Variable selection towards classification of digital images: identification of altered glucose levels in serum

Camilo L.M. Morais,[1*] Kássio M.G. Lima[2] and Francis L. Martin[1*#]

[1]School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, United Kingdom

[2]Biologial Chemistry and Chemometrics, Institute of Chemistry, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil

Corresponding author: cdlmedeiros-de-morai@uclan.ac.uk

[*]Lancashire Teaching Hospitals NHS Trust, Fulwood, Preston, UK.

[#]Biocel Ltd, Hull, Yorkshire, UK.

Abstract: Identification of altered glucose levels in serum is the main indicator for diabetes, where control levels are classed as <100 mg/dL, and altered levels are classified as pre-diabetic (100-125 mg/dL) or diabetic (>125 mg/dL). Herein, we propose a method to identify control, pre-diabetic or diabetic simulated and real-world samples based on their glucose levels using classification-based variable selection algorithms [successive projections algorithm (SPA) or genetic algorithm (GA)] coupled to linear discriminant analysis (SPA-LDA and GA-LDA) towards analysing red-green-blue (RGB) digital images. Images were recorded after glucose enzymatic reaction, whereby 250 μL of reactant content of samples were captured by using a common cell phone camera. Processing was applied to the images at a pixel level, where 72.2% of the pixels were correctly classified as control, 79.2% as pre-diabetic and 90.9% as diabetic using SPA-LDA algorithm; and 76.8% as control, 81.4% as pre-diabetic and 91.7% as diabetic using GA-LDA algorithm in the validation set containing nine simulated samples. Eight real-world samples were measured as an external test set, where an accuracy using GA-LDA was found to be 92%, with sensitivities ranging from 70-100% and specificities ranging from 90-99%. This method shows the potential of variable selection techniques coupled with digital image analysis towards blood glucose monitoring.

Keywords: Glucose; diabetes; digital images; colorimetry; variable selection

Introduction

Image processing is an emerging analytical technique with great potential towards colorimetry. Advantages including low-cost, high portability, sensitivity, fast data acquisition and translation to *in situ* applications make the use of digital images very attractive as a possible alternative method to visible spectroscopy. This is achievable by processing red-green-blue (RGB) channels composed of pixels that combined produce millions of colours with greater details than the human eye can see (Solomon and Breckton 2011). Applications in analytical sciences includes qualitative and quantitative analysis, in fields such as food monitoring (Benedetti et al. 2015; Lyra et al. 2014; dos Santos and Pereira-Filho 2013), environmental analysis (Firdaus et al. 2014; Andrade et al. 2013), forensic investigations (Choodum and Daeid 2011; Choodum et al. 2013; Tosato et al. 2016), and clinical assays (Xia et al. 2015; Morais and Lima 2014; Morais and Lima 2015; Morais et al. 2016a). Despite limitations, the use of cell phones cameras for these types of applications further increases their potential, since with a cell phone one could be potentially carrying a portable spectrometer in their pocket (Scheeline 2016).

For these image processing applications, there are several methodologies that can be used to extract the colour signal and relate it with the substance of interest. Examples include using single channel intensities or absorbances (Morais and Lima 2014; Morais et al. 2016b; Christodouleas et al. 2015; Moraes et al. 2014), sets of colour intensities or absorbances (Morais et al. 2016a; Morais et al. 2018), and colour histograms (Capitán-Vallvey et al. 2015). The use of full images in a pixel perspective can also be used for these applications, where the level of pre-processing is reduced thus keeping the data as original as possible. The computation of such signals are performed by either univariate or multivariate approaches, in which the latter has significant advantages, *e.g.*, being able to work with data in the presence of unknown interferences.

Variable selection is a data reduction method much applied to spectroscopy data, where the original set of variables is reduced to a small number of features representing the most important information in the original data. Successive projections algorithm (SPA) (Soares et al. 2013) and genetic algorithm (GA) (McCall 2005) are two variable selection methods with great potential for such applications. SPA is a forward feature selection method operating by solving co-linearity problems, wherein the variables whose information content is minimally redundant are selected. This is achieved through a series of interactions, starting with one variable (*e.g.*, pixel value), and then incorporating a new one at each interaction until a specified number of variables is reached (Theophilou et al. 2018). As a main advantage, SPA maintains the same variable space, so the selected features have the same physical meaning of the original data. Similarly, GA also reduces the original data to features in the same variable space; however, this is achieved following an evolutionary process mimicking Mendelian genetics. GA is built in an interactive process using combinations, recombination and mutations, to evolve sets of initial variables (chromosomes) though a certain number of generations until the best solution of a problem (*e.g.*, discrimination of two images clusters) is found. The set of variables with a higher chance of achieving this goal is selected (Theophilou et al. 2018).

Despite these obvious advantages of feature selection, there are only a few applications using SPA or GA for processing digital images. Some examples include classification of biodiesel (Costa et al. 2015), edible vegetable oil (Milanez and Pontes 2014), tea (Diniz et al. 2012; Diniz et al. 2015), and coffee (Souto et al. 2015). Herein, we explore the application of SPA-linear discriminant analysis (SPA-LDA) and GA-linear discriminant analysis (GA-LDA) towards detecting altered glucose levels in simulated-diabetic and real-world serum samples. Glucose in serum is the main indicator for diabetes, where levels >99 mg/dL are an indicator of a pre-diabetic state, and >125 mg/dL of diabetes. To the best of our

knowledge, this is the first time that these algorithms have been applied towards detecting altered levels of glucose based on digital images.

Materials and Methods

Samples

Thirty samples were prepared for analysis (22 simulated and 8 real-world samples). Samples assigned as "control" (8 simulated samples, 2 real-world samples) had glucose concentrations ranging from 12.5 to 99 mg/dL; samples assigned as "pre-diabetic" (9 simulated samples, 2 real-world samples) had glucose concentration ranging from 101 to 121 mg/dL; and samples assigned as "diabetic" (5 simulated samples, 4 real-world samples) had glucose concentration ranging from 126 to 200 mg/dL. Simulated samples were prepared based on simulated serum solutions. These samples were composed of glucose diluted with benzoic acid (20.47 mmol/L) as preservative in a buffer solution (pH 7.0), traceable to the National Institute of Standards and Technology (NIST) SRM 917. Real-world samples were obtained from volunteers with informed consent. The colorimetric enzymatic reaction for glucose determination was performed using the Bioclin (Quibasa Química Básica Ltd., Brazil) commercial kit (reference code k.082-2). This reagent is composed of 4-aminoantypyrine (0.3 mmol/L), phenol (10 mmol/L), glucose oxidase (>10,000 U/L), peroxidase (>700 U/L) and sodium azide (15.38 mmol/L) in a buffer solution (pH 7.0). In this reaction, glucose reacts with glucose oxidase in the presence of oxygen forming hydrogen peroxide and δ-D-gluconolactone; then, the hydrogen peroxide reacts, in the presence of peroxidase, with 4-aminoantypyrine and phenol, forming quinoneimine, a cherry chromogen whose colour intensity (λ_{max} = 500-505 nm) is proportional to the glucose concentration (Menezes et al. 2015). This reaction was carried out in microcentrifuge tubes

incubated for 10 min at 37ºC. Reference measurements of glucose concentration were carried out using a ultraviolet-visible (UV-Vis) spectrometer BIO-2000 (Bioplus Ltda, Brazil) with absorbance readings at 505 nm by using a pre-built calibration curve.

Image acquisition

After reaction, 250 μL of sample were transferred to a 96-microwell enzyme-linked immunosorbent assay (ELISA) plate (Fischer Scientific, USA), which acts as a sample holder for image acquisition. The microplate's image was captured by a Sony Xperia C smartphone with camera resolution of 8 megapixels, with a distance of approximately 15 cm from the cell phone to the samples. The image was saved in .JPG format and further processed by cutting the regions of interest (ROI) for each microwell with a size of 63 × 66 pixels by using GIMP 2.10 software (https://www.gimp.org/).

Computational analysis

The ROI images were imported and processed within MATLAB R2014b environment (The MathWorks, Inc., USA) using lab-made routines. Initially, each image with size 66 × 63 × 3 was decomposed into RGB channels, forming a single image with size 66 × 189 (*i.e.*, 66 × 63 times 3). The simulated samples' images were separated into training (n=13) and validation (n=9) sets using the Kennard-Stone sample selection algorithm (Kennard and Stone 1969). The real-world samples' images were assigned to an external test set used to evaluate the model performance towards real-world samples.

SPA-LDA and GA-LDA algorithms were applied to the image data by using SPA and GA for variable selection, followed by a linear discriminant classifier applied based on a Mahalanobis distance calculation (Morais and Lima 2018) between the pixels of each image. SPA and GA reduce hundreds of pixels to a small number of orthogonal variables in the same spatial domain (pixels), where the calculation of inverse matrix operations in LDA can be

achieved with high accuracy. The optimization of SPA-LDA and GA-LDA for selecting variables was performed within the training set according to the lowest risk of misclassification G as follows (Siqueira et al. 2017):

$$G = \frac{1}{N}\sum_{n=1}^{N} g_n \quad (01)$$

in which N is the number of training samples and g_n is defined as follows:

$$g_n = \frac{r^2(x_n, m_{I(n)})}{\min_{I(m)\neq I(n)} r^2(x_n, m_{I(m)})} \quad (02)$$

where $r^2(x_n, m_{I(n)})$ is the squared Mahalanobis distance between pixel x_n (of class index $I(n)$) and the centre of its true class $m_{I(n)}$; and $r^2(x_n, m_{I(m)})$ is the squared Mahalanobis distance between pixel x_n and the centre of the closest wrong class ($m_{I(m)}$). GA was performed though 100 generations, having 200 chromosomes each. Crossover and mutation probability were set to 60% and 1%, respectively. GA was repeated three times, and the best result chosen.

Results and Discussion

The RGB images captured after enzymatic colorimetric reaction were separated into three classes according to their glucose levels: control (glucose <100 mg/dL), pre-diabetic (100 mg/dL ≤ glucose ≤ 125 mg/dL) and diabetic (glucose >125 mg/dL). These images are shown in Figure 1.

[Insert Figure 1 here]

Clearly, there is a colour variation between the images in which the diabetic samples are darker (higher glucose concentration) and the control samples are lighter (lower glucose concentration). The pre-diabetic samples have coloration in between the two classes. This

tendency is confirmed by the colour pixel values for each image in the RGB channels (Figure 2). For the red (Fig. 2a), green (Fig. 2b) and blue (Fig. 2c) channels, the pixels with larger intensity (closer to 1, lighter colour) are from class control, followed by pre-diabetic and diabetic samples (closer to 0, darker colour). Nevertheless, there are some degrees of superposition between the pixels, especially for the pre-diabetic class where pixels are observed superposing the other two classes. Although the colorimetric solution is homogeneous by nature, surface deformities, ambiance lighting, shades and other environmental effects cause colour variations on the image surface giving a heterogeneous multivariate nature to the data. Such variations are depicted by means of the coefficient of variation (CV) across the *x*- and *y*-axis coordinates of the average image for each class (Figure 3). Therefore, multivariate variable selection algorithms could act as tools to remove these effects and filter only chemically-relevant pixel positions.

[Insert Figure 2 here]

[Insert Figure 3 here]

Multivariate analysis was employed by means of SPA-LDA and GA-LDA algorithms for classifying the images pixels into one of the three classes. SPA-LDA selected 11 variables for model construction identified by vertical lines in the RGB images in Figure 4. Most of the variables are in the blue channel (6 variables), followed by the green (3 variables) and red (2 variables) channels. This indicates a higher degree of importance of the blue channel towards glucose concentration. This same trend has been previously observed for glucose based on higher resolution images obtained with a desktop scanner, where the blue channel and glucose concentration exhibited a linear relationship in the range between 12.5 and 100 mg/dL with an R^2 of 0.984 (Morais and Lima 2014). This happens because the absorption centre of the blue channel is at 435.8 nm, which is closer to the absorption region of

quinoneimine formed during glucose enzymatic colorimetric reaction (500-505 nm) (Menezes et al. 2015), and farther from the green (546.1 nm) and red (700.0 nm) channels (Petrou and Petrou 2010). Similarly, the variables selected by GA-LDA exhibit a predominance for the blue channel although this algorithm only selected 4 variables (Figure 5).

[Insert Figure 4 here]

[Insert Figure 5 here]

The classification performance for the simulated samples at a pixel level using SPA-LDA and GA-LDA algorithms are represented by the accuracy values depicted in Table 1. GA-LDA exhibits a better accuracy for the three classes in the validation set in comparison with SPA-LDA, although its performance in the training set is inferior, in particular for control samples (66% accuracy). The SPA-LDA model seems more stable for the simulated samples, where the accuracies for the validation and training sets are more similar. For both algorithms, the diabetes class is the one with higher accuracy, indicating a higher degree of correctly classified pixels. This is related to the larger colour difference between this class in comparison with the others. Figure 6 shows the discriminant function (DF) graphs for SPA-LDA and GA-LDA algorithms, where clear distinction between the images of the three classes are observed at a pixel level.

[Insert Table 1 here]

[Insert Figure 6 here]

Sensitivity and specificity levels for simulated samples were also calculated for SPA-LDA and GA-LDA models. The sensitivity indicates the proportion of positive pixels correctly identified, whereas the specificity highlights the proportion of negative pixels correctly identified (Morais and Lima 2017). Notably, the specificities for control and

diabetes images are the highest (>90%), indicating that these samples are well incorporated into their own clusters; and the pre-diabetic pixels, with lower specificity, are more distributed among the other two classes. This is associated with the larger colour difference between control and diabetic images, whereas for pre-diabetic samples the colour intensities merges with the other two classes.

[Insert Table 2 here]

The classification for simulated samples is presented at a sample level according to the confusion matrices in Table 3. The SPA-LDA model exhibits more consistent results, with only one control sample being incorrectly classified as pre-diabetic in the validation set. All the other samples are correctly assigned to their true classes based on their colour images. On the other hand, the GA-LDA model misclassified two control samples as pre-diabetic and one diabetes sample as pre-diabetic in the training set; and one control sample as pre-diabetic in the validation set. This indicates that SPA-LDA is more stable towards classifying simulated samples.

[Insert Table 3 here]

Finally, an external set of real-world serum samples were introduced as a prediction set in the models previously built using simulated samples. The accuracy for this dataset using SPA-LDA was found to be 76%, with sensitivities ranging from 66-83% and specificities ranging from 82-94% (Table 4). GA-LDA exhibited a better performance with real-world samples, with a total accuracy of 92%, and sensitivities ranging from 70-100% and specificities ranging from 90-99% (Table 4). Overall, the model performance towards real-world samples was inferior than that using simulated samples, which is expected since real-world samples contain more complex matrix effects. GA-LDA shows a better classification performance than SPA-LDA for real-world samples, which indicates that GA-

LDA is more robust, covering more random sources of variations within the real matrix, while SPA-LDA is better well-fitted to the simulated samples only. This reinforces the hypothesis that a well-fitted training model will not necessarily provide good prediction estimates towards external samples.

Conclusion

This paper demonstrates the use of variable selection techniques for processing digital images for classification of simulated and real-world serum samples in three class levels, based on a glucose enzymatic colorimetric reaction: control, pre-diabetic and diabetic. Images were acquired with a cell phone camera and processed by means of SPA-LDA and GA-LDA algorithms. Both algorithms generated high accuracies at a pixel level, especially for diabetic samples, but the GA-LDA model provided more reliable results for real-world samples. In addition, the variables selected by these algorithms suggest the blue colour channel has the most importance for glucose determination, confirming previous findings. These results show the potential of variable selection methods for multivariate classification of digital images, where altered levels of glucose can be easily distinguished based on colour intensities.

Conflict of Interest

There are no conflicts of interest to declare.

Acknowledgments

Camilo L. M. Morais would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for financial support.

References

Andrade, S. I. E., M. B. Lima, I. S. Barreto, W. S. Lyra, L. F. Almeida, M. C. U. Araújo, and E. C. Silva. 2013. A digital image-based flow-batch analyzer for determining Al(III) and Cr(VI) in water. *Microchemical Journal* 109:106–111. doi: 10.1016/j.microc.2012.03.029.

Benedetti, S. P. S., V. B. dos Santos, T. A. Silva, E. Benedetti Filho, V. L. Martins, and O. Fatibello-Filho. 2015. A digital image-based method employing a spot-test for quantification of ethanol in drinks. *Analytical Methods* 7:4138–4144. doi: 10.1039/C5AY00529A.

Capitán-Vallvey, L. F., N. López-Ruiz, A. Martínez-Olmos, M. M. Erenas, and A. J. Palma. 2015. Recent developments in computer vision-based analytical chemistry: A tutorial review. *Analytica Chimica Acta* 899:23–56. doi: 10.1016/j.aca.2015.10.009.

Choodum, A., and N. N. Daeid. 2011. Rapid and semi-quantitative presumptive tests for opiate drugs. *Talanta* 86:284–292. doi: 10.1016/j.talanta.2011.09.015.

Choodum, A., P. Kanatharana, W. Wongniramaikul, and N. N. Daeid. 2013. Using the iPhone as a device for a rapid quantitative analysis of trinitrotoluene in soil. *Talanta* 115:143–149. doi: 10.1016/j.talanta.2013.04.037.

Christodouleas, D. C., A. Nemiroski, A. A. Kumar, and G. M. Whitesides. 2015. Broadly Available Imaging Devices Enable High-Quality Low-Cost Photometry. *Analytical Chemistry* 87 (18):9170–9178. doi: 10.1021/acs.analchem.5b01612.

Costa, G. B., D. D. S. Fernandes, V. E. Almeida, T. S. P. Araújo, J. P. Melo, P. H. G. D. Diniz, and G. Véras. 2015. Digital image-based classification of biodiesel. *Talanta* 139:50–55. doi: 10.1016/j.talanta.2015.02.043.

Diniz, P. H. G. D., H. V. Dantas, K. D. T. Melo, M. F. Barbosa, D. P. Harding, E. C. L. Nascimento, M. F. Pistonesi, B. S. F. Band, and M. C. U. Araújo. 2012. Using a simple digital camera and SPA-LDA modeling to screen teas. *Analytical Methods* 4:2648–2652. doi: 10.1039/C2AY25481F.

Diniz, P. H. G. D., M. F. Pistonesi, M. B. Alvarez, B. S. F. Band, and M. C. U. de Araújo. 2015. Simplified tea classification based on a reduced chemical composition profile via successive projections algorithm linear discriminant analysis (SPA-LDA). *Journal of Food Composition and Analysis* 39:103–110. doi: 10.1016/j.jfca.2014.11.012.

dos Santos, P. M., and E. R. Pereira-Filho. 2013. Digital image analysis – an alternative tool for monitoring milk authenticity. *Analytical Methods* 5:3669–3674. doi: 10.1039/C3AY40561C.

Firdaus, M. L., W. Alwi, F. Trinoveldi, I. Rahayu, L. Rahmidar, and K. Warsito. 2014. Determination of Chromium and Iron Using Digital Image-based Colorimetry. *Procedia Environmental Sciences* 20:298–304. doi: 10.1016/j.proenv.2014.03.037.

Kennard, R. W., and L. A. Stone. 1969. Computer Aided Design of Experiments. *Technometrics* 11 (1):137–148. doi: 10.1080/00401706.1969.10490666.

Lyra, W. S., L. F. de Almeida, F. A. S. Cunha, P. H. G. D. Diniz, V. L. Martins, and M. C. U. de Araujo. 2014. Determination of sodium and calcium in powder milk using digital image-based flame emission spectrometry. *Analytical Methods* 6:1044–1050. doi: 10.1039/C3AY41005F.

McCall, J. 2005. Genetic algorithms for modelling and optimisation. *Journal of Computational and Applied Mathematics* 184:205–222. doi: 10.1016/j.cam.2004.07.034.

Menezes, F. G., A. C. O. Neves, D. F. de Lima, S. D. Lourenço, L. C. da Silva, and K. M. G. Lima. 2015. Bioorganic concepts involved in the determination of glucose, cholesterol and triglycerides in plasma using the enzymatic colorimetric method. *Química Nova* 38 (4):588–594. doi: 10.5935/0100-4042.20150040.

Milanez, K. D. T. M., and M. J. C. Pontes. 2014. Classification of edible vegetable oil using digital image and pattern recognition techniques. *Microchemical Journal* 113:10–16. doi: 10.1016/j.microc.2013.10.011.

Moraes, E. P., N. S. A. da Silva, C. L. M. Morais, L. S. das Neves, and K. M. G. Lima. 2014. Low-Cost Method for Quantifying Sodium in Coconut Water and Seawater for the Undergraduate Analytical Chemistry Laboratory: Flame Test, a Mobile Phone Camera, and Image Processing. *Journal of Chemical Education* 91 (11):1958–1960. doi: 10.1021/ed400797k.

Morais, C. L. M., and K. M. G. Lima. 2014. A colorimetric microwell method using a desktop scanner for biochemical assays. *Talanta* 126:145–150. doi: 10.1016/j.talanta.2014.03.066.

Morais, C. L. M., and K. M. G. Lima. 2017. Comparing unfolded and two-dimensional discriminant analysis and support vector machines for classification of EEM data. *Chemometrics and Intelligent Laboratory Systems* 170:1–12. doi: 10.1016/j.chemolab.2017.09.001.

Morais, C. L. M., and K. M. G. Lima. 2015. Determination and analytical validation of creatinine content in serum using image analysis by multivariate transfer calibration procedures. *Analytical Methods* **7**:6904–6910. doi: 10.1039/C5AY01369K.

Morais, C. L. M., K. M. G. Lima, and F. L. Martin. 2018. Colourimetric Determination of

High-Density Lipoprotein (HDL) Cholesterol Using Red–Green–Blue Digital Colour Imaging. *Analytical Letters* 51 (18):2860–2867. doi: 10.1080/00032719.2018.1453833.

Morais, C. L. M., and K. M. G. Lima. 2018. Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. *Journal of the Brazilian Chemical Society* 29 (3):472–481. doi: 10.21577/0103-5053.20170159.

Morais, C. L. M., A. C. O. Neves, F. G. Menezes, and K. M. G. Lima. 2016a. Determination of serum protein content using cell phone image analysis. *Analytical Methods* 8:6458–6462. doi: 10.1039/C6AY01783E.

Morais, C. L. M., S. R. B. Silva, D. S. Vieira, and K. M. G. Lima. 2016b. Integrating a Smartphone and Molecular Modeling for Determining the Binding Constant and Stoichiometry Ratio of the Iron(II)–Phenanthroline Complex: An Activity for Analytical and Physical Chemistry Laboratories. *Journal of Chemical Education* 93 (10):1760–1765. doi: 10.1021/acs.jchemed.6b00112.

Petrou, M., and C. Petrou. 2010. *Image Processing: The Fundamentals*, 2nd edn. West Sussex: John Wiley & Sons, Ltd.

Scheeline, A. 2016. Cell phone spectrometry: Science in your pocket?. *TrAC Trends in Analytical Chemistry* 85 (Part A):20–25. doi: 10.1016/j.trac.2016.02.023.

Siqueira, L. F. S., R. F. Araújo Júnior, A. A. de Araújo, C. L. M. Morais, and K. M. G. Lima. 2017. LDA vs. QDA for FT-MIR prostate cancer tissue classification. *Chemometrics and Intelligent Laboratory Systems* 162:123–129. doi: 10.1016/j.chemolab.2017.01.021.

Soares, S. F. C., A. A. Gomes, M. C. U. Araujo, A. R. Galvão Filho, and R. K. H. Galvão.

2013. The successive projections algorithm. *TrAC Trends in Analytical Chemistry* 42:84–98. doi: 10.1016/j.trac.2012.09.006.

Solomon, J., and T. Breckon. 2011. *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. West Sussex: John Wiley & Sons, Ltd.

Souto, U. T. C. P., M. F. Barbosa, H. V. Dantas, A. S. de Pontes, W. S. Lyra, P. H. G. D. Diniz, M. C. U. de Araújo, and E. C. da Silva. 2015. Screening for Coffee Adulteration Using Digital Images and SPA-LDA. *Food Analytical Methods* 8 (6):1515–1521. doi: 10.1007/s12161-014-0020-7.

Theophilou, G., C. L. M. Morais, D. E. Halliwell, K. M. G. Lima, J. Drury, P. L. Martin-Hirsch, H. F. Stringfellow, D. K. Hapangama, and F. L. Martin. 2018. Synchrotron- and focal plane array-based Fourier-transform infrared spectroscopy differentiates the basalis and functionalis epithelial endometrial regions and identifies putative stem cell regions of human endometrial glands. *Analytical and Bioanalytical Chemistry* 410 (18): 4541–4554. doi: 10.1007/s00216-018-1111-x.

Tosato, F., T. R. Rosa, C. L. M. Morais, A. O. Maldaner, R. S. Ortiz, P. R. Filgueiras, K. M. G. Lima, and W. Romão. 2016. Direct quantitative analysis of cocaine by thin layer chromatography plus a mobile phone and multivariate calibration: a cost-effective and rapid method. *Analytical Methods* 8:7632–7637. doi: 10.1039/C6AY02126C.

Xia, M., L. Wang, Z. Yang, and H. Chen. 2015. A novel digital color analysis method for rapid glucose detection. *Analytical Methods* **7**:6654–6663. doi: 10.1039/C5AY01233C.

Captions for Figures

Figure 1: Example of ELISA plate image and region of interests (ROI) of glucose images used for analysis, where * represents real-world samples.

Figure 2: Colour intensity of pixels in the (a) red, (b) green and (b) blue channels. Data in blue: “control” images; data in red: “pre-diabetic” images; data in black: “diabetic” images.

Figure 3: Coefficient of variation (CV) on the *x*- and *y*-axis coordinates of the average RGB coloured image for (a) control, (b) pre-diabetic and (c) diabetic samples. R: red channel, G: green channel, B: blue channel.

Figure 4: Mean coloured image, RGB image and selected pixels by SPA-LDA algorithm for “control”, “pre-diabetic” and “diabetic” images.

Figure 5: Mean coloured image, RGB image and selected pixels by GA-LDA algorithm for “control”, “pre-diabetic” and “diabetic” images.

Figure 6: Discriminant function (DF) graphs for (a) SPA-LDA and (b) GA-LDA algorithms in a pixel level.

Table 1: Classification accuracy for SPA-LDA and GA-LDA models applied to simulated samples in a pixel level.

Algorithm	**Class**	**Training**	**Validation**
SPA-LDA	Control	71%	72%
	Pre-diabetic	89%	79%
	Diabetic	91%	91%
GA-LDA	Control	66%	77%
	Pre-diabetic	83%	81%
	Diabetic	80%	92%

Table 2: Sensitivity and specificity for SPA-LDA and GA-LDA models applied to simulated samples in a pixel level.

Algorithm	Class	Sensitivity	Specificity
SPA-LDA	Control	72%	95%
	Pre-diabetic	79%	88%
	Diabetic	91%	94%
GA-LDA	Control	77%	93%
	Pre-diabetic	81%	83%
	Diabetic	92%	94%

Table 3: Confusion matrix of the simulated validation set for SPA-LDA and GA-LDA models in a sample basis. Numbers inside parenthesis are the real number of samples for each class.

SPA-LDA			
Training	Control	Pre-diabetic	Diabetic
Control (5)	5	0	0
Pre-diabetic (5)	0	5	0
Diabetic (3)	0	0	3
Validation	Control	Pre-diabetic	Diabetic
Control (3)	2	1	0
Pre-diabetic (4)	0	4	0
Diabetic (2)	0	0	2
GA-LDA			
Training	Control	Pre-diabetic	Diabetic
Control (5)	3	2	0
Pre-diabetic (5)	0	5	0
Diabetic (3)	0	1	2
Validation	Control	Pre-diabetic	Diabetic
Control (3)	2	1	0
Pre-diabetic (4)	0	4	0
Diabetic (2)	0	0	2

Table 4: Sensitivity and specificity for SPA-LDA and GA-LDA models applied in real-world samples in a pixel level.

Algorithm	Class	Sensitivity	Specificity
SPA-LDA	Control	83%	90%
	Pre-diabetic	66%	82%
	Diabetic	78%	94%
GA-LDA	Control	98%	97%
	Pre-diabetic	70%	99%
	Diabetic	100%	90%

Figure 1

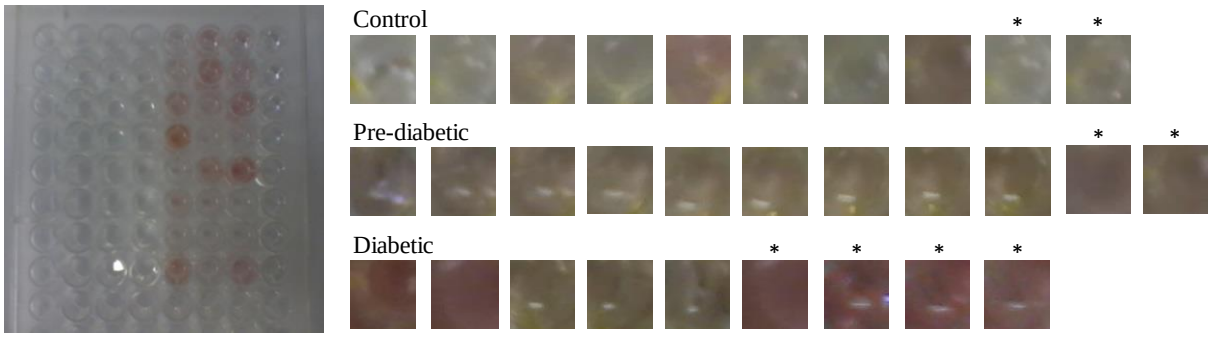


Figure 2

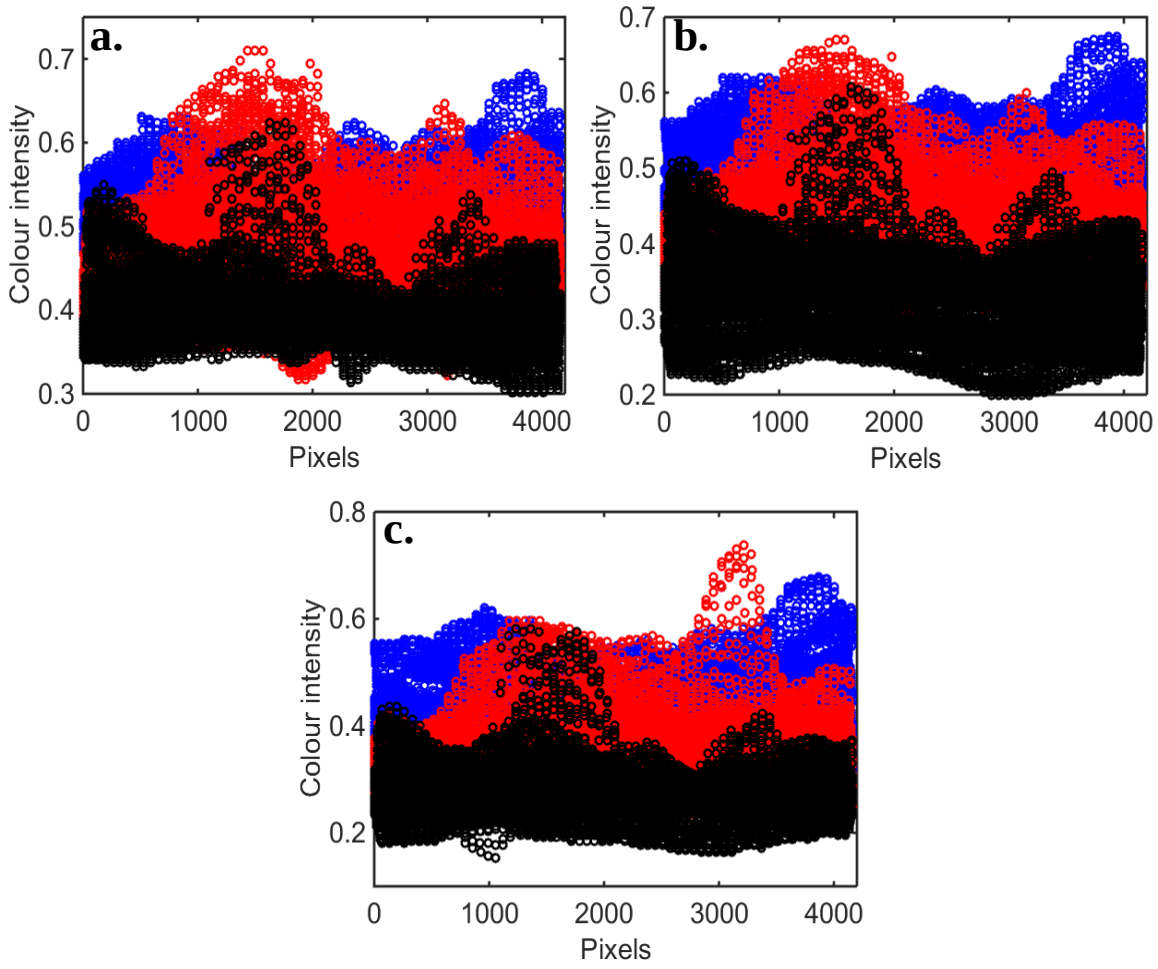

a.
b.
c.
Colour intensity
Pixels
0.7
0.6
0.5
0.4
0.3
0.2
0.8
0
1000
2000
3000
4000

Figure 3

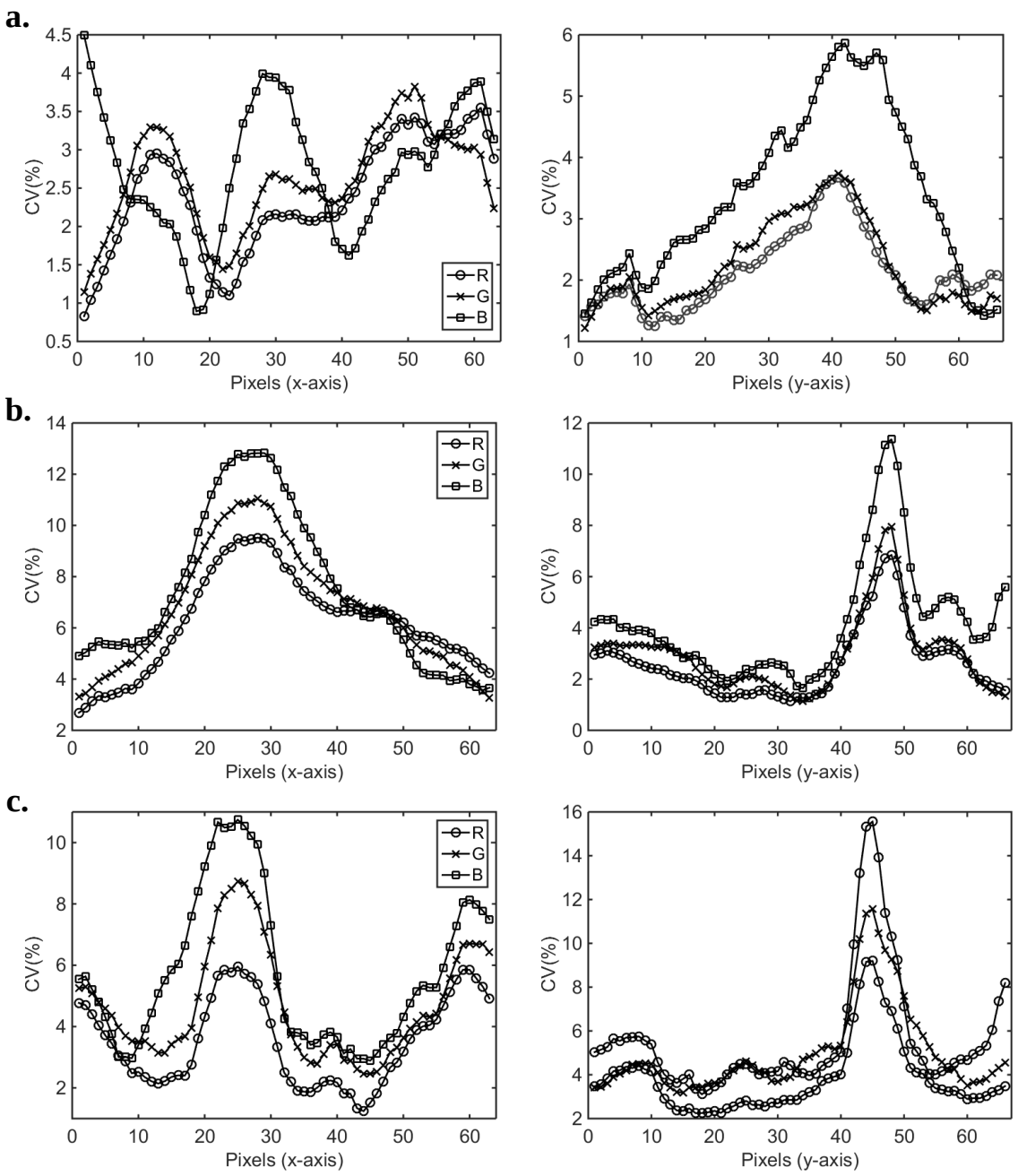

a.
CV(%)
Pixels (x-axis)
Pixels (y-axis)
R
G
B
b.
CV(%)
Pixels (x-axis)
Pixels (y-axis)
R
G
B
c.
CV(%)
Pixels (x-axis)
Pixels (y-axis)
R
G
B

Figure 4

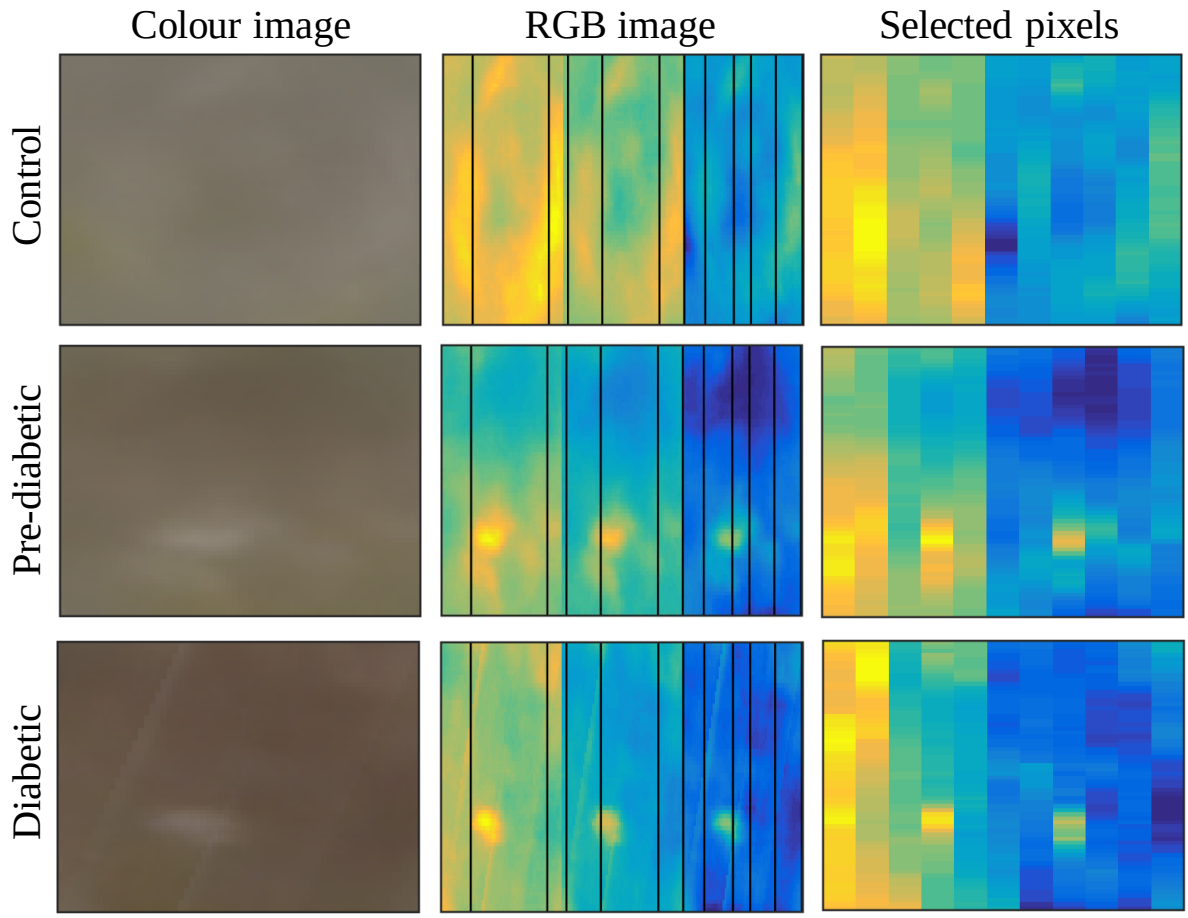

Colour image
RGB image
Selected pixels
Control
Pre-diabetic
Diabetic

Figure 5

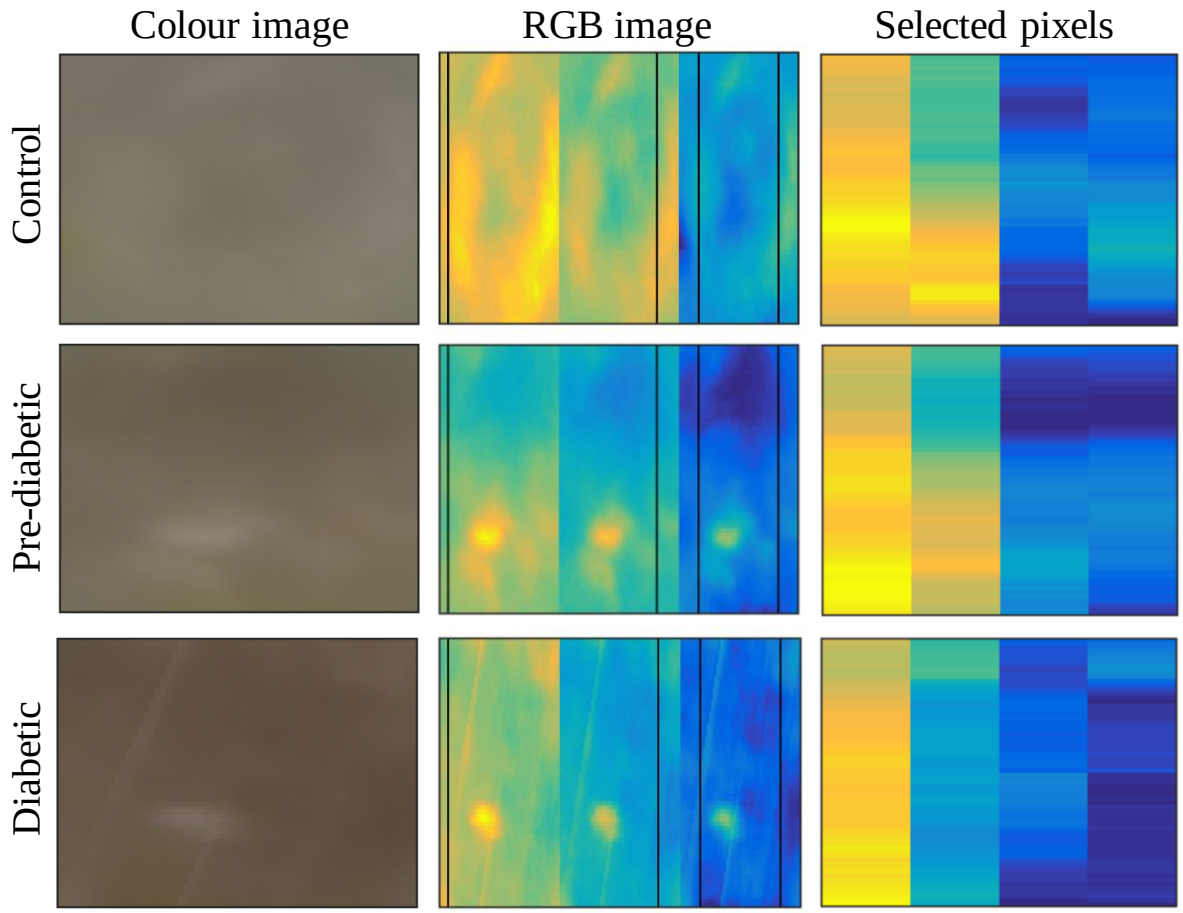
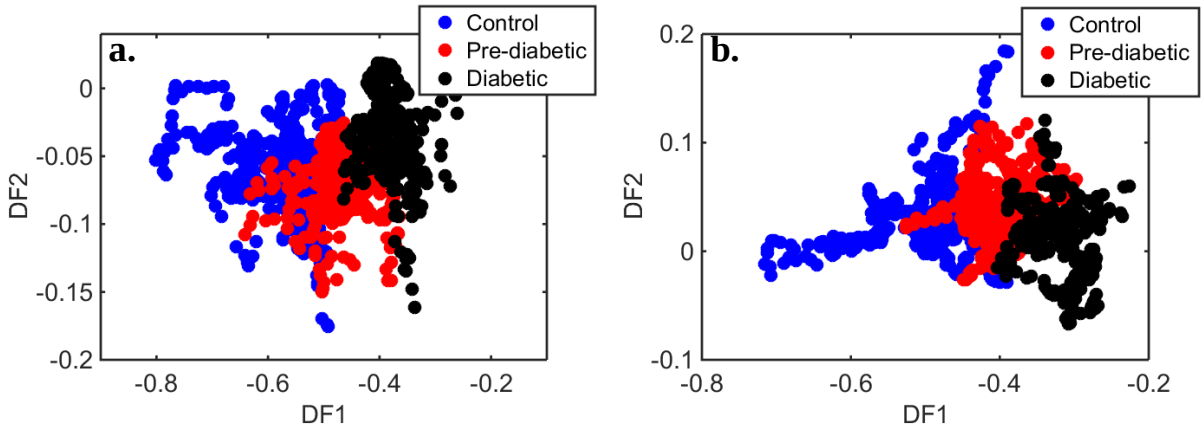

Colour image
RGB image
Selected pixels
Control
Pre-diabetic
Diabetic

Figure 6


a.
Control
Pre-diabetic
Diabetic
DF2
DF1
0
-0.05
-0.1
-0.15
-0.2
-0.8
-0.6
-0.4
-0.2
b.
Control
Pre-diabetic
Diabetic
DF2
DF1
0.2
0.1
0
-0.1
-0.8
-0.6
-0.4
-0.2