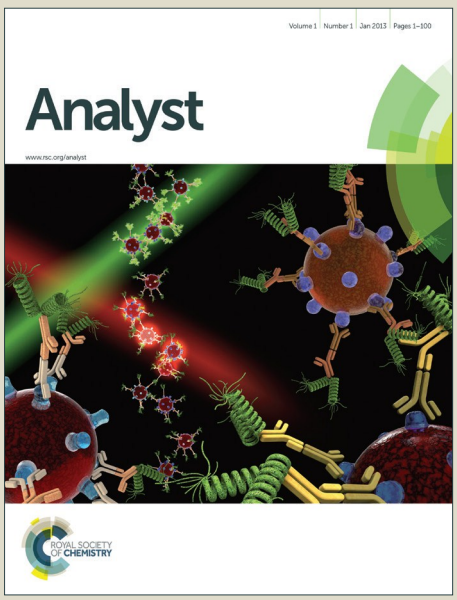
Analyst

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Analyst

RSCPublishing

Competitive evaluation of data mining algorithms for use in classification of leukocyte subtypes with Raman microspectroscopy

Competitive evaluation of data mining algorithms for use in classification of leukocyte subtypes with Raman microspectroscopy

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/analyst

A. Maguire[1,2], I. Vega-Carrascal[2], J. Bryant[2], L. White[2,3], O. Howe[2,3], F. M. Lyng. [1,2] and A. D. Meade.[1, 2*]

Raman microspectroscopy has been investigated for some time for use in label-free cell sorting devices. These approaches require coupling of the Raman spectrometer to complex data mining algorithms for identification of cellular subtypes such as the leukocyte subpopulations of lymphocytes and monocytes. In this study, three distinct multivariate classification approaches, (PCA-LDA, SVMs and Random Forests) are developed and tested on their ability to classify the cellular subtype in extracted peripheral blood mononuclear cells (T-cell lymphocytes from myeloid cells), and are evaluated in terms of their respective classification performance. A strategy for optimisation of each of the classification algorithm is presented with emphasis on reduction of model complexity in each of the algorithms. The relative classification performance and performance characteristics are highlighted, overall suggesting the radial basis function SVM as a robust option for classification of leukocytes with Raman microspectroscopy.

Introduction:

Raman spectroscopy has been used extensively in the analysis of various biological materials, with prevailing issues surrounding appropriate implementation and interpretation of data mining approaches[1–3]. As Raman spectroscopy provides a biochemical fingerprint of the sample and contains multiple overlapping vibrational signals from molecularly distinct biochemical species, spectral decomposition and data mining approaches are required to remove spectral redundancy and maximize the information extracted from the spectral data[4]. Examples of the successes of this approach are demonstrations of the ability of the method to perform classifications of different cell types[5,6] and the creation of diagnostic approaches distinguishing normal from cancer subtypes[7–9] for various cancers including those of the cervix[3,9,10], prostate [8,11], lung[12] and oesophagus[13,14]. Regression algorithms have also demonstrated the ability to predict metabolite concentrations in both blood cells and serum[15,16] delivering advantages in clinical medicine. The modality has also been shown to be capable of screening activated versus non-activated lymphocytes through identification of shifts in spectral bands associated with immunoglobin formation[17]. Coupling of Raman spectroscopy to micro-fluidic platforms and optical trapping has also demonstrated its potential for label-free cell sorting[18].

Development of these types of applications of Raman spectroscopy calls for robust and complex statistical methods to generate classification models with generalizability to unseen test sets. Various approaches are available employing algorithms which differ mainly in the configuration of the separation or classification hyperplane between the classes. Principal component analysis (PCA) -linear discriminant analysis (LDA) is one example of an algorithm which develops a linear classification hyperplane, where pre-processing by PCA is used for dimensionality reduction prior to input of spectral data to the algorithm[8,19,20].

Support vector machines (SVM) is a class of statistical learning algorithm which allows the development of both linear and non-linear classification hyperplanes[21]. A non-linear kernel mapping is applied to the input space in the special case of the development of non-linear classification hyperplanes, where the data points are remapped into feature space in which the data are linearly separable. It is here where the SVM then finds the best separating hyperplane for the classification[22]. Multiple kernel mappings are generally available and are evaluated separately[23].

Random forests are a non-linear classification approach which employ majority voting from the classification outcomes of each individual decision tree to reduce the classification error from any individual classifier. Decision trees are top down classification methods where each attribute of the sampled dataset is tested for its ability to discriminate between target variables. These attributes are ranked and the top ranking attributes are used for the initial decisions with the lower

ranked attributes used for decisions further down the tree. Tree nodes define 'splitting criteria' on which the classifier discriminates classes until finally all data records (spectra) are placed in leaf nodes representing their final class[24,25].

Classifications from PCA-LDA, SVMs and Random forests typically yield high predictability, although all require careful optimisation to prevent over fitting. The classification of haematological cell subtypes using Raman spectroscopy is challenging due to the overlapping nature of spectral band and the similarity in biochemical species seen in each leukocyte cell subclass. Although classification of haematological cell subtypes may be challenging, Bankapur *et al* showed that there were several Raman bands that differed in the spectral fingerprint of granulocytes and lymphocytes, and that could be used to discriminate between both populations of cells, while the spectrum of a red blood cell was drastically different from either white blood cell subtypes[6]. Ramoji *et al* demonstrated that it was possible to discriminate between lymphocytes and neutrophils using Raman spectroscopy coupled with PCA and Hierarchical cluster analysis[5]. Their model achieved an accuracy of 81% when applied to a single completely different donor in the testing set.

Creation of predictive models for the development of clinically relevant applications such as disease detection, diagnosis, estimation of metabolite concentration and identification of cellular subtypes must undergo rigorous procedures prior to acceptance of a technique's validity. Efforts have been made to standardise the procedure in which such applications are developed and validated to a clinical standard. Typically this procedure consists of two stages: exploratory studies and diagnosis studies and are extensively reviewed by Trevisan *et al* in [26]. Baker *et al* describe the development of FTIR spectroscopy for classification or diagnosis of biological materials, while detailing the performance of classification methods on FTIR spectral datasets[27]. The diagnosis of ovarian and endometrial cancers from patient plasma and serum using ATR-FTIR was described by Gajjar *et al*[28]. The authors performed an exhaustive search of classification methods for each cancer type and found that no single classification method performed consistently better across all diagnostic systems.

In this study of Raman spectral data from peripheral blood mononuclear cells (PBMCs) is used for the competitive evaluation of each data-mining model in discriminating a highly pure population of T-cell lymphocytes from other myeloid cells within the PBMCs fraction. The optimisation of each classifier (PCA-LDA, SVMs, and Random Forest) is demonstrated. The classification performance of each of the classifiers is discussed in terms of linearly and non-linearly separable data, with a view to illustrating the need for identifying appropriate classification methods for datasets that may not be linearly separable. The study is an exploratory study that demonstrates that there are fundamental differences in spectral features of myeloid cells and lymphocytes, which are more identifiable by some classification techniques than others. It is a preliminary study that highlights the potential of Raman spectroscopy along with multivariate techniques as a label free method of identification of PBMC subtypes.

Materials and methods:

Peripheral blood lymphocyte and myeloid cell isolation:

Ethics approval was awarded by the Dublin Institute of Technology ethics committee (2012) for the collection of blood donations from volunteers at the Institute for the purposes of this study. Fresh blood was drawn into Li-heparin tubes following consent from each healthy donor. A total of 20ml was collected from each donor. Peripheral blood mononuclear cells were isolated from each donor's whole blood by a density gradient using histopaque and was performed within 4 hours of initial collection. The PBMC layer was removed from the whole blood gradient and was washed three times. The cell pellet was resuspended in 3ml of full media (RPMI+12.5 %(v/v) FBS+2 mM L-glutamine (Sigma)) supplemented with 2.5% (v/v) phytohaemagglutinin (PAA Laboratories). One ml of cell suspension was resuspended in 4 ml of full media in a T25 flask and was incubated for 72 hours at 37°C, 5% CO_2 to allow separation of all other mononuclear cells by plastic adherence. T-cell lymphocytes were obtained from the cells that remained in suspension. Following removal of the lymphocytes from the T25 flasks, the flasks where rinsed in PBS. Cells where removed from the bottom of the flask by using a cell scrapper. These cells where then resuspended in fresh media prior to cell fixation. Population purity was tested for lymphocytes using CD3+ staining by flow cytometry. The Lymphocyte population was found to have a purity of > 85%.

Raman spectroscopic measurements:

Calcium fluoride (CaF_2, Crystran Ltd.) microscope slides were used for mounting of cells for Raman spectroscopy. All Raman spectral measurements were performed using a Horiba Jobin Yvon Labram HR800 UV system. Spectra were collected using a 660nm solid-state diode laser delivering 100mW of power to the sample, a x100 objective with a numerical aperture of 0.9. Spectral resolution was defined by the grating which was ruled with 300 lines/mm, resulting in a spectral resolution of ~2.1cm^{-1}. The confocal hole was set to 150µm and the spectra were recorded with a 20 second integration time averaged over three accumulations. Myeloid spectra were acquired from 7 different healthy donors and lymphocyte spectra were acquired from 14 different healthy donors. Spectra were recorded from each of 20-40 different cells per donor, with a total of 156 myeloid spectra and 463 lymphocyte spectra. Only 7 donors were acquired for myeloid cell spectra due to the difficulty of extraction and isolation of high concentration of myeloid cells from peripheral blood. Each spectrum was recorded by performing a 4x4µm raster scan of the centre of each cell. All cellular spectra from a single donor were recorded on the same day. Multiple spectra of 1, 4-Bis (2-methylstyryl) benzene and NIST SRM 2245 were recorded prior to each group of spectral measurements for calibration purposes. All spectra were recorded within two weeks of slide preparation and slides were stored in a desiccator prior to Raman spectral measurement.

Raman spectral measurement post processing:

Raman spectral post processing was performed in Matlab version 7.9.0 (R2009b) (Mathworks, USA) using the PLS-Toolbox version 6.51 (Eigenvector Research Inc.) and algorithms developed in-house. Spectral calibration was performed using a spectral alignment algorithm which fitted a polynomial to the peak positions of the peaks from the spectrum of 1, 4-Bis (2-methylstyryl) benzene relative to the peak positions of a common reference spectrum of the same material. Calibration of spectral intensity was performed similarly using the spectrum of the standard reference material SRM2245 relative to a common reference spectrum of the same material. Baseline correction was performed with in house algorithms using a nodal point baseline correction with the minimum amount of points required, for minimal spectral

alteration. Savitsky Golay filtering was employed with a 5th order polynomial and a 15 point window. Substrate contributions arising from CaF_2 were subtracted from all spectra and spectra were vector normalized prior to analysis.

Raman data analysis - PCA-LDA:

Principal component analysis is an unsupervised data reduction technique that is extensively used across many disciplines[29–31]. More importantly it is a feature selection process that allows the user to identify variances in the dataset that may be used to classify objects into certain groups. The application has become an important tool in chemometric and spectroscopic analysis.

In the case of Raman spectroscopy PCA is used to reduce the matrix of spectral data in which objects (individual spectra) are measurements of large numbers of variables (wavenumbers). PCA is performed by subtracting the mean of the data set to obtain the mean centered matrix, calculating the covariance matrix of the mean centred matrix and subsequently finding the eiganvectors and eigenvalues of the covariance matrix. The eigenvector with the largest eigenvalue is the first principal component which then describes the largest source of variance across all the spectra[32]. The second principal component is the eigenvector with the next largest eigenvalue, is independent of the first principal component and describes the second largest source of variance. All increasing principal components describe mutually independent sources of variance, and decreasing proportions of the spectral variance in the dataset. Typically in Raman spectroscopy of biological samples, the first 6-10 principal components describe over 99% of the variance or statistical information within the dataset, while beyond this point the principal components are generally noise. A matrix of spectra is decomposed into its scores and loadings according to:

$$\mathbf{X} = \mathbf{P^T T} \qquad \text{Equation 1}$$

where X is the original data set, P is the matrix of Principal components also known as loadings and T is a matrix of scores. Thus any spectrum in X can be reconstructed by the sum of the principal components weighted by the scores for each principal component calculated for each individual spectrum.

LDA is classification method that aims to find one or more linear functions of a dataset with x number of variables that can be used for the purpose of classification[32]. LDA produces a line or hyperplane that results in the maximum separation of two or more classes in a dataset. It has been used in many fields alongside PCA, where LDA uses the PCA scores as latent variables and tries to find the linear hyperplane that discriminates between two or more populations of PCA scores.

Raman data analysis - SVM-Linear and RBF kernel

Support vector machines are statistical learning algorithms that have seen use widely within classification and regression algorithms in data mining [33–35]. As classification algorithms, SVMs are designed to identify the hyperplane or hyperplanes that best separate two or more classes of multivariate data, while at the same time maximising the margin around the hyperplane. SVMs can also employ kernel mappings from a non-linear input space to a new feature space where the SVM searches for the best linear classification hyperplane. As a linear algorithm, the SVM uses the following equation:

$$\boldsymbol{f(x) = < w, x > + b \geq 0} \qquad \text{Equation 2}$$

where x is the input data (in the scope of this article x is spectral data), w is the weight vector and b is the bias. The SVM finds $f(x)$ (the hyperplane) that best discriminates between classes. The instances of x that lie closest to the discrimination hyperplane are called support vectors. There are two main types of SVMs, one which maximises the margin around the discrimination plane with the inclusion of a cost. The cost function allows for misclassification of some instances but incurs some penalty for the misclassification. This type of SV classifier is known as C-SVC. Another type of SVM employes a penalty defined to misclassifications defined by a parameter called ν. This parameter places an upper bound on the fraction of training samples that are misclassified and a lower bound on the fraction of training samples that are support vectors. Unlike linear discriminant analysis and other linear classifiers SVMs can be built to discriminate between both linear and non-linearly separable data. The use of kernel transforms on the input space, mapping the data to a new feature space can allow for discrimination of non-linearly separable data. There are however many forms of transforms and it is sometimes necessary to implement several transforms to identify which one is most capable of separating the data. Radial based functions, polynomials and sigmoid functions are typically applied to the input space prior to identifying the optimal classification hyperplane[23,33].

Raman data analysis - Random Forest

Random forest (RF) classification algorithms are an ensemble method whereby a model consisting of multiple independent decision trees is created. The consensus vote from all the decision trees is then the class determined by the RF algorithm, with overall reduced classification error relative to a single decision tree. Decision trees are a top down method where the tree chooses a series of attributes or variables on which to 'split' such that the class distribution after each node is skewed maximally (i.e. classes are separated). To identify the most important variables for the classification a quantity known as the information gain is used[24]. Information gain is the expected reduction of entropy caused by splitting the data based on a particular variable. Entropy is considered a measure of purity or impurity of a collection of samples. Alternatively an entropy of one represents a collection of samples with an equal number of samples in all classes. An entropy of 0 represents a collection of samples that consist of only a single class. The information gain is calculated for each attribute and the attribute that reduces the entropy (or provides the maximum information gain) is attribute that best classifies the data and is thus used for the initial decision[24]. Each node will continue to split until the entropy of the newly formed nodes become zero. A Random forest is built with multiple decision trees. Each record (spectra) is passed down all of the trees in the forest and the consensus of all the trees in the forest gives the predicted outcome of a particular record.

Model development and parameter optimisation

Careful choice of model training and testing strategy is critical to determining model performance and eliminating over fitting while at the same time preventing penalization of the model performance through supplying it with a small range of training samples [36] This latter point is a key consideration for modelling with small datasets. As the dataset available for modelling decreases in size from hundreds and thousands of examples to tens or less the appropriate model training and testing strategy moves from the holdout method to repeated cross-validation

and bootstrapping [36]. In the latter method it is assumed that the data samples are taken from a normal distribution and therefore could be observed again in the general population were they to be sampled. Therefore each of the donors or patients can be resampled in sequence for both the training and testing sets and the model performance is summarized over all individuals.

In this study, given the size of the dataset, repeated cross-validation and repeated bootstrapping are used. In the first instance, repeated bootstrapping was employed where each of the modelling methods were optimized separately using a training, validation and test set where the spectra were randomized such that resampling of the spectra occurred between each subset. The training data for each of the models was built using 60% of the total dataset while 20% of the data was used as a validation set with the remaining 20% used for testing. Classifications were performed a total of 10 times with randomised training, validation and testing data for each iteration. All classification metrics are averaged over all successive iterations. The parameters for each of the multivariate models were then optimized by choosing the parameters that resulted in the best Matthews correlation coefficient (MCC) for each of the classifications. The MCC is a measure of accuracy which uses a weighted combination of sensitivity and specificity and is suited to datasets with unbalanced class distributions, such as the one used here. Each model was optimised for its respective parameters (number of latent variables (PCA-LDA)) or combination of parameters (cost and γ (SVM), number of trees and number of leafs (RF)). In each case a 10-fold cross-validation was performed to identify the best performing model parameters.

Once a champion modelling approach was obtained from repeated bootstrapping, a second more rigorous evaluation of performance was obtained for the champion model using repeated 7-fold cross-validation. In this instance individuals were randomly sorted to training and testing sets while ensuring spectra from individuals were not resampled to each of the subsets. This process was repeated 10 times and the performance summarized.

PCA-LDA: Optimisation of the linear discriminant model was performed by firstly identifying latent variables from the principal component analysis that resulted in a positive MCC. LDA was performed on each of the principal components scores individually. Latent variables that were found to have an MCC of less than 0 were removed from the LDA classification. After removal of latent variables that did not contribute positively to the classification, LDA was performed on increasing numbers of latent variables. To reduce the complexity of the model, a 4th order polynomial was fitted to the validation set MCC and the second derivative of the polynomial was calculated. The point where the 2nd derivative was found to be zero was chosen as the number of latent variables to use for the classification. Beyond this point, the relative contribution of each additional latent variable to the classification accuracy decreases. The model used to predict the test set was built using the number of latent variables defined by where the 2nd derivative was equal to zero and was constructed using the training data.

SVM: In this article the SVMs that were optimised were the linear and RBF cost dependant SVMs. SVM optimisation was performed by employing a grid search of the penalty parameter C (cost) and the γ parameter. In the case of the linear SVM, γ was varied from $1x10^{-6}$ to 10 while cost was varied from $1x10^{-2}$ to $1x10^{8}$. In the case of the RBF SVM, γ was varied from $1x10^{-3}$ to $1x10^{4}$, while cost was varied from 1 to $1x10^{9}$. The MCC and the support vector (SV) fraction was calculated for each of the classifications and the combination of C and γ that resulted in the maximum difference between MCC and SV fraction was chosen for the model that was used to predict the test set. This resulted in the maximum classification accuracy while minimising the complexity of the model.

Random forest: Random forest optimisation was performed by optimizing training model using the validation set for the size of the leaves per node and the number of trees grown in the classification. Determination of the optimum size of the leaves was performed first, with a fixed number of trees grown (50). The MCC was calculated for the classification of each random forest with leaf nodes sizes of 1, 3, 5, 7, 10, 15, 20, 25, 35 and 50. The best performing leaf size for the validation set was chosen and fixed for the optimisation of the number of trees to be used in the optimisation. The best quality model for the validation set performance was chosen from a number of models, where 20, 30, 40, 50, 80, 100, 150, 200, 300, 400 and 500 trees were grown.

Results:

The mean and standard deviation of unprocessed and processed spectra of lymphocytes and myeloid cells are shown if Figure 1 A and B respectively. The difference spectra of lymphocytes and myeloid cells is plotted in Figure 1 C) and shaded regions represent the regions of the spectrum where the difference in spectra of was significantly different with a significance level of $p<0.001$. Darker regions represent where the spectral intensities were found to be significantly higher in lymphocytes than myeloid cells and lighter shaded regions represent where spectral intensities were significantly lower in lymphocytes than myeloid cells.

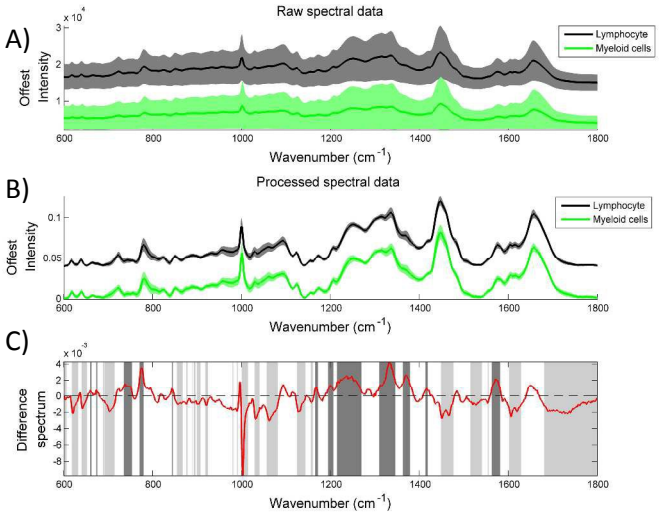


Figure 1: A) The mean and standard deviation of raw spectral data from donor lymphocytes and myeloid cells. B) The mean and standard deviation of processed spectra of lymphocytes and myeloid cells from donors. C) The difference spectrum of lymphocytes and myeloid cells (Shaded regions represent where lymphocytes had significantly higher (dark) or lower (light) spectral intensities).

1. PCA-LDA

Training models were optimized using the validation set performance as described in the methods section of this article,

Figure 2 (A) shows the training and validation set performance for increasing numbers of latent variables. Training and validation was performed on only positively contributing principal components scores. The black line illustrates the number of latent variables that resulted in the best performing training model while minimising the number of latent variables used to classify, tested on the validation set. The number of latent variables that produce the highest MCC while reducing the model complexity was found to be 31. The most accurate model for the validation set was then tested on new data (testing set). The MCC for the most accurate validation model was found to be 0.80. The classifications sensitivity, specificity and MCC for the test set are provided in Table 1. The classification performance was found to be relatively good with a sensitivity and specificity of 0.95 and 0.97 respectively. The MCC coefficient was found to be 0.88.

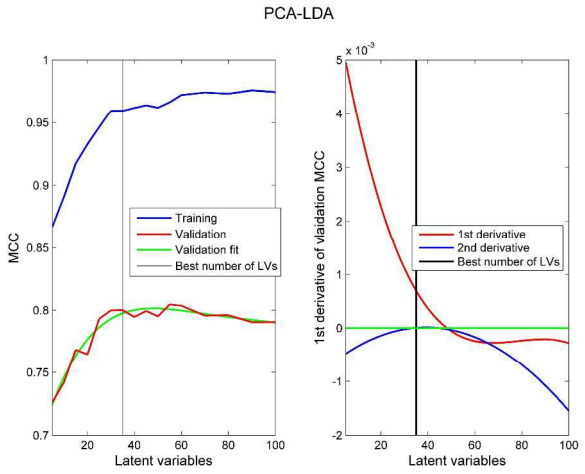


Figure 2: (A) Variation in MCC as a function of the number of latent variables used in the classification for training and validation sets. The green line represents a 4th order polynomial fit of the validation set performance. B) Shows the 1st and 2nd derivatives of the polynomial used to fit the MCC of the validation. The black vertical line illustrates the optimum number of latent variables for the validation set.

Table 1: Sensitivity, specificity and MCC along with their respective standard deviations for the final test set performance following optimization using the validation set.

PCA-LDA		
Sensitivity	Specificity	MCC
0.95±0.03	0.97±0.03	0.88±0.06

2. Random Forests

The random forest algorithm was optimized by using training and validation sets to find the optimal combination of the size of the leaf nodes in the trees of the random forest and number of trees grown in the model. The surface plot of the MCC value as a function of the number of trees grown and the number of leaves per tree for the training set is shown in Figure 3**Error! Reference source not found.** A) while the MCC as a function of number of leaves and the number of trees grown for the validation set are plotted in Figure 3 B). The combination that produced the highest MCC value in the validation set was 50 and 1 for number of trees grown and the number of leaves respectively, and is illustrated by the red dot in the plot. The MCC for the classification was found to be 0.73. These parameters were then used to grow a random forest from the training data and the model was tested on the newly seen test set data. The MCC for the classification of the test data was found to be 0.68. The confusion matrix along with the sensitivity, specificity and MCC are provided in Table 2.

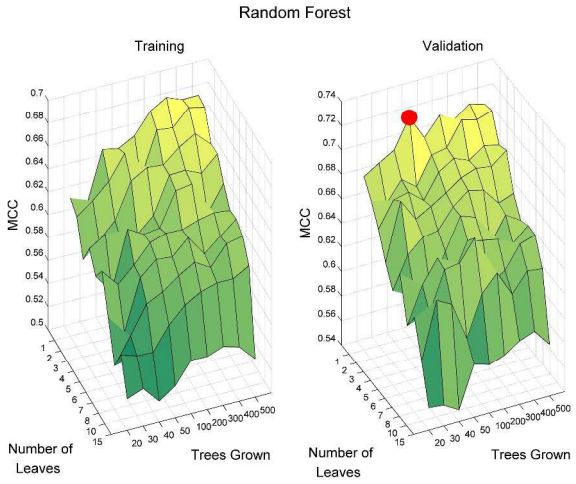


Figure 3: A) Shows the MCC as a function of the leaf size and the number of trees grown in the classification of the training set, B) shows the MCC as a function of leaf size and the number of trees grown in the validation set. The red dot represents the best performing combination of leaf number and number of trees grown for the validation set.

Table 2: Sensitivity, specificity and MCC along with their respective standard deviations for the final test set performance following optimization using the validation set for the random forest classification.

Random Forest		
Sensitivity	Specificity	MCC
0.97±0.01	0.74±0.10	0.68±0.08

3. SVM-(linear and radial based kernel functions)

Linear SVM (C-SVC)

The optimisation of the linear cost dependant SVM is shown in Figure 4. The MCC for each of the values of the γ cost parameter is plotted for both training and validation sets. The value of γ was varied from $1x10^{-6}$ to 10 in uniform log intervals and was found not to affect the outcome of the prediction of the linear SVM and thus was fixed at 0.0001. Cost was varied from $1x10^{-2}$ to $1x10^{8}$ similarly in uniform logarithmic intervals. The black vertical line in Figure 4 represents the value of cost that both maximises the MCC of the validation set and minimises the number of support vectors required by the SVM. The resulting value of the cost parameter was used for the final model, which was used to predict the test set. The value of cost that gave the best prediction in the validation set was $1x10^{7}$ and the SV fraction was 0.21. The MCC for the classification was 0.81. The resulting sensitivity, specificity and MCC for the classification of the test set are provided in Table 3. The MCC for the final test set was 0.84.

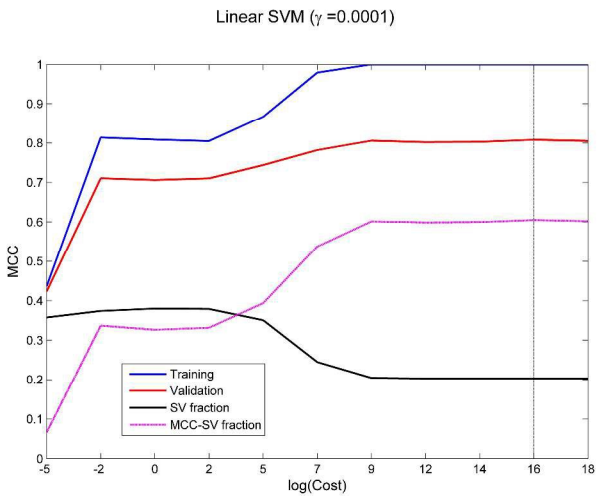


Figure 4: Variation in the MCC for both training and validation sets, the SV fraction and the difference in the MCC of the validation set and the SV fraction, as a function of the log of cost parameter. The black vertical line represents the cost parameter which maximized the performance of the SVM while minimizing the number of SVs required.

RBF SVM (C-SVC)

The cost dependant RBF function was optimised by performing a grid search to find the combination of γ and cost function values that resulted in the highest MCC for both training and validation sets. Figure 5 A) shows the MCC surface plot for the classifications using the training set, with varying values of γ and cost function. γ was varied from $1x10^{-3}$ to $1x10^{4}$, while cost was varied from 1 to $1x10^{9}$ in uniform log intervals. In Figure 5 (B) the surface plot of MCC as a function of cost and γ are shown for the validation set. The SV fraction as a function of cost and γ is plotted in Figure 5 (C) and in Figure 5 (D) the plot of the difference between the validation set MCC and the SV fraction, for each of the combinations of cost and γ is plotted. The highlighted red dots in Figure 5 (C) and (D) show the combination of cost and γ that result in the best performing SVM with the minimal amount of SVs required for the classification. Maximising the difference between the MCC and SV fraction reduces the complexity of the model and results in better performance of the SVM on new data. The combination of γ and cost parameters that resulted in the highest MCC and the minimal amount of SVs required in the validation set were used to build the SVM for the testing set. In this case the values of γ and cost were found to be 10 and $1x10^{4}$ respectively, giving an MCC of 0.92 and a SV fraction of 0.23 in the validation set. The resulting model was then tested on the test set and was found to have an MCC of 0.90. Table 3 shows the sensitivity, specificity and MCC for the final testing set.

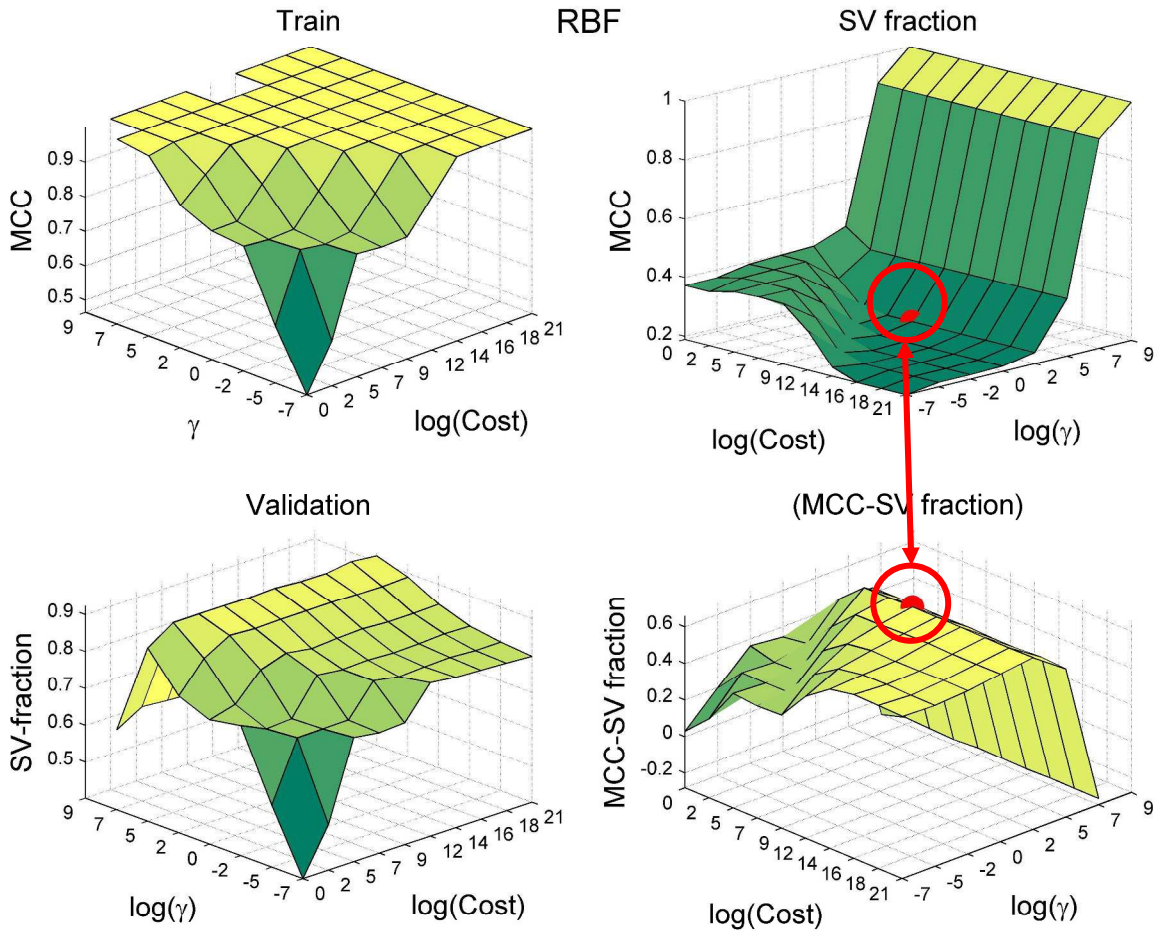


Figure 5: A) and B) Visualisation of the cost-dependant RBF SVM optimization for training and validation sets respectively, MCC is plotted as a function of the log of cost and γ parameter, C) and D) show support vector fraction and the difference between the validation set MCC and SV fraction, with respect to the log of cost and γ. The red dot highlighted in the plots of the SV fraction and the MCC-SV fraction shows the combination of cost and γ that resulted in the best performing SVM with the minimal amount of SVs required.

Table 3: Sensitivities, specificities and MCC along with their respective standard deviations for each of the SVM classification methods, linear and RBF cost dependent SVMs.

C-SVC linear SVM			C-SVC RBF SVM		
Sensitivity	Specificity	MCC	Sensitivity	Specificity	MCC
0.96±0.02	0.91±0.04	0.84±0.06	0.98±0.02	0.92±0.05	0.90±0.06

All of the modelling performances detailed thus far are for repeated boot-strapping of the dataset during training and evaluation. As mentioned in the methods repeated cross-validation was also performed on the best-performing model, the RBF-SVM, using repeated 7-fold cross-validation where donors were randomized for membership of each of the folds. The resulting MCC of the SVM was found to be 0.42±0.27, which corresponds to a specificity and sensitivity of 0.83±0.01 and 0.62±0.05 respectively. This more rigorous approach gives level of reassurance that the first model training and testing approach does not over fit and that the performance statistics are reflective of a performance which would be expected from each model type with a larger training and testing set.

Discussion

Raman spectroscopy has demonstrated its potential in hematology through its ability to discriminate between different cell subtypes and cellular responses to external factors, and further allowing the prediction of concentrations of metabolites found within the blood. In such instances the choice of model and optimisation strategy is key to the development of robust models. Within this consideration, it is critical to consider whether the data can be expected to be linearly or non-linearly separable when choosing a modelling algorithm or approach. Optimisation should then proceed to maximise modelling accuracy while minimizing model complexity and maximizing robustness on unseen testing sets.

The present article demonstrates this approach for three distinct model types; PCA-LDA, SVMs and Random Forest classifiers, applied to a challenging classification problem of subclassification of leukocytes taken from the blood of a population of volunteers. Each model has been optimised on the same data sets and the method of optimisation of each of the models has been presented. All models performed relatively well with MCCs above 0.65 for the test set data.

The performance of the SVMs was found to be the champion out of all three model types with the RBF SVM producing the model with the best classification performance (MCC=0.90). The Random forest classifier performed the worst out of all three classifications resulting in an MCC for the test set of 0.68, with a sensitivity and specificity of 0.97 and 0.74 respectively. The random forest specificity was quite low (0.74) indicating a bias in the classification, where 99% of all lymphocytes were classified correctly but 35% of cells, that were of a different origin, were classified incorrectly.

As the PCA-LDA classifier methodology is somewhat similar to a simplified SVM (linear) and unsurprisingly its validation set performance was similarly high at an MCC of 0.80 similarly to the validation set performance of the linear SVM (MCC=0.80). However the optimal classification for the PCA-LDA classification resulted in a highly complex model using a total of 31 latent variables to perform the classification. The power of PCA-LDA applied to Raman spectroscopy is in allowing the modeller to enquire as to the spectral variables giving origin to the classification through the principal components chosen in the model. However, in a model with such a high level of complexity, although robustly accurate, the advantage of visualisation of the spectral variables disappears.

The SVM classifiers all performed relatively well on the test data in comparison to the random forest classifier. The linear and RBF SVMs support vector fractions used in the test set performance were found to be 0.21 and 0.23 respectively. The RBF SVM performed slightly better than its counterpart linear SVM. This suggests that the data is somewhat non-linearly separable and that the discriminating hyperplane is not strictly linear. Figure 5 (D) shows the importance of increasing the γ-parameter in the RBF SVM, where γ effectively determines the flexibility of the hyperplane. Although this article is meant as an exploratory study, it demonstrates that the choice of multivariate model and the optimisation of that model, is critical to the development of robust, generalizable prediction models based on Raman spectral data. Models should suit the classification problem, providing flexibility in adapting to the dataset and the separation hyperplane and minimizing model complexity. The study demonstrates that there is a fundamental difference in the spectral features of myeloid cells and lymphocytes. Further visualisation of the origin of the classification from the perspective of the spectral variables that are important may be achieved through coupling to variable selection and spectral fitting techniques.

Conclusion

The label-free subclassification of leukocyte subtypes with Raman spectroscopy represents a challenging problem from a technical perspective. Overlapping spectral bands within each leukocyte subtype can reduce the distinct character of each spectral subclass. Presentation of the whole Raman spectrum to a classifier in an unsupervised manner is the most appropriate *a-priori* approach to development of models for classification of these subtypes, although configuration of the modelling parameters and its complexity must be carefully chosen to maximise robustness and accuracy. The present article demonstrates the importance of identifying the best model for classifications and outlines a strategy for optimisation of three distinct modelling approaches. Alternative approaches may be required for other classification algorithms and problems.

Acknowledgements

This work was financially supported by the EU FP7 Network of Excellence DoReMi (Grant Number 249689).

Notes and references

[1]School of Physics, Dublin Institute of Technology, Dublin, Ireland;

[2]Radiation and Environmental Science Centre (RESC), Dublin Institute of Technology, Dublin, Ireland;

[3]School of Biological Sciences, Dublin Institute of Technology, Dublin, Ireland;

Email: aidan.meade@dit.ie

1. Crow, P. *et al.* The use of Raman spectroscopy to differentiate between different prostatic adenocarcinoma cell lines. *Br. J. Cancer* 2005 **92,** 2166–70 (2005).

2. Krafft, C., Steiner, G., Beleites, C. & Salzer, R. Disease recognition by infrared and Raman spectroscopy. *J. Biophotonics* 2009 **2,** 13–28 (2009).

3. Lyng, F. M. *et al.* Vibrational spectroscopy for cervical cancer pathology, from biochemical analysis to diagnostic tool. *Exp. Mol. Pathol.* 2007 **82,** 121–129 (2007).

4. Hedegaard, M. *et al.* Spectral unmixing and clustering algorithms for assessment of single cells by Raman microscopic imaging. *Theor. Chem. Acc.* 2011 **130,** 1249–1260 (2011).

5. Ramoji, A. *et al.* Toward a spectroscopic hemogram: Raman spectroscopic differentiation of the two most abundant leukocytes from peripheral blood. *Anal. Chem.* 2012 **84,** 5335–42 (2012).

6. Bankapur, A., Zachariah, E., Chidangil, S., Valiathan, M. & Mathur, D. Raman tweezers spectroscopy of live, single red and white blood cells. *PLoS One* 2010 **5,** e10427 (2010).

7. Utzinger, U. R. S. *et al.* Near-Infrared Raman Spectroscopy for in Vivo Detection of Cervical Precancers. *Appl. Spectrosc.* 2001 **55,** 955–959 (2001).

8. Matias, R., Silveira, L., Augusto, M. & Silva, R. S. Diagnostic model based on Raman spectra of normal , hyperplasia and prostate adenocarcinoma tissues in vitro. *Spectroscopy* 2011 **25,** 89–102 (2011).

9. Kanter, E. M. *et al.* Multiclass discrimination of cervical precancers using Raman spectroscopy. *J. Raman Spectrosc.* 2009 **40,** 205–211 (2009).

10. Duraipandian, S. *et al.* Simultaneous fingerprint and high-wavenumber confocal Raman spectroscopy enhances early detection of cervical precancer in vivo. *Anal. Chem.* 2012 **84,** 5913–9 (2012).

11. Tollefson, M. *et al.* Raman spectral imaging of prostate cancer: can Raman molecular imaging be used to augment standard histopathology? *BJU Int.* 2010 **106,** 484–8 (2010).

12. Short, M. a *et al.* Development and preliminary results of an endoscopic Raman probe for potential in vivo diagnosis of lung cancers. *Opt. Lett.* 2008 **33,** 711–3 (2008).

13. Shetty, G., Kendall, C., Shepherd, N., Stone, N. & Barr, H. Raman spectroscopy: elucidation of biochemical changes in carcinogenesis of oesophagus. *Br. J. Cancer* 2006 **94,** 1460–4 (2006).

14. Kendall, C. *et al.* Raman spectroscopy, a potential tool for the objective identification and classification of neoplasia in Barrett's oesophagus. *J. Pathol.* 2003 **200,** 602–9 (2003).

15. Poon, K. W. C. *et al.* Quantitative reagent-free detection of fibrinogen levels in human blood plasma using Raman spectroscopy. *Analyst* 2012 **137,** 1807–14 (2012).

16. Rohleder, D., Kiefer, W. & Petrich, W. Quantitative analysis of serum and serum ultrafiltrate by means of Raman spectroscopy. *Analyst* 2004 **129,** 906–11 (2004).

17. Wesełucha-Birczyńska, a, Kozicki, M., Czepiel, J. & Birczyńska, M. Raman micro-spectroscopy tracing human lymphocyte activation. *Analyst* 2013 **138,** 7157–63 (2013).

18. Krafft, C., Dochow, S., Beleites, C. & Jürgen Popp. Cell identification using Raman spectroscopy in combination with optical trapping and microfluidics. *Prog. Biomed. Opt. Imaging - Proc. SPIE* 2014 **8939,** (2014).

19. Notingher, I. *et al.* Discrimination between ricin and sulphur mustard toxicity in vitro using Raman spectroscopy. *J. R. Soc. Interface* 2004 **1,** 79–90 (2004).

20. Gazi, E. *et al.* A correlation of FTIR spectra derived from prostate cancer biopsies with gleason grade and tumour stage. *Eur. Urol.* 2006 **50,** 750–60; discussion 760–1 (2006).

21. Han, J., Kamber, M. & Pei, J. *Data Mining Concepts and Techniques. Third Edition*. 2012 (Morgan Kaufmann).

22. Cristianini, N. & Shawe-Taylor, J. *An introduction to Support Vector Machines and other kernel-based learning methods*. 2000 (Cambridge University Press).

23. Sattlecker, M., Bessant, C., Smith, J. & Stone, N. Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics. *Analyst* 2010 **135,** 895–901 (2010).

24. Mitchell, T. M. *Machine Learning*. 1997 (McGraw-Hill).

25. Menze, B. H., Petrich, W. & Hamprecht, F. a. Multivariate feature selection and hierarchical classification for infrared spectroscopy: serum-based detection of bovine spongiform encephalopathy. *Anal. Bioanal. Chem.* 2007 **387,** 1801–7 (2007).

26. Trevisan, J., Angelov, P. P., Carmichael, P. L., Scott, A. D. & Martin, F. L. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. *Analyst* 2012 **137,** 3202–15 (2012).

27. Baker, M. J. *et al.* Using Fourier transform IR spectroscopy to analyze biological materials. *Nat. Protoc.* 2014 **9,** 1771–91 (2014).

28. Gajjar, K. *et al.* Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer. *Analyst* 2013 **138,** 3917–26 (2013).

29. Das, K., Stone, N., Kendall, C., Fowler, C. & Christie-Brown, J. Raman spectroscopy of parathyroid tissue pathology. *Lasers Med. Sci.* 2006 **21,** 192–7 (2006).

30. Bonnier, F. *et al.* Analysis of human skin tissue by Raman microspectroscopy: Dealing with the background. *Vib. Spectrosc.* 2012 **61,** 124–132 (2012).

31. Lihong, Z. & Zikui, G. Face Recognition Method Based on Adaptively Weighted Block-Two Dimensional Principal Component Analysis. *2011 Third Int. Conf. Comput. Intell. Commun. Syst. Networks* 2011 22–25 (2011). doi:10.1109/CICSyN.2011.18

32. Varmuza, K. & Filzmoser, P. *Introduction to Multivatiate Statistical Analysis in Chemometrics*. 2009 (CRC Press).

33. Ben-Hur, A., Soon Ong, C., Sonnenburg, S., Scholkopf, B. & Ratsch, G. Support Vector Machines and Kernels for Computational Biology. *Plos Comput. Biol.* 2008 **4,** (2008).

34. Zilu, Y. & Guoyi, Z. Facial Expression Recognition Based on NMF and SVM. *2009 Int. Forum Inf. Technol. Appl.* 2009 612–615 (2009). doi:10.1109/IFITA.2009.279

35. Zou, Y., Shi, G., Shi, H. & Zhao, H. Traffic incident classification at intersections based on image sequences by HMM/SVM classifiers. *Multimed. Tools Appl.* 2010 **52,** 133–145 (2010).

36. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. 2013 (Springer).