic accuracy of 99.9%.[11]

If pattern recognition based on Raman spectroscopy is to be translated from the research laboratory to the clinic, its real world

* *Correspondence to: Conrad Bessant, Cranfield University, College Road, Cranfield, Bedfordshire, MK43 0AL, UK. E-mail: c.bessant@cranfield.ac.uk*

a *Cranfield University, College Road, Cranfield, Bedfordshire, MK43 0AL, UK*

b *Biophotonics Research Group, Gloucestershire Royal Hospital, Great Western Road, Gloucester, GL1 3NN, UK*

performance and limitations need to be fully understood. This necessitates rigorous model testing, which goes further than testing a classification model with an unseen testing set. To date, no studies into the robustness of these models and how their performance is impacted by different error sources have been published. This is of special interest since error can be introduced by system-to-system variation, working conditions and by the operator. In order to fully assess the performance of a diagnostic model, it is important to take account of the fact that acquired data might be subject to error from a range of sources.

In this work, we present a series of methods to simulate the effect of error sources on the data set. This includes various impacts, such as linear spectral shifts – which either shift the whole spectra by a wavenumber at a time or modify the intensity of each spectral point linearly; nonlinear spectral shifts – which either result in a stretching or a bending of a spectrum; and random noise – which decreases the signal-to-noise ratio. These are all potential errors that are suspected to occur to differing extents when moving between instruments, modifying the design and changing working conditions or sampling methodologies. In order to assess the robustness of classification models for such errors, classifiers were used to predict the class membership of the corrupted data. For this approach, three different types of classifiers – LDA, partial least square discriminant analysis (PLS-DA) and SVM – were investigated. In previous work,[13] these classification techniques were used to build models for lymph node diagnostics based on Raman spectroscopy. All of these models when independently tested achieved a classification accuracy, specificity and sensitivity beyond 90%. The SVM model even classified 100% of the testing set correctly.

Materials and Methods

Tissue samples and measurement

A total of 43 lymph nodes were collected during sentinel lymph node dissection of breast cancer patients. All samples were approved by the Gloucestershire Research Ethics Committee and used with the consent of fully informed patients. After collection, tissue was snap frozen on acetate paper in order to maintain orientation and sample freshness. Each sample was cut in half, where one half was used for comparative histology [haematoxylin and eosin (H&E) staining], a 7-µm section from the remaining half was cut and mounted on a CaF_2 slide for Raman mapping.

For all Raman measurements a Renishaw System 1000® Raman microspectrometer coupled to a diode laser, a Leica® microscope, a Prior® electronic stage, a video viewer and a desktop computer with customized Grams® software was used. The output of the laser was set to 350 mW and the wavelength to 830 nm in order to minimise the autofluorescence from tissue. Raman mapping was executed in steps of 100 µm in the *x*- and *y*-directions, where, at each point, the spectrum was integrated for a total of 30 s.

Data processing

All data analysis was executed using Matlab (Mathworks, USA) and additional toolboxes, including PLS Toolbox 3.5 (Eigenvector Research) and LIBSVM 2.88.1.[14] The generated Raman maps were first imported into Matlab and converted into false colour images by using the first three principal components (PCs). Using the false colour images and the related H&E staining, homogenous positive or homogenous negative regions were identified for subsequent extraction of spectra. The histopathology classification was confirmed by routine H&E staining and the expert opinion of a consultant histopathologist. In this manner, for each individual sample, multiple spectra were extracted. The resulting data was pre-processed by applying a filtering method that removed outliers and bad-quality spectra for each lymph node sample independently.[13] For this investigation, the removal of bad-quality spectra was of specific importance in order to ensure that the classifiers were not trained to handle bad-quality data. Thus, this allows an accurate assessment of the impact of corruption on the model performance. Finally, the data set was split into a testing and a training set. The training set consisted of 31 samples (9 positive and 22 negative) and the testing set of 12 samples (8 negative and 4 positive).

Diagnostic models

LDA is a frequently applied classification method owing to its simplicity. This classifier produces a linear boundary between classes. For the calculation of the LDA distance to each class, centroid Mahalanobis distance is commonly applied:

$$d_{ig}^2 = (x_i - \overline{x}_g)S_p^{-1}(x_i - \overline{x}_g)' \quad (1)$$

Quite often, principal component analysis (PCA) is executed prior to building an LDA model. The resulting PC scores are then used to generate the LDA model. Using PCs allows simplification of the data by maintaining the overall information content despite using fewer variables. Using a reduced data set is of special importance if the observed data has a higher number of variables than the number of samples because of the fact that Mahalanobis distance fails under these circumstances. In this manner, the optimisation of the LDA model includes the estimation of the ideal number of PCs fed into the LDA. This is commonly done by leave one sample out cross validation (LOOCV), where one sample is left out and the remaining data is used to build a model, which is then used to predict the class membership of the left out sample.

Partial least squares (PLSs) have a long tradition in chemometrics. Similar to PCA, PLS is a data reduction method. The main difference between these two methods is that PLS tries to relate the two types of variables, in this case, the spectral data and the pathology class. Thus, PLS attempts to maximise the covariance between these two building blocks. In order to optimise a PLS-DA model, the number of components (latent variables) must be optimised.

Compared to the two methods presented earlier, SVMs[15] are a relatively new classification method. The basis of SVMs is to separate classes with a hyperplane by maximising the margin between them. For this reason, the measured data is plotted into an *N*-dimensional space (input space), which is, in the simplest approach, equivalent to the measured points in the spectrum. Frequently, classes are not separable in the input space. In order to overcome this problem, data can be plotted into a higher dimensional space (feature space) by a kernel function. There are different types of kernel function used for this purpose; nevertheless, the most frequently applied one is the radial basis function:

$$K(x_i, x_j) = \exp\frac{-||x_i - x_j||^2}{2\sigma^2} \quad (2)$$

In previous work,[13] the classification algorithms described above were used to build diagnostic models for lymph node classification. For the optimisation of these models, the data set was split

Table 1. Summary of all simulated spectral artefacts and potential experimental sources

Spectral artefact	Possible sources
X-shift	Ambient temperature change
–	Calibration error
Constant *y*-shift	Laser intensity variation
Gradient *y*-shift	Stray off-axis light entering the system
–	Ambient light
–	New signals from specimens (fluorescence)
Sine perturbation	Collected light not fully focussed onto charge-coupled device (CCD) detector
Cosine perturbation	Optical artefacts caused by vignetting in the spectrometer
Random Noise	Reduced exposure time
–	Low laser power

randomly into training and testing sets. The training set, consisting of 31 samples (1550 spectra), was used to optimise and build the models. The predictive power of each generated model was finally assessed with an independent testing set, consisting of 12 samples (355 spectra). Among the optimised models, the radial basis function SVM (applied parameters: $\sigma = 2^{-13.25}$, $C = 2^{10}$) performed best by classifying 100% of the testing set correctly. The LDA model (number of PCs = 13) predicted 93.8% (sensitivity: 100%, specificity: 91.9%) of the test data accurately and the PLS-DA model (number of latent variables = 8) predicted 95.2% (sensitivity: 100%, specificity: 93.8%) of the test data correctly. These diagnostic models were the subject of all subsequent investigation of robustness testing.

Simulation of spectral artefacts

In order to assess the model robustness, different types of perturbations were simulated on all spectra of the testing set. The training set was left unmodified – this reflects the possible deployment of a method trained in a control laboratory setting into a clinical setting where sources of error are harder to control. The original models were then used to predict the class membership of the corrupted testing set. For this approach, three general types of perturbation were investigated: linear shifts, nonlinear shifts and random noise. A list of the applied spectral artefacts and their causes is shown in Table 1. For each approach, the perturbation level was increased systematically. It is expected that increased corruption levels result in a loss of predictive power, allowing the assessment of robustness as a function of spectral quality. This allows the comparison of the different types of classifiers, which helps to decide on the type of classifier that would be more sensitive to a specific spectral artefact for a particular spectral data.

Linear shifts

For the investigation of the linear shift, three independent simulations were executed: a constant *x*-shift, a constant *y*-shift and a gradient *y*-shift.

X-shift: Raman shifts can be a result of changes in ambient temperature and poor calibration procedures. In order to evaluate the impact of a varying *x*-shift on the model performance, the first and the last 15 wavenumbers of the training data set were eliminated. The removal of these wavenumbers was necessary in order to gain room for shifting the data set. Thus, the original spectral range of the training set was reduced from 350–1850 to 365–1835 cm^{-1}. The reduced data set was finally used to generate the different types of classification models. The *x*-shift was simulated on the testing data by extracting an alternating spectral range of the data. For instance, an *x*-shift of −15 cm^{-1} was introduced by extracting the spectral range from 350–1820 cm^{-1}. The resulting testing set was then classified by the model.

Constant y-shift: A constant *y*-shift was introduced by adding 0.01 arbitrary intensity units at a time to the original measured intensity of all testing spectra. Thus, for each wavenumber, the intensity was consequently increased for 0.01 arbitrary units. This was executed 50 times up to an intensity increase of all spectra to a maximum of 0.5 arbitrary units.

Gradient y-shift: In order to simulate a linear gradient, a linear function was added to all spectra of the testing data. The gradient of this function was then increased by 0.0001 ranging from negative to positive gradients of 0.0013. The impact of a gradient of 0.0001 on a sample spectrum is illustrated in Fig. 1.

Nonlinear shift

Nonlinear shifts were simulated in two ways. The first was sine based and the second, cosine based. Accordingly, the two functions were used to manipulate all spectra of the testing set:

$$\text{Function 1}: \quad f(x) = a \times 0.5\,(1 + \cos(x)) \tag{3}$$

$$\text{Function 2}: \quad f(x) = a \times 0.5\,(1 + \sin(x - \pi/2)) \tag{4}$$

The impact of the perturbation function is regulated by the amplitude *a*, and for this reason, the amplitude was increased in steps of 0.1 starting from 0.1 up to 30 for both functions. In order to corrupt the data set, the resulting base function was interpolated on the testing data. As Fig. 1 illustrates, a cosine perturbation has a strong impact on the peripheral zones of the spectra. In comparison, a sine perturbation has a higher impact on the centre of a spectrum. However, for Raman measurements, a spectral stretching, as simulated by cosine perturbation, is more likely to occur than a bending, which is simulated by a sine perturbation.

Random noise

Random noise *n* was computed independently for every individual spectral point $s_{(i,j)}$ in the testing set $x_{(i,j)}$. The noise *n* can take any value between −1 and 1. In order to introduce a gradient, only a percentage *p* of the generated noise *n* was added to the original measured intensity:

$$x_{(i,j)} = s_{(i,j)} + s_{(i,j)} \times n \times p$$

In Fig. 1, the impact of the addition of 10% noise on the Raman spectra is illustrated. For each percentage level, 100 models were generated where, every time, a new noise simulation was made for the testing set. The repetitions were executed because a single repetition would not be representative owing to the random nature of the perturbation.

Robustness score

In order to provide a summary of the overall robustness of each classification model, a score was calculated. For this purpose, each

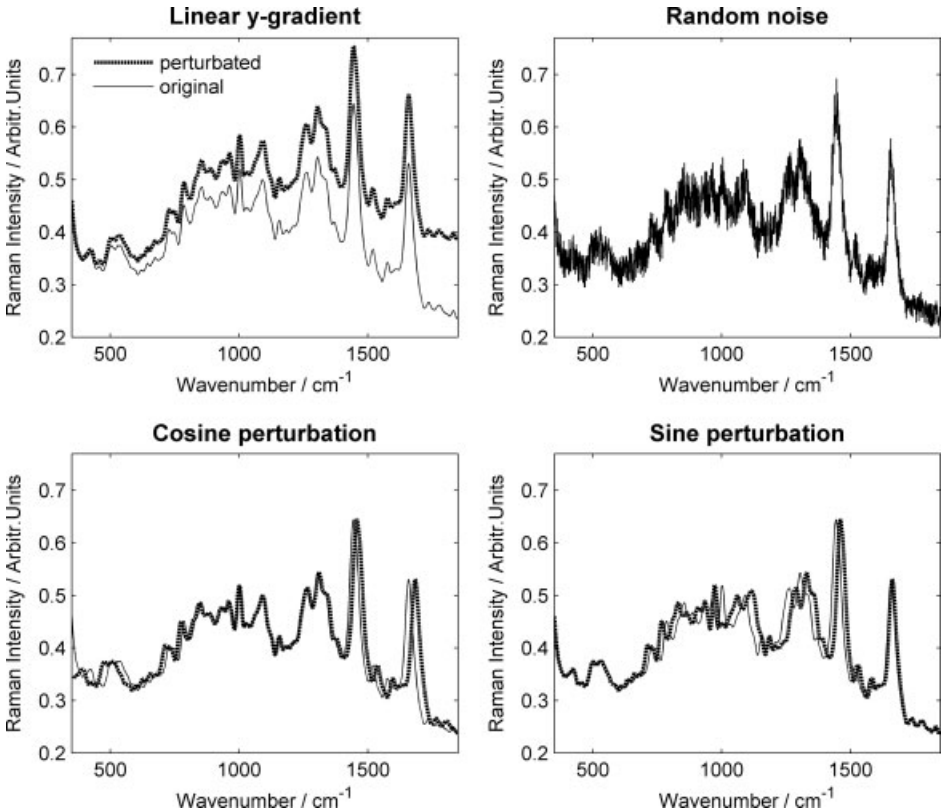

Figure 1. Illustration of the impact that different sources of perturbation have on a sample spectrum. A positive linear gradient, in this illustration, a gradient of 10^{-4}, results in increased intensities towards higher wavenumbers. The random noise plot shows the impact of the addition of 10% noise to the original spectrum. A cosine perturbation impacts the peripheral areas of the spectrum. The sine perturbation affects the centre of the spectrum. Both nonlinear perturbations were generated using the maximum amplitude.

corruption approach was assigned a total of 100 points. In the case that there was a two-way perturbation, for example, the *x*-shift could be either positive or negative, each was assigned 50 points, independently. In this manner, a total of 600 points was theoretically achievable. In addition, an assessment level was set on which the robustness was to be tested. For this approach, the level was initially set for maintenance of 90% accuracy. In this manner, it was possible to investigate how much the data could be corrupted by maintaining a minimum of 90% performance. For instance, in the case of an *x*-shift this was the maximum shift in *n* wavenumbers, which allowed a classification performance of 90%. In order to calculate the score, the proportion of the estimated perturbation limit to the applied maximum perturbation was estimated. This procedure was carried out for each corruption approach and all individual scores were summed. The higher the estimated score, the higher was the robustness of a model. Thus, the score facilitated a numerical comparison of the overall robustness of classification models at a predefined performance level.

Results and Discussion

Linear shift

X-shift

As illustrated in Fig. 2, a negative spectral wavenumber shift has a significantly higher impact on the classification performance than a shift in the positive direction. Among all classification models, the PLS-DA model was most badly affected. A shift of 15 wavenumbers in the negative direction resulted in a reduction to around 45% of prediction accuracy. In comparison, the SVM and the LDA model did not lose more than 20% in prediction accuracy at the maximum negative *x*-shift. Although all models declined in overall accuracy, the sensitivity, as illustrated in Fig. S3 (Supporting Information), was not impacted by a negative *x*-shift. An *x*-shift in the positive direction had a strong impact on the diagnostic sensitivity of the PLS-DA and SVM model. A shift of 12 wavenumbers resulted in a complete loss of sensitivity for the PLS-DA model and a shift of 15 wavenumbers resulted in a diagnostic sensitivity as low as 1.2% for the SVM model. Overall, the LDA model demonstrated to be the most robust model in the presence of an *x*-shift. Further investigations showed that this is due to the fact that only a minimum number of PCs were fed into the LDA. Increasing the number of PCs resulted in a total loss of sensitivity and thus a similar performance loss as for the other classification models. Thus, the previous application of PCA for data reduction and the optimisation of the number of PCs fed into the LDA has a beneficial effect on this model and its robustness. The PLS-DA, which faced the highest performance loss caused by an *x*-shift, must be considered the least robust model for this type of perturbation.

Constant y-shift

As expected, this modification had a severe impact on the model performance, shown in Fig. 2, which mainly resulted in a loss of specificity. This source of perturbation did not impact the

Journal of RAMAN SPECTROSCOPY

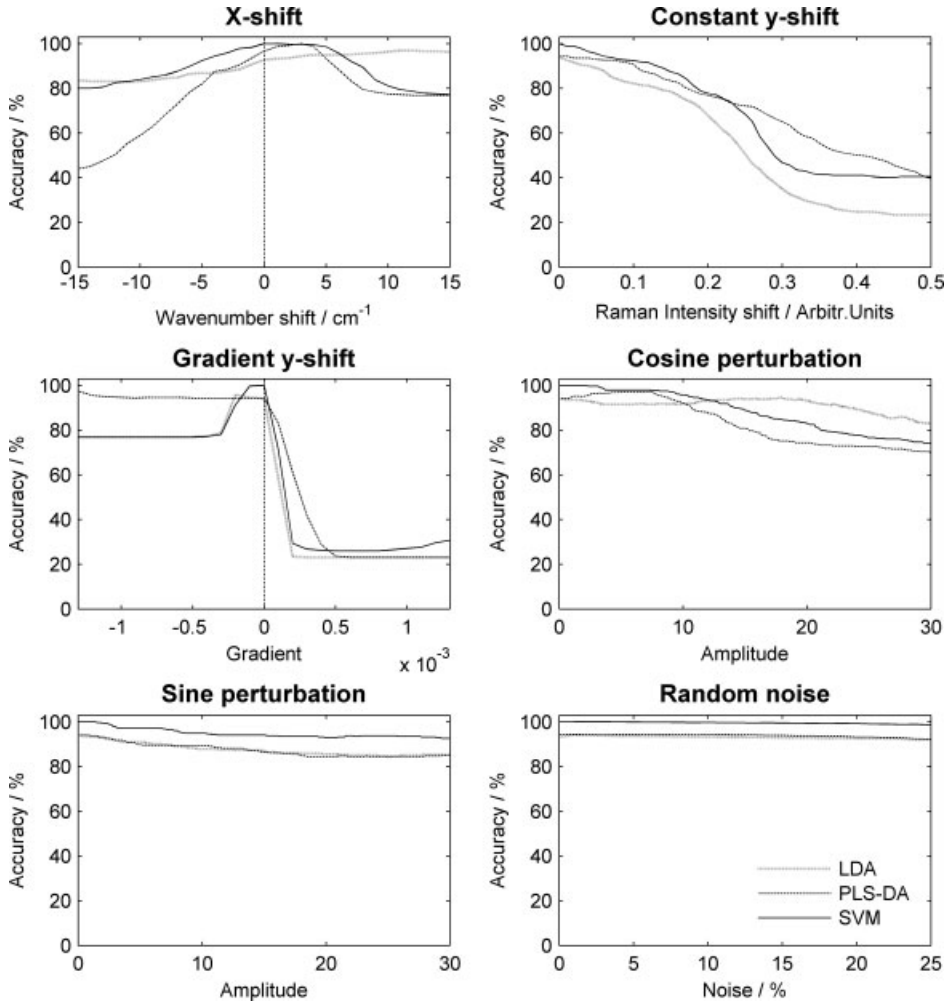

Figure 2. Robustness testing results for all perturbations. The largest loss in predictive accuracy can be seen for the three linear perturbations: *x*-shift, constant *y*-shift and gradient *y*-shift. Additionally, the cosine perturbation demonstrated having a fairly high impact on the classification performance. The sine perturbation and random noise only had a minimum impact on the classification result.

sensitivity of any of these models and therefore, any performance loss was caused by a reduced specificity. The PLS-DA model achieved a classification accuracy of 39.7%, which is equivalent to a specificity of 21.6%, after increasing the intensity of all test spectra by 0.5 arbitrary units. The SVM model performed similarly by achieving a specificity of 22.7% and classifying 40.6% of the testing set correctly. The decline in specificity is illustrated in Fig. S4 (Supporting Information). The major difference in the robustness of these models is that the SVM model loses performance abruptly, whereas the PLS-DA model loses performance more gradually. In comparison, the LDA model only achieved an accuracy of 23.4% and a total loss of specificity. For this reason, it must be considered that the LDA model is most affected by a constant *y*-shift.

Gradient y-shift

A positive gradient drastically impacts the performance of all models as illustrated in Fig. 2. The major reason for the performance decrease is the loss in diagnostic specificity. Thus, all normal lymph nodes were predicted as cancerous lymph nodes. For all models, as shown in Fig. S3 (Supporting Information), a sensitivity of 100% can be maintained up to a *y*-gradient of 0.0013. The most disturbed models were the LDA and the PLS-DA model, both of which only classified 23.1% of all testing spectra correctly with an applied gradient of 0.0013. The relatively low accuracy in comparison to 100% sensitivity can be explained by the fact that there are more negative than positive samples in the testing set. Nevertheless, the SVM model achieved a classification accuracy of 30.7% at the same gradient level. In comparison, a negative gradient results in a total loss of sensitivity for the LDA and the SVM model. The PLS-DA model is not impacted at all and thus maintains the original model performance up to the maximum gradient of −0.0013. Therefore, the PLS-DA can be considered to be the most stable classification model for this type of perturbation.

Nonlinear shift

A cosine perturbation, which has a corrupting effect on the peripheral regions of a spectrum, has a major impact on the model performance as illustrated in Fig. 2. For this perturbation source, it was observed that the sensitivity of all classification models was affected severely, as can be seen in Fig. S3 (Supporting Information). This source of perturbation can be compensated for by each of the classifiers up to a specific level, beyond which the sensitivity drops suddenly from 100% to lower than 40%. The PLS-DA model is the first to lose sensitivity at a cosine amplitude

of 16.8. The SVM model is capable of maintaining sensitivity up to an amplitude of 20.7, and the LDA classifier proved to be the most robust model by maintaining sensitivity up to a cosine amplitude of 30. In comparison, the sine perturbation, which impacts the centre of a spectrum, proved to have a minor impact on the performance of all classifiers. It showed that an applied sine amplitude of 30 does not result in a loss of sensitivity; finally, a sensitivity of 100% is maintained, as is shown in Fig. S3 (Supporting Information). The loss of specificity, illustrated in Fig. S4 (Supporting Information), is marginal for all classification models. The SVM model is the least impacted model for the reason that it maintains a specificity of 90.6%, which is equivalent to an accuracy of 92.7%, at the maximum level of sine perturbation. The LDA and the PLS-DA model perform equally and thus the LDA achieves 85.4% accuracy (81.0% specificity) and the PLS-DA 85.1% accuracy (80.6% specificity) at the maximum sine disruption level. On the basis of these results, the SVM model can be considered as the most robust one for sine perturbation.

Random noise

For each percentage level, noise was randomly added individually to every spectrum of the testing set. This procedure was repeated 100 times and the class membership was then predicted by the classifier. The average result was calculated for each noise level. The addition of random noise proved to have only a minor impact on the performance of all classification models, as illustrated in Fig. 2.

Although, the impact of the noise on to the spectra is high no major loss in classification performance could be observed. For all models, the sensitivity was not affected at all, and thus only a loss of specificity was observed.

Overall robustness

In order to assess the overall robustness for each classification model, a score was calculated, representing how much perturbation each model can compensate until the predictive accuracy drops below 90%. In Table 2, all scores for each individual perturbation source are summarised. The SVM model achieved the highest total score of all models. Nevertheless, it did not demonstrate the highest robustness for each individual perturbation source. The LDA model showed to be more robust towards a positive *x*-shift and cosine perturbation than the other classifiers. The PLS-DA demonstrated superiority in tolerating an increasing negative *y*-gradient. All three models showed almost no tolerance for a positive *y*-gradient. Summarising, the SVM model can be considered as the most robust model since it can cope with a high level of perturbation before dropping below 90% predictive accuracy.

Conclusion

In this paper, we have shown the extent to which various perturbations to Raman spectra would compromise diagnostic systems built around multivariate classification models. Linear perturbations were found to be the most disruptive. Among this group, it was found that a positive linear *y*-gradient had the strongest impact on the model performance. It was observed that even an extremely low positive linear gradient causes a drastic performance loss. Therefore, unexpected spectral features, such as stray light, fluorescence signals in new samples and ambient light signals might have the highest impact on the performance of classification models and, in conclusion, must be considered to be the most disrupting error source when applying Raman spectroscopy for routine diagnostics. Conversely, nonlinear perturbations were found to have negligible impact on the performance of the models. The same was observed for random noise. Since the major cause of random noise is reduced exposure times, these results demonstrate that a reduced exposure time would not impact on the model performance when constructed with high-quality data. This demonstrates that faster spectral measurements are feasible, which is of specific importance for *in vivo* measurements where the minimisation of acquisition times is desirable.

The overall robustness does not vary drastically between the different types of classification methods. Nevertheless, it was shown that each classification method had specific strengths. In relation to the other methods, LDA is less impacted by a positive *x*-shift or cosine perturbation. In comparison, PLS-DA copes better with a linear negative *y*-gradient and SVM with a sine perturbation and random noise. In real clinical use, the most likely differences between newly collected data and data used for training models would be small linear *x*-shifts and cosine shifts. The intensity-related changes can be corrected for by using normalisation methods and/or baseline subtraction. For this purpose, the most robust method would be LDA. However, since these types of

Table 2. Robustness scores for all classification models. Each single score was calculated on the basis of the maximum perturbation that can be tolerated by maintaining 90% of predictive accuracy

		LDA		PLS-DA		SVM	
	Max	Tolerance	Score	Tolerance	Score	Tolerance	Score
Pos. *X*-shift	15	15	50	6	20	8	27
Neg. *X*-shift	−15	−1	3	−2	7	−6	20
Const. *Y*-shift	0.5	0.05	10	0.11	22	0.14	28
Pos. *Y*-gradient	0.0013	1.2×10^{-5}	0	3.8×10^{-5}	2	3.5×10^{-5}	1
Neg. *Y*-gradient	−0.0013	-2.34×10^{-4}	9	−0.0013	50	-2.09×10^{-4}	8
Cosine perturbation	30	23.3	78	10.7	36	14.1	47
Sine perturbation	30	7.0	23	4.7	16	30.0	100
Noise	25	25	100	25	100	25	100
Total score			273		253		331

Pos., positive; Neg., negative; Const., constant

corruption are expected to be small in real applications, the most suitable classification method is SVM. This is due to the fact that it not only achieved the best classification performance on the original data set but it was also not impacted by small *x*-shifts and cosine perturbation. SVM loses predictive power only at very high *x*-shifts and under substantive cosine perturbation. In order to further increase the robustness of the SVM model, it would be required to incorporate imperfect spectra (ideally, from different instruments) into the training data, such that the expected variance is captured in the model. Finally, before attempting to classify spectra, it would be advisable to apply noise reduction methods that, for example, remove fluorescence background.

Acknowledgements

This research was financially supported by Cranfield University and Gloucestershire Hospitals NHS Foundation Trust. Furthermore, Nick Stone is funded by a National Institute of Health Research Senior Research Fellowship. We gratefully thank the technical and administrative staff of the Department of Histopathology, Gloucestershire Royal Hospital, and the medical and administrative staff in the Department of Breast Surgery, Gloucestershire Royal Hospital for their contribution.

Supporting information

Supporting information may be found in the online version of this article.

References

[1] Cancer Research UK, http://info.cancerresearchuk.org/cancerstats/mortality/cancerdeaths/ (accessed 20th May 2010).
[2] C. Kendall, N. Stone, N. Shepherd, K. Geboes, B. Warren, R. Bennett, H. Barr, *J. Pathol.* **2003**, *200*, 602.
[3] E. Montgomery, M. P. Bronner, J. R. Goldblum, J. K. Greenson, M. M. Haber, J. Hart, L. W. Lamps, G. Y. Lauwers, A. J. Lazenby, D. N. Lewin, M. E. Robert, A. Y. Toledano, Y. Shyr, K. Washington, *Hum. Pathol.* **2001**, *32*, 368.
[4] C. Kendall, M. Isabelle, F. Bazant-Hegemark, J. Hutchings, L. Orr, J. Babrah, R. Baker, N. Stone, *Analyst* **2009**, *134*, 1029.
[5] B. W. de Jong, T. C. Schut, K. Maquelin, T. van der Kwast, C. H. Bangma, D. J. Kok, G. J. Puppels, *Anal. Chem.* **2006**, *78*, 7761.
[6] P. R. Jess, D. D. Smith, M. Mazilu, K. Dholakia, A. C. Riches, C. S. Herrington, *Int. J. Cancer* **2007**, *121*, 2723.
[7] S. K. Teh, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, *Br. J. Cancer* **2008**, *98*, 457.
[8] M. Gniadecka, P. A. Philipsen, S. Sigurdsson, S. Wessel, O. F. Nielsen, D. H. Christensen, J. Hercogova, K. Rossen, H. K. Thomsen, R. Gniadecki, L. K. Hansen, H. C. Wulf, *J. Invest. Dermatol.* **2004**, *122*, 443.
[9] S. Sigurdsson, P. A. Philipsen, L. K. Hansen, J. Larsen, M. Gniadecka, H. C. Wulf, *IEEE Trans. Biomed. Eng.* **2004**, *51*, 1784.
[10] S. K. Teh, W. Zheng, D. P. Lau, Z. Huang, *Analyst* **2009**, *134*, 1232.
[11] E. Widjaja, W. Zheng, Z. Huang, *Int. J. Oncol.* **2008**, *32*, 653.
[12] L. Yi, W. Zhi-Ning, L. Long-Jiang, L. Meng-Long, G. Ning, G. Yan-Zhi, *J. Raman Spectrosc.* **2010**, *41*, 142.
[13] M. Sattlecker, C. Bessant, J. Smith, N. Stone, *Analyst* **2010**, *135*, 895.
[14] C.-C. Chang, C.-J. Lin, LIBSVM : a library for support vector machines. **2001**, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.
[15] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer: New York; London, **1995**.