

University of Strathclyde  
Department of Pure and Applied Chemistry

Dealing with challenges related to automated high-  
throughput hyperspectral imaging for assessment  
of tea products

Puneet Mishra

A thesis submitted to the Department of Pure and Applied  
Chemistry, University of Strathclyde, Glasgow, in part  
fulfilment of the regulations for the degree of Doctor of  
Philosophy

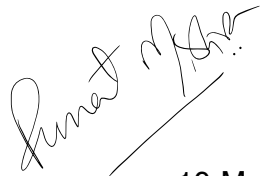
April 2019

# Copyright

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed:

A handwritten signature in black ink, appearing to read 'Sunet M. A.', written over a horizontal line.

Date:

10-May-2019

## Acknowledgments

I would like to express my sincere thanks to my thesis promoter Dr. Alison Nordon for supporting the ideas and research for the past three years and leading to the stage of the thesis defence. Further, I would like to express thanks to the research work collaborators i.e. Prof. Jean Michel Roger, Montpellier France, Prof. Douglas Neil Rutledge, Paris, France and Dr. Azam Karami, Kerman, Iran for their support and suggestion during the development of research. I would like to thank industrial collaborator from Unilever, Colworth Park, the United Kingdom i.e. Ms. Sally Redfern and Prof. Guoping Lian for providing the samples and hosting me for different industrial stays to understand the industrial scenario related to my research. Further, I would also like to thank Bayer Ag, Leverkusen and Bayer CropScience, Frankfurt, Germany for hosting me for industrial experiences related to high-throughput hyperspectral imaging. In particular, at Bayer, I would like to thank Mr. Torsten Feller, Dr. Martin Schmuck, Dr. Andreas Nicol from Bayer AG, Leverkusen for supporting in data acquisition and data analysis.

I would also like to express thanks to CPACT i.e. Ms. Natalie Kerr and Ms. Christine Stevenson and all its member for supporting my three-year research learning experience at University Strathclyde. I would also like to express thank to the European Union's Horizon 2020 research and innovation programme named 'MODLIFE' (Advancing Modelling for Process-Product Innovation, Optimization, Monitoring and Control in Life Science Industries) under the Marie Skłodowska-Curie grant agreement number 675251 for funding this research.

## **Abstract**

Hyperspectral imaging (HSI) is emerging as a potential tool for non-contact, non-destructive high-throughput analysis of products and processes. The main benefit of HSI over point spectroscopy techniques is the possibility to explore both spatial and chemical information present in samples. However, there are challenges associated with using HSI as an automated tool for high-throughput applications.

This study involves the application of near-infrared (NIR) HSI and data analysis methodologies for assessment of tea products. Assessment of tea products involved the development of support vector machine classification models for different tea products sourced from Unilever, UK. The study further involves covering challenges related to handling, managing and processing of HSI data in an automatic sense. Methodologies related to automatic pre-processing, compression and data fusion for assessment of tea products are presented.

The findings showed that NIR HSI can be used for the classification of tea products with an accuracy of more than 97%. The pre-processing methodologies developed provided automatic removal of noise from the HS images. The shearlet-based de-noising method improved the classification accuracy by 11% and 19% compared to Savitzky Golay and median filtering, respectively. The total variation based de-noising method removed the dead strips from HS images, which is not achievable with traditional methods. The compression scheme devised, which utilised 2D wavelets and variance decomposition methods, gave a reduction in the size of HS images by a factor of 40. Finally, a methodology based on grey-level co-occurrence matrices is presented for extracting textural information from HS images. The fusion of textural and NIR information showed an improvement in classification accuracy compared to use of spectral or textural information alone.

The methodologies presented will support implementation of high-throughput HSI for assessment of tea products and processes.

## Table of Contents

<b>List of Figures .....</b>	<b>9</b>
<b>List of Tables.....</b>	<b>16</b>
<b><i>Chapter 1 : Hyperspectral imaging as an on-line monitoring tool and associated major challenges.....</i></b>	<b>17</b>
<b>1.1 Hyperspectral imaging.....</b>	<b>17</b>
<b>1.2 NIR hyperspectral imaging: Instrumentation and theory.....</b>	<b>18</b>
1.2.1 Principal of NIR spectroscopy .....	18
1.2.2 Instrumentation for a typical NIR hyperspectral imaging camera.....	21
1.2.3 On-line implementation of NIR hyperspectral imaging .....	21
<b>1.3 Hyperspectral image processing .....</b>	<b>23</b>
<b>1.4 Hyperspectral imaging as an on-line monitoring tool.....</b>	<b>26</b>
<b>1.5 Scope of hyperspectral imaging for monitoring tea products and processes.....</b>	<b>27</b>
<b>1.6 Challenges associated with hyperspectral imaging as an on-line tool.....</b>	<b>36</b>
<b>1.7 Aims and objectives of the study.....</b>	<b>37</b>
<b>1.9 References.....</b>	<b>40</b>
<b><i>Chapter 2 : Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products.....</i></b>	<b>46</b>
<b>2.1 Abstract .....</b>	<b>47</b>
<b>2.2 Introduction.....</b>	<b>47</b>
<b>2.3 Materials and Methods .....</b>	<b>50</b>
2.3.1 Samples.....	50
2.3.2 Hyperspectral imaging measurements.....	50
2.3.4 Data analysis .....	51
<b>2.4 Results .....</b>	<b>58</b>
2.4.1 Spectral profiles of tea samples.....	58
2.4.2 Visualising high dimensional data .....	60
2.4.3 Support vector machine classification.....	64

<b>2.5 Conclusions.....</b>	<b>67</b>
<b>2.6 References.....</b>	<b>69</b>
<b><i>Chapter 3 : Non-destructive classification of the geographic origin of green teas with near-infrared hyperspectral imaging combined with variable selection.....</i></b>	<b>76</b>
<b>3.1 Abstract .....</b>	<b>76</b>
<b>3.2 Introduction.....</b>	<b>77</b>
<b>3.3 Materials and Method.....</b>	<b>81</b>
3.3.1 Samples.....	81
3.3.2 Hyperspectral imaging measurements.....	81
3.3.3 Data analysis .....	82
<b>3.4 Results .....</b>	<b>86</b>
3.4.1 Spectral profiles .....	86
3.4.3 Support vector machine classification.....	95
3.4.4 Feature selection to identify potential NIR regions .....	97
<b>3.5 Conclusions.....</b>	<b>101</b>
<b>3.6 References.....</b>	<b>102</b>
<b><i>Chapter 4 : Application of a wavelet and variance decomposition-based method for compressing close-range hyperspectral images to support process applications..</i></b>	<b>111</b>
<b>4.1 Abstract .....</b>	<b>111</b>
<b>4.2 Introduction.....</b>	<b>112</b>
<b>4.3 Material and methods .....</b>	<b>115</b>
4.3.1 Samples and hyperspectral imaging measurements.....	115
4.3.2 Spatial decomposition .....	116
4.3.3 Spectral decomposition.....	117
4.3.4 Spatial and spectral Correlation .....	121
4.3.5 Multi-class Support Vector Machine .....	121
<b>4.4 Results .....</b>	<b>123</b>
4.4.1 Image planes .....	123
4.4.2 Spectral profiles .....	127
4.4.3 Classification.....	129

<b>4.5 Conclusions.....</b>	<b>132</b>
<b>4.6 References.....</b>	<b>133</b>
<i>Chapter 5 : Automatic de-noising of close-range hyperspectral images with a wavelength-specific shearlet-based image noise reduction method .....</i>	<i>137</i>
<b>5.1 Abstract .....</b>	<b>138</b>
<b>5.2 Introduction.....</b>	<b>139</b>
<b>5.3 Material and methods .....</b>	<b>142</b>
5.3.1 Samples and imaging sensor .....	142
5.3.2 De-noising methodology .....	144
5.3.3 Savitzky-Golay smoothing .....	148
5.3.4 Median filtering.....	149
5.3.5 De-noising performance .....	149
5.3.6 Multiclass support vector machine classification .....	150
<b>5.4 Results .....</b>	<b>151</b>
5.4.1 Noisy and de-noised spectra from simulated images .....	151
5.4.2. Noisy and de-noised image planes from simulated images.....	155
5.4.3 De-noised VNIR image set spectra .....	161
5.4.4 De-noised VNIR image set image planes.....	163
5.4.5 SVM classification on VNIR image set .....	164
<b>5.5 Conclusions.....</b>	<b>164</b>
<b>5.6 References.....</b>	<b>166</b>
<i>Chapter 6 : Application of a spatio-spectral total variation method for automatic correction for different types of noise in close-range hyperspectral images .....</i>	<i>170</i>
<b>6.1 Abstract .....</b>	<b>170</b>
<b>6.2 Introduction.....</b>	<b>171</b>
<b>6.3 Material and methods .....</b>	<b>173</b>
6.3.1 Samples and imaging measurements.....	173
6.3.2 Data pre-processing .....	174
6.3.3 Spatio-spectral total variation model.....	174
<b>6.3.4 Quality parameters .....</b>	<b>176</b>

<b>6.3.5 Classification with support vector machines.....</b>	<b>177</b>
<b>6.4 Results .....</b>	<b>178</b>
6.4.1 De-noising.....	178
6.4.2 PSNR and SSIM.....	181
6.4.3 Classification.....	182
<b>6.5 Conclusions.....</b>	<b>183</b>
<b>6.6 References.....</b>	<b>184</b>
<i><b>Chapter 7 : Fusing spectral and textural information in near-infrared hyperspectral imaging to improve green tea classification modelling.....</b></i>	<i><b>188</b></i>
<b>7.1 Abstract .....</b>	<b>189</b>
<b>7.2 Introduction.....</b>	<b>189</b>
<b>7.3 Material and methods .....</b>	<b>192</b>
7.3.1 Samples.....	192
7.3.2 Hyperspectral imaging measurements.....	193
7.3.3 Data analysis .....	193
<b>7.4 Results .....</b>	<b>199</b>
<b>7.5 Conclusions.....</b>	<b>205</b>
<b>7.6 References.....</b>	<b>207</b>
<i><b>Chapter 8 : Conclusions and future work.....</b></i>	<i><b>213</b></i>
<b>8.1 Conclusions.....</b>	<b>213</b>
<b>8.2 Suggestions for future work.....</b>	<b>217</b>
<b>Annex 1.....</b>	<b>219</b>
<b>Annex 2.....</b>	<b>221</b>
<b>Annex 3.....</b>	<b>231</b>
<b>Annex 4.....</b>	<b>233</b>
<b>Annex 5.....</b>	<b>235</b>



## List of Figures

<i>Figure 1.1: Schematic of common components in a hyperspectral imaging camera.</i>	21
<i>Figure 1.2: Schematic of the scope of the thesis chapters.</i>	38
<i>Figure 2.1: Processing steps for different tea products starting from fresh green tea leaves to final products.</i>	48
<i>Figure 2.2: Illustrative diagram for the hyperspectral imaging setup used to acquire the images of tea samples.</i>	51
<i>Figure 2.3: Absorbance spectra of pure tea samples of yellow, oolong, green, black, white and Pu-erh. (a). Mean absorbance spectra (n = 200). (b) Mean spectra after pre-processing (SNV and Savitzky- Golay smoothing), and (c) standard deviation of the absorbance spectra and spectra after pre-processing. The vertical green lines denote the positions of the main peaks.</i>	58
<i>Figure 2.4 : 2-Dimensional scatter plots for visualising high dimensional tea data. (a). Principal Component Analysis (PCA), (b). Multidimensional Scaling (MDS), (c). Isometric Mapping (ISOMAP), and (d). t-distributed Stochastic Neighbour Embedding (t-SNE). In all the plots, the first dimension is represented in the x-axis and the second in the y-axis, and the six tea products are coloured as follows: Pu-erh (pink), black (sky blue), oolong (Yellow), green (green), white (blue) and yellow (red).</i>	60
<i>Figure 2.5: Mahalanobis distances between the three different cluster groups obtained using PCA (dark-blue), MDS (sky-blue), ISOMAP (light- green) and t-SNE (yellow).</i>	62
<i>Figure 2.6: (a) Greyscale image constructed from the image plane extracted from the hypercube at 1424 nm, (b) Classification maps obtained from the application of the ECOC-SVM model. From left to right the samples can be understood as yellow (dark blue), oolong (light blue), green (cyan), black (light green), white (orange) and Pu-erh (Yellow). (c) Histograms showing the proportion of pixels attributed to the different tea products for the classification maps in (b).</i>	64
<i>Figure 2.7:(a). Greyscale image at 1424 nm for the sample comprising oolong, black and Pu-erh tea, (b). The classification map for the sample comprising oolong, black</i>	

and Pu-erh tea, (c). Pie chart representing the proportion of pixels belonging to a particular class for the classification map presented in (b), (d). Greyscale image at 1424 nm for a sample containing a mixture of all teas, (e). The classification map for a sample containing a mixture of all teas, and (f) Pie chart representing the proportion of the pixels belonging to a particular class for the classification map presented in (e).

.....66

Figure 3.1: A summary of the geographical origin, by country and continent, of the 16 green tea samples. ....81

Figure 3.2: Mean SNV pre-processed absorbance spectral profiles of tea samples (a) corresponding to 16 products, (b) corresponding to 7 countries of origin, and (c) corresponding to 4 continents of origin.....87

Figure 3.3: Standard deviation calculated for NIR spectra of individual products (blue), countries of origin (red) and continents of origin (Yellow). The vertical green lines denote the assignment of wavelengths to different components present in tea products based on the literature.....88

Figure 3.4 : a) Principal component scores (PC1 v PC8) for visualising data based on continent of origin, b) principal component scores (PC1 v PC8) for visualising data based on country of origin, c) PC1 scores for samples colour coded according to the continent of origin, d) PC1 scores for samples colour coded according to the country of origin, e) PC8 scores for samples colour coded according to the continent of origin, and f) PC8 scores for samples colour coded according to the country of origin. The ellipses in c) to f) are to aid visualisation of samples from different continents and countries.....92

Figure 3.5: Loading vectors corresponding to (a) PC1 and (b) PC8. ....94

Figure 3.6: Classification maps for the 16 green tea samples obtained using classification models corresponding to a) different samples, b) country of origin and c) continent of origin. Class 0 presents the dark blue background and does not correspond to any tea product.....95

Figure 3.7: Summary of misclassification between different samples based on: (a) continents and (b) countries. ....96

<i>Figure 3.8: (a). ReliefF threshold for weights and the model accuracy obtained from using features corresponding to a particular threshold. The black horizontal line represents a model accuracy of 90 %. The two vertical dashed lines represent the final threshold used for the three different classification cases. (b) The evolution of the misclassification rate (SFS criterion) with the number of selected wavelengths.....</i>	<i>97</i>
<i>Figure 3.9: Wavelength selection (vertical green lines) performed with ReliefF for 90% validation accuracy. Selected wavelengths for (a) 16 sample classification problem, (b) 7 country classification problem and (c) 4 continent classification problem. Selected wavelengths with sequential forward selection (SFS) for (d) 16 sample classification problem, (e) 7 country classification problem and (f) 4 continent classification problem.....</i>	<i>98</i>
<i>Figure 4.1: Scheme for 2D-DWT decomposition of a hypercube <math>i</math> to high and low pass filtered components.....</i>	<i>116</i>
<i>Figure 4.2: The scheme for performing PCA decomposition and reconstruction for 3D hypercube dataset generated with HSI.....</i>	<i>118</i>
<i>Figure 4.3: The scheme for performing Tucker decomposition and reconstruction of the 3D hypercube dataset generated with HSI.....</i>	<i>120</i>
<i>Figure 4.4: Greyscale images produced using the image plane at 1267 nm for the original and reconstructed data. (a) Original data (approx. 40 MB), (b) data reconstructed using the 2D-DWT and PCA (approx. 5 MB), and (c) data reconstructed using the 2D-DWT and Tucker (approx. 1 MB).....</i>	<i>123</i>
<i>Figure 4.5: Spatial correlation across the entire spectral range for data reconstructed using the 2D-DWT and PCA (blue) and the 2D-DWT and Tucker (red).....</i>	<i>124</i>
<i>Figure 4.6: Greyscale images produced from the image plane at 1736 nm for the original and reconstructed data. (a) Original data, (b) data reconstructed using the 2D-DWT and PCA, and (c) data reconstructed using the 2D-DWT and Tucker.....</i>	<i>126</i>
<i>Figure 4.7: Spectral profiles corresponding to the pixel with <math>x</math> and <math>y</math> co-ordinates of 50 and 300, respectively, obtained from the original data (blue), data reconstructed using the 2D-DWT and PCA (red), and data reconstructed using the 2D-DWT and Tucker (yellow).....</i>	<i>127</i>

<i>Figure 4.8: Spectral correlation calculated for each pixel of the reconstructed data with respect to the original data. (a) Data reconstructed using the 2D-DWT and PCA and (b) data reconstructed using the 2D-DWT and Tucker.....</i>	<i>128</i>
<i>Figure 4.9: Classification accuracy of the original data, the data reconstructed using the 2D-DWT and PCA and the data reconstructed using the 2D-DWT and Tucker. ....</i>	<i>129</i>
<i>Figure 4.10: Classification maps for the original data, data reconstructed using PCA, and data reconstructed using Tucker. The red circles highlight example regions where reconstruction using Tucker gives fewer misclassifications between the classes compared to the original data and reconstruction using PCA. ....</i>	<i>130</i>
<i>Figure 4.11: Percent of pixels belonging to each class in the classification maps (see Figure 4.10) using the original data, data reconstructed using the 2D-DWT and PCA, and data reconstructed using the 2D-DWT and Tucker. ....</i>	<i>131</i>
<i>Figure 5.1: Multi-scale and multi-dimensional three-scale decomposition with a non-subsampled shearlet transform (NSST) where <math>A</math> and <math>D</math> signifies the low and high frequency sub images, respectively, and <math>l = 16, 8</math> and <math>4</math> at scales 1, 2 and 3, respectively.....</i>	<i>147</i>
<i>Figure 5.2: A single spectrum at pixel location (370, 135) extracted from the following image sets: (a). VNIR, (b). VNIR + Gaussian noise, and (c). VNIR + Mixed noise. 20 random wavelengths were used to simulate Gaussian and mixed noise images. ....</i>	<i>151</i>
<i>Figure 5.3: Noisy and de-noised spectra from pixel location (370, 135), the red line shows the clean spectrum, the blue line shows the noisy spectrum and the yellow line shows the spectrum after de-noising. (a). SAVGOL smoothing filter applied to the spectrum with mixed noise added, (c). Median filtering applied to spectrum with Gaussian noise added, (d). Median filtering applied to spectrum with mixed noise added, (e). de-noising with shearlet-based method for spectrum with Gaussian noise added, and (f). de-noising with shearlet-based method for spectrum with mixed noise added.....</i>	<i>152</i>

Figure 5.4: Correlation for the spectra extracted from the pixel at location (370, 135) from noisy and de-noised data with clean data. The blue line dashed shows the Gaussian noise and the red line shows the mixed noise. .... 154

Figure 5.5: The image planes corresponding to six different tea samples (oolong, black, green, yellow, Pu-erh and white (from left to right)) and the simulated noisy and de-noised image planes for data with Gaussian noise added. (a). Reflectance image plane (588 nm), (b) SAVGOL smoothed image plane, (c) Median filtered image plane, (d) Shearlet de-noised image plane, (e). Clean adjacent reflectance image plane (586 nm), (f). Adjacent image plane after SAVGOL smoothing, (g) Adjacent image plane after median filtering, and (h). Adjacent image plane after shearlet de-noising. .... 155

Figure 5.6: The image planes corresponding to six different tea samples (oolong, black, green, yellow, Pu-erh and white (from left to right)) and the simulated noisy and de-noised image planes for data with mixed noise added. (a). Reflectance image plane (588 nm), (b) SAVGOL smoothed image plane, (c). Median filtered image plane, (d) Shearlet de-noised image plane, (e). Clean adjacent reflectance image plane (586 nm), (f) Adjacent image plane after SAVGOL smoothing, (g). Adjacent image plane after median filtering, and (h). Adjacent image plane after shearlet de-noising. .... 156

Figure 5.7: Peak signal-to-noise ratio (PSNR) for the 20 wavelengths with added noise: purple dashed (SAVGOL smoothed data with mixed noise), thick blue dashed (SAVGOL smoothed data with Gaussian noise), thick red dashed and dot (median filtered data with Gaussian noise), green dashed and dot (median filtered data with mixed noise), thick yellow (shearlet de-noised data with Gaussian noise) and dashed sky-blue (shearlet de-noised data with mixed noise). .... 158

Figure 5.8: A single spectrum at pixel location (370,135) extracted from the VNIR hypercube, blue represents the raw reflectance signal, and red represents the de-noised spectra. (a). SAVGOL smoothing, (b). Median filtering, and (c). Shearlet-based de-noising method. .... 161

Figure 5.9: Noisy and de-noised image planes (405 nm) extracted from real hyperspectral data from a VNIR camera. (a). VNIR noisy image plane, (b). SAVGOL

<i>de-noised image plane, (c). Median filter de-noised image plane, and (d) Shearlet de-noised image plane. ....</i>	<i>163</i>
<i>Figure 6.1: Greyscale images produced using the image plane at 1035 nm containing different types of noise: (a) Gaussian, (b) salt and pepper, (c) dead stripe, and (d) a combination of Gaussian, salt and pepper, and dead stripe noise. In each case, the original, noisy and de-noised images are given which can be understood as mentioned at the top in the figure. ....</i>	<i>178</i>
<i>Figure 6.2: Sample spectra extracted from the original, mixed noise and de-noised images from the pixel with x and y co-ordinates of 108 and 29, respectively. ....</i>	<i>180</i>
<i>Figure 6.3: PSNR for noisy and de-noised images. (a) Gaussian and salt and pepper noise, and (b) dead stripe and mixed noise. ....</i>	<i>181</i>
<i>Figure 6.4: SSIM for noisy and de-noised images. (a) Gaussian and salt and pepper noise, and (b) dead stripe and mixed noise. ....</i>	<i>181</i>
<i>Figure 6.5: Classification accuracies of SVM models for noisy and de-noised data. ....</i>	<i>182</i>
<i>Figure 7.1: Schematic of the window operation performed for extracting textural features. ....</i>	<i>196</i>
<i>Figure 7.2: Schematic for raw data-level and feature-level fusion. ....</i>	<i>197</i>
<i>Figure 7.3: Criteria used for selection of the best image plane on which to perform sharpening and textural analysis: a) SSIM and b) PSNR for all image planes in the range 967.11 – 1700 nm. ....</i>	<i>199</i>
<i>Figure 7.4: Greyscale images produced using the image plane at 1381 nm (a) without and (b) with sharpening. ....</i>	<i>200</i>
<i>Figure 7.5: Mean classification accuracies (in percent) of the 16 green tea products obtained for the calibration samples (pixels) using models built with raw data and PCA features. In both cases, models were built using spectral information alone, textural information alone and fused spectral and textural information. The error bars denote <math>\pm 1</math> standard deviation (<math>n = 100</math>). ....</i>	<i>201</i>

*Figure 7.6: Classification maps for the 16 green tea products obtained from SVM modelling of (a). raw spectral information, (b). raw textural information, and (c). concatenated raw spectral and textural information.....202*

*Figure 7.7: Classification maps for the 16 green tea products obtained from SVM modelling of (a). PCA features extracted from spectral information, (b). PCA features extracted from textural information, and (c). concatenated PCA features from spectral and textural information. ....203*

*Figure 7.8: Classification maps for the 16 green tea products obtained from decision-level data fusion, using a majority voting scheme, of the six classification maps obtained from SVM modelling of spectral information, textural information, and spectral and textural information using raw data (Figure 7.6) and PCA features (Figure 7.7). ....203*

*Figure 7.9: Percentage of pixels correctly identified in the classification maps for the 16 green tea products obtained using six different SVM models and decision-level fusion by majority voting.....205*

## List of Tables

<i>Table 1.1: Recent works utilising NIRS for estimation of various chemical parameters responsible for tea quality.....</i>	<i>30</i>
<i>Table 3.1: A summary of the spectral regions and bands selected from the range 950 – 1700 nm using ReliefF and SFS. ....</i>	<i>99</i>
<i>Table 5.1: Details of image sets. ....</i>	<i>144</i>
<i>Table 5.2: SVM classification model accuracies (%) obtained using different denoising methods with the simulated and real datasets (see Table 5.1). The value given is the mean <math>\pm</math> one standard deviation resulting from 100 iterations of a 10-fold cross validated model. ....</i>	<i>160</i>



# Chapter 1 : Hyperspectral imaging as an on-line monitoring tool and associated major challenges

## 1.1 Hyperspectral imaging

Close-range hyperspectral imaging (HSI) is now a well-established technique for non-destructive, non-contact, rapid exploration of spatially-resolved spectral properties of materials. Application of hyperspectral imaging can be found in assessments related to foods [1,2], pharmaceuticals [3], plants [4], fabrics [5], forensics [6] and many more [7]. HSI is the integration of traditional spectroscopic techniques with the imaging modality. Traditional spectroscopy is really limited to the study of homogeneous samples. However, due to the additional imaging modality, HSI provides the possibility to explore spatially inhomogeneous samples. HSI can be explored across different spectroscopic domains such as visible (VIS), near-infrared (NIR), ultra-violet (UV), fluorescence, and Raman, and also at different spatial resolution ranging from remote sensing [8] applications to the imaging of nanomaterials [9].

The acquisition of HS images is typically possible in four different ways i.e., point scanning, line scanning (push broom), wavelength scanning and snapshot scanning [10]. Point scanning uses a linear array detector to record a single spectrum one at a time and then consecutively measures the spectral information over the spatial scene. Line scanning involves utilising a 2D array to collect one line of the imaging scene, which is then repeated over the spatial dimension to generate the complete cube. Wavelength scanning includes scanning the complete imaging scene consecutively for all wavelengths and lastly, snapshot scanning acquires complete images with a single exposure. Of all the configurations, push broom is the most commonly used and is the commercially available configuration for close-range HSI.

A typical push-broom HSI instrument consists of an objective lens, spectrograph, camera, acquisition system, translation stage, illumination and computer system [1]. All these components are integrated together based on the samples to be imaged and

the experimental settings required for imaging. A major part of HSI is related to the processing of the images, which involves tasks such as segmentation, exploration and classification utilising advanced machine learning techniques as well as specific chemometric techniques developed for spectral data processing.

## 1.2 NIR hyperspectral imaging: Instrumentation and theory

The following subsections introduce the theory behind NIR spectroscopy and a description of the instrumentation typically employed for NIR HSI.

### 1.2.1 Principal of NIR spectroscopy

NIR spectroscopy is the spectroscopic method that utilises the near-infrared part of electromagnetic radiation (780 – 2500 nm) for the study of materials in terms of reflection, absorption and transmission. Assuming the chemical bonds to be like a weak spring holding together two or more atoms which vibrate naturally, when more energy (infrared radiation) is added to this system the vibration will become more energetic. However, due to restriction from quantum mechanics the atoms in the molecules are allowed only a few specific energy levels. In more detail, if there are only two atoms the only vibration noticeable will be stretching whereas when the number of atoms increases the vibration can arise due to bending phenomenon giving rise to a whole series of different vibrations. Different chemical bonds vary in strength and so the energy requirement for the bond vibration to move from one level to the next. This energy variation is usually represented as a spectrum presenting the absorptions at different wavelengths due to molecular vibrational transitions. These molecular vibrations can be defined in two mechanical models i.e. harmonic oscillator and anharmonic oscillator.

#### *1.2.1.1 Harmonic oscillator*

The molecular vibration in this model are described using the classical physics model of the harmonic diatomic oscillator as the simplest vibrating system. Here, the diatomic molecules are considered as two spherical masses connected with a spring with a given force constant. Any vibration in the masses connected with a spring leads

to changes in the internuclear distance. The energy of the system can be explained using the Hooke law as:

$$E = \frac{h}{2\pi} \sqrt{\frac{k}{\mu}} \quad (1.1)$$

where  $h$  is the Planck constant,  $k$  is the force constant of the bond,  $\mu$  is the reduced mass of the bonded atoms. Combining the Hooke law with the Newton's force law as in equation 1.2 leads to the equation 1.3 as:

$$E = h\nu \quad (1.2)$$

$$\nu_{vib} = \frac{1}{2\pi} \sqrt{\frac{k}{\mu}} \quad (1.3)$$

Equation 1.3 explains that the frequency of vibration is the function of the bond strength. The vibrations frequencies are the key to the essential information about the structure of the studies compound.

The molecular vibration can be defined using a harmonic oscillator where the potential energy ( $V$ ) can be defined as the function of the displacement of the atoms ( $x$ ) as equation 1.4:

$$V = \frac{1}{2} k(r - r_e)^2 = \frac{1}{2} kx^2 \quad (1.4)$$

where  $V$  is the potential energy,  $r$  is the internuclear distance,  $r_e$  is the internuclear distance at equilibrium and  $x$  is the displacement of vibrating atoms.

Quantum mechanical considerations show that the vibrational energy for the harmonic oscillator has no continuum for vibrational energy levels, but only discrete energy levels defined by  $v$  (0,1,2,3...) and can be understood in equation 1.5 as:

$$E_{vib} = \left(v + \frac{1}{2}\right) \frac{h}{2\pi} \sqrt{\frac{k}{\mu}} \quad (1.5)$$

According to the equation 1.5, transitions between neighbouring and equidistant energy levels are allowed in the harmonic oscillator. The distribution of the molecules among the allowed vibrational levels are defined by Boltzmann distribution. At the room temperature most molecules reside at the ground vibrational level  $v=0$ , hence the first fundamental transition to  $v=1$  dominates the spectrum. Absorption of a light photon has precondition that the frequency of the light photons equals the energy difference between two vibrations states in the bond. The energy absorption can be defined according to the resonance theory which stated that the energy transfer occurs through the molecular polarity defined as dipole moment. The degree of dipole

moment defines the extent of light absorption, further, only heteronuclear, diatomic molecules shows the transitions between the photons of the light and the molecule vibrations.

#### 1.2.1.2 Anharmonic oscillator

The main consideration in the harmonic oscillator involves considering the potential energy curve as parabolic (equation 1.4) and only transitions between consecutive energy levels that change the dipole moment are possible. According to the harmonic oscillator a bond can store infinite amount of energies without bond breaking. However, in practicality, all the bond disassociates when enough energy is applied and vibrating bond is extended, further, there are strong repelling forces when the atoms are forced together. These mechanical anharmonicity thus modifies the classical ball and spring model and the potential energy equation can now be understood as equation 1.6 where the quadratic function of the displacement is extended by adding higher order displacement terms:

$$V = \frac{1}{2}kx^2 + \frac{1}{2}k_a x^3 + \frac{1}{2}k_b x^3 + \dots \quad (1.6)$$

Furthermore, the Schrödinger equation can now be understood as:

$$E_{vib} = h\nu \left( v + \frac{1}{2} \right) - y \left( v + \frac{1}{2} \right)^2 \quad (1.7)$$

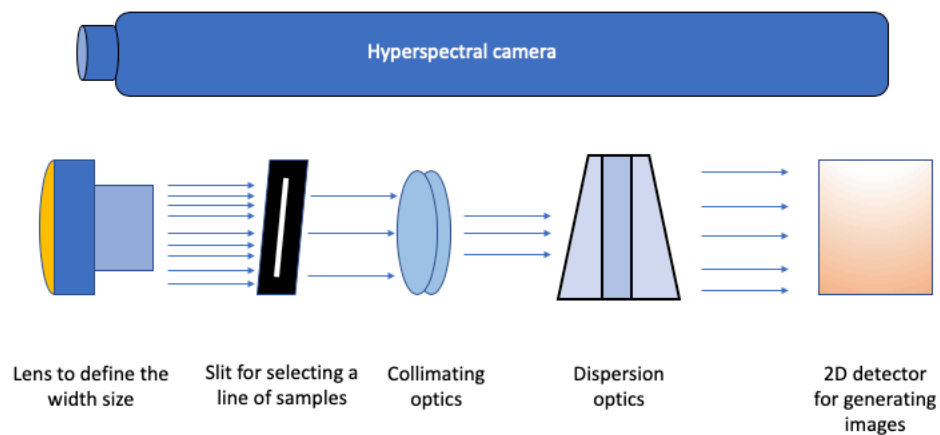
In equation 1.7,  $y$  represents the anharmonicity constant. The anharmonicity now allows transitions between the noncontiguous energy states furthermore, the energy levels are non-longer equidistant and energy differences decrease with increasing quantum number  $v$ .

Absorption in the NIR region is the result of overtones and combinations of fundamental vibrations. Overtone transitions that correspond to quantum numbers greater than one and that appear as multiples of the fundamental vibrational frequency are reflected in the region of 780-2000 nm. Combinations modes that usually present in polyatomic molecules emerge between 1900-2500 nm further they can be understood as the sum of the multiples of each interacting frequency. The number of fundamental absorptions in the molecules is limited to a few, however, due to these multiple combinations and overtones, NIR spectra are usually broad leading to

complex spectra which cannot be directly interpreted to a specific chemical component. To perform the interpretation multivariate calibration techniques are usually required.

### 1.2.2 Instrumentation for a typical NIR hyperspectral imaging camera

The typical NIR instrumentation involves a source of NIR radiation, a detector and a dispersive element such as a diffraction grating to allow the intensity at different wavelengths to be recorded. NIR spectrometers allow the recording of reflectance and transmission depending on the sample to be analysed. The typical light source used for generating broadband NIR radiation is a halogen light bulb. Further, the type of detector used is based on the range of wavelengths to be measured. Silicon based CCDs are suitable for the short end of the NIR range, however they are not very sensitive at wavelengths greater than 1000 nm. For the complete NIR range, InGaAs and PbS devices are more suitable but they are less sensitive than CCDs. Typically for HSI, a 2D array detector with an acousto-optic tuneable filter is used providing multiple images to be recorded sequentially at different narrow wavelength bands. A typical diagram representing the main component of the HSI camera are presented in Figure 1.1.



*Figure 1.1: Schematic of common components in a hyperspectral imaging camera.*

### 1.2.3 On-line implementation of NIR hyperspectral imaging

On-line implementation of HSI requires continuous monitoring of process and product lines. Typically for that two major types of hyperspectral acquisition techniques i.e. line scan and snapshot cameras can be used. The line scan modality is particularly important when the HSI needs to be implemented over a continuous moving process line. In that case, the speed of acquisition is determined by the integration time and the acquisition speed of the camera. From an efficiency point of view, the speed of acquisition should be fast and should match the speed of the conveyor belt. For a typical acquisition for on-line application the camera is placed at an appropriate distance to achieve a relevant field of view (FOV) covering the complete area of interest. To obtain an image with an accurate aspect ratio, the pixel dimensions along the scanning directions need to equal the pixel size along the FOV. Considering a sample of length  $a$ , the number of lines the camera needs to acquire to scan it completely is given by equation 1.8 as

$$\frac{a}{\text{pixel size}}, \quad (1.8)$$

which can be rewritten as 1.9

$$\frac{a \times \text{number of pixels}}{FOV} \quad (1.9)$$

Now, considering the integration time of the camera is given by  $t$ , the maximum frame rate is  $\frac{1}{t}$ . Thus, the final scanning time can be calculated as 1.10

$$\frac{a \times \text{number of pixels}}{FOV \times \text{frame rate}} \quad (1.10)$$

Depending on the imaging scene, the speed can be increased or decreased based on the integration time, where a low integration time will increase the acquisition speed at the expense of the signal-to-noise ratio (SNR) and a too high integration time will saturate the detector and the resulting signal. Finally, the acquisition for the line scan systems can be synchronised with the processing and product lines. However, to make them superfast like real-world video cameras, developing multispectral systems based on the wavelengths identified with a HS camera is of particular interest. Recently, snapshot HS cameras are also emerging which provides a high acquisition speed of almost 200 ms at 7-megapixel resolution [12] Further, HS video cameras are also emerging which provides a continuous video acquisition of the scene.

### 1.3 Hyperspectral image processing

HSI records the information from three dimensions ( $n \times p \times q$ ) of which two are spatial ( $n \times p$ ) and one spectral ( $q$ ) resulting in a dataset commonly known as a hypercube [11]. The images can also be understood as thousands of spectra distributed over the imaged scene as pixels. Typically, the mathematical operations performed over the HSI data are performed either by selecting specific pixels as training samples for any supervised modelling or unfolding the 3D cube ( $n \times p \times q$ ) into a long 2D matrix ( $np \times q$ ) for unsupervised exploration of data. Finally, after processing, the data are refolded to represent the spatial distribution of modelled spectral properties. There are different toolboxes available to perform HSI processing such as HYPER-Tools [13] and the Multi-variate Image Analysis add-on in PLS Toolbox [14]. Some key steps of HSI processing include:

- 1. Radiometric flat field calibration:** The HSI camera usually provides the output as a radiance image. For a radiance image to be converted to reflectance or absorbance requires white and dark reference images. In a typical experiment, the white and dark images are obtained simultaneously with the sample images, and used later for the radiometric calibration of the images as the first step in HSI processing. Typically, the white reference images are obtained from diffuse reflectance standard such a spectralon. The spectralon is a fluoropolymer and has the highest diffuse reflectance of any known material over the UV, VIS and NIR region of the EMR. The dark reference is obtained by putting a cap in front of the camera lens to block the all incoming light and then recording the image. The white and dark reference images are used as 1.11:

$$I_R(i,j,k) = \frac{I_{raw(i,j,k)} - I_{dark(i,j,k)}}{I_{white(i,j,k)} - I_{dark(i,j,k)}} \quad (1.11)$$

where  $I_R$  is the calibrated reflectance,  $I_{raw}$  is the raw intensity measured from the test sample,  $I_{dark}$  the intensity of the dark response,  $I_{white}$  is the intensity of the uniform white reference, and  $i$  and  $j$  are spatial coordinates and  $k$  is the wavelength in the image.

- 2. Pre-processing:** After radiometric calibration, the images are corrected for any noise present. The pre-processing of the images is performed both in the spatial and spectral domains. Spatial filtering mainly aims to compensate for any noise present in the spatial domain through, for example, cropping of the edges and performing filtering to predict missing pixel values. Spectral pre-processing involves dealing with the noise present in the spectral domain such as correcting any kind of discontinuity by performing operations such as smoothing and filtering. Two common approaches to deal with the pre-processing in the spectral and spatial domain are smoothing and filtering where the smoothing involves utilising the polynomials and the filtering based approaches utilising criteria of measure of central tendency to approximate the corrupted information. In, the present work inside each chapter the pre-processing performed on the dataset is well explained inside the material and method sections.
  
- 3. Spectral normalisation:** Radiometric calibration usually compensates for differences in the illumination. However, since a flat white spectral reference is mostly used as a reference it does not compensate for any inhomogeneity at the surface of the samples. Surface inhomogeneity results in the scattering of light, which further modifies the final signal recorded. The scattering can be minimal in the case of samples with a completely flat surface, but can be maximal and consist of multiple scattering in the imaging of powder samples. To correct for these scattering effects, various spectral normalisation and scaling techniques are used. Two common method for reducing the effects of scattering in NIRS are standard normal variate (SNV) normalisation and the multiplicative scatter correction (MSC) methods. The SNV has the benefit over the MSC that it does not require the reference spectrum to perform the correction and is totally based on the mean and the standard deviation of the individual spectrum. In the present work to correct for the scatter effect the SNV normalisation was used as there is need to acquire the reference spectrum.



- 4. Segmentation:** Once the effects of any noise or scattering are corrected for, the task is to extract the subset of pixels that are of interest in the image. The subset of pixels can be used to perform the modelling and later the model can be applied to the complete image to represent the modelled case. Segmentation is the task performed to identify and remove different uninteresting regions from the images. The segmentation task is often carried out utilising the masking techniques. The masks are 2D binary images where the 0 value pixels represent uninteresting pixels and the 1 value pixels represent a pixel of interest. There are different ways of masking, which can be divided into supervised and unsupervised. Supervised techniques can range from identifying manually the region of interest to setting up threshold values by visualising the histogram of the intensities of the pixels in the imaging scene. Unsupervised techniques can range from automatically identifying shapes and edges over the image to utilising advanced classification techniques such as Otsu and K-means clustering.
- 5. Explorative analysis:** Once the images are pre-processed and the relevant pixels are extracted from the images, primary explorative analysis is usually performed. The most commonly used techniques utilise variance-based methods such as principal component analysis (PCA), and independent component analysis (ICA) to enhance the contrast between spectrally different objects present in the imaging scene. The explorative analysis also provides insight into the type of analysis method that should be used in future data processing. Details regarding the explorative analysis performed in this work are presented inside each chapter accompanied with mathematical explanation.
- 6. Predictive analysis:** After the explorative analysis and through enhanced understanding of the data, the images can be processed to perform predictive analysis based on the experimental requirements. Commonly used predictive analysis methods include regression and classification

models. Regression analysis includes training models to perform the prediction of continuous variables. Classification analysis includes performing prediction of discrete variables or classes. Details regarding the classification analysis performed in this work are presented inside each chapter accompanied with mathematical explanation.

- 7. Reconstruction of predicted images after data processing:** Once the predictive analysis is performed on the small training set, the developed models are used to predict the values or classes for every pixel over the image. If the analysis is performed by unfolding the image into 2D matrices, the 1D vector resulting from the model application is reshaped to the spatial dimension as a final output.
  
- 8. Image analysis (any morphological analysis):** The final images (prediction maps/score images/classification maps) can further be processed utilising classical image processing techniques such as contrast enhancement, sharpening, and deblurring for output enhancement. In the present work the texture analysis is performed as an extra step and the details regarding the methodology can be found in chapter 7.

To understand all of the steps involved in HS image processing and related techniques, readers are encouraged to read [11,13].

#### 1.4 Hyperspectral imaging as an on-line monitoring tool

The use of HSI for close-range analysis of materials is now gaining popularity as a potential tool for on-line monitoring of processes and production lines [1]. The aim of utilising HSI as an on-line tool is to design, analyse and control critical quality and performance parameters of materials and manufacturing processes in real-time to ensure process integrity and product quality. In the process control domain, HSI has been popularly known as ‘chemical imaging’ but has always been side-lined compared

to other spectral techniques. However, due to the Food and Drug Administration's (FDA's) recent process analytical technology (PAT) initiative, support for spectral imaging techniques to enable detailed understanding of product and processes has grown [15]. HSI has not only gained interest as a tool for analysis of pharmaceuticals, it has spread to many domains including food engineering. The main benefits of the use of HSI as an on-line monitoring tool can be seen when the product monitoring or testing task are labour intensive and time consuming, when there are opportunities for error leading to an increased risk of product contamination, and when individual samples are not representative of the process and so multi-sample measurements are needed. In such cases, HSI can be a non-invasive and rapid means for acquisition of meaningful information for process and product control. HSI has been used for on-line monitoring of an industrial polymer process [16], a pharmaceutical coating process [17], meat production [18], and fresh-cut lettuce [19], with the number of reported applications continually increasing.

Utilisation of HSI for on-line monitoring involves placing a push-broom camera at a fixed location in the line and moving the sample past the camera. Movement of the samples can be via a conveyor belt for continuous on-line monitoring of processes and production lines. In cases related to foods, on-line monitoring is usually needed to meet the hazard analysis and critical control points (HACCP) requirement of the industrial production process, to control the composition of valuable food components and to ensure the absence of any contamination resulting from microbiological impurities, allergen contaminants and hazardous foreign material [20].

### 1.5 Scope of hyperspectral imaging for monitoring tea products and processes

Tea is the most widely consumed drink in the world after water [21]. Most of the tea products available in the market are obtained by application of different processing conditions to the hand plucked tender leaves of the plant *Camellia sinensis* [22]. The main processing step in the processing of tea is fermentation, which results in the enzymatic oxidation of polyphenols in tea causing the formation of theaflavins and

thearubigins. Based on the extent of fermentation, tea can be classified as three main products: green (unfermented), oolong (semi-fermented), and black (fermented) [23]. The most important chemical constituents that influence the final taste and flavour of tea products are polyphenols, flavonols, caffeine, sugars, organic acids, amino acids, and volatile flavour compounds [24,25]. However, the final quality of processed tea products also depends on factors like genetic strain, climatic conditions, soil, growth altitude and horticultural practices, the plucking season, sorting of the leaves, the processing, and storage. The quality of tea products is of critical importance for an industrial brand to provide its customer with a satisfactory product and to continue to stay in the market. The quality of tea can be expressed in many ways. Sensorially, tea infusion is assessed by its appearance, flavour, and aroma, and these characteristics are reflected in the price of the final tea products. Big industries usually do not grow and process all the tea on their own, they are dependent on various tea auctions around the world to buy tea products. In such a global tea market, assuring the quality of products helps to make buying decisions as the final tea product made for the market has to be a blend of different teas from different sources. Typically, industry uses sensory evaluation experts known as Tea Tasters to make buying decisions during auctions and for any tea quality evaluation in later stages. However, this evaluation method is subjective, inconsistent and vulnerable to human error. Further, in an industrial scenario, the Tea Taster has to analyse a large number of samples putting them under business stress. Also, the tea tasters cannot be used for on-line monitoring of continuous product lines in an industrial scenario. Hence there is a requirement to have a robust and accurate method of tea quality evaluation.

Near-infrared spectroscopy (NIRS) in recent years has gained popularity for rapid and non-destructive assessment of quality parameters in food and agricultural products. NIRS contains abundant chemical information, which can support estimating quality parameters quickly and non-invasively. In particular, a wide range of works has been performed for quantitative estimation of the chemical constituents that are present in tea products (Table 1.1). In addition, imaging and image processing have emerged as potential tools to enable the analysis of the physical properties of tea products such as shape, texture and other visual bulk properties [26,27,28]. However, a combination of

NIRS and imaging is still in its initial stages for assessment of tea products and hence, is the focus of this work.

*Table 1.1: Recent works utilising NIRS for estimation of various chemical parameters responsible for tea quality.*

<i>Spectral range (nm)</i>	<i>Sample state</i>	<i>Type of tea</i>	<i>Task</i>	<i>Parameters</i>	<i>Type of modelling</i>	<i>Model accuracy</i>	<i>Cross Validation</i>	<i>References</i>
909 -2631	Powder (final product)	Green, Black and Oolong	Biochemistry quantification	Caffeine and total polyphenols	Partial least square regression (PLSR)	Caffeine prediction: $R^2=0.9688$ Polyphenol: $R^2=0.9299$	Leave one sample out	(29)

1100-2500	Leaf and powder (better calibration for leaves)	Green	Biochemistry quantification	Caffeine, epigallocatechin gallate (EGCG), epicatechin (EC), Total antioxidant capacity (TAC)	PLSR	Caffeine: $R^2=0.96$ TAC: $R^2=0.90$ EGCG: $R^2=0.83$ EC: $R^2=.44$	Leave one sample out	(30)

1100-2500	Leaves (final product)	Green	Biochemistry quantification	Total antioxidant capacity	Principal component regression (PCR)	$R^2=0.92$	Leave one sample out	(31)
1108-2490	Leaves (final product)	Green	Biochemistry quantification	Gallic acid (GA), EC, epigallocatechin (EGC), epicatechin gallate (ECG), EGCG, caffeine, theobromine, theogallin	PLSR	GA: $R^2= 0.89$ EC: $R^2= 0.97$ EGC: $R^2=0.85$ EGCG: $R^2= 0.93$ ECG: $R^2= 0.95$ Caffeine: $R^2=0.97$ Theobro: $R^2=0.86$ Theogallin: $R^2= 0.94$	Not mentioned but performed	(32)
1000-2500	Powder (final product)	Green	Biochemistry quantification	Total polyphenol	PLSR	$R^2=0.95$	Leave one sample out	(33)
	Leaves (final product)	Black	Biochemistry quantification	Theaflavin and moisture content		Not able to access article		(34)



909-2631	Powder (final product)	Green	Biochemistry quantification	Caffeine, total polyphenols	PLSR	Caffeine: $R^2=0.9688$ Poly: $R^2=0.9299$	Leave one sample out	(35)
900-1700	Leaves (final product)	Green	Biochemistry quantification	Polyphenols	PLSR	$R^2=0.925$	Leave one sample out	(36)
1000-2500	Leaves (final product)	Green	Biochemistry quantification	EGCG	PLSR	$R^2=0.97$	Not mentioned	(37)
1100-2500	Leaves (final product)	Green and black	Biochemistry quantification	Polyphenols, Caffeine, Amino acids, Lignin	Multiple linear regression (MLR)	Polyphenols: $R^2=0.94$ Caffeine: $R^2=0.91$ Amino acids: $R^2=0.98$ Lignin: $R^2=0.97$	Not mentioned	(38)
1000-2500	Powder (final product)	Black	Biochemistry quantification	Amino acid, Caffeine, theaflavins and water extract	PLSR	Amino acid: $R^2=0.9396$ Caffeine: $R^2=0.9195$ Theaflavins: $R^2=0.9056$	k-fold	(39)

						Water extract: R <sup>2</sup> =0.8886		
1000-2500	Powder (final product)	Black, oolong, green	Biochemistry quantification	Total polyphenols, caffeine, Free amino acids	CARS	Total polyphenols: R <sup>2</sup> =0.994 Caffeine: R <sup>2</sup> =0.986 Free amino acids: R <sup>2</sup> =0.993	k-fold	(40)
800-2500	Powder (final product)	Black	Biochemistry quantification	Caffeine, water extract, total polyphenols, free amino acids	PLSR	Caffeine: R <sup>2</sup> =0.983 water extract: R <sup>2</sup> =0.977 total polyphenols: R <sup>2</sup> =0.975 free amino acids: R <sup>2</sup> =0.943	Leave one sample out	(41)
325-1075	Leaves (final product)	Green	Biochemistry quantification	Moisture content	Wavelet	R <sup>2</sup> =0.991	Not mentioned	(42)

408-1117 (HSI)	Leaves (final product)	Green	Classification	Different grades	SVM	95%	Not mentioned	(43)
-------------------	------------------------------	-------	----------------	------------------	-----	-----	------------------	------

## 1.6 Challenges associated with hyperspectral imaging as an on-line tool

Although HSI systems are now readily available on the market, utilising HSI as an on-line monitoring tool comes with different challenges. One of the major challenges is related to the handling and interpretation of the huge volume of data generated during the high-throughput measurements [44]. The high-throughput implementation of HSI can be understood in two ways, one in which the HSI is continuously implemented as an online tool to generate a continuous stream of data where as another where the HSI generate discrete but a huge number of HSI such as in a implementation of high-throughput plant phenotyping. In both high-throughput cases whether continuous or discrete image generation, HSI generate hundreds of MBs of images in a couple of seconds. Some common challenges can be understood below:

- A challenge from an on-line application point of view is the continuous accumulation of huge volumes of data, which requires the use of compression methods for efficient storage of data for future records and use.
- The other challenge comes from the point of view of automating the analysis of data and removing human subjectivity in tasks such as pre-processing. The information generated in HS images is often accompanied by noise, which can arise from detector sensitivity, illumination conditions (e.g. the choice of the light source) and experimental conditions (e.g. interference from other light sources). The typical way of removing noise is to visually check to see if the images contain noise, then the user based on his/her experience will utilise different methods such as smoothing/filtering to remove the noise. However, in an on-line implementation or in an automated scenario, it will not be possible for the user to check and correct every image. Therefore, there is a need for automated methods which can intelligently identify the noise and correct for it based on the identification of the type of noise present in the images.
- The third challenge comes from the point of view of combining the spectral and spatial information obtained from the spectroscopic and imaging modalities of HSI, respectively. The use of spatial information is of particular importance to tea products as differently processed tea products exhibit

different textural properties. The complementary information about the texture can, therefore, enhance the modelling task performed with only spectral information.

### 1.7 Aims and objectives of the study

This research performed in this thesis links to the ModLife project, a H2020 ITN (<http://www.modlife.eu/>) which aims to develop advanced model-based optimisation, monitoring and control as enabling technologies for bioprocess-product development and innovation tailored for the needs of the life science industries. The ModLife ITN aims to develop the next generation of high-performance computing tools and in-situ measurements for increasing the efficiency, innovation and competitiveness of Europe's life sciences and processing industries.

This thesis contributes to the ModLife ITN through evaluation of optical and imaging techniques for biological process related to tea products and processes. Tea product analysis was of major interest to the industrial partner 'Unilever' in the project. The thesis further represents working of the industrial and the academic partners together to solve a problem which is of key importance for a European project like ModLife. Further, in the thesis different data analysis methods are explored and methods for handling large volumes of HSI data are presented. These data analysis methodologies link directly to the ModLife ITN aim to develop novel data modelling methodologies. The methodologies are currently being explored for implementation at Bayer Ag for HSI based assessment of biological plants in high-throughput phenotyping setup.

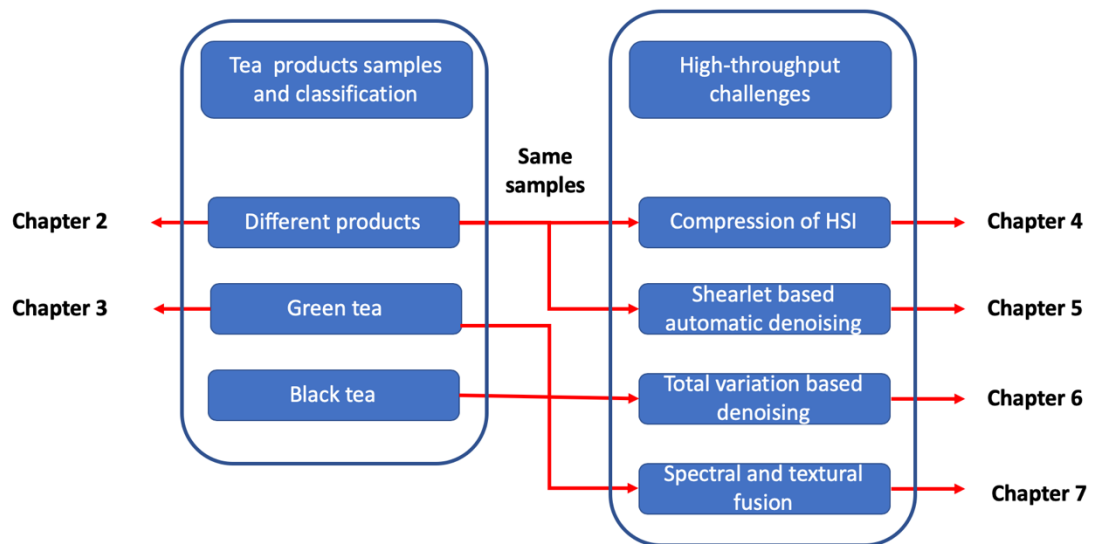


Figure 1.2: Schematic of the scope of the thesis chapters.

The overall aims of this study were to investigate the feasibility of using NIR HSI for assessment of tea products and to address some of the key challenges related to automated handling, managing and processing of HSI data that are needed for on-line application to be fully realised. The work has been performed in collaboration with the ModLife project partner, Unilever, based at Colworth Park in the United Kingdom. A summary of the chapter are presented in figure 1.2. The specific objectives of the study were as follows:

1. To evaluate the feasibility of using NIR HSI for assessment of a wide range of tea products (Chapters 2 and 3).
2. To utilise advanced data visualisation, classification techniques and variable selection techniques to process the NIR HSI data related to tea products (Chapters 2 and 3).
3. To deal with the huge volumes of HS images by utilising data compression methods (Chapter 4).
4. To automate the pre-processing of HS images to identify and correct for various types of noises present (Chapters 5 and 6).
5. To extract and utilise the textural (spatial) information present in images and to fuse textural and spectral information to enhance the performance of data modelling (Chapter 7).

The following six chapters (Chapters 2 to 7) of the thesis are written in the format of journal publications having their own introduction, material and methods, results and discussion, and conclusions sections. Chapter 8 gives the overall conclusions of the research described in this thesis, along with suggestions for future work. Some results related to the tea analysis were confidential to be presented in the thesis, however, these results once cleared from the industrial partners will be published as journal articles and conference proceeding. The readers are advised to follow up the thesis in the coming months and years for more information related to the thesis.

## 1.9 References

1. A.A. Gowen, C.P. O'Donnell, P.J. Cullen, G. Downey, J.M. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, *Trends Food Sci. Technol.* 18 (2007) 590–598. doi:<https://doi.org/10.1016/j.tifs.2007.06.001>.
2. Y. Liu, H. Pu, D.-W. Sun, Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications, *Trends Food Sci. Technol.* 69 (2017) 25–35. doi:<https://doi.org/10.1016/j.tifs.2017.08.013>.
3. S.J. Hamilton, R.A. Lodder, Hyperspectral imaging technology for pharmaceutical analysis, in: *SPIE 2002*: pp. 4612–4626. <https://doi.org/10.1117/12.472076>.
4. P. Mishra, M.S.M. Asaari, A. Herrero-Langreo, S. Lohumi, B. Diezma, P. Scheunders, Close range hyperspectral imaging of plants: A review, *Biosyst. Eng.* 164 (2017) 49–67. doi:<https://doi.org/10.1016/j.biosystemseng.2017.09.009>.
5. J.A. Herweg, J.P. Kerekes, E.J. Ientilucci, M.T. Eismann, Spectral variations in HSI signatures of thin fabrics for detecting and tracking of pedestrians, in: *SPIE 2011*: pp. 80400G–8040–11. <https://doi.org/10.1117/12.883421>.
6. G.J. Edelman, E. Gaston, T.G. van Leeuwen, P.J. Cullen, M.C.G. Aalders, Hyperspectral imaging for non-contact analysis of forensic traces, *Forensic Sci. Int.* 223 (2012) 28–39. doi:<https://doi.org/10.1016/j.forsciint.2012.09.012>.
7. B. Boldrini, W. Kessler, K. Rebner, R.W. Kessler, Hyperspectral Imaging: A Review of Best Practice, Performance and Pitfalls for in-line and on-line Applications, *J. Near Infrared Spectrosc.* 20 (2012) 483–508. <http://www.osapublishing.org/jnirs/abstract.cfm?URI=jnirs-20-5-483>.



8. T. Adão, J. Hruška, L. Pádua, J. Bessa, E. Peres, R. Morais, J.J. Sousa, Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry, *Remote Sens.* 9(11) (2017) 1110. doi:10.3390/rs9111110.
9. G.A. Roth, S. Tahiliani, N.M. Neu-Baker, S.A. Brenner, Hyperspectral microscopy as an analytical tool for nanomaterials, *Wiley Interdiscip. Rev. Nanomedicine Nanobiotechnology.* 7 (2015) 565–579. doi:10.1002/wnan.1330.
10. L. Gao, R.T. Smith, Optical hyperspectral imaging in microscopy and spectroscopy - a review of data acquisition, *J. Biophotonics.* 8 (2015) 441–456. doi:10.1002/jbio.201400051.
11. J.M. Amigo, H. Babamoradi, S. Elcoroaristizabal, Hyperspectral image analysis. A tutorial, *Anal. Chim. Acta.* 896 (2015) 34–51. doi:<https://doi.org/10.1016/j.aca.2015.09.030>.
12. <https://www.imec-int.com/en/hyperspectral-imaging> accessed on 1 March 2019.
13. N. Mobaraki, J.M. Amigo, HYPER-Tools. A graphical user-friendly interface for hyperspectral image analysis, *Chemom. Intell. Lab. Syst.* 172 (2018) 174–187. doi:<https://doi.org/10.1016/j.chemolab.2017.11.003>.
14. [http://www.eigenvector.com/software/mia\\_toolbox.htm](http://www.eigenvector.com/software/mia_toolbox.htm) accessed on 04 January 2019.
15. <https://www.fda.gov/downloads/drugs/guidances/ucm070305.pdf> accessed on 05 January 2019.
16. D. Ishikawa, T. Nishii, F. Mizuno, H. Sato, S.G. Kazarian, Y. Ozaki, Potential of a Newly Developed High-Speed Near-Infrared (NIR) Camera (Compovision) in Polymer Industrial Analyses: Monitoring Crystallinity and Crystal Evolution of Polylactic Acid (PLA) and Concentration of PLA in PLA/Poly-(R)-3-Hydroxybutyrate (PHB) Blends, *Appl. Spectrosc.* 67 (2013) 1441–1446. doi:10.1366/13-07103.

17. O. Daikos, K. Heymann, T. Scherzer, Monitoring of thickness and conversion of thick pigmented UV-cured coatings by NIR hyperspectral imaging, *Prog. Org. Coatings*. 125 (2018) 8–14. doi:<https://doi.org/10.1016/j.porgcoat.2018.08.022>.
18. M. Kamruzzaman, Y. Makino, S. Oshita, Online monitoring of red meat color using hyperspectral imaging, *Meat Sci.* 116 (2016) 110–117. doi:<https://doi.org/10.1016/j.meatsci.2016.02.004>.
19. C. Mo, G. Kim, J. Lim, Online hyperspectral imaging system for evaluating quality of agricultural products, in: *SPIE 2017*: pp. 103293G–10329–7. <https://doi.org/10.1117/12.2269895>.
20. E. Ostertag, R.W. Kessler, Hyperspectral imaging for process analytical technology, *Chim. Oggi-Chemistry Today*. 29 (2011) 2.
21. S. Sang, Tea: Chemistry and Processing, in: B. Caballero, P.M. Finglas, F.B.T.-E. of F. and H. Toldrá (Eds.), Academic Press, Oxford, 2016: pp. 268–272. doi:<https://doi.org/10.1016/B978-0-12-384947-2.00685-1>.
22. H. Lv, Y. Zhang, Z. Lin, Y. Liang, Processing and chemical constituents of Pu-erh tea: A review, *Food Res. Int.* 53 (2013) 608–618. doi:<https://doi.org/10.1016/j.foodres.2013.02.043>.
23. K. Chang, 2015. World tea production and trade: current and future development, Food and Agriculture Organisation (FAO), United Nations (UN), Rome. <http://www.fao.org/3/a-i4480e.pdf> (last accessed: 14 September 2017)
24. J. Ruan, R. Haerdter, J. Gerendás, Impact of nitrogen supply on carbon/nitrogen allocation: a case study on amino acids and catechins in green tea (*Camellia sinensis* (L.) O. Kuntze) plants, *Plant Biol.* 12 (2010) 724–734. doi:[10.1111/j.1438-8677.2009.00288.x](https://doi.org/10.1111/j.1438-8677.2009.00288.x).
25. B. Ozturk, F. Seyhan, I.S. Ozdemir, B. Karadeniz, B. Bahar, E. Ertas, S. Ilgaz, Change of enzyme activity and quality during the processing of Turkish green tea, *LWT - Food Sci. Technol.* 65 (2016) 318–324. doi:<https://doi.org/10.1016/j.lwt.2015.07.068>.

26. Q. Chen, J. Zhao, J. Cai, Identification of Tea Varieties Using Computer Vision, *Transactions of the ASABE*. 51 (2008) 623-628. doi:10.13031/2013.24363.
27. X. Wang, J. Huang, W. Fan, H. Lu, Identification of green tea varieties and fast quantification of total polyphenols by near-infrared spectroscopy and ultraviolet-visible spectroscopy with chemometric algorithms, *Anal. Methods*. 7 (2015) 787–792. doi:10.1039/C4AY02106A.
28. H. Zhu, Y. Ye, H. He, C. Dong, Evaluation of green tea sensory quality via process characteristics and image information, *Food Bioprod. Process*. 102 (2017) 116–122. doi:https://doi.org/10.1016/j.fbp.2016.12.004.
29. Q. Chen, J. Zhao, H. Zhang, X. Wang, Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration, *Anal. Chim. Acta*. 572 (2006) 77–84. doi:http://dx.doi.org/10.1016/j.aca.2006.05.007.
30. J. Luypaert, M.H. Zhang, D.L. Massart, Feasibility study for the use of near infrared spectroscopy in the qualitative and quantitative analysis of green tea, *Camellia sinensis* (L.), *Anal. Chim. Acta*. 478 (2003) 303–312. doi:https://doi.org/10.1016/S0003-2670(02)01509-X.
31. M.H. Zhang, J. Luypaert, J.A. Fernández Pierna, Q.S. Xu, D.L. Massart, Determination of total antioxidant capacity in green tea by near-infrared spectroscopy and multivariate calibration, *Talanta*. 62 (2004) 25–35. doi:https://doi.org/10.1016/S0039-9140(03)00397-7.
32. H. Schulz, U.H. Engelhardt, A. Wegent, H.-H. Drews, S. Lapczynski, Application of Near-Infrared Reflectance Spectroscopy to the Simultaneous Prediction of Alkaloids and Phenolic Substances in Green Tea Leaves, *J. Agric. Food Chem*. 47 (1999) 5064–5067. doi:10.1021/jf9813743.
33. Q. Chen, J. Zhao, M. Liu, J. Cai, J. Liu, Determination of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms., *J. Pharm. Biomed. Anal*. 46 (2008) 568–573. doi:10.1016/j.jpba.2007.10.031.

34. M.N. Hall, A. Robertson, C.N.G. Scotter, Near-infrared reflectance prediction of quality, theaflavin content and moisture content of black tea, *Food Chem.* 27 (1988) 61–75. doi:[https://doi.org/10.1016/0308-8146\(88\)90036-2](https://doi.org/10.1016/0308-8146(88)90036-2).
35. Q. Chen, J. Zhao, X. Huang, H. Zhang, M. Liu, Simultaneous determination of total polyphenols and caffeine contents of green tea by near-infrared reflectance spectroscopy, *Microchem. J.* 83 (2006) 42–47. doi:<http://dx.doi.org/10.1016/j.microc.2006.01.023>.
36. A.K. Hazarika, S. Chanda, S. Sabhapondit, S. Sanyal, P. Tamuly, S. Tasrin, D. Sing, H. Naskar, B. Tudu, R. Bandyopadhyay, On-site estimation of total polyphenol in fresh tea leaf using near-infrared spectroscopy, *NIR News.* 29 (2018) 9–14. doi:[10.1177/0960336017750326](https://doi.org/10.1177/0960336017750326).
37. Z. Guo, Q. Chen, L. Chen, W. Huang, C. Zhang, C. Zhao, Optimization of Informative Spectral Variables for the Quantification of EGCG in Green Tea Using Fourier Transform Near-Infrared (FT-NIR) Spectroscopy and Multivariate Calibration, *Appl. Spectrosc.* 65 (2011) 1062–1067. doi:[10.1366/11-06287](https://doi.org/10.1366/11-06287).
38. S.H. Yan, Evaluation of the composition and sensory properties of tea using near infrared spectroscopy and principal component analysis, *J. Near Infrared Spectrosc.* 13 (2005) 313–325.
39. M. Zareef, Q. Chen, Q. Ouyang, F. Kutsanedzie, M.M. Hassan, V. Annavaram, A. Wang, Prediction of amino acids, caffeine, theaflavins and water extract in black tea by FT-NIR spectroscopy coupled chemometrics algorithms, *Anal. Methods.* (2018) 3023-3031. doi:[10.1039/C8AY00731D](https://doi.org/10.1039/C8AY00731D).
40. J. Wang, Y. Wang, J. Cheng, J. Wang, X. Sun, S. Sun, Z. Zhang, Enhanced cross-category models for predicting the total polyphenols, caffeine and free amino acids contents in Chinese tea using NIR spectroscopy, *LWT.* 96 (2018) 90–97. doi:<https://doi.org/10.1016/j.lwt.2018.05.012>.
41. G. Ren, S. Wang, J. Ning, R. Xu, Y. Wang, Z. Xing, X. Wan, Z. Zhang, Quantitative analysis and geographical traceability of black tea using Fourier

transform near-infrared spectroscopy (FT-NIRS), *Food Res. Int.* 53 (2013) 822–826. doi:<https://doi.org/10.1016/j.foodres.2012.10.032>.

42. Li, X., Xie, C., He, Y., Qiu, Z., Zhang, Y., 2012. Characterizing the Moisture Content of Tea with Diffuse Reflectance Spectroscopy Using Wavelet Transform and Multivariate Analysis. *Sensors* 12(7) (2012) 9847-9861. <https://doi.org/10.3390/s120709847>
43. J. Zhao, Q. Chen, J. Cai, Q. Ouyang, Automated tea quality classification by hyperspectral imaging, *Appl. Opt.* 48 (2009) 3557–3564. doi:10.1364/AO.48.003557.
44. C. Ferrari, G. Foca, A. Ulrici, Handling large datasets of hyperspectral images: Reducing data size without loss of useful information, *Anal. Chim. Acta.* 802 (2013) 29–39. doi:<https://doi.org/10.1016/j.aca.2013.10.009>.

## Chapter 2 : Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products

Puneet Mishra<sup>1</sup>, Alison Nordon<sup>1</sup>, Julius Tschannerl<sup>2</sup>, Guoping Lian<sup>3,4</sup>, Sally Redfern<sup>3</sup>, Stephen Marshall<sup>2</sup>

<sup>1</sup>*WestCHEM, Department of Pure and Applied Chemistry and Centre for Process Analytics and Control Technology, University of Strathclyde, 295 Cathedral Street, Glasgow, G1 1XL, United Kingdom*

<sup>2</sup>*Hyperspectral Imaging Centre, Department of Electronic and Electrical Engineering, University of Strathclyde, 204 George Street, Glasgow, G1 1XW, United Kingdom*

<sup>3</sup>*Unilever R&D Colworth, Colworth House, Sharnbrook, Bedford, MK44 1LQ, United Kingdom*

<sup>4</sup>*Department of Chemical and Process Engineering, University of Surrey, Guildford GU2 7XH, United Kingdom*

**Paper Published as:** P. Mishra, A. Nordon, J. Tschannerl, G. Lian, S. Redfern, S. Marshall, Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products, *J. Food Eng.* 238 (2018) 70-77. doi:<https://doi.org/10.1016/j.jfoodeng.2018.06.015>.

### **Contribution:**

- Performed HSI NIR measurements
- Conducted chemometrics analysis of data

- Wrote the first draft of article
- Corrected article based on comments of co-authors

## 2.1 Abstract

Tea is the most consumed manufactured drink in the world. In recent years, various high-end analytical techniques such as high-performance liquid chromatography have been used to analyse tea products. However, these techniques require complex sample preparation, are time consuming, expensive and require a skilled analyst to carry out the experiments. Therefore, to support rapid and non-destructive assessment of tea products, the use of near infrared (NIR) (950 - 1760 nm) hyperspectral imaging (HSI) for classification of six different commercial tea products (oolong, green, yellow, white, black and Pu-erh) is presented. To visualise the HSI data, linear (principal component analysis (PCA) and multidimensional scaling (MDS)) and non-linear (t-distributed stochastic neighbour embedding (t-SNE) and isometric mapping (ISOMAP)) data visualisation methods were compared. t-SNE provided separation of the six commercial tea products into three groups based on the extent of processing: minimally processed, oxidised and fermented. To perform the classification of different tea products, a multi-class error-correcting output code (ECOC) model containing support vector machine (SVM) binary learners was developed. The classification model was further used to predict classes for pixels in the HSI hypercube to obtain the classification maps. The SVM-ECOC model provided a classification accuracy of  $97.41 \pm 0.16\%$  for the six commercial tea products. The methodology developed provides a means for rapid, non-destructive, *in situ* testing of tea products, which would be of considerable benefit for process monitoring, quality control, authenticity and adulteration detection.

**Keywords:** Imaging spectroscopy, hypercube, multivariate, data visualisation, neighbourhood methods

## 2.2 Introduction

Being the oldest beverage, tea is the most consumed drink in the world [1]. Different

tea products exist due to different processes for freshly harvested tea leaves [2]. There are six main types of tea products, i.e. oolong, green, yellow, white, black and Pu-erh [3], which differ in terms of processing (see Figure 2.1). Green, yellow and white tea products undergo minimal processing, oolong and black tea products have been oxidised while Pu-erh tea has been fermented. The chemical composition of fresh tea (*Camellia sinensis*) leaves is a complex mixture of caffeine, polyphenols, polysaccharides and nutrients such as protein, amino acids, lipids, and vitamins [4]. Typically, during the processing of fresh tea leaves, such as oxidation and fermentation, they undergo chemical compositional changes. Free amino acids, total tea polyphenols, soluble sugars, and caffeine are the four major chemical components that determine the nature and quality of the final tea products [5].

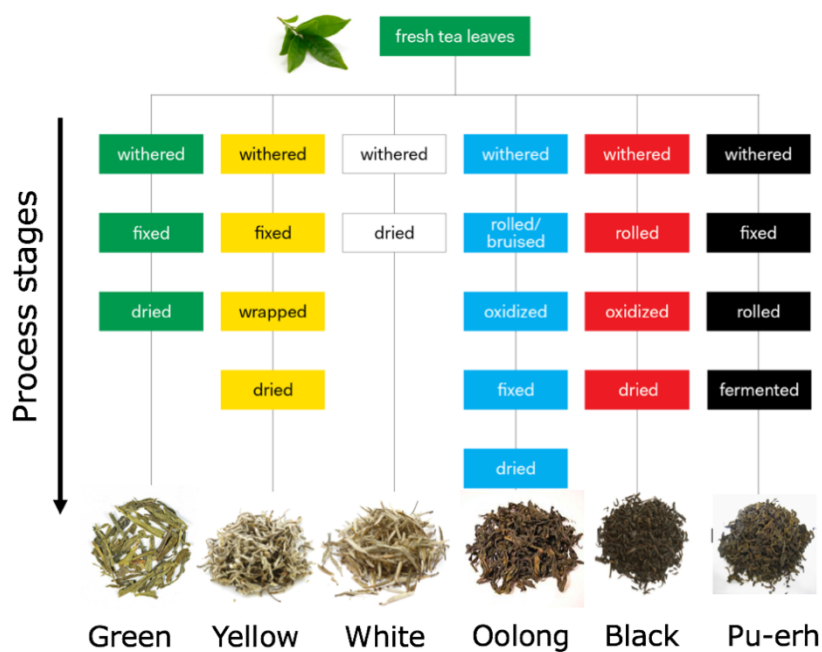


Figure 2.1: Processing steps for different tea products starting from fresh green tea leaves to final products.

Analytical methods used to measure chemical constituents as quality indicators of plant-based products include high-performance liquid chromatography (HPLC) [6], liquid chromatography/mass spectrometry (LC/MS) [7], gas chromatography/mass spectrometry (GC/MS) [8] and electrochemical systems [9,10]. However, these methods have complex sample preparation, are time consuming, expensive and require a skilled analyst to carry out the experiments [11]. A non-destructive technique that



has been used for analysis of tea processes and quality monitoring is e-nose [12,13]. E-nose devices usually include an array of metal oxide sensors which respond to the number of biochemical volatiles coming into contact with the corresponding sensor surface to explain the chemical profile [14]. However, a major disadvantage of e-nose sensors is that they are affected by environmental conditions such as temperature and humidity, which leads to sensor drift [15].

In recent years, there has been increasing interest in the use of optical spectroscopic techniques for rapid, non-destructive assessment of food products. NIR spectroscopy is particularly attractive for this purpose, where changes in the NIR spectral profiles can be correlated to perform qualitative and quantitative analysis of food products [16,17]. NIR spectroscopy has been explored for discrimination [18,19], identification [20, 21] and quality assessment [22] of tea products. Also reported for non-destructive tea analysis are emerging studies utilising imaging techniques for the identification [23], classification [24] and for evaluation of sensory quality [25] of tea products. Integration of spectroscopy and imaging is known as hyperspectral imaging (HSI) and use of NIR-HSI still seems unexplored in its application to the analysis of tea products.

HSI has been widely used in remote sensing for military applications [26], but it is now popular in scientific domains such as forensics [27], medical [28], food [29], pharmaceutical [30] and plants [31]. There are reports of the use of HSI for the understanding of different food products such as coffee [32], tobacco [33], and seeds of vegetable and fruits [34,35]. Some applications of HSI of tea have been reported but these studies only considered a single variety of tea and measured the visible and very near infrared (VNIR) range (around 400 - 1000 nm), which is dominated by the pigments and physical characteristics of the samples [36,37]. In comparison to the VNIR region, the NIR region provides more detailed chemical information such as overtones resulting from the molecular vibration of O-H, C-H, N-H bonds and their combinations, which can support a better classification system based on the chemistry of the samples [38].

The aim of the present work is to demonstrate the use of NIR (950 - 1760 nm) HSI for rapid, non-destructive classification of six different commercial tea products (oolong, green, yellow, white, black and Pu-erh). The study investigates and compares four

different dimensionality reduction techniques (linear and non-linear) to visualise the high dimensional HSI tea data. Furthermore, multi-class support vector machine (SVM) modelling has been performed to generate spatial classification maps of tea products.

## 2.3 Materials and Methods

### 2.3.1 Samples

Six commercial tea samples were obtained from the local market (Glasgow, United Kingdom). The samples were obtained in airtight sealed packaging and stored at ambient temperature. All samples of tea were in loose-leaf form. Black, green and white tea were from Vahdam Teas (New Delhi, India), oolong tea was from Yamamotoyama (California, USA), Pu-erh tea was from The Tea Makers of London (London, United Kingdom) and yellow tea was of an unspecified Chinese origin. The six tea products can also be broadly grouped as minimally processed (green, white and yellow), oxidised (black and oolong tea) and fermented (Pu-erh tea). The samples for each imaging experiment were transferred on the day of analysis into a black plastic circular container (diameter = 3.3 cm, depth = 1.3 cm). A different cap was used for each tea to avoid any cross-contamination.

### 2.3.2 Hyperspectral imaging measurements

Imaging was performed with a push-broom line scan HSI camera (*Model name: RedEye 1.7*) from INNO-SPEC (Nurnberg, Germany). The camera has an InGaAs sensor and generates a spatial map of 320 x 256 pixels in the spectral range of 950 - 1760 nm. The pixel size was 30 x 30  $\mu\text{m}^2$  and the spectral resolution was 3.2 nm. The camera communicated with the computer via a gigabit Ethernet connection. The lighting was provided by two halogen light sources 50 W each and the integration time used was 300 ms. Imaging was performed by placing the samples over the translation stage which was controlled by an independent stage motor connected to the computer system (Zolix TSA 200 BF). The speed of the translation stage was optimised before image acquisition to avoid any distortion in the shape of the image arising from the overlapping of the spectral information in the adjacent pixels. The image acquisition and management of settings (integration time) were performed using the software

interface called SiCAP provided with the camera by INNO-SPEC. Images were first acquired of six different tea samples placed adjacent to each other in their respective sample containers in the field of view of the camera. An image was then acquired of black, Pu-erh and oolong teas where each tea occupied approximately a third of the volume of the sample container; the teas were not physically mixed. Finally, equal proportions of all six tea samples were mixed, by manually shaking the different tea products in a container, and an image of the mixture was acquired. One image was acquired of each sample, with each image comprising more than 2000 pixels (spectra) for the individual tea samples and more than 11200 pixels for the samples containing more than one type of tea. An illustration of the HSI setup conFIGured for imaging of tea samples can be found in Figure 2.2.

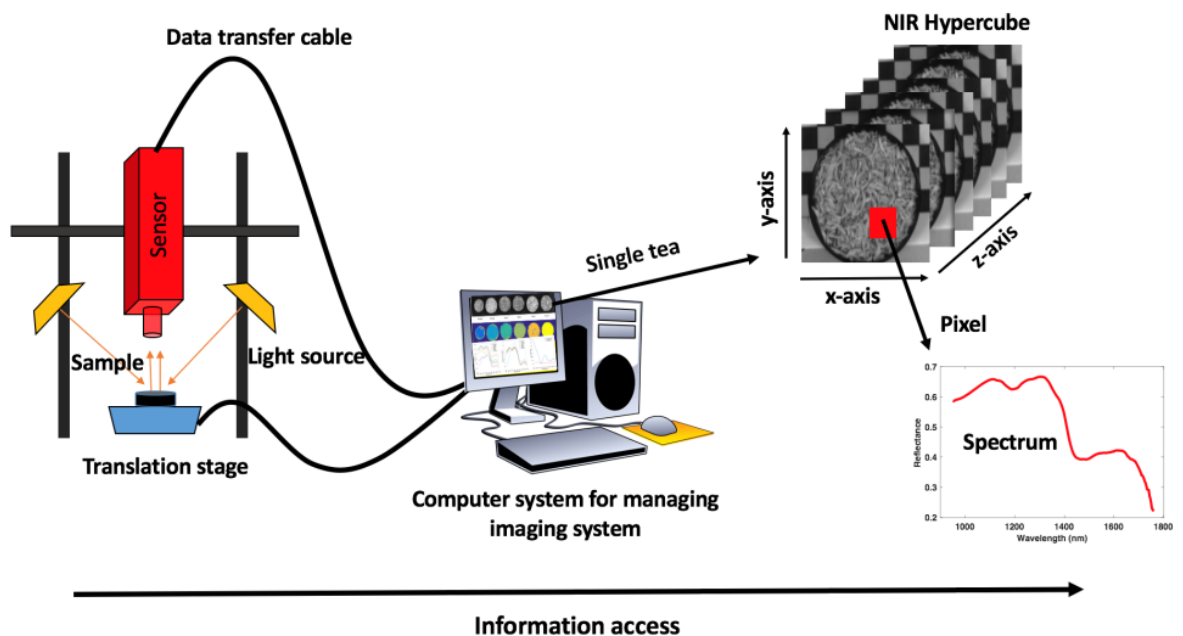


Figure 2.2: Illustrative diagram for the hyperspectral imaging setup used to acquire the images of tea samples.

### 2.3.4 Data analysis

#### 2.3.4.1 Pre-processing of HSI data

The data cubes not only contain information about the samples imaged but also consist of different unwanted influences in signal resulting from factors such as illumination intensity, the detector sensitivity and transmission properties of the optics. The effects

resulting from these factors are both wavelength dependent and independent. To correct for these effects, radiometric calibration was performed using dark and white reference images acquired along with the samples. The correction was performed for every pixel in the HS image according to equation 2.1:

$$I_{R(i,j,k)} = \frac{I_{raw(i,j,k)} - I_{dark(i,j,k)}}{I_{white(i,j,k)} - I_{dark(i,j,k)}} \quad (2.1)$$

where  $I_R$  is the calibrated reflectance,  $I_{raw}$  is the raw intensity measured from the test sample,  $I_{dark}$  the intensity of the dark response,  $I_{white}$  is the intensity of the uniform white reference, and  $i$  and  $j$  are spatial coordinates and  $k$  is the wavelength in the image.

Often, the radiometric correction is sufficient to remove the effects of illumination inhomogeneity from the spectral data, however, when the sample surfaces are not uniform, as in the case of samples of loose tea leaves, the light scattering during diffuse reflection causes additive and multiplicative effects [38]. These scattering effects lead to baseline shifts in the spectrum and variation in the global intensity, which is again dependent on the wavelength. Standard normal variate (SNV) is a very common technique used in NIR spectroscopy to remove these effects [39]. In SNV, the mean and standard deviation of each spectrum for each pixel are calculated, the mean is subtracted, and the standard deviation is used to normalise the difference. This transformation normalises each spectrum to zero mean and unit standard deviation. Before applying the SNV transform, the spectral range was reduced from 950 - 1760 nm to 967 nm - 1700 nm, to remove the noisy regions at the edges of the spectral range, and converted to absorbance. Further, the spectral absorbance profiles were smoothed with a Savitzky-Golay filter (15-point width and second order polynomial) [40]. The *savgol* and *snv* functions from PLS Toolbox (version 8.11, Eigenvector Research Inc., USA) were used. All visualisation and classification analysis were performed on the pre-processed spectra. The pre-processed pure spectra of six pure tea samples were extracted using Matlab's (R2016b, Mathworks, USA) *roipoly* function. The *roipoly* function provides a graphical user interface in Matlab to extract the information from each image over the manually selected locations.

### 2.3.4.2 Principal Component Analysis

Principal component analysis (PCA) introduced by Pearson in 1901 belongs to the family of linear methods for visualising high dimension data [41]. In PCA, a set of observations containing correlated variables is orthogonally transformed to linearly uncorrelated variables defined as principal components (PCs). In PCA, the transformation is performed to retain the major amount of variability in the dataset.

The PCA decomposition model for a given observation data matrix  $X$  can be understood as equation 2.2:

$$X = TW^T \quad (2.2)$$

where  $T$  is the score in the lower dimension explained by the number of PCs specified and  $W$  is a  $l \times l$  ( $l$  denotes number of variables) matrix whose columns are the eigenvectors of  $X^T X$ .

In the case of dimensionality reduction, the aim is to preserve the maximum amount of meaningful variation present in the dataset. The extracted PCs define a new orthonormal basis set which can be used to transform the data from a high dimension space to the lower space explained by the PCs. PCA from a dimensionality reduction perspective can be understood as minimising the squared reconstruction error as given in equation 2.3.

$$\min ||TW^T - T_r W_r^t||^2 \quad (2.3)$$

where,  $TW$  and  $T_r W_r$  are the reconstructed original dataset in higher and lower dimensional space, respectively. Minimisation of the reconstruction error results in the maximisation of the information that was present in the higher dimensional space when defined in the lower dimensional space given by the significant number of PCs. To interpret the data in two- or three-dimensional plots, the respective PCs can be selected and used for transformation to the orthogonal axes represented by the PCs. Transformation from a higher dimension to a lower dimension can be performed as in equation 2.4.

$$\hat{X}_r = XW \quad (2.4)$$

### 2.3.4.3 Multi-Dimensional Scaling

Multi-dimensional scaling (MDS) is a linear method for visualising high dimensional data [42]. MDS performs a transformation by preserving the between object distances from the higher dimension to lower dimension. The MDS utilises calculation of the Euclidean distances for each data point in the multidimensional space to capture the pattern. The distances are defined as a symmetric distance matrix ( $D$ ). MDS attempts to find data points in a specified ( $d$ -dimensional) space such that the Euclidean distance between data points ( $\hat{D}$ ) is similar to the distance in higher dimensional space. The minimisation function can be understood as equation 2.5:

$$\min \sum_i \sum_j \|d_{ij} - \hat{d}_{ij}\|^2 \quad (2.5)$$

where,  $D = d_{ij} = \|x_i - x_j\|^2$  and  $\hat{D} = \hat{d}_{ij} = \|y_i - y_j\|^2$  explaining the Euclidean distance between points in high ( $x_i, x_j$ ) and low dimensional space ( $y_i, y_j$ ), respectively.  $i, j$  denotes the specific position of a point.

### 2.3.4.4 Isometric Mapping

Isometric mapping (ISOMAP) belongs to the family of non-linear techniques for visualising high dimensional data [43,44]. ISOMAP can be understood as a generalised non-linear form of MDS which utilises the geodesic space accounting for the non-linearity in the high dimensional data manifold. The geodesic distance is defined as the shortest distance between two data points on a curved surface of a non-linear manifold. As a first step, ISOMAP approximates a neighbourhood graph by identifying  $k$  nearest neighbours (kNNs) or selecting neighbourhood data points based on any other condition for all data points. The geodesic distance is then approximated for all the pairs of data points on the neighbourhood graph. Finally, the distance data obtained from the graph is embedded to a lower dimension Euclidean space using MDS as shown in equation (2.6).

$$\min \sum_i \sum_j ||D_G - D_E||^2 \quad (2.6)$$

where,  $D_G$  and  $D_E$  explain the geodesic and Euclidean distance between points in high and low dimensional space, respectively.

#### 2.3.4.5 *t-Distributed Stochastic Neighbour Embedding*

t-distributed stochastic neighbour embedding (t-SNE) is a non-linear technique used to visualise high dimensional data in two- or three-dimensional scatter plots [45]. Unlike PCA, t-SNE is based on utilising local relationships between data points to create a low dimensional mapping which allows to capture non-linear relationships. t-SNE further utilises the local relationship between data point to create a probability distribution utilising Gaussian distribution which defines the relationship between respective data points in high-dimensional space. t-SNE then uses a t-distribution in the low dimensional space which explain the probabilities. The optimisation step in final try to make the distribution in low dimensional as similar to that in the high dimension. The optimisation process actually make the distribution in the low dimension learn the pattern from the high dimensional data. The main objective of t-SNE is to model similar points using nearby points (small pairwise distance) and dissimilar points using distant points (large pairwise distances). As a first step, to represent the similarity, the t-SNE converts high-dimensional Euclidean distances between data points into conditional probabilities using a Gaussian distribution. The joint probability for a data point  $\mathbf{x}_j$  to  $\mathbf{x}_i$  can be calculated with equation (2.7):

$$p_{j|i} = \frac{\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i^2}\right)}{\sum_k \sum_{k \neq i} \exp\left(-\frac{\|\mathbf{x}_k - \mathbf{x}_i\|^2}{2\sigma_i^2}\right)} \quad (2.7)$$

The conditional probability represents the probability that  $\mathbf{x}_i$  will pick  $\mathbf{x}_j$  as a neighbour based on the proportion of probability density under a Gaussian centred at  $\mathbf{x}_i$ . If the points are near then the value of  $p_{j|i}$  will be higher compare to the points far away. Furthermore, the conditional probabilities are symmetrised to reduce the effects of outliers by setting (2.8):

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \quad (2.8)$$

To represent joint probabilities in the low dimensional map  $q_{ij}$ , t-SNE utilises a heavy tailed Student t-distribution. The benefit of using a heavy tailed distribution is that it makes the joint probabilities invariant to changes in the scale of the map. The joint probabilities  $q_{ij}$  can be estimated by (2.9):

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_k \sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \quad (2.9)$$

Finally, the t-SNE minimises a single Kullback-Leibler (KL) divergence between a joint probability distribution, P, in the high-dimensional space and a joint probability distribution, Q, in the low-dimensional space as can be understood from equation (2.10):

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (2.10)$$

The minimisation of the KL divergence is performed using a gradient descent algorithm with respect to the locations of the points in the map  $y_i$ .

All the data visualisation methods (PCA, MDS, ISOMAP and t-SNE) were implemented in Matlab using the Toolbox for Dimensionality Reduction (<https://lvdmaaten.github.io/drtoolbox/>) [45,46]. The Mahalanobis distance [47] was used to assess the separation of the clusters identified with the different data visualisation methods.

#### 2.3.4.6 Support vector machines for multi-class classification

Support vector machines (SVMs) are supervised non-probabilistic learning models which utilise hyperplanes to define the decision boundaries for performing classification [48]. The SVM algorithms are usually developed to perform a binary classification, however, SVM can be used for multi-class classification problems by utilising several independent binary classifiers. This can be performed by combining it with ensemble methods such as error correcting output codes (ECOC). The ECOC deals with the multi-class classification problem by converting it into several



independent binary classification problems. A wide range of applications of SVM to process HS images can be seen [49,50].

In the present work, the ECOC-SVM algorithm available in Matlab's Statistics and Machine Learning Toolbox (R2016b) was implemented to perform the classification utilising the classification learner application. ECOC-SVM uses a one-versus-all coding design, in which for each binary learner one class is assigned a positive value and all others are assigned negative values. To map the data to the higher dimension, a radial basis function (RBF) kernel (scale parameter = 10) was used. The RBF kernel has the benefit of non-linearly mapping the sample to the higher dimensional space for dealing with a non-linear relationship between observation and classes. For every pure tea sample, spectra (967 - 1700 nm) were extracted from 200 pixels, which were selected at random from the image collected, leading to 1200 spectra in total for calibration of the classification model. Validation of the model was performed with a 10-fold cross-validation method. Furthermore, to have confidence in the model accuracy, the model was recalibrated with 1200 iterations and the mean and standard deviation were noted. The trained classifier was further used to generate the classification maps of the HS images. The HSI cubes were first unfolded from a 3D map ( $n \times p \times q$ ) to a 2D matrix ( $np \times q$ ) and then the class of every row of the matrix (representing the pixel) was predicted, where  $n$ ,  $p$ , and  $q$  define the x, y and z dimensions of the data, respectively. After prediction, the matrix ( $np \times 1$ ) was reshaped to the original image dimension ( $n \times p$ ).

## 2.4 Results

### 2.4.1 Spectral profiles of tea samples

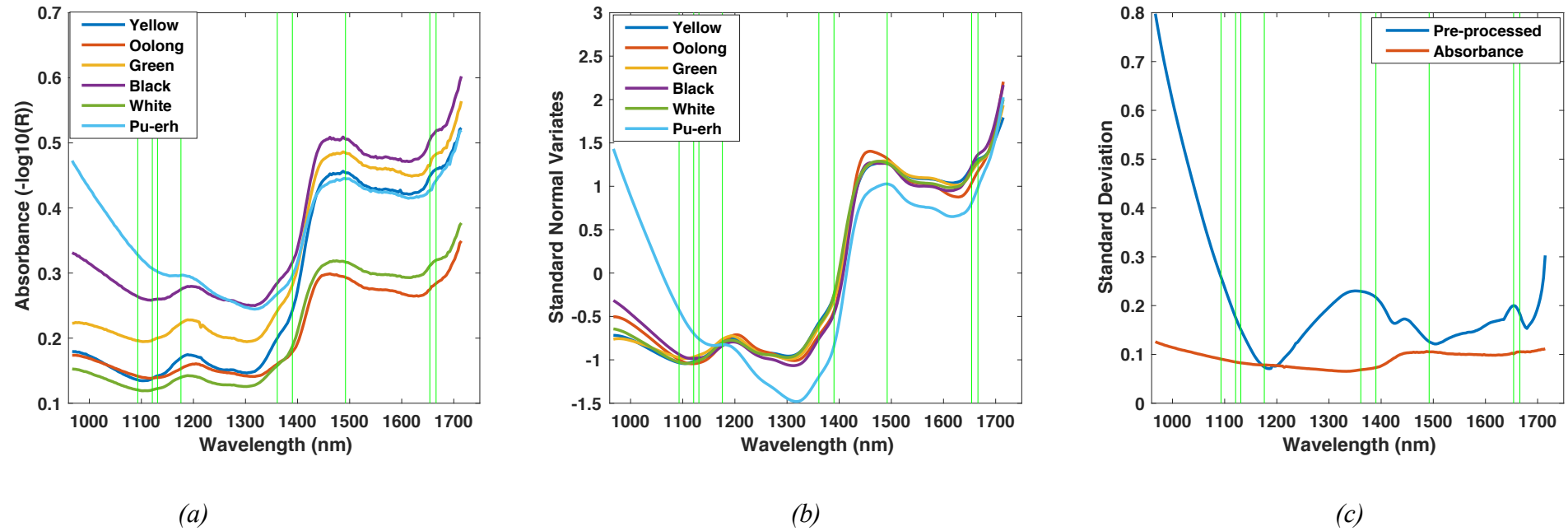


Figure 2.3: Absorbance spectra of pure tea samples of yellow, oolong, green, black, white and Pu-erh. (a). Mean absorbance spectra ( $n = 200$ ). (b) Mean spectra after pre-processing (SNV and Savitzky- Golay smoothing), and (c) standard deviation of the absorbance spectra and spectra after pre-processing. The vertical green lines denote the positions of the main peaks.

Figure 2.3 presents the spectral profiles of individual tea samples. Figure 2.3(a) presents the mean absorbance spectra calculated from the 200 spectra extracted for each of the six tea samples (yellow, oolong, green, black, white and Pu-erh), Figure 2.3(b) presents the mean spectra after pre-processing with Savitzky-Golay filtering followed by SNV (a plot along with the standard deviation for each samples can be find in Annex 1 (Figure A1.1)), and Figure 2.3(c) presents the standard deviation of the spectra before and after pre-processing. From Figure 2.3(a), it can be seen that the absorbance spectra of different tea samples contain scattering effects leading to baseline shifts. These effects can also be seen in the standard deviation plot in Figure 2.3(c) for the absorbance spectra (red), where the standard deviation over the entire spectral range is approximately constant. These scattering effects can bias modelling of the data, therefore, they were removed via pre-processing. In Figure 2.3(b), it can be seen that after pre-processing, differences in spectra at various wavelengths have emerged, and so spectral differences corresponding to different teas can be noted. Scattering effects arise in the imaging experiments as the inhomogeneity in the size of the loose leaves does not get compensated for by the flat surface of the white reflectance standard used for radiometric calibration.

In Figure 2.3(c), it can be noted that the pre-processing reveals the spectral variation arising from differences in the tea, which was previously dominated by the effects of light scattering. In Figure 2.3(b), various peaks (depicted by the green vertical lines) can be identified at representative wavelengths. In previous works, the peaks at 1131, 1654 and 1666 nm were found to be representative of the total tea polyphenols [51,52], 1361 nm is representative of moisture content [22], 1093-1121 nm for thearubigin components of TRS1 [22], 1492 nm corresponds to free amino acids [52], 1176 nm is a second overtone C-H [53] and 1390 nm for the CH<sub>2</sub> overtone [54].

## 2.4.2 Visualising high dimensional data

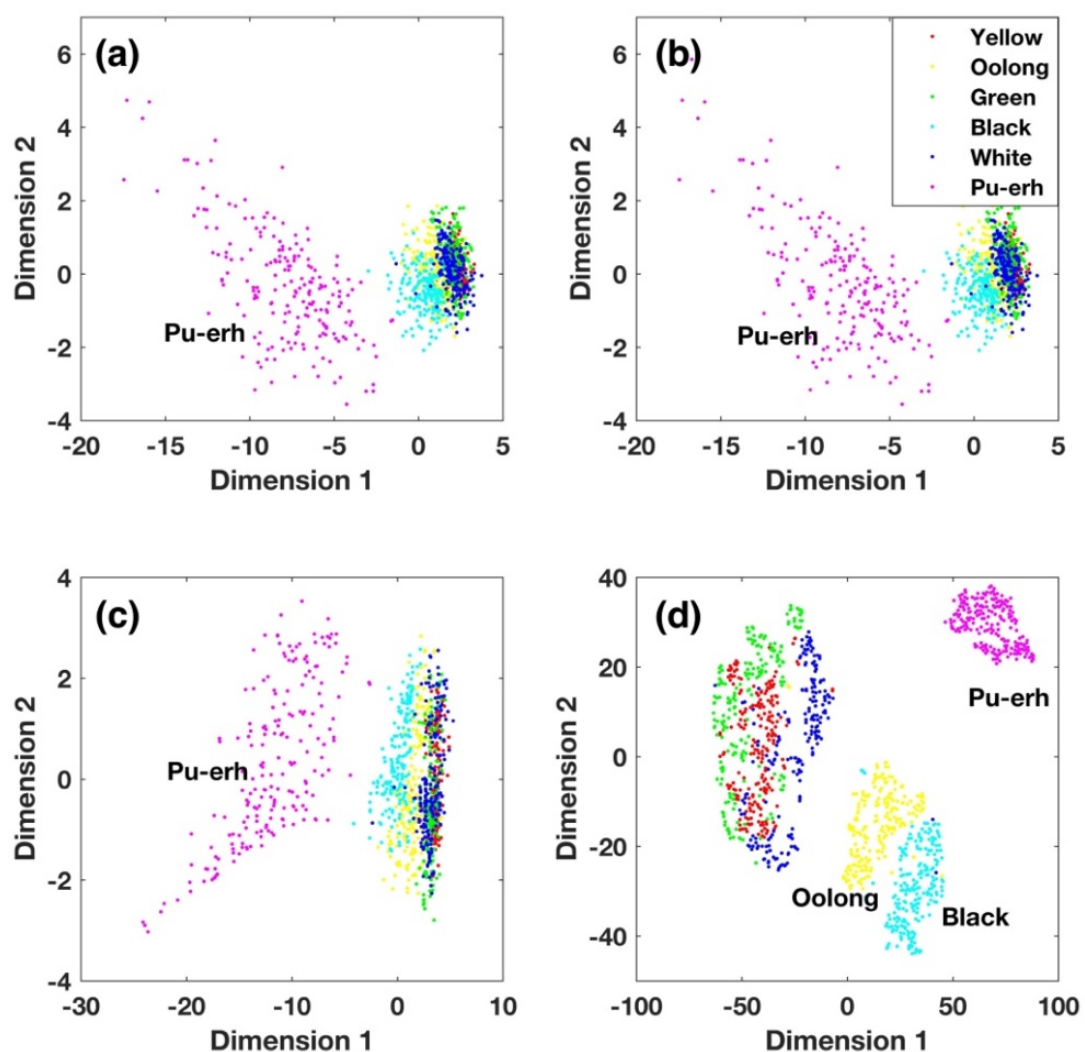


Figure 2.4 : 2-Dimensional scatter plots for visualising high dimensional tea data. (a). Principal Component Analysis (PCA), (b). Multidimensional Scaling (MDS), (c). Isometric Mapping (ISOMAP), and (d). *t*-distributed Stochastic Neighbour Embedding (*t*-SNE). In all the plots, the first dimension is represented in the *x*-axis and the second in the *y*-axis, and the six tea products are coloured as follows: Pu-erh (pink), black (sky blue), oolong (Yellow), green (green), white (blue) and yellow (red).

To visualise the high dimensional data in the lower dimension, the 256-dimensional HSI data were transformed to 2-dimensional plots using PCA, MDS, ISOMAP and *t*-

SNE as shown in Figure 2.4. The loading related to PCA can be found in Annex 1 (Figure A1.2). It can be seen clearly in Figure 2.4 that the t-SNE (Figure 2.4(d)) outperforms PCA, MDS and ISOMAP (Figures 2.4(a), 2.4(b) and 2.4(c), respectively) regarding identification of the maximum number of separate clusters. These separate clusters correspond to different tea products and their representation as separate clusters in the plots signifies that the visualisation method is able to preserve the structure of the data on transformation from a high dimensional space to a lower dimensional space. In general, all the methods were able to separate the Pu-erh tea (pink) from all other tea samples. The reason for this can be seen in Figure 2.3(b) where Pu-erh tea (sky blue) has a very different spectral signature compared to the other tea samples. This is likely to be because the Pu-erh tea undergoes very different processing, which includes microbial fermentation of sun-dried leaves [2], compared to the other teas.

It can be seen in Figure 2.4((a), (b) and (c)) that with the exception of Pu-erh tea, all other types of tea samples are mixed and their clear distinction is not possible. In comparison, black and oolong tea are identified as separate clusters with t-SNE. However, while t-SNE was not able to separate the green, yellow and white tea, it still provided better separation of these three teas as shown in Figure 2.4(d). Green, yellow and white teas appear in the same cluster as they have similar spectral signatures (see Figure 2.3(b)). This may arise from the fact that these teas are most similar in terms of processing conditions; they are subjected to either limited or no oxidation. In comparison, oolong and black teas undergo oxidation during their manufacturing. This may be why these two teas lie in two adjacent clusters that are far away from the cluster containing green, yellow and white teas. However, further information is required to identify the exact source of the spectral differences observed.

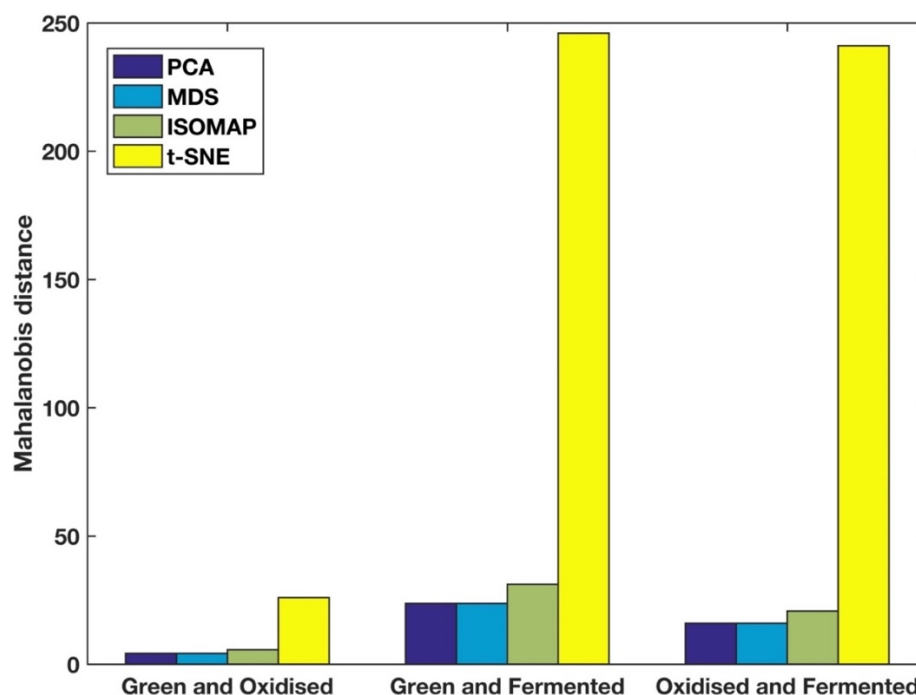


Figure 2.5: Mahalanobis distances between the three different cluster groups obtained using PCA (dark-blue), MDS (sky-blue), ISOMAP (light-green) and t-SNE (yellow).

To assess further the separation of clusters with each method, the Mahalanobis distance between the clusters was calculated. Figure 2.5 presents the Mahalanobis distance estimated for the three major clusters identified in Figure 2.4. The three major cluster can be understood as the group of minimally processed tea products available on the market (denoted the green group), the teas subjected to oxidation (oxidised group) and those that have been subjected to microbial fermented (fermented group). The x-axis in Figure 2.5 presents the pairwise groups used for estimating the distance and the y-axis gives the respective Mahalanobis distance obtained from the different data visualisation methods. It can be seen that the t-SNE (yellow) was superior to all other methods followed by the ISOMAP (light green), and then PCA (dark blue) and MDS (sky blue) for separating all three groups in the data-visualisation plots.

From a statistical perspective, a better visualisation of separate clusters corresponding to different tea products with t-SNE could be due to its ability to capture the non-linearity present in the data set and consideration of neighbourhood information. This supports the modelling of both distant and nearby points [45]. Often, in high dimensional space when the data lies near, or in a non-linear manifold, linear methods

like PCA and MDS fail to preserve the structure of data in the lower dimension space. This is because with linear methods like PCA and MDS, the aim is to keep the distant object far apart; no consideration is given to utilising the information about the neighbouring data points [46].

It can be seen in Figure 2.4(c) that ISOMAP provides a little insight on differences in the classes belonging to black and oolong teas compared to what was achieved with PCA (Figure 2.4(a)) and MDS (Figure 2.4(b)). However, ISOMAP was not able to provide a clear separation of the two teas as was obtained with t-SNE. A reason for the poor performance of ISOMAP compared to t-SNE could be due to its weakness in dealing with the holes and non-convex nature of the data manifold in the higher dimension [43]. Another important weakness of ISOMAP is its topological instability, which leads to a short-circuiting problem in the neighbourhood graph and results in its poor performance [44].

### 2.4.3 Support vector machine classification

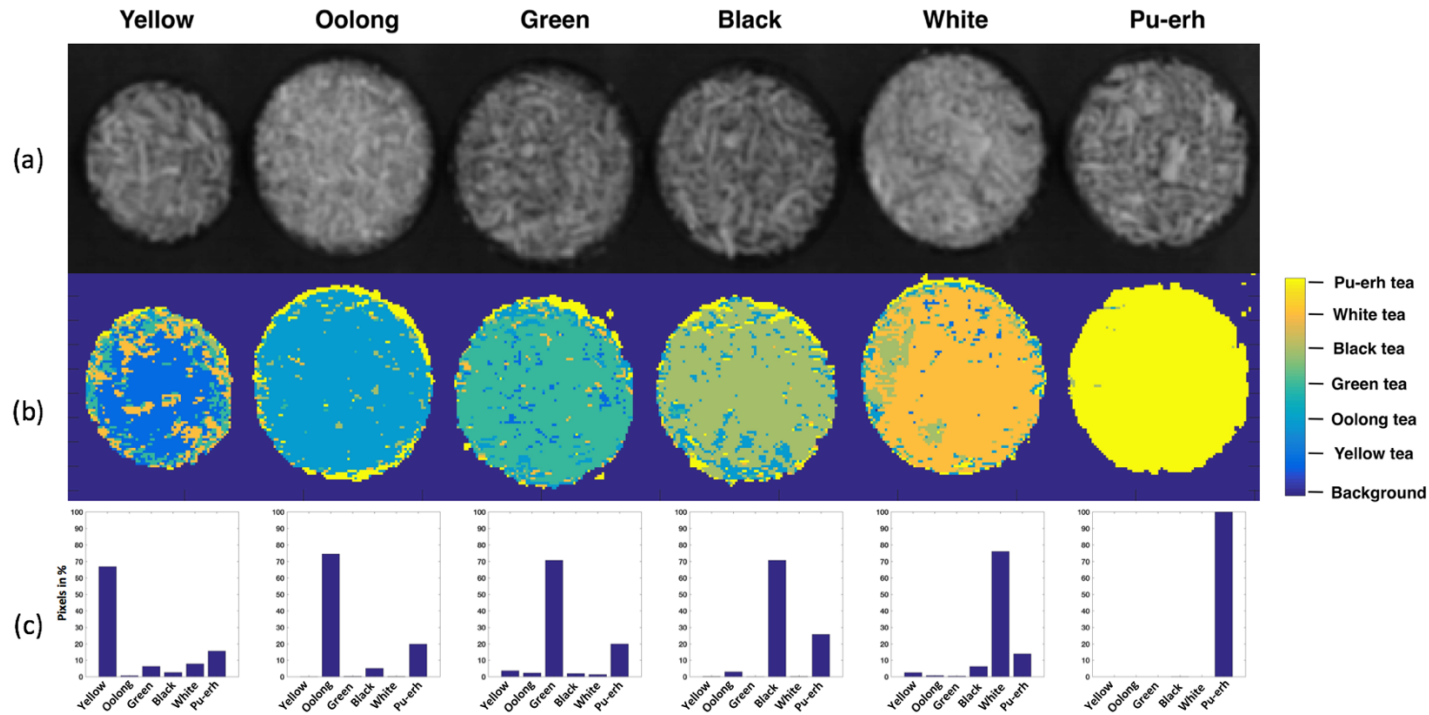


Figure 2.6: (a) Greyscale image constructed from the image plane extracted from the hypercube at 1424 nm, (b) Classification maps obtained from the application of the ECOC-SVM model. From left to right the samples can be understood as yellow (dark blue), oolong (light blue), green (cyan), black (light green), white (orange) and Pu-erh (Yellow). (c) Histograms showing the proportion of pixels attributed to the different tea products for the classification maps in (b).



The results from the application of the ECOC-SVM multi-class classification model are presented as classification maps in Figures 2.6 and 2.7. Figure 2.6(b) presents the classification maps of pure tea samples, from left to right, the samples can be understood as yellow, oolong, green, black, white and Pu-erh. For comparison, a greyscale image was also produced (Figure 2.6(a)) using the image plane corresponding to 1424 nm; this wavelength was selected merely to allow visualisation of the data hypercube. It can be seen from Figure 2.6(b) that all six teas were classified into their respective individual classes. However, there are some pixels that were misclassified; Figure 2.6(c) shows the proportion of pixels attributed to the different tea products for the classification maps in Figure 2.6(b). The misclassification was most dominant at the edges owing to signal from the circular sample container; such pixels (approximately 20%) were misclassified as Pu-erh. When these pixels were excluded, an overall accuracy of  $97.41 \pm 0.16\%$  was obtained for cross-validated samples using 1200 iterations.

Apart from the edges, a reason for the misclassification between different teas can be attributed to their spectral similarity. When visualising the data with t-SNE (see Figure 2.4(d)), green, white and yellow tea were found to be lying near in the same cluster, and black and oolong were near to each other due to their spectral similarity. Hence, the classification map for the yellow tea (dark blue) has some misclassified pixels that have been attributed to either white (orange) or green tea (cyan). For black and oolong teas, it can be noted that there are some pixels in the classification map for black tea (light green) that were misclassified as oolong (light blue class) and vice-versa. Another possible reason for misclassification could arise from the purity of the tea; for example, a minimally processed tea (e.g. white) may contain small amounts of oxidised product (e.g. black tea).

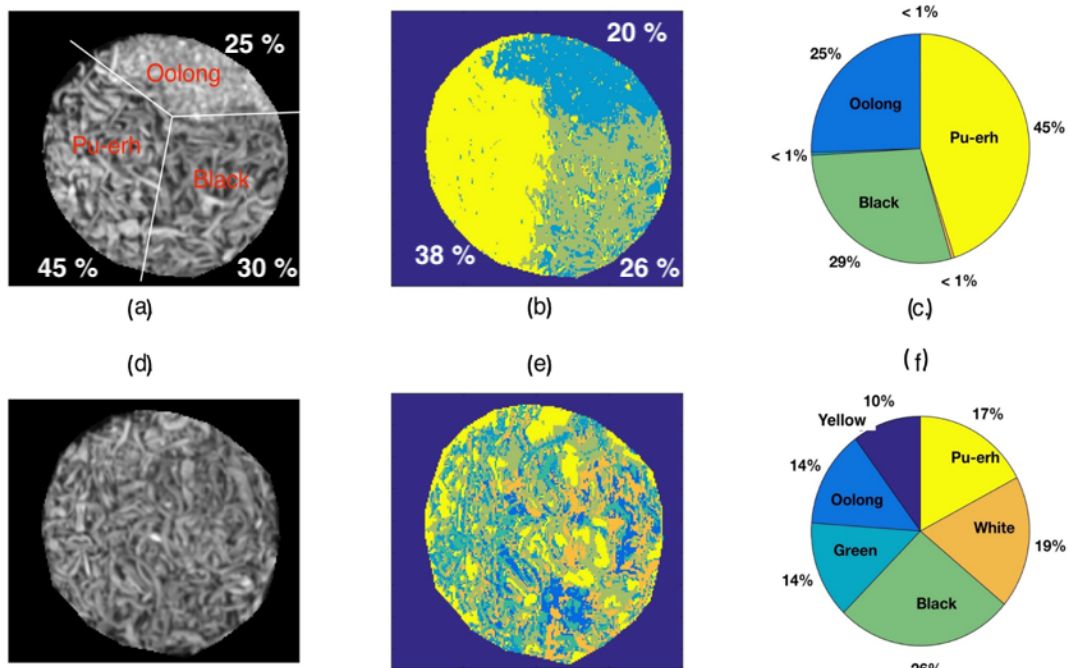


Figure 2.7:(a). Greyscale image at 1424 nm for the sample comprising oolong, black and Pu-erh tea, (b). The classification map for the sample comprising oolong, black and Pu-erh tea, (c). Pie chart representing the proportion of pixels belonging to a particular class for the classification map presented in (b), (d). Greyscale image at 1424 nm for a sample containing a mixture of all teas, (e). The classification map for a sample containing a mixture of all teas, and (f) Pie chart representing the proportion of the pixels belonging to a particular class for the classification map presented in (e).

Figure 2.7 presents the classification maps for the HS images acquired for samples comprising mixtures of teas. This analysis was performed to assess the feasibility of using the methodology developed to classify different tea samples when more than one tea is present. Figure 2.7(a) presents the image plane corresponding to 1424 nm for a sample containing oolong, black and Pu-erh teas (not mixed) in roughly equal portions. These three teas were selected as there is an oxidation stage in their manufacturing. The location of the different teas in Figure 2.7(a) can be identified with the red markers. As can be seen from Figure 2.7(b), the model provided a clear classification of the three teas into their respective classes. However, some misclassification can be seen at the interface between different types of teas; individual pixels will detect the presence of more than one type of tea at these locations. Furthermore, Figure 2.7(c) provides insight into the proportion of pixels belonging to each class. It can be seen that the pie chart is mainly dominated by the proportion of oolong, black and Pu-erh tea and contains a very small portion (<1%) of pixels classified as green, white and yellow.

The methodology developed was also tested for a mixture of all six tea samples. The result for classification of the sample containing a mixture of all six types of tea is presented in Figure 2.7(e). The classification map shown in Figure 2.7(e) can be interpreted in conjunction with the pie chart (Figure 2.7(f)) representing the proportion of pixels classified belonging to different classes. The pie chart shows that the presence of all the classes can be detected with the classification model and the portion of each type of tea ranged from 10 – 26%. However, it was not possible to validate the classification result of the mixture image because it is not known if the sample was a homogenous mixture of the six types of teas and hence, the exact composition of the upper surface of the sample is unknown. In addition, there may be some misclassification of pixels that detect more than one type of tea.

## 2.5 Conclusions

NIR HSI has been used to classify six different types of commercial tea samples. Before any data modelling, the spectral imaging data from tea products should be pre-processed to reduce the effects of light scattering arising from the inhomogeneous and

uneven leaf surface. Four different types of linear and non-linear dimensionality reduction methods were compared for visualisation of imaging data. The non-linear method, t-SNE, gave better separation of the different tea products than classical linear techniques such as PCA and MDS. This is because t-SNE uses information from neighbouring data points in the high dimensional space to preserve the structure in the low dimensional representation. It was possible to classify the tea according to product type using a ECOC-SVM multi-class classification model constructed using the NIR HSI data. Therefore, NIR HSI in conjunction with machine learning could be a potential tool for classification of different types of tea products. The source of spectral differences is assumed to arise from the different processing steps that are involved in the manufacture of various types of tea. However, there could be other sources, e.g. geographical, that contribute to spectral differences and hence, this requires further investigation.

## 2.6 References

1. S. Sang, Tea: Chemistry and Processing, in: B. Caballero, P.M. Finglas, F.B.T.-E. of F. and H. Toldrá (Eds.), Academic Press, Oxford, 2016: pp. 268–272. doi:<https://doi.org/10.1016/B978-0-12-384947-2.00685-1>.
2. H. Lv, Y. Zhang, Z. Lin, Y. Liang, Processing and chemical constituents of Pu-erh tea: A review, Food Res. Int. 53 (2013) 608–618. doi:<https://doi.org/10.1016/j.foodres.2013.02.043>.
3. K. Chang, 2015. World tea production and trade: current and future development, Food and Agriculture Organisation (FAO), United Nations (UN), Rome. <http://www.fao.org/3/a-i4480e.pdf> (last accessed: 14 September 2017)
4. J. Ruan, R. Haerdter, J. Gerendás, Impact of nitrogen supply on carbon/nitrogen allocation: a case study on amino acids and catechins in green tea [*Camellia sinensis* (L.) O. Kuntze] plants, Plant Biol. 12 (2010) 724–734. doi:[10.1111/j.1438-8677.2009.00288.x](https://doi.org/10.1111/j.1438-8677.2009.00288.x).
5. B. Ozturk, F. Seyhan, I.S. Ozdemir, B. Karadeniz, B. Bahar, E. Ertas, S. Ilgaz, Change of enzyme activity and quality during the processing of Turkish green tea, LWT - Food Sci. Technol. 65 (2016) 318–324. doi:<https://doi.org/10.1016/j.lwt.2015.07.068>.
6. C.-H. Nieh, B.-C. Hsieh, P.-C. Chen, H.-Y. Hsiao, T.-J. Cheng, R.L.C. Chen, Potentiometric flow-injection estimation of tea fermentation degree, Sensors Actuators B Chem. 136 (2009) 541–545. doi:<https://doi.org/10.1016/j.snb.2008.09.024>.
7. J. Tan, W. Dai, M. Lu, H. Lv, L. Guo, Y. Zhang, Y. Zhu, Q. Peng, Z. Lin, Study of the dynamic changes in the non-volatile chemical constituents of black tea during fermentation processing by a non-targeted metabolomics approach, Food Res. Int. 79 (2016) 106–113. doi:<https://doi.org/10.1016/j.foodres.2015.11.018>.

8. J. Jing, Y. Shi, Q. Zhang, J. Wang, J. Ruan, Prediction of Chinese green tea ranking by metabolite profiling using ultra-performance liquid chromatography–quadrupole time-of-flight mass spectrometry (UPLC–Q-TOF/MS), *Food Chem.* 221 (2017) 311–316. doi:<https://doi.org/10.1016/j.foodchem.2016.10.068>.
9. A.S. Kumar, R. Shanmugam, S. Nellaiappan, R. Thangaraj, Tea quality assessment by analyzing key polyphenolic functional groups using flow injection analysis coupled with a dual electrochemical detector, *Sensors Actuators B Chem.* 227 (2016) 352–361. doi:<https://doi.org/10.1016/j.snb.2015.12.072>.
10. I. Domínguez, A. Doménech-Carbó, Screening and authentication of tea varieties based on microextraction-assisted voltammetry of microparticles, *Sensors Actuators B Chem.* 210 (2015) 491–499. doi:<https://doi.org/10.1016/j.snb.2015.01.009>.
11. J. Li, B. Fu, D. Huo, C. Hou, M. Yang, C. Shen, H. Luo, P. Yang, Discrimination of Chinese teas according to major amino acid composition by a colorimetric {IDA} sensor, *Sensors Actuators B Chem.* 240 (2017) 770–778. doi:<http://dx.doi.org/10.1016/j.snb.2016.09.019>.
12. I. Yaroshenko, D. Kirsanov, L. Kartsova, N. Bhattacharyya, S. Sarkar, A. Legin, On the application of simple matrix methods for electronic tongue data processing: Case study with black tea samples, *Sensors Actuators B Chem.* 191 (2014) 67–74. doi:<https://doi.org/10.1016/j.snb.2013.09.093>.
13. P. Sharma, A. Ghosh, B. Tudu, S. Sabhapondit, B.D. Baruah, P. Tamuly, N. Bhattacharyya, R. Bandyopadhyay, Monitoring the fermentation process of black tea using QCM sensor based electronic nose, *Sensors Actuators B Chem.* 219 (2015) 146–157. doi:<https://doi.org/10.1016/j.snb.2015.05.013>.

14. N. Bhattacharyya, S. Seth, B. Tudu, P. Tamuly, A. Jana, D. Ghosh, R. Bandyopadhyay, M. Bhuyan, Monitoring of black tea fermentation process using electronic nose, *J. Food Eng.* 80 (2007) 1146–1156. doi:<https://doi.org/10.1016/j.jfoodeng.2006.09.006>.
15. E.A. Baldwin, J. Bai, A. Plotto, S. Dea, Electronic noses and tongues: Applications for the food and pharmaceutical industries, *Sensors (Basel)*. 11 (2011) 4744–4766. doi:10.3390/s110504744.
16. J.-H. Qu, D. Liu, J.-H. Cheng, D.-W. Sun, J. Ma, H. Pu, X.-A. Zeng, Applications of near-infrared spectroscopy in food safety evaluation and control: A review of recent research advances, *Crit. Rev. Food Sci. Nutr.* 55 (2015) 1939–1954. doi:10.1080/10408398.2013.871693.
17. X. Fu, Y. Ying, Food safety evaluation based on near infrared spectroscopy and imaging: A review, *Crit. Rev. Food Sci. Nutr.* 56 (2016) 1913–1924. doi:10.1080/10408398.2013.807418.
18. Y. He, X. Li, X. Deng, Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model, *J. Food Eng.* 79 (2007) 1238–1242. doi:<https://doi.org/10.1016/j.jfoodeng.2006.04.042>.
19. Q. Chen, J. Zhao, H. Lin, Study on discrimination of Roast green tea (*Camellia sinensis* L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition, *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* 72 (2009) 845–850. doi:<http://dx.doi.org/10.1016/j.saa.2008.12.002>.
20. Q. Chen, J. Zhao, C.H. Fang, D. Wang, Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM), *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* 66 (2007) 568–574. doi:<https://doi.org/10.1016/j.saa.2006.03.038>.
21. S. Wang, X. Yang, Y. Zhang, P. Phillips, J. Yang, T. Yuan, Identification of green, oolong and black teas in China via wavelet packet entropy and fuzzy support vector machine, *Entropy*, (2015) 6663–6682. doi:10.3390/e17106663.

22. N. Panigrahi, C.S. Bhol, B.S. Das, Rapid assessment of black tea quality using diffuse reflectance spectroscopy, *J. Food Eng.* 190 (2016) 101–108. doi:<https://doi.org/10.1016/j.jfoodeng.2016.06.020>.
23. Q. Chen, J. Zhao, J. Cai, Identification of tea varieties using computer vision, *Transactions of the ASABE*. 51 (2008). doi:10.13031/2013.24363.
24. X. Wang, J. Huang, W. Fan, H. Lu, Identification of green tea varieties and fast quantification of total polyphenols by near-infrared spectroscopy and ultraviolet-visible spectroscopy with chemometric algorithms, *Anal. Methods*. 7 (2015) 787–792. doi:10.1039/C4AY02106A.
25. H. Zhu, Y. Ye, H. He, C. Dong, Evaluation of green tea sensory quality via process characteristics and image information, *Food Bioprod. Process.* 102 (2017) 116–122. doi:<https://doi.org/10.1016/j.fbp.2016.12.004>.
26. A.F. Goetz, G. Vane, J.E. Solomon, B.N. Rock, Imaging spectrometry for Earth remote sensing., *Science*. 228 (1985) 1147–1153. doi:10.1126/science.228.4704.1147.
27. G.J. Edelman, E. Gaston, T.G. van Leeuwen, P.J. Cullen, M.C.G. Aalders, Hyperspectral imaging for non-contact analysis of forensic traces, *Forensic Sci. Int.* 223 (2012) 28–39. doi:<https://doi.org/10.1016/j.forsciint.2012.09.012>.
28. G. Lu, B. Fei, Medical hyperspectral imaging: a review., *J. Biomed. Opt.* 19 (2014) 10901-1-23. doi:10.1117/1.JBO.19.1.010901.
29. Y.-Y. Pu, Y.-Z. Feng, D.-W. Sun, Recent progress of hyperspectral imaging on quality and safety inspection of fruits and vegetables: A Review, *Compr. Rev. Food Sci. Food Saf.* 14 (2015) 176–188. doi:10.1111/1541-4337.12123.
30. L.M. Kandpal, J. Tewari, N. Gopinathan, P. Boulas, B.-K. Cho, In-process control assay of pharmaceutical microtablets using hyperspectral imaging coupled with multivariate analysis, *Anal. Chem.* 88 (2016) 11055–11061. doi:10.1021/acs.analchem.6b02969.



31. P. Mishra, M.S.M. Asaari, A. Herrero-Langreo, S. Lohumi, B. Diezma, P. Scheunders, Close range hyperspectral imaging of plants: A review, *Biosyst. Eng.* 164 (2017) 49–67. doi:<https://doi.org/10.1016/j.biosystemseng.2017.09.009>.
32. C. Nansen, K. Singh, A. Mian, B.J. Allison, C.W. Simmons, Using hyperspectral imaging to characterize consistency of coffee brands and their respective roasting classes, *J. Food Eng.* 190 (2016) 34–39. doi:<https://doi.org/10.1016/j.jfoodeng.2016.06.010>.
33. P.B. Garcia-Allende, O.M. Conde, J. Mirapeix, A. Cobo, J.M. Lopez-Higuera, Quality control of industrial processes by combining a hyperspectral sensor and Fisher's linear discriminant analysis, *Sensors Actuators B Chem.* 129 (2008) 977–984. doi:<https://doi.org/10.1016/j.snb.2007.09.036>.
34. S. Shrestha, M. Knapič, U. Žibrat, L.C. Deleuran, R. Gislum, Single seed near-infrared hyperspectral imaging in determining tomato (*Solanum lycopersicum* L.) seed quality in association with multivariate data analysis, *Sensors Actuators B Chem.* 237 (2016) 1027–1034. doi:<https://doi.org/10.1016/j.snb.2016.08.170>.
35. L.M. Kandpal, S. Lohumi, M.S. Kim, J.-S. Kang, B.-K. Cho, Near-infrared hyperspectral imaging system coupled with multivariate methods to predict viability and vigor in muskmelon seeds, *Sensors Actuators B Chem.* 229 (2016) 534–544. doi:<https://doi.org/10.1016/j.snb.2016.02.015>.
36. J. Zhao, Q. Chen, J. Cai, Q. Ouyang, Automated tea quality classification by hyperspectral imaging, *Appl. Opt.* 48 (2009) 3557–3564. doi:[10.1364/AO.48.003557](https://doi.org/10.1364/AO.48.003557).
37. C. Xie, X. Li, Y. Shao, Y. He, Color Measurement of tea leaves at different drying periods using hyperspectral imaging technique, *PLoS One.* 9 (2015) 1–15. doi:[10.1371/journal.pone.0113422](https://doi.org/10.1371/journal.pone.0113422).
38. P. Mishra, C.B.Y. Cordella, D.N. Rutledge, P. Barreiro, J.M. Roger, B. Diezma, Application of independent components analysis with the JADE algorithm and

- NIR hyperspectral imaging for revealing food adulteration, *J. Food Eng.* 168 (2016) 7–15. <https://doi.org/10.1016/j.jfoodeng.2015.07.008>
39. R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, *Appl. Spectrosc.* 43 (1989) 772–777. doi:10.1366/0003702894202201.
40. A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures., *Anal. Chem.* 36 (1964) 1627–1639. doi:10.1021/ac60214a047.
41. S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemom. Intell. Lab. Syst.* 2 (1987) 37–52. doi:[https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9).
42. T.F. Cox, M.A.A. Cox, Multidimensional scaling, CRC press, 22(16) (2000) 609–623, doi: 10.1016/S0169-7161(03)22018-6
43. J.B. Tenenbaum, Mapping a manifold of perceptual observations, in: M.I. Jordan, M.J. Kearns, S.A. Solla (Eds.), *Adv. Neural Inf. Process. Syst.* 10, MIT Press, 1998: pp. 682–688. <http://papers.nips.cc/paper/1332-mapping-a-manifold-of-perceptual-observations.pdf>.
44. M. Balasubramanian, E.L. Schwartz, The Isomap algorithm and topological stability, *Science*, 295(2002) 7. doi: 10.1126/science.295.5552.7a
45. L. van der Maaten, G. Hinton, Visualising high-dimensional data using t-SNE. *J. Mach. Learn. Res.* 9 (2008), 2579–2605.
46. L. van der Maaten, E. Postma, J. Van den Herik, Dimensionality reduction: a comparative review. Tilburg University Technical Report (2009), TICC-TR 2009-005.
47. P.C. Mahalanobis, On the generalised distance in statistics, in: *Proceedings National Institute of Science, India*, (1936) pp. 49–55.
48. V.N. Vapnik, V. Vapnik, *Statistical learning theory*. Wiley New York (1998).

49. G. Mountrakis, J. Im, C. Ogole, Support vector machines in remote sensing: A review, *ISPRS J. Photogramm. Remote Sens.* 66 (2011) 247–259. doi:<https://doi.org/10.1016/j.isprsjprs.2010.11.001>.
50. P.B. Garcia-Allende, F. Anabitarte, O.M. Conde, J. Mirapeix, F.J. Madruga, J.M. Lopez-Higuera, Support vector machines in hyperspectral imaging spectroscopy with application to material identification, *SPIE* (2008), pp. 69661V–6966–11. <http://dx.doi.org/10.1117/12.770306>.
51. Q. Chen, J. Zhao, X. Huang, H. Zhang, M. Liu, Simultaneous determination of total polyphenols and caffeine contents of green tea by near-infrared reflectance spectroscopy, *Microchem. J.* 83 (2006) 42–47. doi:<http://dx.doi.org/10.1016/j.microc.2006.01.023>.
52. M. Bian, A.K. Skidmore, M. Schlerf, T. Wang, Y. Liu, R. Zeng, T. Fei, Predicting foliar biochemistry of tea (*Camellia sinensis*) using reflectance spectra measured at powder, leaf and canopy levels, *ISPRS J. Photogramm. Remote Sens.* 78 (2013) 148–156. doi:<https://doi.org/10.1016/j.isprsjprs.2013.02.002>.
53. S. M. Tan, R. M. Luo, Y. P. Zhou, H. Gong, Z. Tan, Rapid and non-destructive discrimination of tea varieties by near infrared diffuse reflection spectroscopy coupled with classification and regression trees, *African J. Biotechnol.* 11 (2012), 2303–2312.
54. M.-S. Lee, Y.-S. Hwang, J. Lee, M.-G. Choung, The characterization of caffeine and nine individual catechins in the leaves of green tea (*Camellia sinensis* L.) by near-infrared reflectance spectroscopy., *Food Chem.* 158 (2014) 351–357. doi:[10.1016/j.foodchem.2014.02.127](https://doi.org/10.1016/j.foodchem.2014.02.127).

## Chapter 3 : Non-destructive classification of the geographic origin of green teas with near-infrared hyperspectral imaging combined with variable selection

Puneet Mishra<sup>1</sup>, Alison Nordon<sup>1</sup>, Sally Redfern<sup>2</sup>, Guoping Lian<sup>2</sup>

<sup>1</sup>*WestCHEM, Department of Pure and Applied Chemistry and Centre for Process Analytics and Control Technology, University of Strathclyde, 295 Cathedral Street, Glasgow, G1 1XL, United Kingdom*

<sup>2</sup>*Unilever, Colworth, MK44 1LQ, United Kingdom*

Paper Published as: To be submitted to Journal of Food Engineering

Contribution:

- Performed HSI NIR measurements
- Conducted chemometrics analysis of data
- Wrote the first draft of article
- Corrected article based on comments of co-authors

### 3.1 Abstract

The distinct flavour and aroma of different tea products are dependent on the climate and soil conditions in which the plants are grown. Therefore, there is an increasing demand for tea products originating from specific geographical locations. In addition, it is important to be able to identify the geographic origin of tea products within the context of food traceability. Traditionally, discrimination of tea products is performed via sensory analysis, which requires an expert human panel. More recently techniques such as high-performance liquid chromatography have been employed, but these are destructive, time-consuming and require complex sample preparation. Therefore, in

this work, near-infrared hyperspectral imaging (NIR-HSI) has been used for rapid and non-destructive classification of sixteen different green tea products corresponding to four different continents and seven different countries of origin. To perform the multi-class classification of the different tea products, support vector machine (SVM) binary learners were combined with the error-correcting output codes (ECOC) ensemble technique. Three different classification models were constructed considering sixteen, seven and four classes based on the samples, countries and continents of origin, respectively. The 10-fold cross validated SVM-ECOC model provided a classification accuracy of  $97.01 \pm 0.17\%$ ,  $96.36 \pm 0.17\%$  and  $97.77 \pm 0.16\%$  for prediction of product type, and country and continent of origin, respectively. The classification models were used to produce spatial classification maps by predicting the class of pixels in the hypercube. Further, to identify the important region of the spectral range responsible for the success of each classification model, two different feature selection methods, ReliefF and sequential forward selection (SFS), were explored. SVM-ECOC models for the ReliefF were constructed using the selected features so as to achieve a predictive accuracy of 90%. The selected features comprised 7.5 – 22.5% and 8.0 – 11.4% of the full spectral range for ReliefF and SFS, respectively, for prediction of product type, and country and continent of origin. The methodology developed enables rapid, non-destructive *in situ* analysis of green tea products originating from different geographic locations, which could be valuable for counterfeit detection, quality control and traceability of green tea products.

**Keywords:** Traceability, classification, geographic origin, non-destructive, NIRS, feature selection

### 3.2 Introduction

Beverages prepared from the infusion of tea leaves are the most consumed manufactured drinks in the world. Different tea products, such as green, oolong and black, are obtained from the *Camellia sinensis* plant after a series of processing stages such as withering, drying, oxidation and fermentation. In recent decades, tea products have gained the attention of the world's population owing to the health benefits associated with incorporating tea in the daily diet. Of all the tea products, the

minimally processed green tea has potential health benefits as it contains natural polyphenolic antioxidants [1]. The reported benefits of bioactive polyphenols such as epigallocatechin gallate (EGCG) and theaflavin (TF) include potential anti-cancer effects [2], influence on the human brain function [3], protection from environmental toxins [4], anti-angiogenesis effects [5], managing oxidative stress [6] and many others [7].

Within the globalised tea market, there is an increasing demand for high-quality tea products of specific geographical origin. There is also a requirement to be able to identify the geographic origin of food products to support international food traceability [8]. High-quality tea products are mainly characterised by the flavour that they impart, which involves two primary sensory perceptions, i.e. taste and aroma. The distinct taste and aroma of any tea product are derived from its geographical origin as they are unique to the climate and soil conditions in which the plants were grown [9]. Typically, discrimination of green tea products via sensory analysis is performed using an expert human panel. Sensory analysis involves assessment of tea products by appearance, colour, aroma and taste, along with the overall quality of the samples. However, distinguishing tea products based on sensory analysis is a time-consuming and expensive task as it requires an expert human panel. Furthermore, sensory analysis is subjective, and it can be inconsistent and unpredictable owing to physiological and psychological differences between tasters [10]. One more limitation is that the expert panel cannot be used as an on-line technique for the grading of tea products [11].

In recent years, different analytical techniques have been explored for assessment of tea products to work independently or partially with the sensory panel to support the discrimination of tea products. Different high-end analytical techniques such as high-performance liquid chromatography (HPLC) [12], gas chromatography (GC) [13] and isotope analysis [14] have been explored for assessment of tea products. However, these techniques are time-consuming, expensive, destructive and require trained experts to perform the experiments. Two other possible techniques for the assessment of green tea products, which mimic the taste and odour assessment by the human test panel, are the e-nose and the e-tongue. The e-nose has been used for identification [15], discrimination [16, 17], quality evaluation [18] and classification [19] of green

teas, while the e-tongue has been used for multicomponent analysis [20], discrimination [21] and classification [11]. However, it is extremely difficult to use the e-nose and the e-tongue to discriminate between similar complex mixtures because of the non-specificity of the sensor arrays [10]. Further to that, the performance of both sensors is affected by environmental conditions, such as humidity and temperature [22].

Applications of electrochemical methods have been reported for the quality assessment [23], authentication [24] and estimation of biochemistry [25, 26] of tea products. A methodology based on calorimetric indicator displacement assays was recently used to discriminate teas of different geographical origin [27]. The methodology showed good performance for discriminating tea samples based on quality grade and geographical origin. However, it was destructive and required extraction of liquor from the tea leaves to record the calorimetric response. On the other hand, optical spectroscopic techniques can be used for non-destructive analysis of a wide range of food products and beverages. For example, near-infrared spectroscopy (NIRS) has been used to discriminate between spices [28], coffee [29], oils [30] and alcohols [31] of different geographical origin. Some initial work to discriminate between different tea products has also been reported; Chen et al. used NIRS to discriminate between green teas originating from four different geographical locations in China [9]. However, the study was limited to samples that originated from a single country (China). Computer vision techniques have been used to grade tea products on the basis of colour and texture [32]. Hyperspectral imaging (HSI) combines the benefits of spectroscopy and imaging, producing spatial maps of spectral variation [33], and has recently been used to discriminate between different types of tea products [34]. However, NIR HSI has been used to discriminate food products, on the basis of geographic origin, such as wolfberries [35], Jatropha seeds [36], rice [37] and fish species [38].

NIR HSI cameras typically capture a broad spectral range (e.g. 900 – 1700 nm). However, the chemical information related to vibrations and overtones of the bonds is mainly limited to particular regions within the overall range [39]. Identification of distinct NIR regions can be performed manually by the spectroscopist, utilising their

prior knowledge of band assignments. However, feature selection methods can be used to identify the most important NIR regions in an automated way. The goal of feature selection methods is to identify a reduced number of optimal wavelengths (variables), which provide comparable model classification/regression prediction accuracies to models developed with the full spectral range [40]. There are different methods for variable selection ranging from fast-filter based approaches to computationally intensive wrapper-based methods [41]. Furthermore, feature selection for NIR HSI brings further benefits such as supporting the development of cheap multi-spectral sensors through identification of unimportant variables. This enables more efficient storage of data and decreases the computational intensity of data processing algorithms [42]. In the present work, the interest in utilising feature selection methods is to identify the NIR wavelengths responsible for distinction between green tea samples of different geographical origin.

The aim of the present work is to demonstrate the use of NIR-HSI, in conjunction with multi-class SVM modelling and variable selection, for classification of green tea samples based on geographic origin (country and continent). In addition, the use of two different feature selection methods, ReliefF and sequential forward selection (SFS), was explored as a means of identifying the most important wavelengths for successful classification with a smaller subset of features.



### 3.3 Materials and Method

#### 3.3.1 Samples

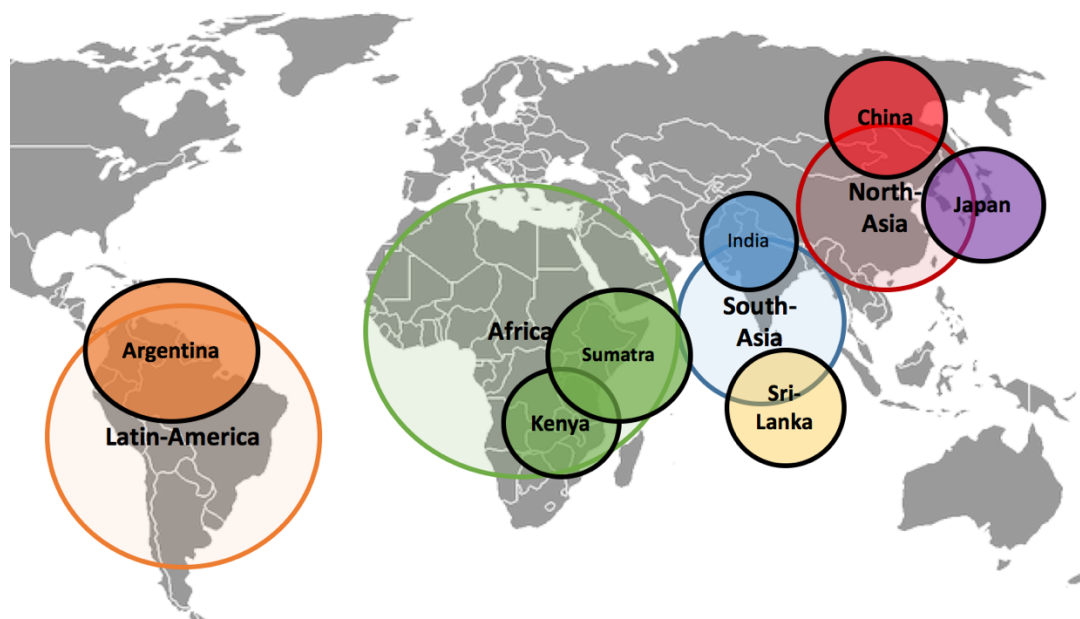


Figure 3.1: A summary of the geographical origin, by country and continent, of the 16 green tea samples.

Sixteen green tea samples differing in geographical origin were obtained from Unilever, Colworth Park, UK. The 16 samples originated from 7 different countries: Argentina (1), India (5), Sri Lanka (2), China (2), Japan (2), Kenya (3) and Sumatra (1). Further, it can be seen that the samples originated from three continents Asia, Africa and Latin America. To extend the model, the Asian continent was sub-divided into North and South-Asia, thus separating China and Japan from India and Sri Lanka. A summary of the geographic origin of the 16 samples is given in Figure 3.1. All samples of tea were provided in loose-leaf form in sealed packaging, and were stored at ambient temperature. Each sample was transferred into a black plastic circular container (diameter = 3.3 cm, depth = 1.3 cm) for analysis using HSI. A different cap was used for each tea sample to avoid any possible cross-contamination.

#### 3.3.2 Hyperspectral imaging measurements

Images were acquired with a NIR push-broom line scan HSI camera (*Model name:*

RED EYE 1.7) from INNO-SPEC (Nurnberg, Germany) over a spectral range of 950 - 1760 nm. The camera has an InGaAs sensor and generates a spatial map of 320 x 256 pixels, and has a pixel size of 30 x 30  $\mu\text{m}^2$ . Images were acquired over the spectral range 950 – 1765 nm with a spectral resolution of 3.2 nm. The camera communicated with the computer via a gigabit Ethernet connection. Two halogen light sources of 50 W each were used to illuminate the samples and an integration time of 300 ms was used. Imaging was performed by placing the samples over the translation stage, which was controlled by an independent stage motor system (Zolix TSA 200 BF). The speed of the translation stage, 2.5 mm s<sup>-1</sup>, was optimised before image acquisition to avoid any distortion in the shape of the image arising from the overlapping of the spectral information in adjacent pixels. The distance from the lens to the translation stage was 15 cm. The image acquisition and management of settings (integration time) were performed using the software interface provided with the camera by INNO-SPEC. Prior to acquisition of an image, a set of white (Spectralon diffuse reflectance standard) and dark reference were recorded for radiometric calibration. An image was acquired of the 16 different tea samples placed adjacent to each other in their respective sample containers in the field of view of the camera. The image comprised more than 2000 pixels (spectra) for each of the individual green tea samples.

### 3.3.3 Data analysis

#### 3.3.3.1 Pre-processing of HSI

Variations in signal resulting from illumination intensity; the detector sensitivity and the transmission properties of the optics were corrected by radiometric calibration using dark and white reference images. The correction was performed for every pixel in the HS image according to equation 3.1:

$$I_{R(i,j,k)} = \frac{I_{raw(i,j,k)} - I_{dark(i,j,k)}}{I_{white(i,j,k)} - I_{dark(i,j,k)}} \quad (3.1)$$

where  $I_R$  is the calibrated reflectance,  $I_{raw}$  is the raw intensity measured from the test sample,  $I_{dark}$  the intensity of the dark response,  $I_{white}$  is the intensity of the uniform white reference, and  $i$  and  $j$  are spatial coordinates and  $k$  is the wavelength in the image.

After radiometric calibration, the spectral range of the hypercube was reduced from

950 - 1765 nm to 967.11 - 1700 nm to remove noise. The data were converted to absorbance by calculating the negative logarithm (base 10) of reflectance. A moving window Savitzky-Golay (SAVGOL) filter [43] (15-point width and second order polynomial) was applied to each pixel of the image to remove random noise from spectra. To reduce light scattering effects arising from inhomogeneity in the sample surface, the spectra were normalised using the standard normal variate (SNV) [44].

### *3.3.3.2 Principal component analysis*

Data visualisation was performed using the scores obtained from principal component analysis (PCA). Data were mean centred prior to analysis and PCA decomposition was performed in Matlab utilising PLS\_Toolbox (version 8.11, Eigenvector Research Inc., USA).

### *3.3.3.3 Classification with support vector machines*

Classification of the 16 green tea products on the basis of individual samples (16 classes), country of origin (7 classes) and continent of origin (4 classes) was performed using multi-class error correcting output code (ECOC) models containing SVM binary learners, which used a one-versus-one coding design. A total of  $k$  number of learners was utilised in combination in the coding design. A quadratic kernel (scale parameter = 5) was used to map the data to a higher dimension. Pre-processed spectra of the 16 green tea samples were extracted. For every pure green tea sample, pre-processed spectra were extracted from 200 pixels, which were selected at random from the image collected, using the 'roipoly' function in MATLAB (R2016b, Mathworks, Natwick, USA). This gave a total of 3200 spectra for the calibration of the classification models. The models were cross-validated with a 10-fold cross-validation method. Furthermore, this whole calibration procedure was performed with 100 iterations and the mean validation accuracy and the standard deviation were recorded. The trained classifiers were later used to generate the classification maps of the tea samples based on sample type, and the country and continent of origin. The ECOC-SVM models were implemented in MATLAB using the Statistics and Machine Learning Toolbox (R2016b). The predictive accuracy of the ECOC-SVM model built using the full (967.11 - 1700 nm) spectral range was used as the performance criterion for the feature

selection methods.

#### *3.3.3.4 Feature selection to identify potential NIR regions*

In the present work, two different feature selection methods, ReliefF and SFS, were explored. The two methods are distinct in their statistical backgrounds; ReliefF is based on a fast filter-based approach while SFS is a computationally-intensive wrapper-based method. The choice of these two methods was also made based on their level of interaction with the classifiers; ReliefF is independent of the classifier and selects the features before developing the classifier, while the SFS is dependent on the classifier to select the subset of features. The methods are briefly explained in the following sections.

##### *3.3.3.4.1 ReliefF*

ReliefF is a filter-based approach for feature selection and belongs to the family of Relief methods, which were designed to estimate attributes based on how well their values distinguish between the instances that are near to each other [45]. The Relief methods, like typical filter-based methods, use the variable ranking technique as the criterion for feature selection. The initial version of Relief was only capable of dealing with two class problems. However, the extension ReliefF can deal with multi-class problems including incomplete and noisy data [46]. The key idea behind ReliefF is the same as Relief and starts with searching for the nearest hit and miss, however, ReliefF searches for  $k$  of its nearest neighbours from the same class as the nearest hit and another  $k$  nearest neighbours from different classes as the nearest miss. ReliefF updates the weights for variables using the average contribution of all the hits and all the misses. Further, the contribution for each class of the misses is estimated from the prior probability for that particular class. The normalisation of prior probabilities of classes is also performed to deal with the influence of classes with a small number of cases for  $k$  near misses from each different class. Further, like Relief, ReliefF is also repeated  $m$  number of times.

In the present work, the ReliefF algorithm was implemented in Matlab using the Statistics and Machine Learning Toolbox. Three different ReliefF feature selections

were performed for the 16, 7 and 4 class classification cases corresponding to samples, countries and continents, respectively. Since ReliefF is a filter-based approach, a threshold is required to select the features. In our case, we present a methodology to automatically decide on the threshold for selection of the relevant features. The methodology partitions the range of ReliefF weights into 100 steps and a classification model is developed for every threshold value, and the validation accuracy recorded. The threshold was selected based on a model accuracy greater than 90% and the corresponding features were retained.

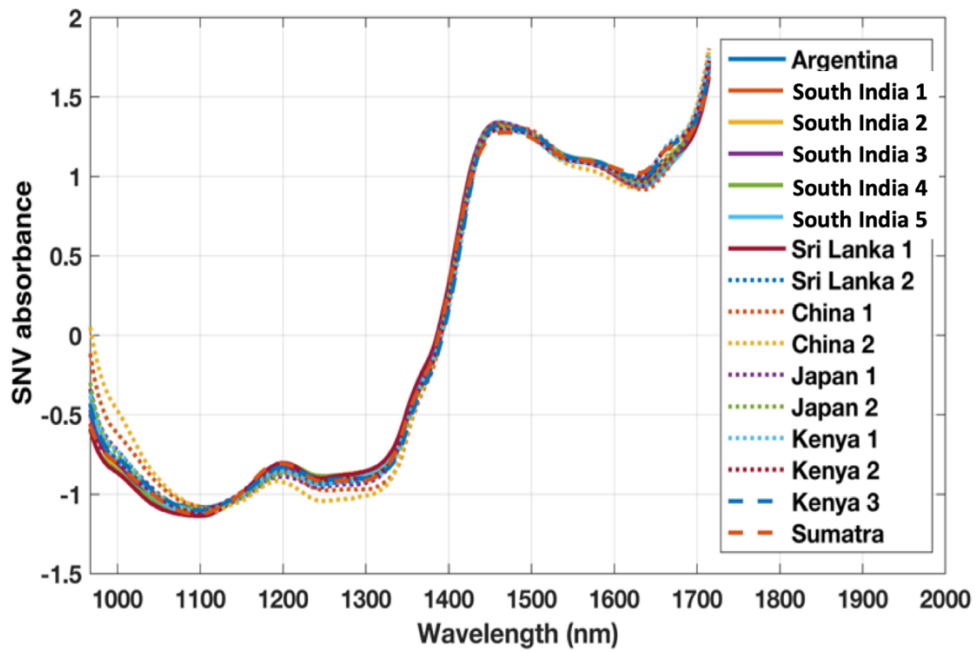
#### 3.3.3.4.2 Sequential forward selection

SFS belongs to the family of wrapper-based methods, which aim to find the subset of features through exhaustive search [41]. The exhaustive search involves assessing all of the possible combinations of subsets of features against a particular criterion. The criterion can be explored for both regression or classification based on the need. Further, the selection criteria for the features involves minimisation or maximisation of criterion such as reduction of the predictive error for regression and minimisation of the misclassification rate for classification. Typically, the forward selection case starts with an empty feature set and the criterion is assessed by involving a feature in the modelling. If inclusion of the feature in the model satisfies the criterion selected, the feature set is updated by adding the new feature. This procedure is repeated until all relevant subsets of features are selected. In the present work, the criterion chosen was misclassification accuracy of the ten-fold cross validated multi-class ECOC-SVM classification models for the 16, 7 and 4 class classification cases corresponding to samples, countries and continents, respectively. Therefore, three different SFS feature selections were performed for classification of green tea products.

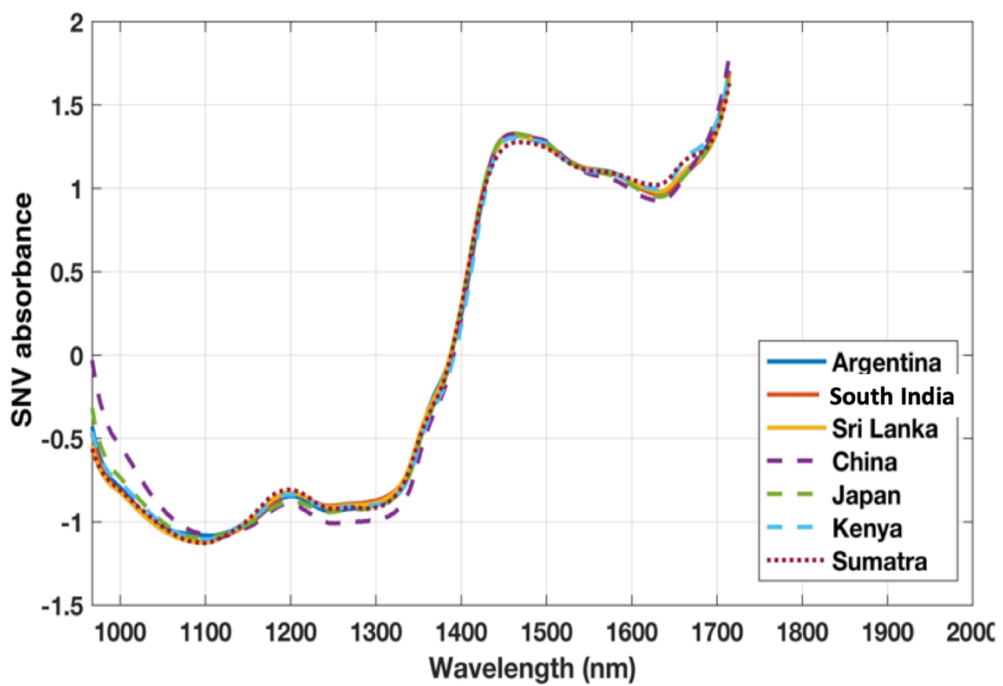
In the present work, the SFS algorithm was implemented in Matlab using the Statistics and Machine Learning Toolbox.

### 3.4 Results

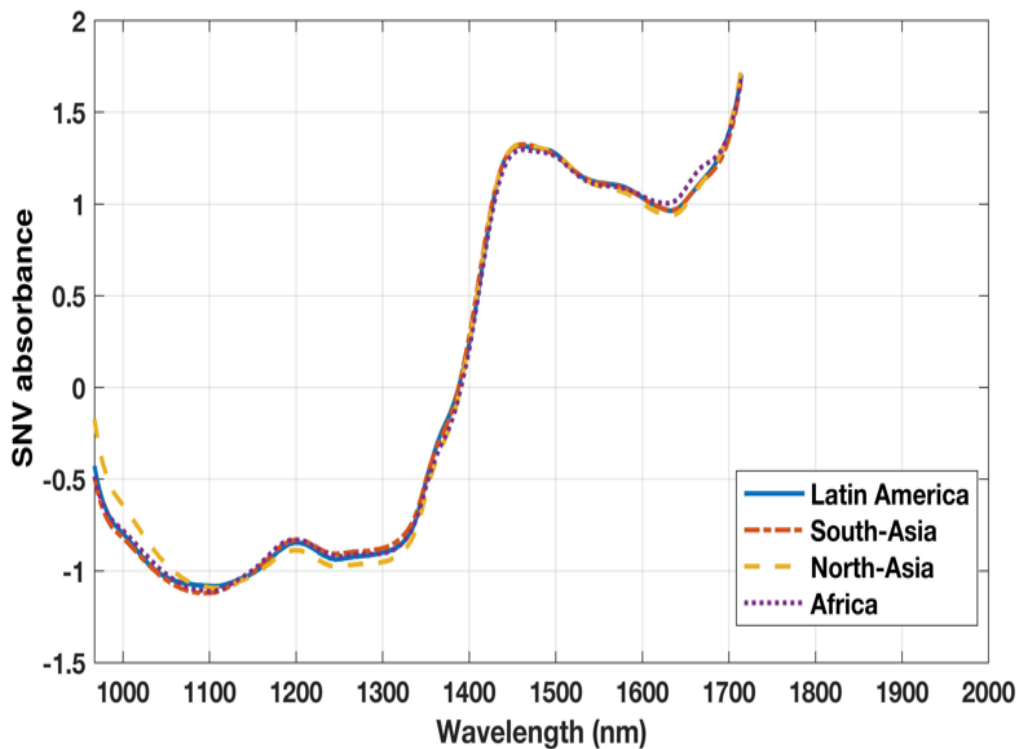
#### 3.4.1 Spectral profiles



(a)



(b)



(c)

*Figure 3.2: Mean SNV pre-processed absorbance spectral profiles of tea samples (a) corresponding to 16 products, (b) corresponding to 7 countries of origin, and (c) corresponding to 4 continents of origin.*

Figure 3.2 (a, b and c) presents the mean SNV pre-processed absorbance spectral profiles of the green tea samples corresponding to 16 individual products, 7 countries of origin and 4 continents of origin. The spectra were pre-processed to reduce the light scattering effects originating from inhomogeneous surface of the tea products as recommended in [34]. In Figure 3.2 (a, b and c) it can be noted that variability in spectra is present at various regions over the complete range for NIR spectra recorded. Such variability can be noted higher at specific locations compared to others. To understand this variability and the specific NIR regions the standard deviation was calculated and presented in Figure 3.3.

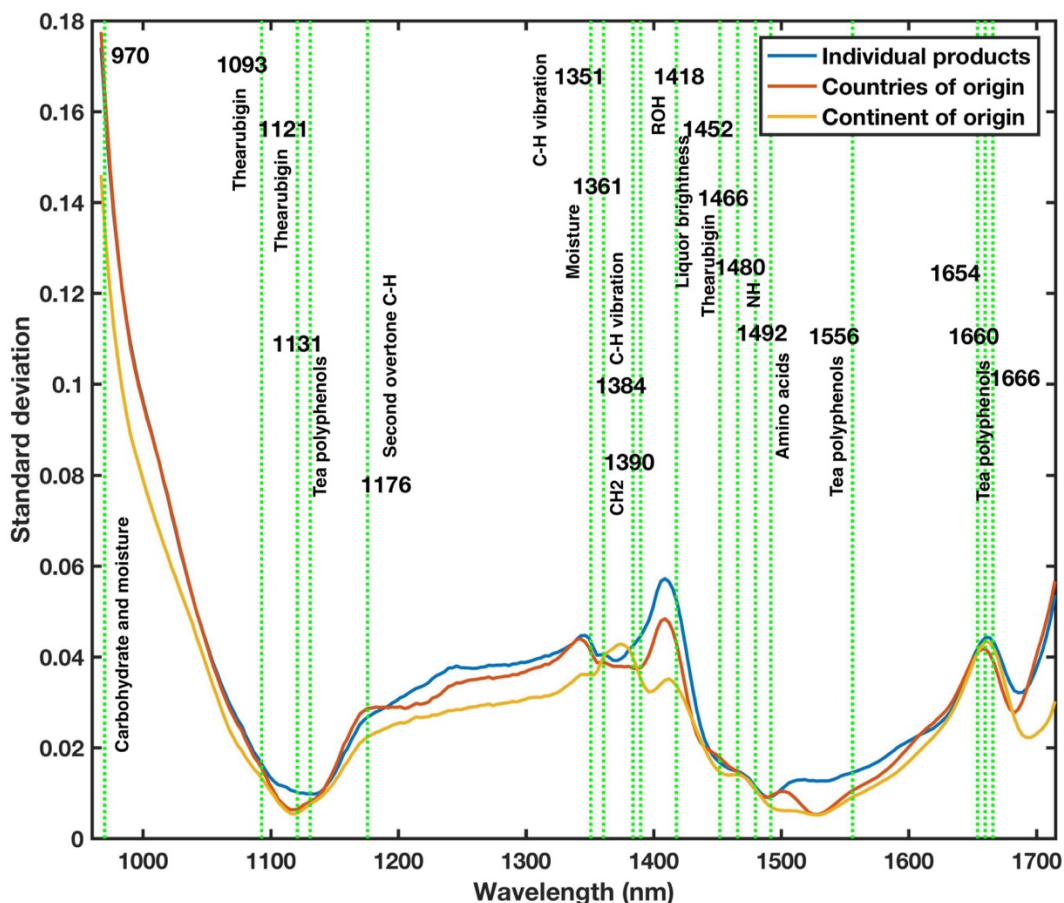


Figure 3.3: Standard deviation calculated for NIR spectra of individual products (blue), countries of origin (red) and continents of origin (Yellow). The vertical green lines denote the assignment of wavelengths to different components present in tea products based on the literature.

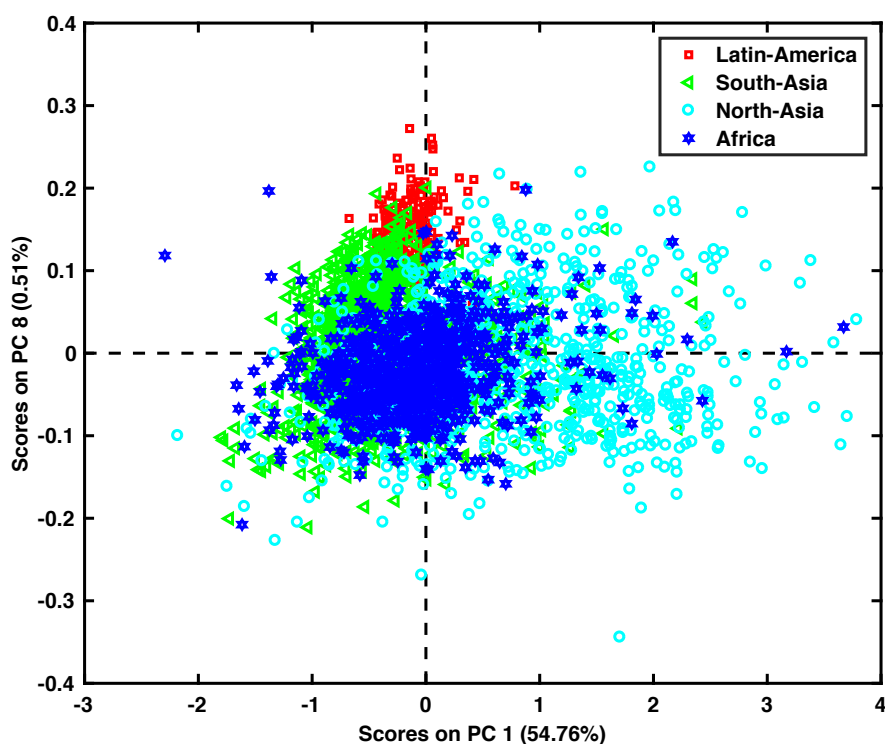
Figure 3.3 presents the standard deviation calculated for mean spectra of individual products (in blue), mean spectra based on countries of origin (in red) and mean spectra based on continents of origin (in yellow). The vertical lines presented in Figure 3.3 denote the assignment of wavelengths to different components present in tea products based on the literature. Using previous works, the following assignments can be made: carbohydrates and moisture (970 nm) [52]; Thearubigin components of TRS1 (1093 - 1121 nm) [49]; total tea polyphenols (1131, 1654 and 1666 nm) [47, 48], a second overtone C-H (1176 nm) [51], Theaflavin (1218 nm) [49], a C-H vibration (1351 – 1384 nm) [53], moisture (1361 nm) [49], CH<sub>2</sub> (1390 nm) [50], ROH (1418 nm) [50],



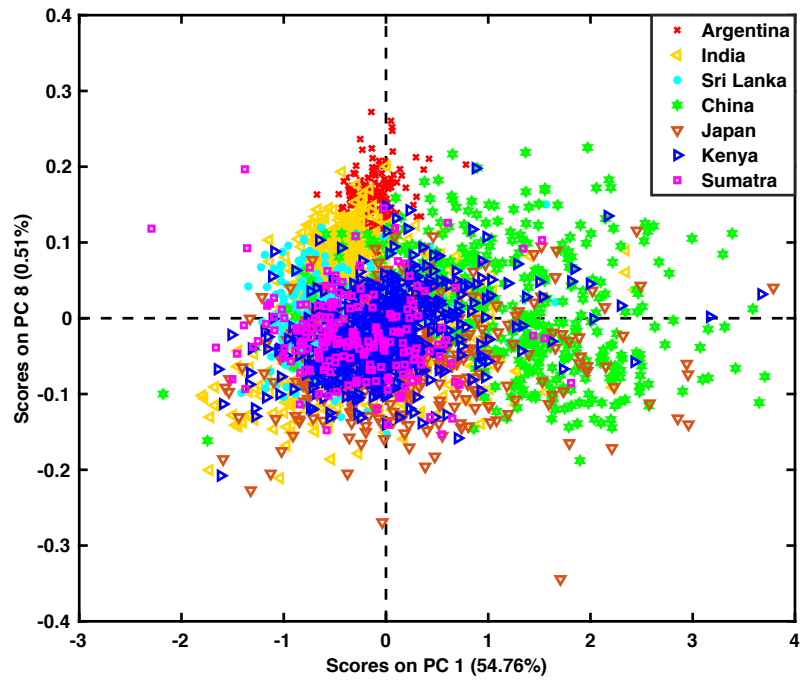
liquor brightness (1452 nm) [49], Thearubigin components of TRS1 (1416 and 1466 nm) [49], NH (1480 nm) [50], free amino acids (1492 nm) [47] and tea polyphenols (1556-1660 nm) [48].

### 3.4.2 Principal component analysis

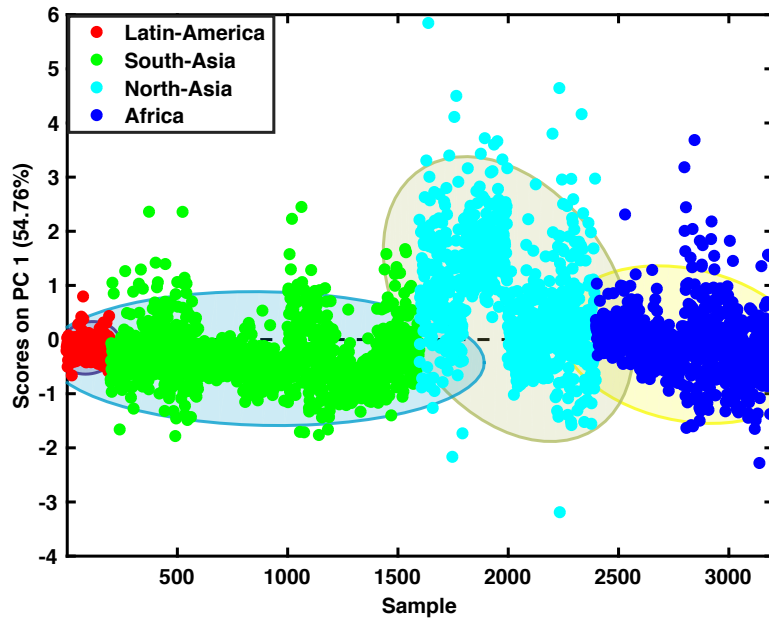
PCA was performed on the NIR spectra of tea products and the results obtained for PC1 (54.76%) and PC8 (0.51%) are shown in Figure 3.4. Figures 3.4(a) and 3.4(b) present the PC1 v PC8 scores biplots with samples highlighted on the basis of continent and country, respectively. The scores biplots for PC1 v PCs 2 – 7, colour coded on the basis of continent, are presented in Figures A2.1 to A2.6 in Annex 2. In Figure 3.4(a), the samples corresponding to Latin-America, South-Asia, North-Asia and Africa can be identified with red squares, green triangles, sky blue circles and dark blue stars, respectively.



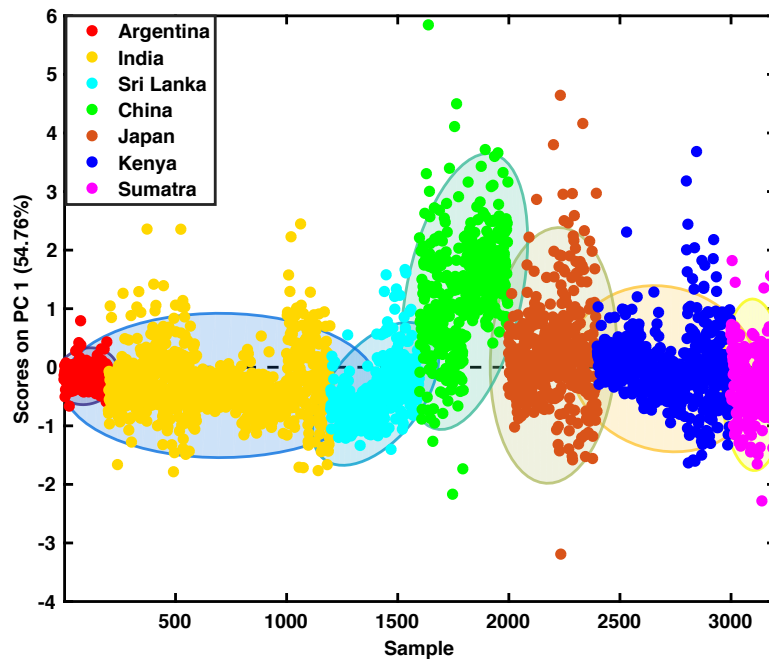
(a)



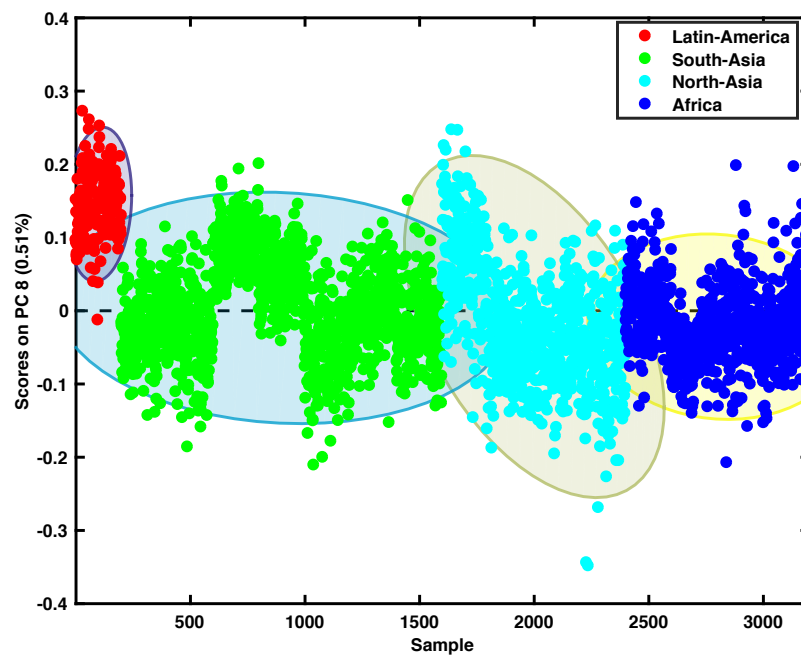
(b)



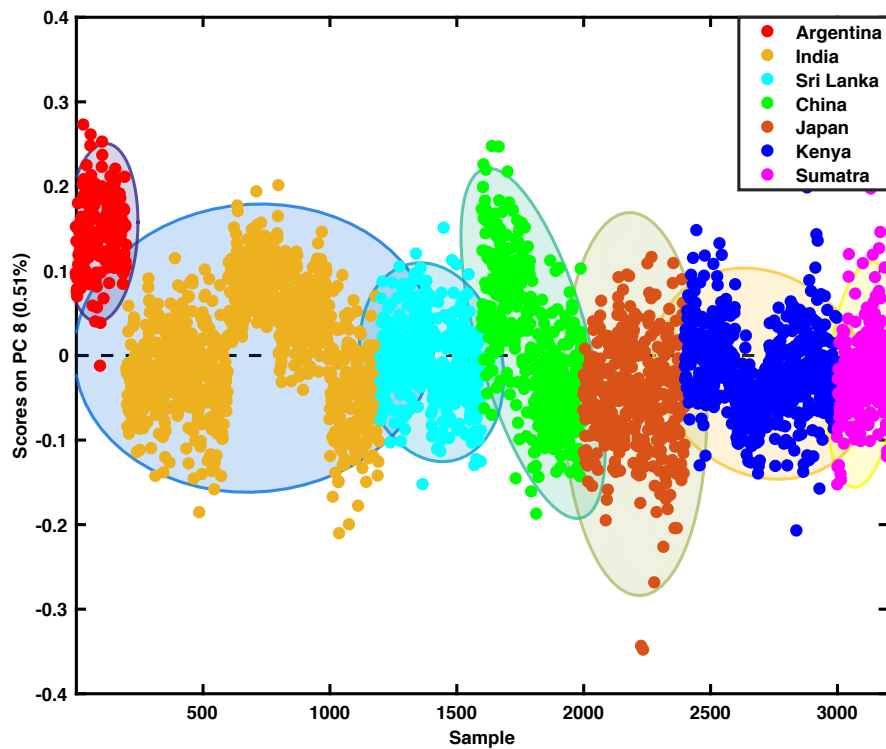
(c)



(d)



(e)

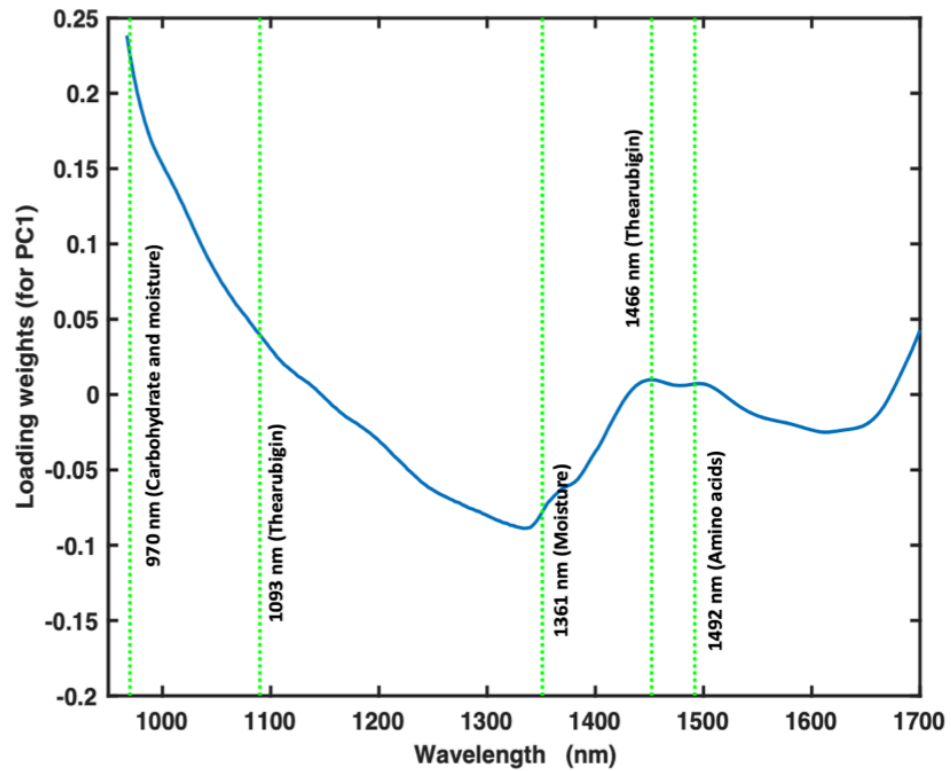


(f)

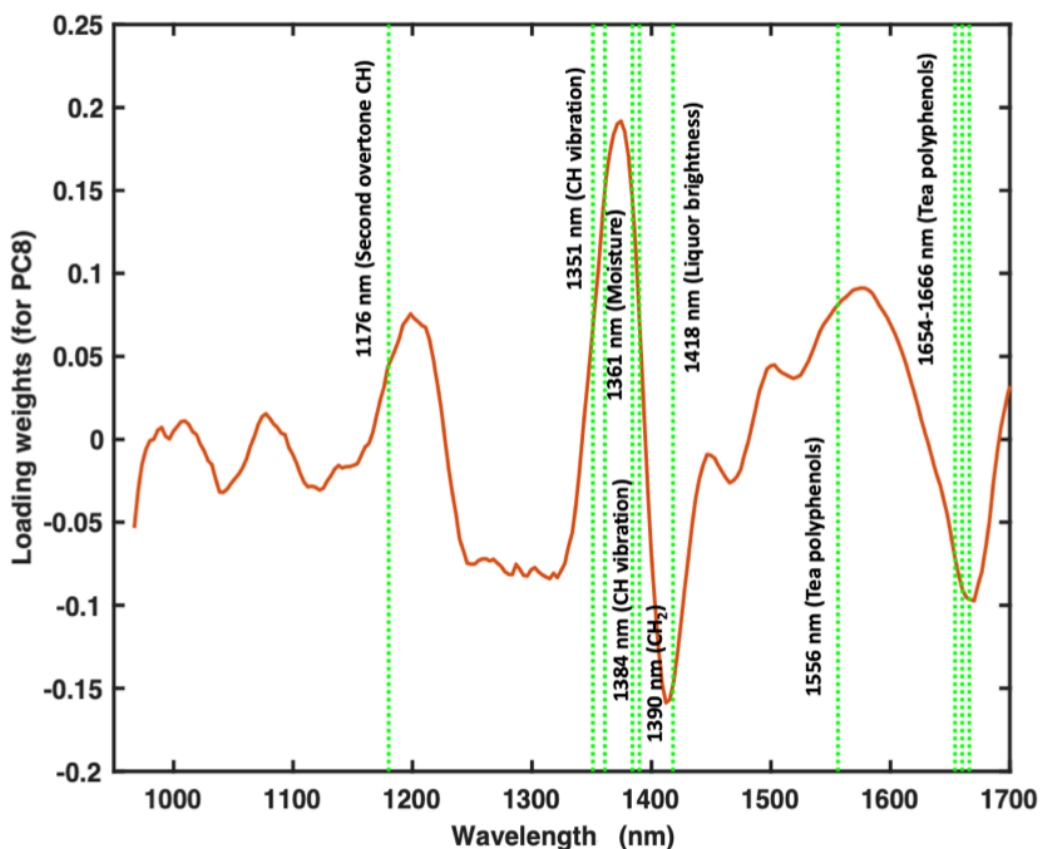
Figure 3.4 : a) Principal component scores (PC1 v PC8) for visualising data based on continent of origin, b) principal component scores (PC1 v PC8) for visualising data based on country of origin, c) PC1 scores for samples colour coded according to the continent of origin, d) PC1 scores for samples colour coded according to the country of origin, e) PC8 scores for samples colour coded according to the continent of origin, and f) PC8 scores for samples colour coded according to the country of origin. The ellipses in c) to f) are to aid visualisation of samples from different continents and countries.

In Figure 3.4(b), the samples corresponding to Argentina, India, Sri-Lanka, China, Japan, Kenya and Sumatra can be understood as red crosses, yellow triangles, sky-blue circles, green stars, orange triangles, dark-blue triangles and pink squares, respectively. It can be seen in Figures 3.4(a) and 3.4(b) that the samples are overlapped as the spectral profiles of different green tea products are very similar. No information

was available on any potential differences between samples originating from the same country. However, it can be seen from Figures 3.4(a) and 3.4(e) that the samples from Latin-America, in this case Argentina, have a high PC8 score making them distinct from other samples. In addition, the samples from North Asia (sky blue circles in Figure 3.4(a)) and specifically from China (sky blue circles in Figure 3.4(c)) have, in general, a higher PC1 score.



(a)



(b)

Figure 3.5: Loading vectors corresponding to (a) PC1 and (b) PC8.

The corresponding loadings for PC1 (a) and PC8 (b) are presented in Figure 3.5. It can be seen in Figure 3.5(a) that the loading weights for PC1 are higher for the spectral range of 950 to 1100 nm compared to the rest of the spectrum. Potential assignment of regions with the highest loadings include carbohydrates (970 nm) and moisture (970 and 1361 nm), Thearubigin (1093 and 1466 nm) and amino acids (1492 nm). PC8, on the other hand, has higher positive loading weights for peaks at around 1200 nm, 1350 – 1390 and 1585 nm. These peaks can be attributed to CH vibrations at 1176, 1351, 1384 and 1390 nm. The PC8 loadings plot (Figure 3.5(b)) also contains some negative peaks, which can be attributed to liquor brightness (1418 nm) and tea polyphenols (1654 – 1666 nm).

### 3.4.3 Support vector machine classification

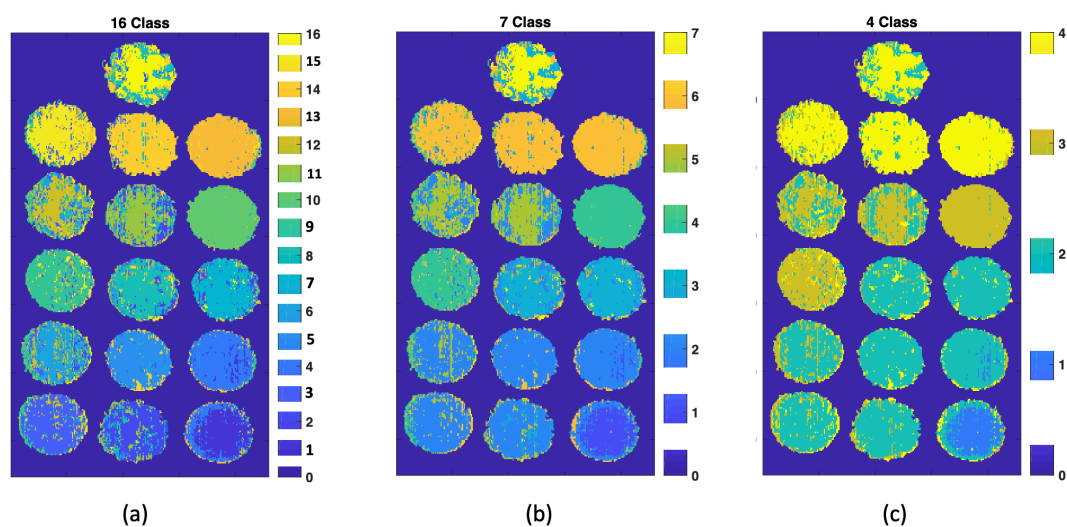


Figure 3.6: Classification maps for the 16 green tea samples obtained using classification models corresponding to a) different samples, b) country of origin and c) continent of origin. Class 0 presents the dark blue background and does not correspond to any tea product.

Figure 3.6 presents the classification maps obtained from the 10-fold cross-validated multiclass SVM-ECOC models for the sixteen (Figure 3.6(a)), seven (Figure 3.6(b)) and four (Figure 3.6(c)) class problems explaining classification based on the samples, country of origin, and continent of origin, respectively. The classification accuracy obtained for prediction of individual products, country of origin, and continent of origin were  $97.01 \pm 0.17\%$ ,  $96.36 \pm 0.17\%$  and  $97.77 \pm 0.16\%$ , respectively. Confusion matrices showing classification accuracies for individual classes obtained using continent, country and product SVM models are given in Figures A2.7 to A2.9, respectively, of Annex 2. In Figure 3.6, the presence of misclassification can be seen for all the tea samples. The source of the misclassification can be understood from

Figure 3.7, where misclassification of samples originating from different continents (Figure 3.7a) and different countries (Figure 3.7b) are presented.

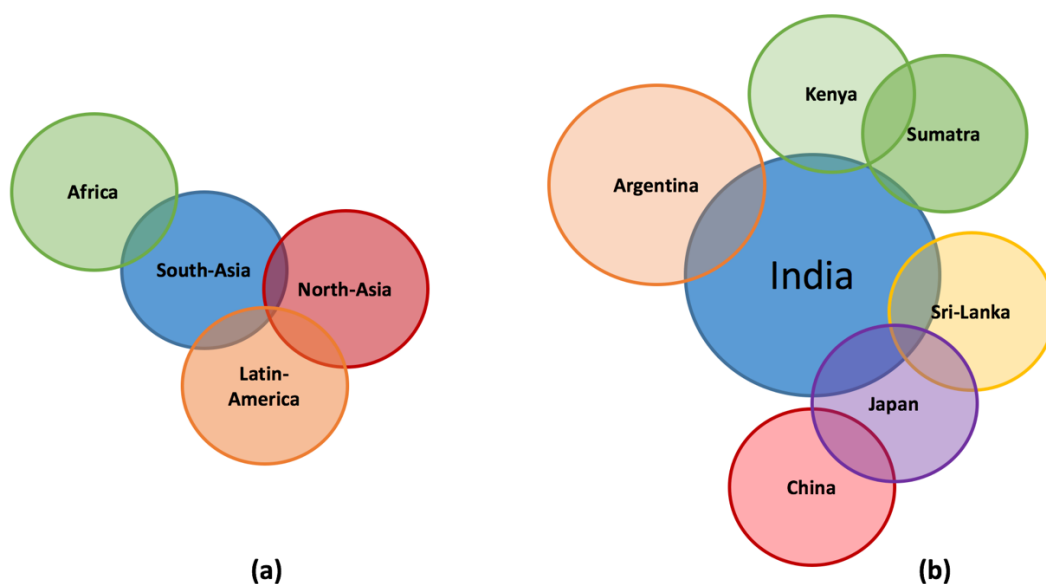


Figure 3.7: Summary of misclassification between different samples based on: (a) continents and (b) countries.

It can be seen from Figure 3.7(a) that samples from Africa, North-Asia and Latin-America are misclassified as samples from South-Asia. Further, the North-Asian samples and Latin-American samples have some misclassification with each other. However, the North-Asian and Latin-American samples do not have any misclassification with the samples originating from Africa. Further detailed misclassification insight is presented in Figure 3.7(b). It can be seen that all of the samples except those from China have some misclassification with the samples of Indian origin. The samples of Chinese origin have some misclassification with the samples of Japanese origin. Furthermore, the samples of Japanese origin have some misclassification with the samples from Sri-Lanka. The samples from Sumatra and Kenya have some misclassification with each other but not with the samples from Sri-Lanka and Japan.



### 3.4.4 Feature selection to identify potential NIR regions

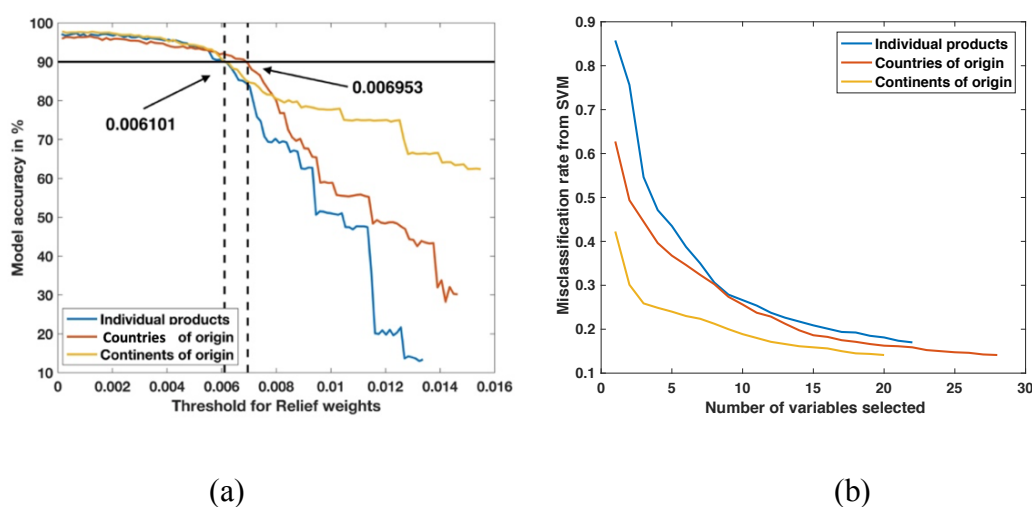
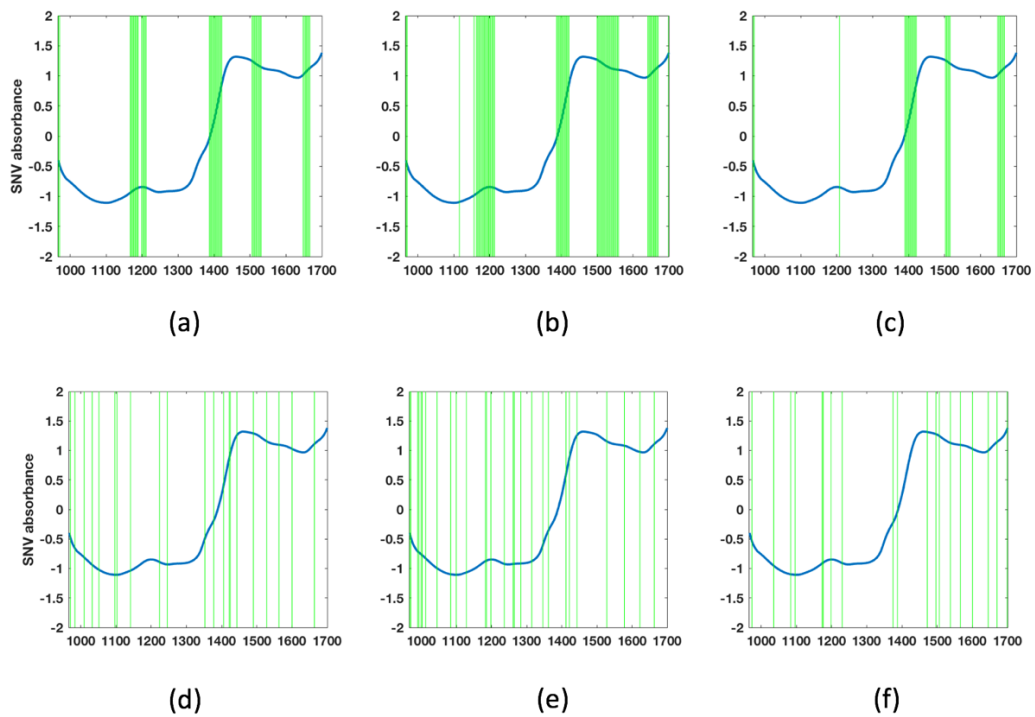


Figure 3.8: (a). ReliefF threshold for weights and the model accuracy obtained from using features corresponding to a particular threshold. The black horizontal line represents a model accuracy of 90 %. The two vertical dashed lines represent the final threshold used for the three different classification cases. (b) The evolution of the misclassification rate (SFS criterion) with the number of selected wavelengths.

Figure 3.8 presents the evolution of the criterion of the feature selection methods used. Figure 3.8(a) presents the evolution of the classification model accuracy with increasing threshold for relief weights. Figure 3.8(b) presents the decrease of misclassification rate as the number of features added was increased. In Figure 3.8, the three different classification models corresponding to samples, countries and continents of origin can be identified with blue, red and yellow coloured lines, respectively. Figure 3.8(a) for the ReliefF also presents two dashed vertical lines highlighting selection of a subset of features that can provide a classification accuracy of at least 90%. It was hypothesised that a 90% accurate model would assist in identifying the key important wavelengths. Maintaining the same classification accuracy as that obtained with the entire spectral range was not expected to result in a significant reduction in the number of wavelengths. This is because the tea sample spectra are very similar and peaks are not evolving with time. Usually in process

monitoring applications, it is possible to identify a much smaller set of wavelengths that gives a similar model accuracy as that attainable using the full spectrum as there are specific peaks evolving or shifting. A different number of features were obtained for the three different classification problems based on the threshold (see Figure 3.9). In Figure 3.8(b), the SFS combined with the EVM-ECOC classifier provided a decreasing misclassification rate for the three cases, and the corresponding extracted features are presented in Figure 3.9. In Figure 3.8(b), it can be seen that for the sixteen-class classification case the misclassification rate starts to decrease from a very high value, followed by the seven class and four class classification case.



*Figure 3.9: Wavelength selection (vertical green lines) performed with ReliefF for 90% validation accuracy. Selected wavelengths for (a) 16 sample classification problem, (b) 7 country classification problem and (c) 4 continent classification problem. Selected wavelengths with sequential forward selection (SFS) for (d) 16 sample classification problem, (e) 7 country classification problem and (f) 4 continent classification problem.*

Figure 3.9 presents the features selected with ReliefF (Figure 3.9 (a,b,c)) and SFS (Figure 3.9 (d,e,f)). The features selected are shown with the vertical green lines. In Figure 3.9 (a,b,c), the features selected with ReliefF are located very near to each other compared to the features selected with SFS (Figure 3.9 (d,e,f)). A reason for this is that ReliefF extracts correlated features, which is not the case with SFS. The classification maps corresponding to selected variables can be found in Annex 2 (Figure A2.1). It is easier to identify the important regions that are responsible for discrimination between different classes using ReliefF. In comparison, with SFS the features extracted have different wavelengths and no such concentrated information about the features can be extracted. Both ReliefF and SFS share some common features too, however, for different classification models different spectral regions were identified, which is more visually apparent with ReliefF (Figure 3.9 (a,b,c)). The different spectral regions and spectral bands identified by ReliefF and SFS, respectively, are presented in Table 3.1.

*Table 3.1: A summary of the spectral regions and bands selected from the range 950 – 1700 nm using ReliefF and SFS for classification of samples with a predictive accuracy of 90% on the basis of product, country and continent.*

		Basis of classification		
		Product type	Country of origin	Continent of origin
ReliefF	% of variables selected from full spectrum	16.4%	22.5%	7.5%
	Selected wavelengths (nm) and main assignments	967.1 (Carbohydrate and moisture), 1170 – 1211 (second overtone CH),	967.1 (Carbohydrate and moisture), 1116 (Thearubigin), 1157 – 1214 (Second overtone	970.4 (Carbohydrate and moisture), 1208 (second overtone CH), 1390 – 1421 (CH <sub>2</sub> and ROH),

		1387 – 1421 (ROH), 1509 – 1531, 1651 – 1667 (tea polyphenols)	CH), 1387 – 1421 (CH <sub>2</sub> and ROH), 1500 – 1559 (tea polyphenols), 1641 – 1670 (tea polyphenols)	1506 - 1515, 1651 – 1667 (tea polyphenols)
SFS	% of variables selected from full spectrum	9.2%	11.4%	8.0%
	Selected wavelengths (nm) and main assignments	970.4 (Carbohydrate and moisture), 983.4, 1009, 1032, 1052, 1096 (thearubigin), 1103, 1141, 1224, 1246, 1353 (CH vibration), 1378, 1406, 1421, 1425, 1443, 1490 (NH), 1528, 1563 (Tea polyphenols), 1600, 1664 (Tea polyphenols), 1702	970.4 (Carbohydrate and moisture), 990, 993.2, 999.7, 1003, 1013, 1045, 1084 (thearubigin), 1100 (thearubigin), 1128 (thearubigin), 1183, 1186 (Second overtone CH), 1198, 1236, 1261, 1265, 1284, 1315, 1346 (CH vibration), 1362 (moisture), 1412 (ROH), 1421 (ROH), 1443 (Liquor brightness), 1528, 1578, 1622 (tea polyphenols),	973.6 (Carbohydrate and moisture), 1035, 1084, 1096 (therubigin), 1173 (second overtone CH), 1176 (second overtone CH), 1198, 1230, 1374, 1387 (CH vibration), 1472 (therubigin), 1497 (amino acids), 1506, 1537, 1566, 1600, 1645, 1670, 1699

			1664 (tea polyphenols)	
--	--	--	------------------------	--

### 3.5 Conclusions

Non-destructive and rapid classification of green tea products of different geographical origin is crucial for international market and trade. NIR-HSI and multi-class SVM were successfully used to perform geographical classification of green tea products based on country and continent of origin. The SVM-ECOC model provided a classification accuracy of  $97.01 \pm 0.17\%$ ,  $96.36 \pm 0.17\%$  and  $97.77 \pm 0.16\%$  for the sixteen, seven and four class problems explaining the classification based on the samples, countries and continent of origin, respectively. The tea samples considered were from India, Sri-Lanka, China, Japan, Kenya, Sumatra and Argentina. PCA provided primary visualisation of tea samples. ReliefF and SFS identified the subset of spectral bands which can be used to obtain classification models with a classification accuracy of 90%. The selected spectral bands can also be used to compress the images by removing the unimportant spectral bands. By identifying the correlated features, ReliefF identified spectral regions whereas SFS identified spectral wavelengths spread all over the spectral range, making interpretation more difficult compared to the results obtained using ReliefF. However, implementation of ReliefF is fast, computationally light, it avoids overfitting and has generalisation ability. On the other hand, SFS has the advantage of interaction with the classifier and explanation of the feature dependencies. The drawback of ReliefF is the need to select the threshold whereas the drawbacks of SFS includes high computational cost, the risk of overfitting and classifier dependent feature selection.

### 3.6 References

1. D. Del Rio, L. Calani, F. Scazzina, L. Jechiu, C. Cordero, F. Brighenti, Bioavailability of catechins from ready-to-drink tea, *Nutrition*. 26 (2010) 528–533. doi:<https://doi.org/10.1016/j.nut.2009.06.013>.
2. S. Sur, C.K. Panda, Molecular aspects of cancer chemopreventive and therapeutic efficacies of tea and tea polyphenols, *Nutrition*. 43-44 (2017) 8–15. doi:<https://doi.org/10.1016/j.nut.2017.06.006>.
3. E. Mancini, C. Beglinger, J. Drewe, D. Zanchi, U.E. Lang, S. Borgwardt, Green tea effects on cognition, mood and human brain function: A systematic review, *Phytomedicine*. 34 (2017) 26–37. doi:<https://doi.org/10.1016/j.phymed.2017.07.008>.
4. L. Chen, H. Mo, L. Zhao, W. Gao, S. Wang, M.M. Cromie, C. Lu, J.-S. Wang, C.-L. Shen, Therapeutic properties of green tea against environmental insults, *J. Nutr. Biochem*. 40 (2017) 1–13. doi:<https://doi.org/10.1016/j.jnutbio.2016.05.005>.
5. B. Rashidi, M. Malekzadeh, M. Goodarzi, A. Masoudifar, H. Mirzaei, Green tea and its anti-angiogenesis effects, *Biomed. Pharmacother*. 89 (2017) 949–956. doi:<https://doi.org/10.1016/j.biopha.2017.01.161>.
6. S. Roychoudhury, A. Agarwal, G. Virk, C.-L. Cho, Potential role of green tea catechins in the management of oxidative stress-associated infertility, *Reprod. Biomed. Online*. 34 (2017) 487–498. doi:<https://doi.org/10.1016/j.rbmo.2017.02.006>.

7. S. Bansal, S. Choudhary, M. Sharma, S.S. Kumar, S. Lohan, V. Bhardwaj, N. Syan, S. Jyoti, Tea: A native source of antimicrobial agents, *Food Res. Int.* 53 (2013) 568–584. doi:<https://doi.org/10.1016/j.foodres.2013.01.032>.
8. R. Badia-Melis, P. Mishra, L. Ruiz-García, Food traceability: New trends and recent advances. A review, *Food Control.* 57 (2015) 393–401. doi:<https://doi.org/10.1016/j.foodcont.2015.05.005>.
9. Q. Chen, J. Zhao, H. Lin, Study on discrimination of Roast green tea (*Camellia sinensis* L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition, *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* 72 (2009) 845–850. doi:<http://dx.doi.org/10.1016/j.saa.2008.12.002>.
10. D. Huo, Y. Wu, M. Yang, H. Fa, X. Luo, C. Hou, Discrimination of Chinese green tea according to varieties and grade levels using artificial nose and tongue based on colorimetric sensor arrays, *Food Chem.* 145 (2014) 639–645. doi:<https://doi.org/10.1016/j.foodchem.2013.07.142>.
11. W. He, X. Hu, L. Zhao, X. Liao, Y. Zhang, M. Zhang, J. Wu, Evaluation of Chinese tea by the electronic tongue: Correlation with sensory properties and classification according to geographical origin and grade level, *Food Res. Int.* 42 (2009) 1462–1467. doi:<https://doi.org/10.1016/j.foodres.2009.08.008>.
12. X. He, J. Li, W. Zhao, R. Liu, L. Zhang, X. Kong, Chemical fingerprint analysis for quality control and identification of Ziyang green tea by HPLC, *Food Chem.* 171 (2015) 405–411. doi:<https://doi.org/10.1016/j.foodchem.2014.09.026>.

13. S. Lv, Y. Wu, J. Zhou, M. Lian, C. Li, Y. Xu, S. Liu, C. Wang, Q. Meng, The Study of Fingerprint Characteristics of Dayi Pu-Erh Tea Using a Fully Automatic HS-SPME/GC–MS and Combined Chemometrics Method, *PLoS One*. 9 (2015) e116428. <https://doi.org/10.1371/journal.pone.0116428>.
14. T.S. Pilgrim, R.J. Watling, K. Grice, Application of trace element and stable isotope signatures to determine the provenance of tea (*Camellia sinensis*) samples, *Food Chem.* 118 (2010) 921–926. doi:<https://doi.org/10.1016/j.foodchem.2008.08.077>.
15. H. Yu, J. Wang, H. Zhang, Y. Yu, C. Yao, Identification of green tea grade using different feature of response signal from E-nose sensors, *Sensors Actuators B Chem.* 128 (2008) 455–461. doi:<https://doi.org/10.1016/j.snb.2007.07.048>.
16. H. Yu, J. Wang, Discrimination of LongJing green-tea grade by electronic nose, *Sensors Actuators B Chem.* 122 (2007) 134–140. doi:<https://doi.org/10.1016/j.snb.2006.05.019>.
17. Q. Chen, J. Zhao, Z. Chen, H. Lin, D.-A. Zhao, Discrimination of green tea quality using the electronic nose technique and the human panel test, comparison of linear and nonlinear classification tools, *Sensors Actuators B Chem.* 159 (2011) 294–300. doi:<https://doi.org/10.1016/j.snb.2011.07.009>.
18. H. Yu, J. Wang, H. Xiao, M. Liu, Quality grade identification of green tea using the eigenvalues of PCA based on the E-nose signals, *Sensors Actuators B Chem.* 140 (2009) 378–382. doi:<https://doi.org/10.1016/j.snb.2009.05.008>.



19. P. Pławiak, W. Maziarz, Classification of tea specimens using novel hybrid artificial intelligence methods, *Sensors Actuators B Chem.* 192 (2014) 117–125. doi:<https://doi.org/10.1016/j.snb.2013.10.065>.
20. L. Lvova, A. Legin, Y. Vlasov, G.S. Cha, H. Nam, Multicomponent analysis of Korean green tea by means of disposable all-solid-state potentiometric electronic tongue microsystem, *Sensors Actuators B Chem.* 95 (2003) 391–399. doi:[https://doi.org/10.1016/S0925-4005\(03\)00445-3](https://doi.org/10.1016/S0925-4005(03)00445-3).
21. P. Ivarsson, S. Holmin, N.-E. Höjer, C. Krantz-Rülcker, F. Winqvist, Discrimination of tea by means of a voltammetric electronic tongue and different applied waveforms, *Sensors Actuators B Chem.* 76 (2001) 449–454. doi:[https://doi.org/10.1016/S0925-4005\(01\)00583-4](https://doi.org/10.1016/S0925-4005(01)00583-4).
22. E.A. Baldwin, J. Bai, A. Plotto, S. Dea, Electronic Noses and Tongues: Applications for the Food and Pharmaceutical Industries, *Sensors (Basel)*. 11 (2011) 4744–4766. doi:10.3390/s110504744.
23. H. Horie, S. Fukatsu, T. Mukai, T. Goto, M. Kawanaka, T. Shimohara, Quality evaluation on green tea, *Sensors Actuators B Chem.* 13 (1993) 451–454. doi:[https://doi.org/10.1016/0925-4005\(93\)85424-9](https://doi.org/10.1016/0925-4005(93)85424-9).
24. I. Domínguez, A. Doménech-Carbó, Screening and authentication of tea varieties based on microextraction-assisted voltammetry of microparticles, *Sensors Actuators B Chem.* 210 (2015) 491–499. doi:<https://doi.org/10.1016/j.snb.2015.01.009>.
25. S. Masoum, M. Behpour, F. Azimi, M.H. Motaghedifard, Potentiality of chemometric approaches for the determination of (+)-catechin in green tea leaves at the surface of multiwalled carbon nanotube paste electrode, *Sensors*

- Actuators B Chem. 193 (2014) 582–591.  
doi:<https://doi.org/10.1016/j.snb.2013.12.022>.
26. A.S. Kumar, R. Shanmugam, S. Nellaiappan, R. Thangaraj, Tea quality assessment by analyzing key polyphenolic functional groups using flow injection analysis coupled with a dual electrochemical detector, *Sensors Actuators B Chem.* 227 (2016) 352–361.  
doi:<https://doi.org/10.1016/j.snb.2015.12.072>.
27. J. Li, B. Fu, D. Huo, C. Hou, M. Yang, C. Shen, H. Luo, P. Yang, Discrimination of Chinese teas according to major amino acid composition by a colorimetric {IDA} sensor, *Sensors Actuators B Chem.* 240 (2017) 770–778.  
doi:<http://dx.doi.org/10.1016/j.snb.2016.09.019>.
28. P. Wang, Z. Yu, Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: A review, *J. Pharm. Anal.* 5 (2015) 277–284. doi:<https://doi.org/10.1016/j.jpha.2015.04.001>.
29. D.F. Barbin, A.L. de S.M. Felicio, D.-W. Sun, S.L. Nixdorf, E.Y. Hirooka, Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview, *Food Res. Int.* 61 (2014) 23–32.  
doi:<https://doi.org/10.1016/j.foodres.2014.01.005>.
30. P. Lin, Y. Chen, Y. He, Identification of Geographical Origin of Olive Oil Using Visible and Near-Infrared Spectroscopy Technique Combined with Chemometrics, *Food Bioprocess Technol.* 5 (2012) 235–242.  
doi:[10.1007/s11947-009-0302-z](https://doi.org/10.1007/s11947-009-0302-z).
31. V. Uričková, J. Sádecká, Determination of geographical origin of alcoholic beverages using ultraviolet, visible and infrared spectroscopy: A review,

- Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 148 (2015) 131–137.  
doi:<https://doi.org/10.1016/j.saa.2015.03.111>.
32. G.S. Gill, A. Kumar, R. Agarwal, Monitoring and grading of tea by computer vision – A review, J. Food Eng. 106 (2011) 13–19.  
doi:<https://doi.org/10.1016/j.jfoodeng.2011.04.013>.
33. A.A. Gowen, Y. Feng, E. Gaston, V. Valdramidis, Recent applications of hyperspectral imaging in microbiology, Talanta. 137 (2015) 43–54.  
doi:<https://doi.org/10.1016/j.talanta.2015.01.012>.
34. P. Mishra, A. Nordon, J. Tschannerl, G. Lian, S. Redfern, S. Marshall, Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products, J. Food Eng. 238 (2018) 70–77.  
doi:<https://doi.org/10.1016/j.jfoodeng.2018.06.015>.
35. W. Yin, C. Zhang, H. Zhu, Y. Zhao, Y. He, Application of near-infrared hyperspectral imaging to discriminate different geographical origins of Chinese wolfberries, PLoS One. 12 (2017) e0180534.  
<https://doi.org/10.1371/journal.pone.0180534>.
36. J. Gao, X. Li, F. Zhu, Y. He, Application of hyperspectral imaging technology to discriminate different geographical origins of *Jatropha curcas* L. seeds, Comput. Electron. Agric. 99 (2013) 186–193.  
doi:<https://doi.org/10.1016/j.compag.2013.09.011>.
37. J. Sun, X. Lu, H. Mao, X. Jin, X. Wu, A Method for Rapid Identification of Rice Origin by Hyperspectral Imaging Technology, J. Food Process Eng. 40 (2017) e12297–n/a. doi:10.1111/jfpe.12297.

38. H. Lv, W. Xu, J. You, S. Xiong, Classification of freshwater fish species by linear discriminant analysis based on near infrared reflectance spectroscopy, *J. Near Infrared Spectrosc.* 25 (2017) 54–62. <http://www.osapublishing.org/jnirs/abstract.cfm?URI=jnirs-25-1-54>.
39. J. Burger, A. Gowen, Data handling in hyperspectral image analysis, *Chemom. Intell. Lab. Syst.* 108 (2011) 13–22. doi:<https://doi.org/10.1016/j.chemolab.2011.04.001>.
40. T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, Chemometrics and Intelligent Laboratory Systems A review of variable selection methods in Partial Least Squares Regression, *Chemom. Intell. Lab. Syst.* 118 (2012) 62–69. doi:[10.1016/j.chemolab.2012.07.010](https://doi.org/10.1016/j.chemolab.2012.07.010).
41. G. Chandrashekar, F. Sahin, A survey on feature selection methods, *Comput. Electr. Eng.* 40 (2014) 16–28. doi:<https://doi.org/10.1016/j.compeleceng.2013.11.024>.
42. J. M. Amigo, I. Marti, A. Gowen, Hyperspectral imaging and chemometrics: a perfect combination for the analysis of food structure, composition and quality, *Data Handling in Science and Technology* 28 (0) (2013) 343–370.
43. A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures., *Anal. Chem.* 36 (1964) 1627–1639. doi:[10.1021/ac60214a047](https://doi.org/10.1021/ac60214a047).
44. R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra, *Appl. Spectrosc.* 43 (1989) 772–777. doi:[10.1366/0003702894202201](https://doi.org/10.1366/0003702894202201).

45. I. Kononenko, E. Šimec, M. Robnik-Šikonja, Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF, *Appl. Intell.* 7 (1997) 39–55. doi:10.1023/A:1008280620621.
46. V. Bolón-Canedo, N. Sánchez-Marroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, *Knowl. Inf. Syst.* 34 (2013) 483–519. doi:10.1007/s10115-012-0487-8.
47. M. Bian, A.K. Skidmore, M. Schlerf, T. Fei, Y. Liu, T. Wang, Reflectance spectroscopy of biochemical components as indicators of tea (*Camellia sinensis*) quality, *Photogramm. Eng. Remote Sens.* 76 (2010) 1385–1392.
48. M. Bian, A.K. Skidmore, M. Schlerf, T. Wang, Y. Liu, R. Zeng, T. Fei, Predicting foliar biochemistry of tea (*Camellia sinensis*) using reflectance spectra measured at powder, leaf and canopy levels, *ISPRS J. Photogramm. Remote Sens.* 78 (2013) 148–156. doi:https://doi.org/10.1016/j.isprsjprs.2013.02.002.
49. N. Panigrahi, C.S. Bhol, B.S. Das, Rapid assessment of black tea quality using diffuse reflectance spectroscopy, *J. Food Eng.* 190 (2016) 101–108. doi:https://doi.org/10.1016/j.jfoodeng.2016.06.020.
50. M.-S. Lee, Y.-S. Hwang, J. Lee, M.-G. Choung, The characterization of caffeine and nine individual catechins in the leaves of green tea (*Camellia sinensis* L.) by near-infrared reflectance spectroscopy., *Food Chem.* 158 (2014) 351–357. doi:10.1016/j.foodchem.2014.02.127.
51. S.-M. Tan, R.-M. Luo, Y.-P. Zhou, H. Gong, Z. Tan, Rapid and non-destructive discrimination of tea varieties by near infrared diffuse reflection spectroscopy

coupled with classification and regression trees, *African J. Biotechnol.* 11 (2012) 2303–2312.

52. X. Li, Y. He, C. Wu, D.-W. Sun, Nondestructive measurement and fingerprint analysis of soluble solid content of tea soft drink based on Vis/NIR spectroscopy, *J. Food Eng.* 82 (2007) 316–323. doi:<https://doi.org/10.1016/j.jfoodeng.2007.02.042>.
53. X. Li, L. Luo, Y. He, N. Xu, Determination of dry matter content of tea by near and middle infrared spectroscopy coupled with wavelet-based data mining algorithms, *Comput. Electron. Agric.* 98 (2013) 46–53. doi:<https://doi.org/10.1016/j.compag.2013.07.014>.

## Chapter 4 : Application of a wavelet and variance decomposition-based method for compressing close-range hyperspectral images to support process applications

Puneet Mishra<sup>1\*</sup>, Alison Nordon<sup>1\*</sup>, Azam Karami<sup>2</sup>, Douglas N. Rutledge<sup>3</sup>, Jean-Michel Roger<sup>4</sup>

<sup>1</sup>*WestCHEM, Department of Pure and Applied Chemistry and Centre for Process Analytics and Control Technology, University of Strathclyde, 295 Cathedral Street, Glasgow, G1 1XL, United Kingdom*

<sup>2</sup>*Faculty of Physics, Shahid Bahonar University of Kerman, 7616-914111, Kerman, Iran*

<sup>3</sup>*UMR Ingénierie Procédés Aliments, AgroParisTech, Inra, Université Paris-Saclay, F-91300 Massy, France*

<sup>4</sup>*ITAP, Irstea, Montpellier SupAgro, University Montpellier, Montpellier, France*

**Paper Published as:** To be submitted to Applied Spectroscopy

### **Contribution:**

- Performed HSI NIR measurements
- Conducted data analysis
- Wrote the first draft of article
- Corrected article based on comments of co-authors

### 4.1 Abstract

A two-level data decomposition scheme is applied for compressing the hyperspectral imaging (HSI) data for rapid transfer and easy storage to support process applications. At the first level, the decomposition of the spatial information was performed with 2D-

Discrete Wavelet Transform (2D-DWT). At the second stage, two different variance-based data decomposition methods (i.e. principal components analysis (PCA) and Tucker) were applied separately to decompose the spectral information. The potential of the methods was judged by visual inspection of reconstructed data cubes and spectral profiles, by spatial and spectral correlation, and by multi-class support vector machine (SVM) classification performance. The results showed that both methods gave a significant reduction in the size of data cubes and provided a comparable reconstruction of the original data by providing very high spatial and spectral correlation. Furthermore, by eliminating some unwanted variance, the reconstructed data cubes resulted in the reduction of noise in initially noisy bands. The de-noising step also increased the classification accuracy of the SVM model and further improved the classification maps. The results showed that based on 2D-DWT and variance decomposition techniques, the HSI data cubes could be decomposed and reconstructed for easy transfer and storage.

**Keywords:** data reduction, storage, imaging spectroscopy, hypercube, multivariate data analysis

## 4.2 Introduction

In recent years, close-range hyperspectral imaging (HSI) has become a potential non-destructive tool for rapidly exploring spatially resolved light absorbing and scattering properties of different materials. Based on the chemical properties of the material, HSI can be exploited across different spectral modalities such as ultraviolet-visible (UV-VIS), near-infrared (NIR), mid-infrared (MIR), Raman, fluorescence and terahertz (THz) [1] ranges. Applications of HSI are popular in different domains such as food science [2] [3], forensics [4] [5], plant science [6] [7], microbiology [8] [9], medicine [10] [11] and pharmaceutical science [12] [13] [14]. Most of the applications of HSI presented in the literature are limited to the laboratory scale, where the first feasibility studies were performed, and no subsequent integration of the methodologies for real-world process and industrial applications was reported. One of the main factors limiting the application of HSI for real world process and industrial applications is the huge volume of data generated during the experiments which is complex to manage.



Typical hyperspectral cameras recorded binary images with the size directly proportions to spatial and spectral dimensions.

HSI combines two sensor modalities - spectroscopy and imaging. The data generated by HSI is typically in the form of data cubes of size  $n \times p \times q$ , where the first two dimensions ( $n \times p$ ) represent the spatial information obtained with the imaging modality of HSI. The third dimension ( $q$ ) represents the spectral information obtained with the spectroscopic modality of HSI. The main advantage of HSI is the complementary information generated with the imaging and spectroscopic modalities, allowing to spatially resolve spectral properties of materials. However, the data generated with HSI is huge as it contains redundant and correlated information in both the spatial and spectral domains [15]. Depending on the spatial and spectral resolution of the camera used, the typical size of the data array can range from a few hundred megabytes to gigabytes. The difficulty in handling such a large amount of data has already been highlighted in [16] [17], where it was emphasised that to extract information from such data is computationally expensive, complicating the development of fast and efficient applications of HSI.

Process applications of HSI can range from simple monitoring of fruits and vegetables for quality control [18] [19] to process analytical technology (PAT) for monitoring pharmaceutical tablet manufacturing [12] [13]. Typically, in process applications, the data is acquired continuously to gain insight into the process for efficient control of the process. Therefore, when data are acquired with HSI, the data cubes are generated continuously throughout the process and so keep on accumulating. There are two ways to deal with this type of real-time HSI data: firstly, if the computer server system is present at the experimental site, the data can be processed in real-time and results displayed; secondly, if HSI is employed remotely, the acquired data needs to be transferred rapidly to the server where the processing is performed. In both the cases, the later storage of data is also an important concern. Sometimes the storage of the data is for future reference and recall, and sometimes it is because of strict data protection requirements of industries such as for pharmaceutical manufacturing.

One major benefit of the redundant/correlated information present in the HSI is the

possibility to obtain compressed representations of the original data. Since HSI is the combination of two different modalities, i.e., imaging and spectroscopy, redundant information can be found in both the spatial and spectral domains. Therefore, to perform this compression, both the spatial and spectral information can be utilised. In the first step the spatial information can be pre-conditioned with the help of wavelet transforms as performed in classical image compression methods [20]. The spectral or spatio-spectral information can then be compressed by identifying the latent subspaces, with methods based on maximising a criterion such as the variance is of particular interest. This is because a small number of important variance-based subspaces can be retained in the datasets and can later be used to reconstruct the original data. Selection of such a small number of important subspaces also allows elimination of the subspace that does not contribute significantly to explaining the data cube. Subspace elimination can also reduce the amount of noise present in the reconstructed data cubes, improving the quality of the dataset for future data analysis using such as regression or classification. Principal component analysis (PCA), a variance decomposition technique, has been of wide interest for exploration of HSI data, and some applications of PCA for visualising big HSI data sets have been reported in [16][17][21]. Also, there are application where PCA has been implemented to process continuous streams of the datasets and later using the set of scores and loading vectors to reconstruct the data [22]. However, PCA performs the decomposition of the spectral domain once the 3D cubes have been unfolded into 2D matrices, and do not consider the spatial correlation. There are different advantages associated with processing data with unfolded analysis such as it can make the exploration of data easy, however, it comes with the cost of throwing away the information presented in 3D. Further, unfolded models are complex, sometimes difficult due to interpret directly as the third dimension is lost and might accompany with the risk of fitting more noise to the structural model leading to poor predictive power [23]. However, Tucker, an extension of PCA for trilinear data decomposition, deals directly with the 3D cube of data generated by HSI [24]. Tucker performs simultaneous decomposition of both the spatial and spectral domains and application of Tucker for HS image decomposition and compression can be found in [25]. The spatial and spectral decomposition can be combined to result in a higher quality

compression of HS images [20]. To the best of our knowledge, the performance of the 2D-discrete wavelet transforms (2D-DWT) in combination with variance decomposition methods (PCA and Tucker) have never been presented qualitatively and quantitatively as a means for storage and transfer of close-range HS images.

The present work aims to apply a two-level data decomposition scheme [25] for compressing the HSI data to perform rapid transfer and easy storage. At the first level, the decomposition of spatial information was performed with 2D-Discrete Wavelet Transform (2D-DWT). At the second stage, two different variance-based data decomposition methods (i.e. principal components analysis (PCA) and Tucker) were applied separately to decompose the spectral information. The two methods differ in the way they decompose the data cube; while PCA decomposes the data after first unfolding the 3D data cube to the 2D matrix, Tucker decomposes the 3D data cube directly.

### 4.3 Material and methods

#### 4.3.1 Samples and hyperspectral imaging measurements

Four different types of tea (oolong (Yamamotoyama, California, USA), black, green and white (all Vahdam Teas, New Delhi, India)), all in loose-leaf form, were obtained from a local market (Glasgow, UK). The samples were obtained in airtight sealed packaging and stored at ambient temperature. Each tea sample was transferred to a black plastic circular container (diameter = 3.3 cm and depth = 1.3 cm) for analysis. Imaging was performed using a push-broom line scan HSI camera (Model name: RED EYE 1.7) from INNO-SPEC (Nurnberg, Germany). The camera has an InGaAs sensor and generates a spatial map of 320 x 256 pixels in the spectral range of 950 - 1700 nm. The pixel size is 30 x 30  $\mu\text{m}^2$  and the spectral resolution is 3.2 nm. The NIR camera was controlled with a computer via a gigabit Ethernet connection. Lighting was provided by two 50 W halogen light sources and an integration time of 300 ms was used. Image acquisition was performed by placing the samples over the translation stage controlled by an independent stage motor connected to the computer system (Zolix TSA 200 BF). The speed of the translation stage was optimised before image analysis to avoid any distortion in the shape of the image arising from the overlap of

the spectral and spatial information. Image acquisition and management of settings (integration time) for the NIR camera were performed using a software interface provided by the camera manufacturer. An image was acquired of the four different tea samples placed adjacent to each other in their respective containers in the field of view of the camera. Each image comprised more than 2000 pixels (spectra) for the individual tea samples. Prior to analysis, images were radiometrically calibrated using white (Spectralon diffuse reflectance standard) and dark references. The correction was performed for every pixel in the HS image according to equation 4.1:

$$I_R(i,j,k) = \frac{I_{raw}(i,j,k) - I_{dark}(i,j,k)}{I_{white}(i,j,k) - I_{dark}(i,j,k)} \quad (4.1)$$

where  $I_R$  is the calibrated reflectance,  $I_{raw}$  is the raw intensity measured from the test sample,  $I_{dark}$  the intensity of the dark response,  $I_{white}$  is the intensity of the uniform white reference, and  $i$  and  $j$  are spatial coordinates and  $k$  is the wavelength in the image.

#### 4.3.2 Spatial decomposition

##### 4.3.2.1 2D discrete wavelet transform

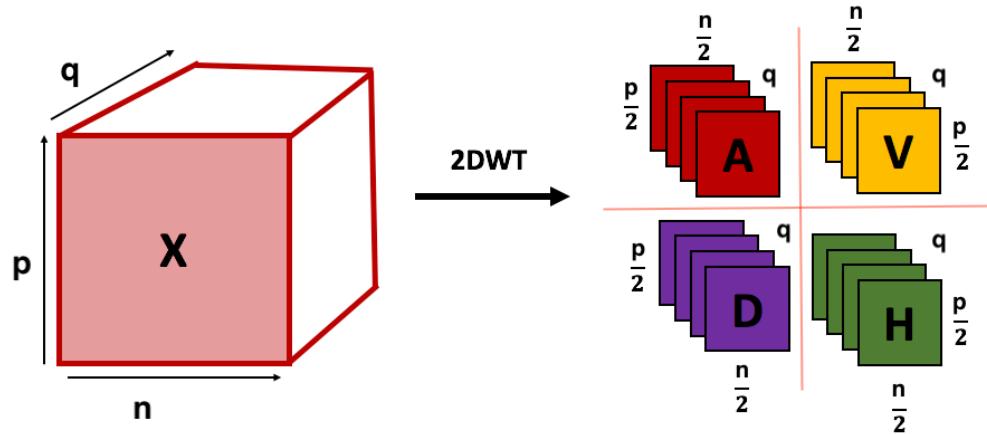


Figure 4.1: Scheme for 2D-DWT decomposition of a hypercube  $i$  to high and low pass filtered components.

The DWT provides an efficient decomposition of signals into low resolution components and detailed information. Many successful applications of the DWT can be found in the image processing domain such as for noise reduction, edge detection and compression [26]. From a deterministic point of view, the DWT can be understood as successive low and high pass filtering of the discrete time-domain signal. At each level, the high-pass filter produces detailed information (horizontal (H), vertical(V) and diagonal (D)), while the low-pass filter associated with the scaling function produces coarse approximations (A). In this work, the 2D-DWT was performed on each band of the HS images, as presented in [25]. A scheme depicting the decomposition is shown in Figure 4.1, where each image plane with  $n$  rows and  $p$  columns, was decomposed into four sub-band images (A, H, V and D), each with  $\frac{n}{2}$  rows and  $\frac{p}{2}$  columns.

In the present work, a single level 2D-DWT was performed and later for retaining the important information in the four sub-images, variance-based data decomposition was employed. In the present compression scheme, a single level 2D-DWT was performed and all the decomposed sub-images (H, D, V and A) were retained. The wavelet filter used was the ‘biorthogonal’ from the Wavelet Toolbox in Matlab (R2016b, (Mathworks Inc., Natick, USA). Biorthogonal wavelets are the standard wavelets used in image compression application and are also used in JPEG2000 compression.

#### 4.3.3 Spectral decomposition

For retention of the important details in the four sub-images, variance-based data decomposition was employed and more than 99.5% of variance was retained. The four data cubes generated from the 2D-DWT were decomposed using two different variance-based decomposition methods. The first one involved performing PCA on the unfolded data cubes and the other was by direct decomposition of the 3D cubes using Tucker. The PCA and Tucker decomposition methods are explained further in the following sections.

### 4.3.3.1 Principal component analysis

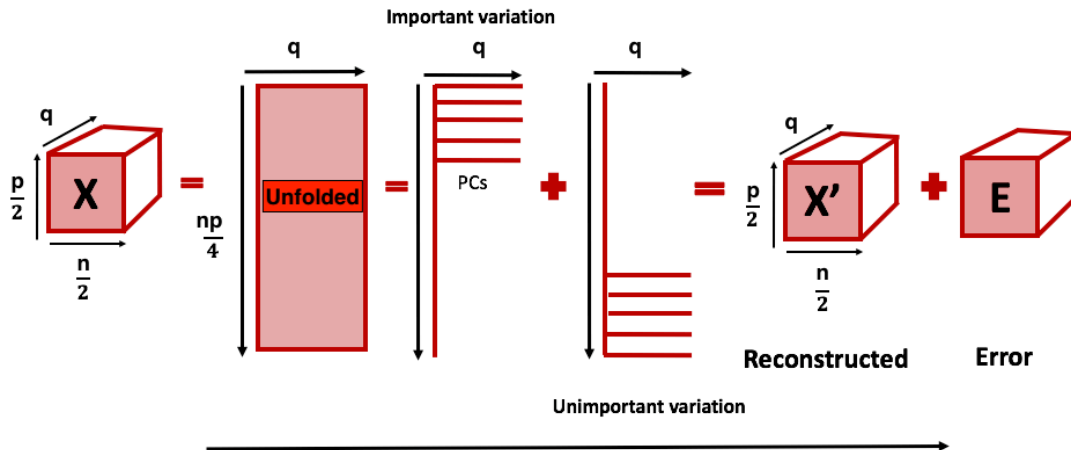


Figure 4.2: The scheme for performing PCA decomposition and reconstruction for 3D hypercube dataset generated with HSI.

PCA, as introduced by Pearson in 1901, is a linear variance decomposition method and has been widely employed for identifying the variance maximising orthogonal subspaces for a range of multivariate data generated in spectroscopy [27]. The scheme for performing PCA on the 3D data cubes generated from HSI is given in Figure 4.2. In the present case, the 3D cube to be decomposed is a sub-image obtained from the 2D-DWT. For example, to perform PCA decomposition on the sub-image formed from the horizontal details ( $n/2 \times p/2 \times q$ ), the corresponding 3D cube must first be unfolded into a 2D matrix ( $np/4 \times q$ ). PCA decomposes the matrix into two sub-matrices of scores and loading. PCA decomposition performs a linear transformation of the original variables to give uncorrelated orthogonal variables called principal components (PCs). The PCA transformation is performed to retain the maximum amount of variability in the dataset. Furthermore, the first PCs retain the maximum amount of variability, and the subsequent PCs contain the highest amount of remaining variability while being orthogonal to the previous PCs. A small number of PCs can, therefore, be retained to keep the useful details and the rest can be discarded to remove the unwanted variability in the dataset. The PCs that are retained can be used to reconstruct a less noisy approximation of the original data matrix. Finally, the 2D

matrix can be reshaped into the original 3D data cube.

The PCA decomposition model for a given observation data matrix  $H$  (horizontal details unfolded cube) is given by 4.2:

$$H = TP^T \quad \dots(4.2)$$

where  $T$  is the scores matrix for a given number of principal components and  $P$  is a  $q \times q$  loadings matrix whose columns are the eigenvectors of  $H^T H$ .

For successful data reconstruction, the maximum amount of significant variation present in the dataset should be preserved. PCA from a data reconstruction perspective can be understood as minimising the square reconstruction error as given in equation 4.3:

$$\min ||TP^T - T_r P_r^T||^2 \quad \dots(4.3)$$

where  $TP$  and  $T_r P_r$  are the original and reconstructed datasets, respectively. The minimisation of the reconstruction error results in maximisation of the information that was present in the data reconstructed by the significant PCs. In the present work, PCA was employed for retaining the details in the four sub-images obtained from the 2D-DWT. PCA was performed separately on the unfolded matrices of all four sub-images. To reconstruct the sub-images, the scores and loadings matrices were multiplied. Furthermore, the inverse 2D-DWT (i2D-DWT) was used to reconstruct an approximation of the original HS image from the sub-images. PCA decomposition was performed in Matlab using the PLS Toolbox (version 8.11, Eigenvector Research Inc., USA).

### 4.3.3.2 Tucker decomposition

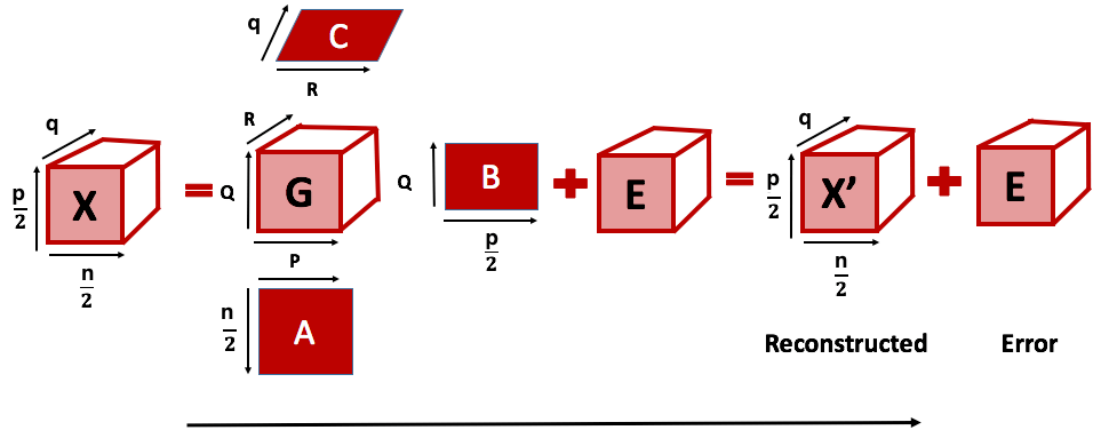


Figure 4.3: The scheme for performing Tucker decomposition and reconstruction of the 3D hypercube dataset generated with HSI.

Tucker is a popular data analysis tool, widely used for multi-way analysis in chemometrics for the analysis of multi-way spectral data [28]. Tucker can be understood as a generalisation of PCA to higher-order data. The scheme for Tucker decomposition and reconstruction of a 3D cube, which in this case is a sub-image obtained from the 2D-DWT, is given in Figure 4.3. It can be seen that Tucker decomposes the 3D data cube of a sub-image, e.g., the horizontal detailed sub-image, H, into a core tensor multiplied by a matrix along each of the 3 modes. Thus, the decomposition for a 3D cube  $H \in \mathbb{R}^{n/2 \times p/2 \times q}$  is given by equation (4.4):

$$H \approx G \times_1 A \times_2 B \times_3 C = \sum_{x=1}^P \sum_{y=1}^Q \sum_{z=1}^R g_{xyz} a_x \circ b_y \circ c_z \quad \dots(4.4)$$

Here,  $A \in \mathbb{R}^{n/2 \times P}$ ,  $B \in \mathbb{R}^{P \times Q}$ , and  $C \in \mathbb{R}^{Q \times R}$  are the factor matrices and can be understood as the principal components in each mode.  $\times_1$ ,  $\times_2$  and  $\times_3$  defined the outer product in three different modes of the HS data. The tensor  $G \in \mathbb{R}^{P \times Q \times R}$  is called the core tensor and its entries show the level of interaction between the different components.

Elementwise, the Tucker decomposition can be understood from equation (4.5):



$$x_{ijk} \approx \sum_{x=1}^P \sum_{y=1}^Q \sum_{z=1}^R g_{xyz} a_{ix} b_{jy} c_{kz},$$

for  $i = 1, \dots, n/2, \quad j=1, \dots, p/2, \quad k=1, \dots, q \quad \dots(4.5)$

where P, Q and R are the number of components in the factor matrices A, B, and C, respectively. P, Q and R are smaller than  $n/2$ ,  $p/2$  and  $q$ , and, therefore, the core tensor, G, can be thought of as a compressed version of H. Typically, the decomposed version of the tensor is significantly smaller than the original tensor. In this case, the 3D sub-images obtained from the 2D-DWT were decomposed individually into a core array and factor matrix to support storage. Furthermore, to reconstruct the sub-images, the core array can be multiplied by the factor matrices in each mode. The i2D-DWT can be used to reconstruct an approximation of the original HS image from the sub-images. Tucker decomposition was performed in Matlab utilising PLS Toolbox.

#### 4.3.4 Spatial and spectral Correlation

To comment on the quality of reconstructed data cubes, spectral and spatial correlations were measured. Both the spatial and spectral correlations were calculated with respect to the original data cube. The spatial and spectral correlations were estimated using the *corr2* and *corr* functions in Matlab respectively.

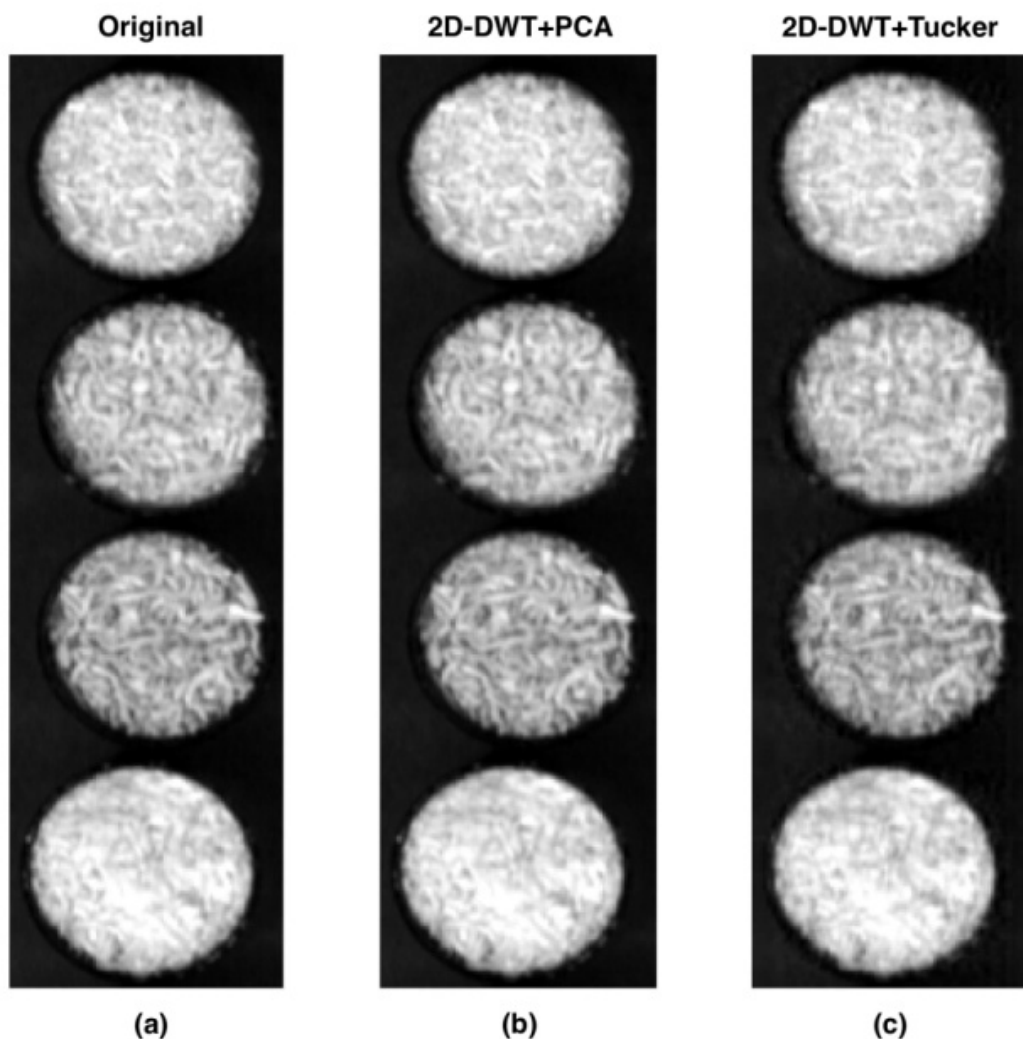
#### 4.3.5 Multi-class Support Vector Machine

To test the data reconstruction efficacy of PCA and Tucker, classification was performed with a multi-class support-vector machine (SVM) classifier. Three different SVM classifiers were developed corresponding to the original data, data reconstructed using the 2D-DWT and PCA, and data reconstructed using the 2D-DWT and Tucker. SVM utilises hyperplanes to define the decision boundaries to perform classification. Furthermore, to deal with the data complexity, the SVM performs high dimensional mapping of the data using kernel functions. Mapping to a higher dimension is usually performed to select the hyperplane that provides the best separation of the classes. As there are four different classes corresponding to four different tea products in this case, the traditional SVM binary classifier combined with the error correcting output code (ECOC) ensemble method was used. The ECOC deals with the multi-class problem by developing several independent binary classification models. The SVM along with

ECOC was implemented in Matlab using the Statistics and Machine Learning Toolbox (R2016b). The coding design used was one-vs.-all where one class was assigned as positive and all other classes were assigned negative. In total, the coding design utilises all combinations of class pair assignments with  $K$  (number of classes) number of learners. High dimensional mapping of the data was performed with a gaussian kernel function. The spectra from the images for the four different classes were selected in a supervised way using the '*roipoly*' function in Matlab. The '*roipoly*' function provides a graphical interface to manually extract the information from the hypercube. For each individual tea product, spectra were extracted from 200 pixels, which were selected at random from the image collected, leading to a total of 800 spectra for calibration of the classification model. Validation of the model was performed with a 10-fold cross-validation method. This was repeated 100 times, and the average prediction accuracy was recorded.

## 4.4 Results

### 4.4.1 Image planes



*Figure 4.4: Greyscale images produced using the image plane at 1267 nm for the original and reconstructed data. (a) Original data (approx. 40 MB), (b) data reconstructed using the 2D-DWT and PCA (approx. 5 MB), and (c) data reconstructed using the 2D-DWT and Tucker (approx. 1 MB).*

Greyscale images produced using the image plane corresponding to 1267 nm for the original, data reconstructed using the 2D-DWT and PCA, and data reconstructed using the 2D-DWT and Tucker are presented in Figure 4.4; 1267 nm was selected merely to allow visualisation of the data hypercubes. The four circular portions in the image represent the four different tea samples. It can be seen in Figure 4.4 that reconstruction

of the data with PCA and Tucker provided image planes of comparable quality to those in the original dataset. However, the size of the data was significantly smaller for the reconstructed data cubes (5 and 1 MB for PCA and Tucker, respectively) compared to the original data cube (40 MB). Furthermore, to understand the reconstruction quality across the entire spectral range of the camera, the spatial correlation for the data reconstructed using the 2D-DWT and PCA, and the 2D-DWT and Tucker was calculated with respect to the original dataset (see Figure 4.5).

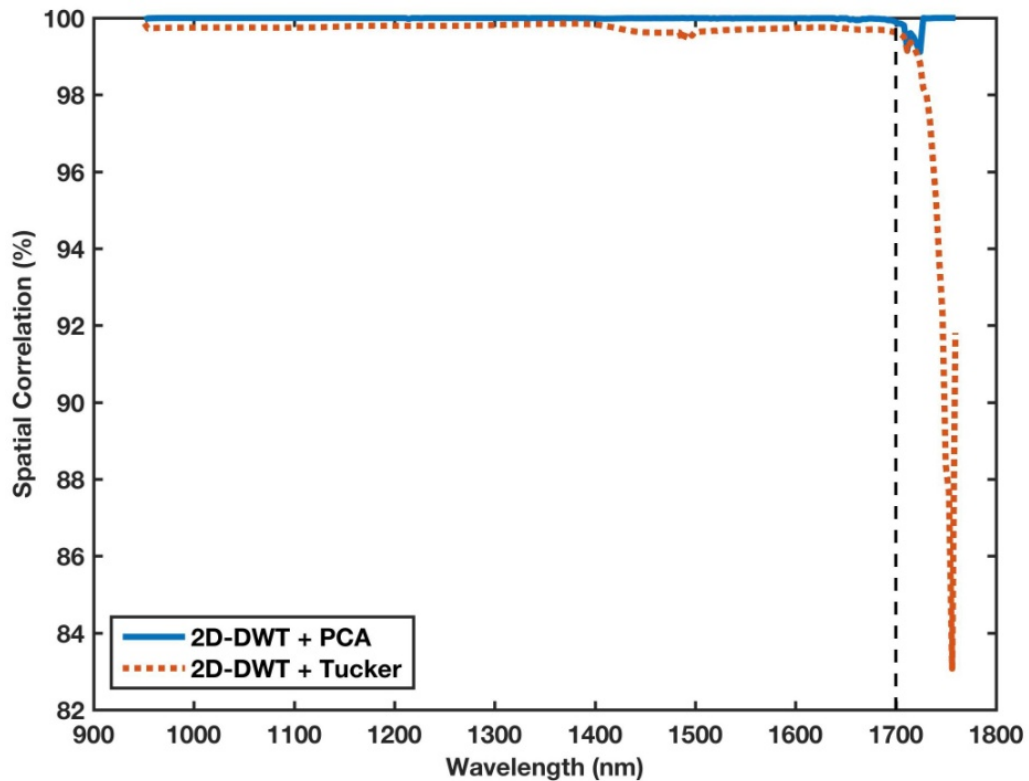
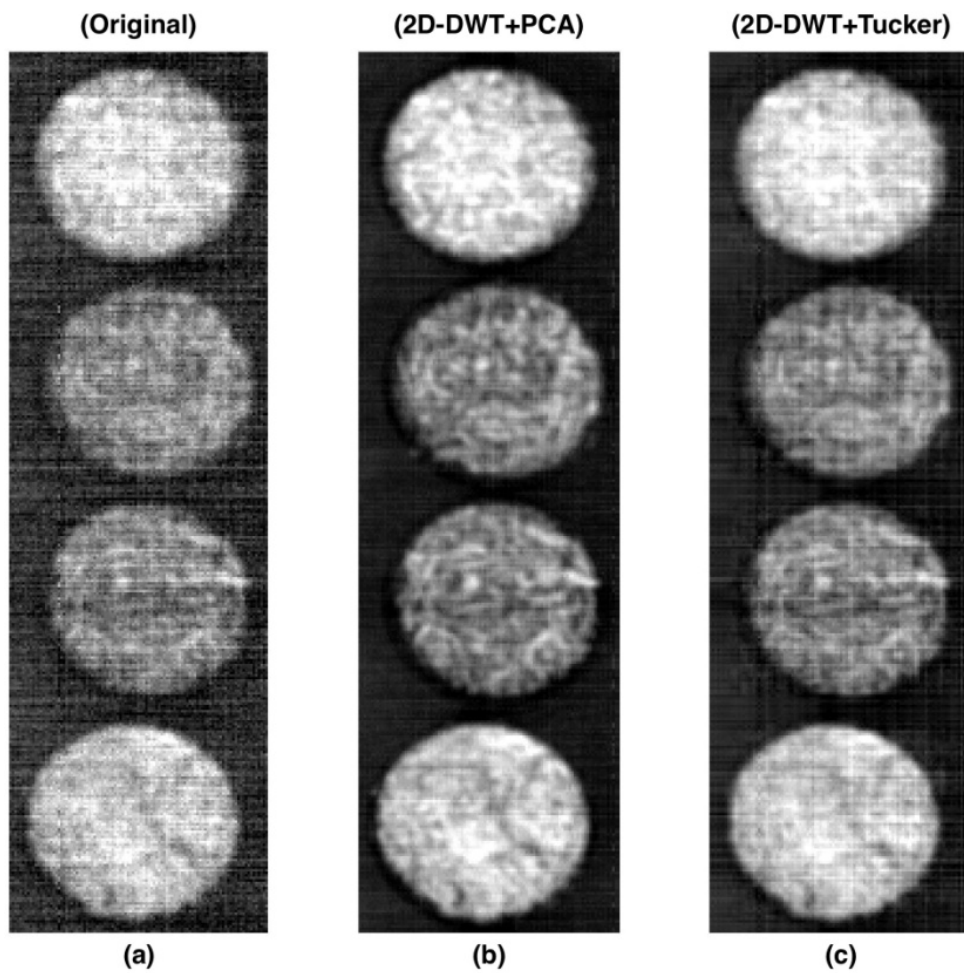


Figure 4.5: Spatial correlation across the entire spectral range for data reconstructed using the 2D-DWT and PCA (blue) and the 2D-DWT and Tucker (red).

It can be seen from Figure 4.5 that the spatial correlation was almost 100% for data reconstructed using both the 2D-DWT and PCA, and the 2D-DWT and Tucker. However, a slightly higher spatial correlation was obtained for the 2D-DWT and PCA method compared to that obtained for the 2D-DWT and Tucker method. Furthermore, at wavelengths greater than 1700 nm, the spatial correlation for data reconstructed using the 2D-DWT and PCA, and the 2D-DWT + Tucker decreases. This can be

attributed to the removal of the noise from the data set through limiting the details in the sub-band images (obtained after the 2D-DWT) during PCA and Tucker decomposition of the data. Evidence for this can be seen in Figure 4.6, where greyscale images produced using the image plane at 1736 nm are presented for the original data, data reconstructed using the 2D-DWT and PCA, and data reconstructed using the 2D-DWT and Tucker. It can be seen in Figure 4.6 that the greyscale image produced from the original data are very noisy at 1736 nm; this wavelength was chosen to allow visualisation of a noisy region in the data hypercube. However, the images produced from the reconstructed data using the 2D-DWT and PCA, and the 2D-DWT and Tucker are clearer. The spatial correlation for the data reconstructed using the 2D-DWT and Tucker decreased more compared to that for the 2D-DWT and PCA. This indicates that a higher amount of noise was removed utilising the 2D-DWT and Tucker decomposition and reconstruction and it is because of this that the spatial correlation decreases.



*Figure 4.6: Greyscale images produced from the image plane at 1736 nm for the original and reconstructed data. (a) Original data, (b) data reconstructed using the 2D-DWT and PCA, and (c) data reconstructed using the 2D-DWT and Tucker.*

#### 4.4.2 Spectral profiles

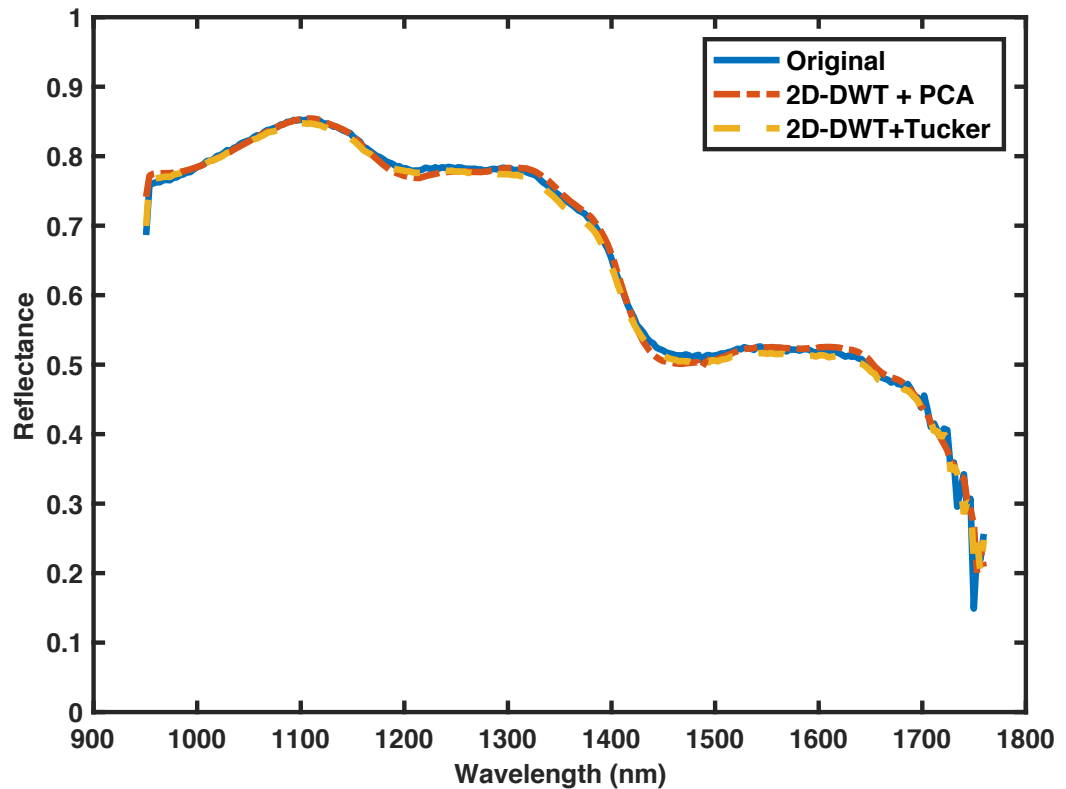


Figure 4.7: Spectral profiles corresponding to the pixel with  $x$  and  $y$  co-ordinates of 50 and 300, respectively, obtained from the original data (blue), data reconstructed using the 2D-DWT and PCA (red), and data reconstructed using the 2D-DWT and Tucker (yellow).

Figure 4.7 presents the spectra for the pixel with  $x$  and  $y$  co-ordinates of 50 and 300, respectively, extracted from the original data (blue), data reconstructed using the 2D-DWT and PCA (red), and data reconstructed using the 2D-DWT + Tucker (orange). It can be seen that the profiles from the reconstructed data are similar to those extracted from the original dataset. However, at wavelengths greater than 1700 nm, there is a slight difference in the spectral profiles from the reconstructed and original data. This is because the sensitivity of the camera drops significantly above 1700 nm and this results in a noisy spectral profile. However, when reconstructing the data with the 2D-DWT and PCA and the 2D-DWT and Tucker, this noise was reduced (as previously

shown in Figure 4.6) resulting in a difference in the spectral profile above 1700 nm.

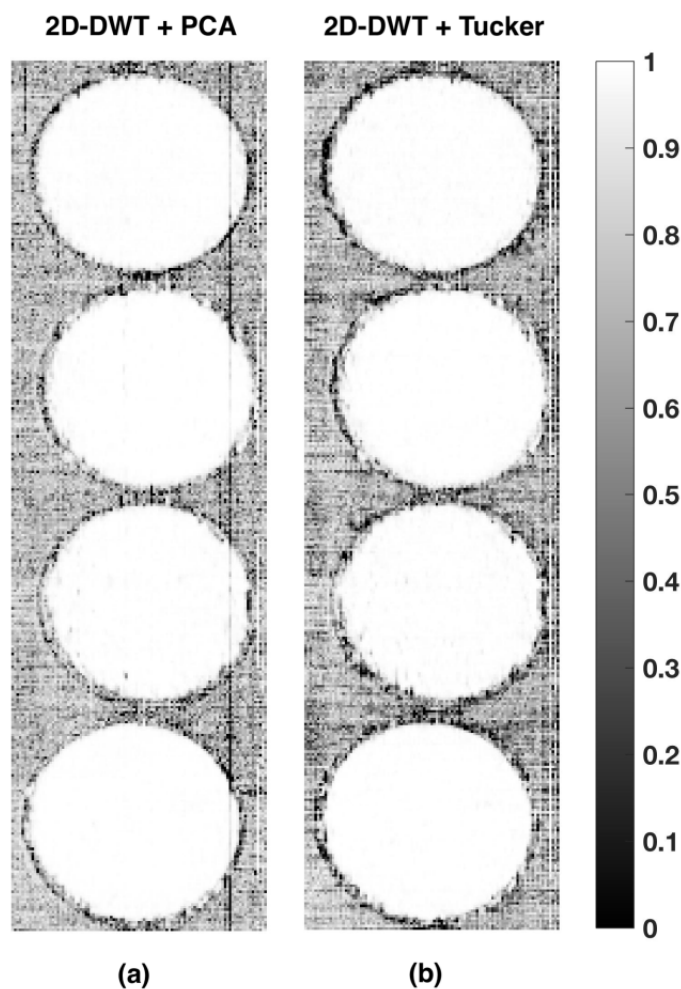


Figure 4.8: Spectral correlation calculated for each pixel of the reconstructed data with respect to the original data. (a) Data reconstructed using the 2D-DWT and PCA and (b) data reconstructed using the 2D-DWT and Tucker.

To estimate the quality of the spectral profile for the complete image, the spectral correlation was calculated between each pixel in the reconstructed data and the same pixel in the original dataset. The spectral correlation plot for the complete image is presented in Figure 4.8. It can be seen that the area representing the four tea samples gave very high spectral correlations, while the edges and background gave slightly lower spectral correlation. However, in this case, the important part of the image is



that representing the four tea samples. The background used was just a black board leading to a very low intensity spectral signature, and therefore, the background was removed prior to further analysis of the image.

#### 4.4.3 Classification

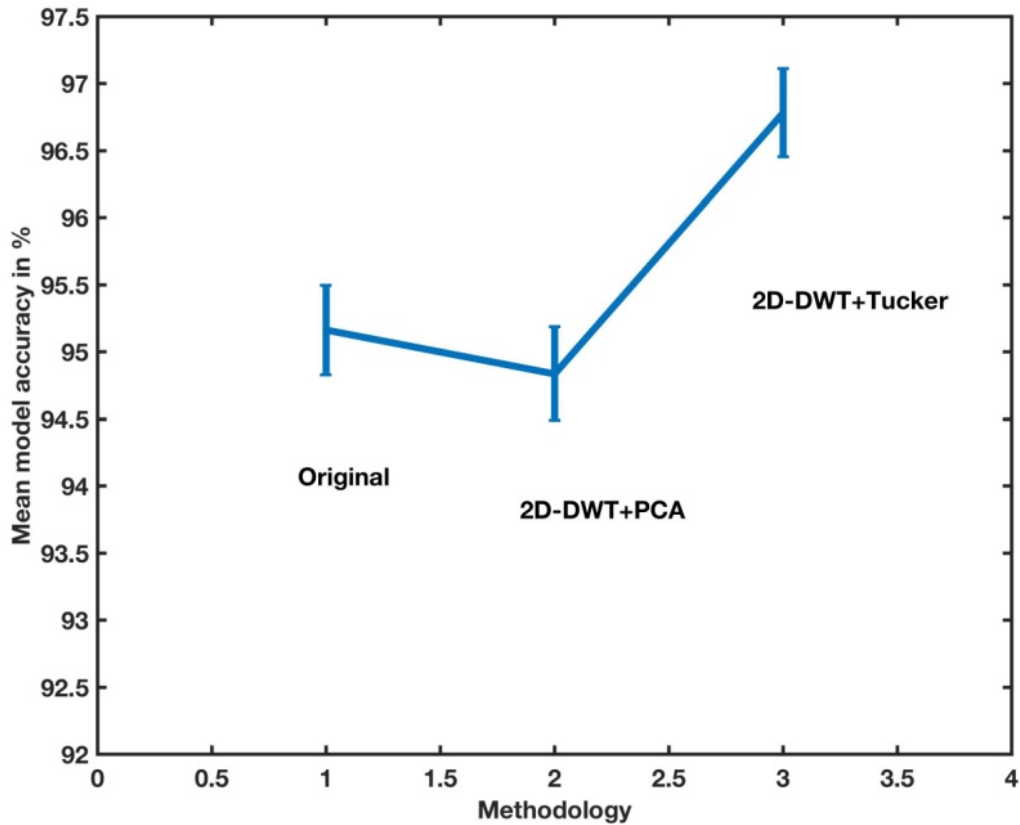


Figure 4.9: Classification accuracy of the original data, the data reconstructed using the 2D-DWT and PCA and the data reconstructed using the 2D-DWT and Tucker.

The classification accuracy of the multi-class SVM model is presented in Figure 4.9. It can be seen that the classification accuracy was highest for the model built using data reconstructed using the 2D-DWT and Tucker, followed by the original data then data reconstructed using the 2D-DWT and PCA. This increase in the classification accuracy of the model built using data reconstructed using the Tucker method compared to that built using the original data is due to the removal of noise from the dataset during the decomposition and reconstruction steps. The superior performance

of models built from data reconstructed using the Tucker method, compared to the PCA method, is due to differences in the way that Tucker and PCA decompose the dataset. In the Tucker method, the spatial dimensions along with the spectral dimension are used to perform the decomposition whereas in the PCA method only the spectral dimension is decomposed. However, the classification accuracy obtained with the model built using data reconstructed using the 2D-DWT and PCA was slightly lower compared to those built using the original data and the Tucker reconstructed dataset. A reason for this could be due to discarding of a very small amount of important variability present in the data while retaining a limited number of components from the PCA.

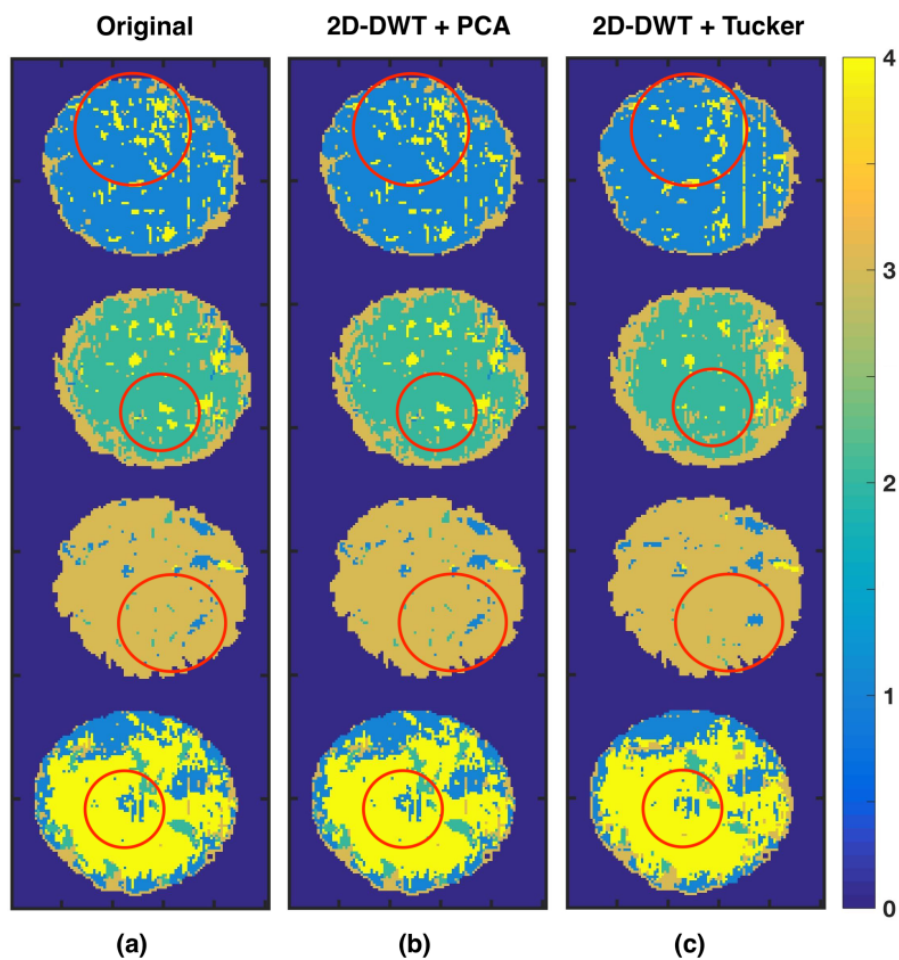


Figure 4.10: Classification maps for the original data, data reconstructed using PCA, and data reconstructed using Tucker. The red circles highlight example regions where

reconstruction using Tucker gives fewer misclassifications between the classes compared to the original data and reconstruction using PCA.

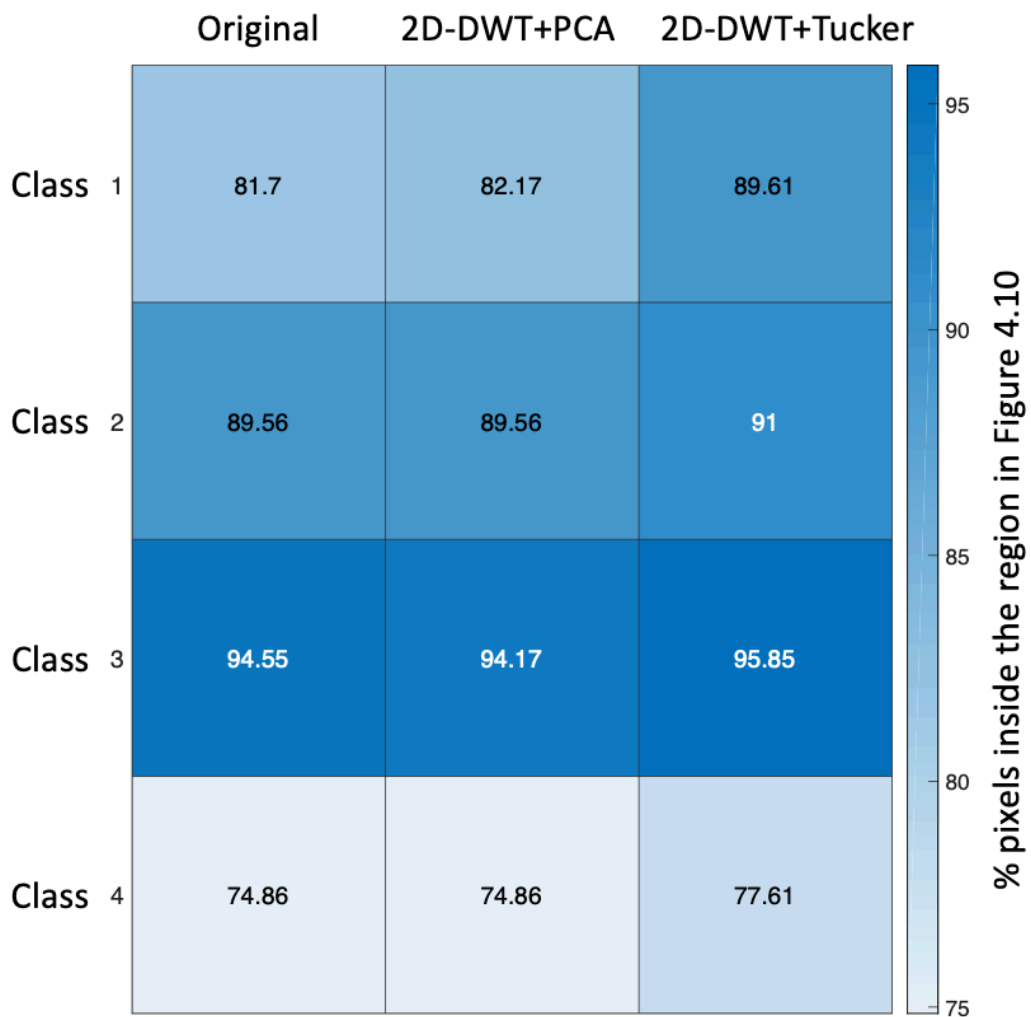


Figure 4.11: Percent of pixels belonging to each class in the classification maps (see Figure 4.10) using the original data, data reconstructed using the 2D-DWT and PCA, and data reconstructed using the 2D-DWT and Tucker.

The classification maps obtained for the multi-class SVM are presented in Figure 4.10. It can be seen that the quality of the classification maps obtained from data reconstructed using the 2D-DWT and PCA, and the 2D-DWT and Tucker are similar to those obtained using the original dataset. Furthermore, the Tucker reconstruction method seems to enhance the classification maps by reducing the number of misclassifications between the classes, as illustrated in the circled regions in Figure

4.10. The percent of pixels identified inside the markers (red circular regions) in Figure 4.10 are presented in Figure 4.10. In Figure 4.11, it can be seen that 2D-DWT+Tucker based decomposition and reconstruction provided a clear improvement for modelling each class.

#### 4.5 Conclusions

Process applications of HSI can be characterised by the acquisition and accumulation of large amounts of HSI data. The data generated needs to be transferred as well as stored for efficient management of the information generated. Due to the redundant and correlated information present in both the spatial and spectral domains, HSI data can be decomposed in both domains prior to transfer and storage and can be reconstructed later for data analysis. The two-step compression scheme based on the 2D-DWT and variance decomposition methods presented in this work showed that the reconstructed data were comparable to the original data but were much smaller in size. A further advantage of variance-based decomposition methods in combination with the 2D-DWT is that important details in the data can be represented by a small number of latent spaces leading to reduced noise in the reconstructed data. In the present work, the decomposition of the variance of the 2D-DWT sub images was faster with PCA compared to with Tucker. However, the quality of the reconstructed images was higher with Tucker owing to the simultaneous decomposition of both the spatial and spectral domains. Variance-based decomposition requires selection of the number of latent variables to be retained; the number retained affects the compression rate and the quality of the spectral image. Future work will involve implementation of these methodologies for rapid transfer and storage of remotely acquired HSI data of the processing of tea, which is being collected for process understanding and monitoring.

## 4.6 References

1. J.M. Amigo, H. Babamoradi, S. Elcoroaristizabal, “Hyperspectral image analysis. A tutorial”, *Anal. Chim. Acta.* 896 (2015) 34–51. doi:<https://doi.org/10.1016/j.aca.2015.09.030>.
2. L. Zhang, Y. Peng, “Noninvasive Qualitative and Quantitative Assessment of Spoilage Attributes of Chilled Pork Using Hyperspectral Scattering Technique”, *Appl. Spectrosc.* 70 (2016) 1309–1320. doi:[10.1177/0003702816654060](https://doi.org/10.1177/0003702816654060).
3. P. Mishra, C.B.Y. Cordella, D.N. Rutledge, P. Barreiro, J.M. Roger, B. Diezma, “Application of independent components analysis with the JADE algorithm and NIR hyperspectral imaging for revealing food adulteration”, *J. Food Eng.* 168 (2016) 7–15.
4. C.S. Silva, M.F. Pimentel, J.M. Amigo, R.S. Honorato, C. Pasquini, “Detecting semen stains on fabrics using near infrared hyperspectral images and multivariate models”, *TrAC Trends Anal. Chem.* 95 (2017) 23–35. doi:<https://doi.org/10.1016/j.trac.2017.07.026>.
5. F. Zapata, F.E. Ortega-Ojeda, C. García-Ruiz, “Revealing the location of semen, vaginal fluid and urine in stained evidence through near infrared chemical imaging”, *Talanta.* 166 (2017) 292–299. doi:<https://doi.org/10.1016/j.talanta.2017.01.086>.
6. P. Mishra, M.S.M. Asaari, A. Herrero-Langreo, S. Lohumi, B. Diezma, P. Scheunders, “Close range hyperspectral imaging of plants: A review”, *Biosyst. Eng.* 164 (2017) 49–67. doi:<https://doi.org/10.1016/j.biosystemseng.2017.09.009>.
7. S. Higa, H. Kobori, S. Tsuchikawa, “Mapping of Leaf Water Content Using Near-Infrared Hyperspectral Imaging”, *Appl. Spectrosc.* 67 (2013) 1302–1307.

doi:10.1366/13-07028.

8. A.A. Gowen, Y. Feng, E. Gaston, V. Valdramidis, “Recent applications of hyperspectral imaging in microbiology”, *Talanta*. 137 (2015) 43–54.  
doi:<https://doi.org/10.1016/j.talanta.2015.01.012>.
9. H.-J. He, D.-W. Sun, “Hyperspectral imaging technology for rapid detection of various microbial contaminants in agricultural and food products”, *Trends in Food Science & Technology* 46 (1) (2015) 99–109.  
doi: <https://doi.org/10.1016/j.tifs.2015.08.001>.
10. Q. Li, M. Zhou, H. Liu, Y. Wang, F. Guo, “Red Blood Cell Count Automation Using Microscopic Hyperspectral Imaging Technology”, *Appl. Spectrosc.* 69 (2015) 1372–1380.  
doi:10.1366/14-07766.
11. Q. Wang, Q. Li, M. Zhou, L. Sun, S. Qiu, Y. Wang, “Melanoma and Melanocyte Identification from Hyperspectral Pathology Images Using Object-Based Multiscale Analysis”, *Appl. Spectrosc.* (2018) 0003702818781352.  
doi:10.1177/0003702818781352.
12. L.M. Kandpal, J. Tewari, N. Gopinathan, P. Boulas, B.-K. Cho, “In-Process Control Assay of Pharmaceutical Microtablets Using Hyperspectral Imaging Coupled with Multivariate Analysis”, *Anal. Chem.* 88 (2016) 11055–11061.  
doi:10.1021/acs.analchem.6b02969.
13. L.M. Kandpal, J. Tewari, N. Gopinathan, J. Stolee, R. Strong, P. Boulas, B.-K. Cho, “Quality assessment of pharmaceutical tablet samples using Fourier transform near infrared spectroscopy and multivariate analysis”, *Infrared Phys. Technol.* 85 (2017) 300–306. doi:<https://doi.org/10.1016/j.infrared.2017.07.016>.

14. P.R. Wahl, I. Pucher, O. Scheibelhofer, M. Kerschhaggl, S. Sacher, J.G. Khinast, “Continuous monitoring of API content, API distribution and crushing strength after tableting via near-infrared chemical imaging”, *Int. J. Pharm.* 518 (2017) 130–137. doi:<https://doi.org/10.1016/j.ijpharm.2016.12.003>.
15. J. Burger, A. Gowen, “Data handling in hyperspectral image analysis”, *Chemom. Intell. Lab. Syst.* 108 (2011) 13–22. doi:<https://doi.org/10.1016/j.chemolab.2011.04.001>.
16. C. Ferrari, G. Foca, A. Ulrici, “Handling large datasets of hyperspectral images: Reducing data size without loss of useful information”, *Anal. Chim. Acta.* 802 (2013) 29–39. doi:<https://doi.org/10.1016/j.aca.2013.10.009>.
17. C. Ferrari, G. Foca, R. Calvini, A. Ulrici, “Fast exploration and classification of large hyperspectral image datasets for early bruise detection on apples”, *Chemom. Intell. Lab. Syst.* 146 (2015) 108–119. doi:<https://doi.org/10.1016/j.chemolab.2015.05.016>.
18. A.A. Gowen, C.P. O’Donnell, P.J. Cullen, G. Downey, J.M. Frias, “Hyperspectral imaging – an emerging process analytical tool for food quality and safety control”, *Trends Food Sci. Technol.* 18 (2007) 590–598. doi:<https://doi.org/10.1016/j.tifs.2007.06.001>.
19. Y.-Y. Pu, Y.-Z. Feng, D.-W. Sun, “Recent Progress of Hyperspectral Imaging on Quality and Safety Inspection of Fruits and Vegetables: A Review”, *Compr. Rev. Food Sci. Food Saf.* 14 (2015) 176–188. doi:10.1111/1541-4337.12123.
20. J.S. Walker, T.Q. Nguyen, “Wavelet-based image compression”, Sub-chapter of CRC Press book : *Transform and Data Compression* (2001).

21. A.A. Gowen, C.P. O'Donnell, M. Taghizadeh, P.J. Cullen, J.M. Frias, G. Downey, "Hyperspectral imaging combined with principal component analysis for bruise damage detection on white mushrooms (*Agaricus bisporus*)", *J. Chemom.* 22 (2008) 259–267. doi:10.1002/cem.1127.
22. R. Vitale, A. Zhyrova, J.F. Fortuna, O.E. de Noord, A. Ferrer, H. Martens, On-The-Fly Processing of continuous high-dimensional data streams, *Chemom. Intell. Lab. Syst.* 161 (2017) 118–129. doi:https://doi.org/10.1016/j.chemolab.2016.11.003.
23. J. Huang, H. Wium, K.B. Qvist, K.H. Esbensen, Multi-way methods in image analysis—relationships and applications, *Chemom. Intell. Lab. Syst.* 66 (2003) 141–158. doi:https://doi.org/10.1016/S0169-7439(03)00030-3.
24. N. Renard, S. Bourennane, "Dimensionality reduction based on tensor modeling for classification methods", *IEEE Trans. Geosci. Remote Sens.* 47 (2009) 1123.
25. A. Karami, M. Yazdi, G. Mercier, "Compression of hyperspectral images using discrete wavelet transform and tucker decomposition", *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 5 (2012) 444–450.
26. M. Stéphane, "CHAPTER 10 – Compression", in: M.B.T.-A.W.T. of S.P. (Third E. Stéphane (Ed.), Academic Press, Boston, 2009: pp. 481–533. doi:https://doi.org/10.1016/B978-0-12-374370-1.00014-8.
27. S. Wold, K. Esbensen, P. Geladi, "Principal component analysis", *Chemom. Intell. Lab. Syst.* 2 (1987) 37–52. doi:https://doi.org/10.1016/0169-7439(87)80084-9.
28. T. Kolda, B. Bader, "Tensor Decompositions and Applications", *SIAM Rev.* 51 (2009) 455–500. doi:10.1137/07070111X.



## Chapter 5 : Automatic de-noising of close-range hyperspectral images with a wavelength-specific shearlet-based image noise reduction method

Puneet Mishra<sup>1\*</sup>, Azam Karami<sup>2</sup>, Alison Nordon<sup>1\*</sup>, Douglas N. Rutledge<sup>3</sup>, Jean-Michel Roger<sup>4</sup>

<sup>1</sup>*WestCHEM, Department of Pure and Applied Chemistry and Centre for Process Analytics and Control Technology, University of Strathclyde, Glasgow, G1 1XL, United Kingdom*

<sup>2</sup>*Faculty of Physics, Shahid Bahonar University of Kerman, 7616-914111, Kerman, Iran*

<sup>3</sup>*UMR Ingénierie Procédés Aliments, AgroParisTech, Inra, Université Paris-Saclay, F-91300 Massy, France*

<sup>4</sup>*ITAP, Irstea, Montpellier SupAgro, University Montpellier, Montpellier, France*

**Paper Published as:** P. Mishra, A. Karami, A. Nordon, D.N. Rutledge, J.-M. Roger, Automatic de-noising of close-range hyperspectral images with a wavelength-specific shearlet-based image noise reduction method, *Sensors Actuators B Chem.* 281 (2019) 1034–1044. doi:<https://doi.org/10.1016/j.snb.2018.11.034>.

### **Contribution:**

- Performed HSI VNIR measurements
- Conducted data analysis
- Wrote the first draft of article
- Corrected article based on comments of co-authors

## 5.1 Abstract

Hyperspectral imaging (HSI) has become an essential tool for exploration of different spatially-resolved properties of materials in analytical chemistry. However, due to various technical factors such as detector sensitivity, choice of light source and experimental conditions, the recorded data contain noise. The presence of noise in the data limits the potential of different data processing tasks such as classification and can even make them ineffective. Therefore, reduction/removal of noise from the data is a useful step to improve the data modelling. In the present work, the potential of a wavelength-specific shearlet-based image noise reduction method was utilised for automatic de-noising of close-range HS images. The shearlet transform is a special type of composite wavelet transform that utilises the shearing properties of the images. The method first utilises the spectral correlation between wavelengths to distinguish between levels of noise present in different image planes of the data cube. Based on the level of noise present, the method adapts the use of the 2-D non-subsampled shearlet transform (NSST) coefficients obtained from each image plane to perform the spatial and spectral de-noising. Furthermore, the method was compared with two commonly used pixel-based spectral de-noising techniques, Savitzky-Golay (SAVGOL) smoothing and median filtering. The methods were compared using simulated data, with Gaussian and Gaussian and spike noise added, and real HSI data. As an application, the methods were tested to determine the efficacy of a visible-near infrared (VNIR) HSI camera to perform non-destructive automatic classification of six commercial tea products. De-noising with the shearlet-based method resulted in a visual improvement in the quality of the noisy image planes and the spectra of simulated and real HSI. The spectral correlation was highest with the shearlet-based method. The peak signal-to-noise ratio (PSNR) obtained using the shearlet-based method was higher than that for SAVGOL smoothing and median filtering. There was a clear improvement in the classification accuracy of the SVM models for both the simulated and real HSI data that had been de-noised using the shearlet-based method. The method presented is a promising technique for automatic de-noising of close-range HS images, especially when the amount of noise present is high and in consecutive wavelengths.

**Keywords:** pre-processing; noise reduction; tea classification; VNIR; non-subsampled shearlet transform

## 5.2 Introduction

Close-range hyperspectral imaging (HSI) and image processing techniques are popular analytical tools in many scientific domains and are used in applications such as the exploration of food properties [1], pharmaceutical product characterisation [2, 3], forensics analysis [4, 5], exploration of plant traits for phenotype studies [6, 7], and microbiology [8]. The major advantage of HSI over other conventional analytical techniques is its non-invasive and non-destructive nature which is further complemented by rapid data acquisition.

HSI combines two sensor modalities that are spectroscopy and imaging, where the spectroscopy provides the chemical information about the samples and the imaging adds a complementary domain of spatial information [9]. The data generated by HSI can be understood as spatial maps of spectral variation arranged in 3-D cubes ( $n \times p \times q$ ). The first two dimensions ( $n \times p$ ) of the cubes are usually the spatial dimensions, and the third dimension ( $q$ ) contains the spectral information. To extract the meaningful information from HS images, different data processing steps such as exploration, regression and classification are often performed. However, before any data processing, as a standard first step, the cubes are usually pre-processed to remove various types of noise from the data so as to increase the signal-to-noise ratio (SNR) [10].

The information generated in HSI is often accompanied by noise, which can arise from detector sensitivity, illumination conditions (e.g. the choice of light source) and experimental conditions (e.g. interference from other light sources). In general term, noise can be understood as unwanted modification a signal may suffer during capture, storage, transmission, processing or conversion. The types of noise in HSI can range from small signal-independent noise such as low-level Gaussian to high-level mixed noise such as Gaussian, Poisson and spike [11]. Since the noise is in the acquired signals, it can be observed in each individual spectrum. However, it can also be observed as pixel-to-pixel intensity variations in each spatial plane. For this reason,

the noise is visible in both the spatial and the spectral domains. The need for methodologies to deal with the noise present in the data cubes generated by close-range HSI has already been highlighted in [12]. A typical approach to deal with noisy signals in the chemometrics field is to remove the affected spectral range from the dataset. This approach can be seen very often when the extreme wavelengths of images are noisy, and the easiest option is to remove that part of the spectrum. However, the downside of this approach is that with the removal of noisy wavelengths, relevant information in the data might also be removed. The other common approach to remove noise from the signal (spectra) while keeping the information is to apply smoothing filters. Two commonly used filter methods are Savitzky-Golay (SAVGOL) smoothing [12, 13,14] and median filtering. SAVGOL smoothing and median filtering can be used alone and independently for each spectrum corresponding to a pixel of the HS image. However, if the level of noise is too high in the spectrum, the use of SAVGOL smoothing and median filtering can become tedious because of the need to determine the optimum window width for smoothing. Also, if noise is present in successive wavelengths, SAVGOL and median filters can result in a deformed spectral profile. The deformation mainly occurs when a large number of noisy wavelengths are present inside the smoothing window, dominating the normal wavelengths. To deal with this, the filter-based methods e.g. median filtering are currently applied after removing the high noise wavelengths [15, 16]. Furthermore, the main drawback of both of these methods is that they can only be used to deal independently with the noise in each spectrum of the HS image and it is not possible to consider the spatial relations between the spectra of the pixels. Without removing the noise from the spatial domain, the scores maps resulting from classification and regression procedures can become noisy (misclassified pixels) leading to inefficient data modelling.

Methods like SAVGOL and median filtering require testing and optimisation of parameters such as the window size, order of derivative etc., which often requires expertise and visualisation skills to decide on efficacy. However, the use of HSI for process analysis, where real-time data processing is required, means that there is a need for automatic de-noising methods. To the best of our knowledge, there is no existing automatic method that deals simultaneously with both the spatial and spectral noise in the data generated with close-range HSI. However, in the field of remote

sensing, the problem of automatic de-noising of HS images is well understood. There are three main families of methods that are used for automatic de-noising of HSI data. The first is the family of methods that utilise the sparse representation of spatial planes such as wavelets but do not consider the spectral noise [17]. The second is the family of methods that combine decorrelation techniques such as principal components analysis (PCA) with the sparse techniques [18]. However, these methods deal with the spatial and spectral noise separately. The third is the family of methods, such as tensor decomposition methods [19], that utilise the spatial and spectral noise together and are based on the 3-D representation of data. However, the major drawback of such tensor approaches is that the spectral and spatial dimensions are treated equally whereas typically in HSI, spectral correlation is far higher than spatial correlation. Also, the type and amount of noise ranges from low signal-independent noise to mixed Gaussian, Poisson and spike noise for different wavelengths. Therefore, a method utilising both the spatial and spectral information together, and based on the type of noise present in the data would be of great use for de-noising HS images.

Recently, a wavelength-specific shearlet-based image noise reduction method was proposed for de-noising of HS images in the remote sensing domain [11]. The method perfectly fits the needs of HS image de-noising by considering both the spatial and spectral correlations and also considering the types of noise present in different image planes. The method first identifies the type of noise present in the image planes via measurement of spectral correlation. Based on the spectral correlation, the method categorises the noise into low-level Gaussian noise or high levels of mixed noise. After identification of the type of noise, the non-subsampled shearlet transform (NSST) is then performed on each image plane. Later, to de-noise the low-level Gaussian noise wavelengths, the method assumes an additive noise model and performs spatial de-noising using the BayesShrink threshold method [20]. To de-noise the high-level mixed noise wavelengths, the method utilises the NSST information from the neighbouring low-level Gaussian noise wavelengths. The shearlet coefficients of adjacent low-level Gaussian noise wavelengths are fused with the details of mixed noise wavelengths utilising a weighted linear combination criterion, which results in spectral de-noising. Finally, after de-noising, the inverse of the NSST is applied to reconstruct the image planes.

The aim of the present work is to apply a wavelength-specific shearlet-based image noise reduction method [11] for HS image de-noising and to test its potential for de-noising close-range HS images. Furthermore, the method was compared with two pixel-based spectral smoothing techniques, i.e., SAVGOL and median filtering. The potential of the method was tested using three different sets of HS images. The first two image sets comprised simulated supervised images containing known amounts of Gaussian and mixed noise. The third image set was a real VNIR HSI dataset generated for the classification of six commercial tea products (oolong, black, green, yellow, Pu-erh and white). The performance of the de-noising techniques was evaluated through visual inspection of denoised image, spectral correlation between denoised and original image, peak signal-to-noise ratio (PSNR), and through classifications performed with a multi-class support vector machine (SVM). The PSNR was used to quantify the improvement in the spatial domain and the spectral correlation was used to quantify the similarity of the spectra after de-noising with the corresponding spectra in the absence of noise, i.e. the clean spectra (spectral domain).

### 5.3 Material and methods

#### 5.3.1 Samples and imaging sensor

De-noising and classification experiments were performed with visible-near infrared (VNIR) hyperspectral images of six different commercial tea products, which were purchased from a local market (Glasgow, United Kingdom). The samples were obtained in airtight sealed packaging and stored at ambient temperature. All samples of tea were in loose-leaf form. Black, green and white tea were from Vahdam Teas (New Delhi, India), oolong tea was from Yamamotoyama (California, USA), Pu-erh tea was from The Tea Makers of London (London, UK) and yellow tea was of an unspecified Chinese origin. Each tea sample was transferred to a black plastic circular container (diameter = 3.3 cm, depth = 1.3 cm) for analysis. The six samples were placed adjacent to each other on the translation stage so that all six samples were imaged in a single measurement. Imaging was performed using a push-broom line scan HSI system comprised of a V10E spectrograph from SPECIM (Oulu, Finland) and a CCD camera (C8484-05C, Hamamatsu Photonics, UK). The HSI system was used to acquire spatial maps consisting of  $1350 \times 256$  pixels over the spectral range 383 –

1000 nm with a spectral resolution of 2.45 nm. The pixel size of the CCD camera is  $6.45 \times 6.45 \mu\text{m}^2$ . Lighting was provided by two 20 W halogen light sources. The distance from the lens to the translation stage was 30 cm, and the stage was controlled by an independent stage motor connected to the computer system (Zolix TSA 200 BF). The speed of the translation stage,  $\sim 3 \text{ mm s}^{-1}$ , was optimised using a checkerboard to avoid any distortion in the shape of the image arising from the overlapping of the spectral and spatial information. A single image, comprising more than 2000 pixels per tea sample, was acquired of the six tea samples using a frame rate of 21 fps and an exposure time of 5 ms.

The acquisition and management of data were performed using in-house code developed in Matlab (R2016b, Mathworks Inc., Natick, United States). Before data analysis, the radiometric calibration of images was performed using white and dark references. The correction was performed for every pixel in the HS image according to equation 5.1,

$$I_{R(i,j,k)} = \frac{I_{raw(i,j,k)} - I_{dark(i,j,k)}}{I_{white(i,j,k)} - I_{dark(i,j,k)}} \quad (5.1)$$

where  $I_R$  is the calibrated reflectance,  $I_{raw}$  is the raw intensity measured from the test sample,  $I_{dark}$  the intensity of the dark response,  $I_{white}$  is the intensity of the uniform white reference, and  $i$  and  $j$  are spatial coordinates and  $k$  is the wavelength in the image.

To demonstrate the effectiveness of the HSI de-noising method, two more sets of HS images were simulated by adding different types of noise to the VNIR images. The simulation was performed by manually reducing the VNIR hypercubes to the cleanest (smoothest) spectral profile range (546 – 791 nm). Of the two sets of simulated images, one set was simulated with a known amount of Gaussian noise (zero mean and 0.03 variance), and the other was simulated with mixed noise comprising a combination of Gaussian (zero mean and 0.03 variance) and spike noise (density of 0.08) at 20 randomly selected image planes. The term density of 0.08 here means that the reflectance will change from zero to one with a probability of 0.08. In the following text, mixed noise (MN) will be used to represent the combination of Gaussian and spike noise and Gaussian noise (GN) will be used to represent Gaussian noise. A

summary of the sets of images analysed is presented in Table 5.1.

*Table 5.1: Details of image sets.*

<b>Image set</b>	<b>Dimension (height × width × wavelength)</b>	<b>Wavelength range/nm</b>
VNIR	1350 × 287 × 256	383-1000
VNIR + Gaussian Noise	1350 × 287 × 101	546-791
VNIR + Mixed Noise	1350 × 287 × 101	546-791

### 5.3.2 De-noising methodology

The de-noising methodology has three main steps in its implementation. The first step is to identify the type of noise present (low-level Gaussian noise or mixed noise) to choose the de-noising techniques for that particular wavelength. Different techniques here signify the different ways of using the shearlet coefficients for de-noising. The second step is to perform the sub-sampled shearlet transform on each image plane to capture the shearlet coefficients. The third step is to utilise the shearlet transform coefficients to perform de-noising individually for each wavelength based on the type of noise identified. The detailed methodology is explained in the following sections.

#### *5.3.2.1 Noise characterisation*

In HSI, the noise varies from wavelength to wavelength and can range from simple low-level Gaussian noise to high-level mixed noise resulting from a combination of Gaussian and spike noise. The typical additive noise model for any image plane of a data cube ( $n \times p \times q$ ) can be understood from equation 5.2:

$$Y = X + N \quad (5.2)$$

where Y is the recorded image plane containing the useful informative signal part (X)



and the noise part (N). This assumed model is usually correct if the noise present in the plane is limited to Gaussian white noise, but this is not always the case. In the methodology presented here, the nature of the noise is assumed to be unknown. To find the type of noise present in image planes, the method utilises the correlation coefficient,  $R$ , between two image planes,  $Y_k$  and  $Y_{k+r}$ , as in equation 5.3:

$$R(Y_k, Y_{k+r}) = \frac{cov(Y_k, Y_{k+r})}{\sqrt{var(Y_k)var(Y_{k+r})}} \quad k = 1, 2 \dots q - r \quad (5.3)$$

From the correlation measurement between two image planes, it can be understood that if the two image planes are very similar, then they will have a very high correlation coefficient. However, in the presence of noise, the correlation between the image planes will be significantly reduced. Furthermore, the greater the noise, the more the correlation will decrease. To differentiate between the low-level Gaussian noise and the high-level mixed noise, the threshold for the mean correlation between the image plane and its neighbouring image planes was set. The mean correlation was obtained by choosing a window,  $w_1$ , containing 10% of the total number of wavelengths centred around the wavelength considered, as in equation 5.4.

$$\bar{R}(Y_k) = mean_{r \in w_1}(R(Y_k, Y_{k+r})) \quad (5.4)$$

The values obtained for the mean correlation for low-levels of noise will be very high. Depending on the amount and the complexity of the noise, the correlation will decrease. Therefore, the values will span a heavy-tail distribution for the  $\bar{R}$ . For such a distribution, the median is already known to be the best estimator to represent the central tendency of the distribution [21]. In the present methodology, the median estimated from the distribution of the correlation coefficients was chosen to be the threshold and to classify the image plane as either a low- or high-level noise image plane.

### 5.3.2.2 Shearlet transform

After the classification of the image planes as low-level Gaussian noise or high-level and/or mixed noise, the NSST coefficients for each image plane were calculated independently. NSST is a special type of discrete shearlet transform that provides an

additional feature of invariance to the shift of the input signal [22]. The shearlet transform is a special type of composite wavelet transform in which the mother wavelet matrix is an anisotropic dilation matrix along with the shear matrix, compared to the dilated matrix associated with the scale transformation and directional transformation in the composite wavelet transform. The composite wavelet function can be understood as equation 5.5.

$$\psi_{j,l,k}(x) = |\det M|^{j/2} \psi(S^l M^j(x - k)) \quad (j, l \in \mathbb{Z}, k \in \mathbb{Z}^2) \quad (5.5)$$

where  $\psi$  is the mother wavelet,  $M$  is an anisotropic dilation matrix,  $S$  is a shear matrix and  $j$ ,  $l$  and  $k$  are scale, directional and shift parameters, respectively. In this work, we have limited the explanation to the NSST only, however, more detailed information on the composite wavelet transform and shearlet transform can be found in [22] [23] [24].

The implementation of NSST to decompose the image planes requires two steps. The first is the application of non-subsampled pyramid (NSP) filter banks and the other is the use of non-subsampled shearing (NSS) filter banks. A non-subsampled filter bank has no shift variant issues as there is no down- or up-sampling during the decomposition. Furthermore, the NSP filter gives the multiscale decomposition of the original image into high- and low-frequency sub-images of the same size as the original image. The NSS part of the NSST performs directional filtering in the spatial domain and decomposes the high-frequency sub-images into directional sub-images. For a typical application, the filter banks are applied in an iterative way where the low-frequency sub-images obtained are again decomposed to lower scale high- and low-frequency sub-images, resulting in a multi-scale and multi-directional decomposition. An example of the multi-scale and multi-directional decomposition performed by the NSST can be understood with the three-scale decomposition shown in Figure 5.1. In the present case, a three-level shearlet decomposition with 16, 8 and 4 shearing directions at scales of 1, 2 and 3, respectively, was performed.

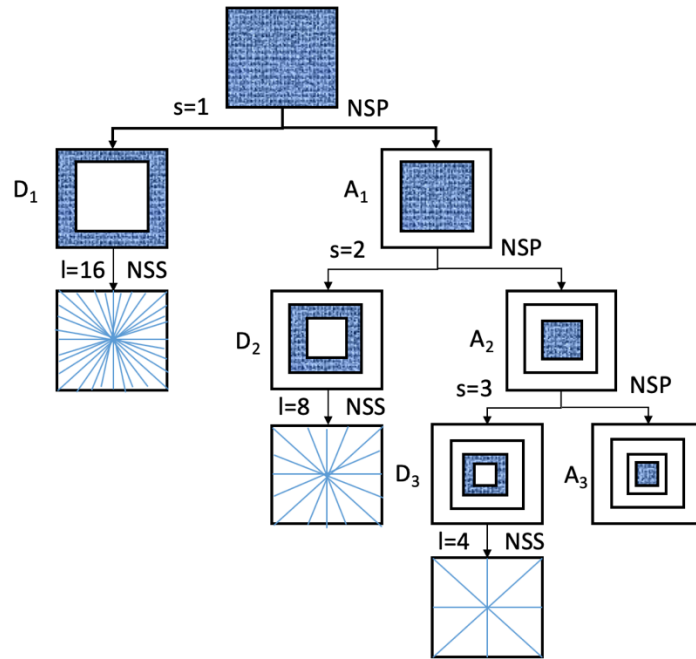


Figure 5.1: Multi-scale and multi-dimensional three-scale decomposition with a non-subsampled shearlet transform (NSST) where  $A$  and  $D$  signifies the low and high frequency sub images, respectively, and  $l = 16, 8$  and  $4$  at scales  $1, 2$  and  $3$ , respectively.

### 5.3.2.3 De-noising image planes

The coefficients obtained from the NSST of the image are mostly very small and close to zero. But due to the presence of the noise, the sparsity of the matrix of NSST coefficients is greatly reduced. Therefore, to perform the de-noising with NSST, the aim is to re-attain the sparsity of the matrix of NSST coefficients. To do this, a threshold is used to distinguish the coefficients corresponding to noise from the coefficients containing signal information. Different techniques were used for de-noising the image planes identified with low and high levels of noise. For the low-level noisy image planes, the threshold for the shearlet coefficients,  $T_{s,t}$ , was determined assuming an additive noise model and utilising the BayesShrink method [20], as shown in equation 5.6:

$$T_{s,t} = \frac{\sigma_{N,s,t}^2}{\sigma_{s,t}} \quad (5.6)$$

where the noise variance at scale  $s$  ( $s = 1 \dots j$ ) and direction  $t$  ( $t = 1 \dots l$ ) is given by  $\sigma_{N,s,t}^2$  and is estimated as mentioned in [25]. The standard deviation ( $\sigma_{s,t}$ ) of the signal measured from the sub-image  $Y_{s,t}$  at scale  $s$  and direction  $t$  is estimated as in equation 5.7:

$$\sigma_{s,t} = \sqrt{\max((\sigma_{Y_{s,t}}^2 - \sigma_{N,s,t}^2), 0)} \quad (5.7)$$

where  $\sigma_{Y_{s,t}}^2 = \frac{1}{N^2} \sum_{i,j=1}^N Y_{s,t}^2(i, j)$

$i$  and  $j$  define the spatial coordinates and  $N$  is the maximum value of  $(n, p)$ .

Once the image planes with a low level of noise were de-noised, then to de-noise the high and/or mixed noise image planes, the shearlet coefficients of the adjacent low noise level image planes were fused to the details of the mixed noise sub-images. The shearlet coefficients were used to replace the details of the sub-images of the mixed noise level image planes by the weighted average of the sub-image details of the 10% closest low noise level image planes. The weights used were inversely proportional to the distance between the neighbouring wavelengths as explained in equation 5.8:

$$\sum_{r \in w_2} w_r = 1 \quad \widehat{D}_{MN}(k) = \sum_{r \in w_2} w_r \widehat{D}_{LGN}(k + r) \quad (5.8)$$

where  $w_2$  are the adjacent LGN image planes, and  $\widehat{D}_{MN}$  and  $\widehat{D}_{LGN}$  correspond to de-noised mixed noise and low-level Gaussian noise image planes, respectively. To reconstruct the de-noised image plane, the inverse of the NSST was applied to the coefficients. Any further classification analysis was performed on the resulting de-noised images.

### 5.3.3 Savitzky-Golay smoothing

SAVGOL smoothing is a window-based technique that utilises different polynomial functions to smooth signals [13]. To perform smoothing with SAVGOL, a window of fixed size is chosen, centred on the signal point to be smoothed, and a polynomial is

fitted to the variables within the window. The value of the central variable is replaced by the value calculated by the polynomial function. The window is moved point-by-point over the signal to perform the smoothing on the complete spectrum. The window size and the polynomial function are usually chosen manually, and the optimum choice is based on visual inspection of the spectral profile. For the present work, a second order polynomial and 15-point window were used. SAVGOL smoothing was performed using the PLS Toolbox (version 8.11, Eigenvector Research Inc., USA).

#### 5.3.4 Median filtering

Median filtering belongs to the family of non-linear signal filtering techniques and is often used to deal with high levels of noise such as spikes in the data. Median filtering in the spectral domain can be understood as a moving window that replaces each observation with the median value of the observations inside the window. When the number of observations inside the window is odd the median is a single value, however when the number of observations is even, then the median is the average of the two middle values. In the present work, the median filter was employed by unfolding the  $(n \times p \times q)$  HSI array to give a  $(np \times q)$  matrix, performing the median filtering with a 4-point window and later reshaping the matrix back to the cube. To perform the filtering, the *meadfilt1* Matlab function was used.

#### 5.3.5 De-noising performance

Spectral correlation and peak signal-to-noise ratio (PSNR) were used to quantify the performance of the de-noising methods in the spectral and spatial domains, respectively.

Spectral correlation provides a measure of the similarity of spectra after de-noising with the corresponding spectra in the absence of noise (i.e., the clean spectra), and was estimated via calculation of the correlation coefficient between the de-noised and clean spectra utilising the *corr* function in Matlab. The PSNR was calculated as shown in equation 5.9:

$$PSNR = 10 \log_{10} \left( \frac{peakvalue^2}{MSE} \right) \quad (5.9)$$

where the mean square error is given by  $MSE = \frac{\sum_{i,j=1}^N [Y_1(i,j) - Y_2(i,j)]^2}{N^2}$

and  $Y_1$  and  $Y_2$  are the two image planes to compare (i.e., the image plane after denoising and the corresponding image plane in the absence of noise, respectively) and peakvalue is either specified by the user or selected from a range that is dependent on the image datatype (e.g. 255 for a uint8 image). The PSNR was calculated utilising the PSNR function in Matlab.

### 5.3.6 Multiclass support vector machine classification

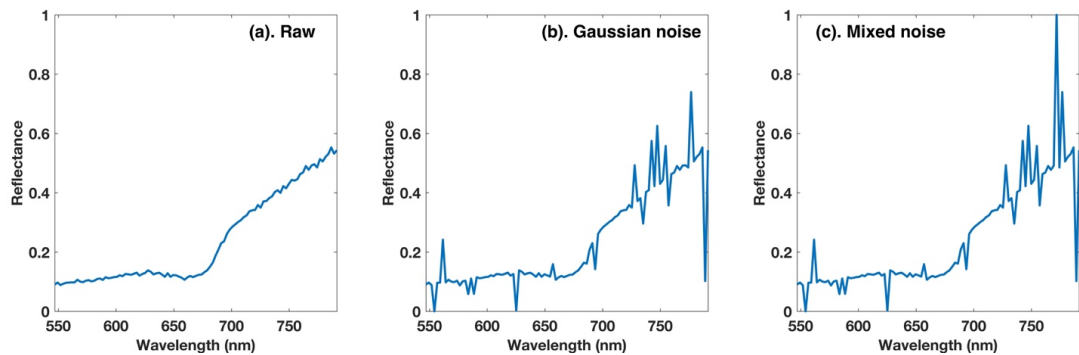
To perform the classification experiment on the HSI data sets, support vector machine (SVM) classifiers were developed. SVM utilises the hyperplanes to define the decision boundaries to perform classification. Furthermore, to deal with the data complexity, the SVM performs high dimensional mapping of the data using kernel functions. Mapping to higher dimensions is usually carried out to make the data linearly separable. Furthermore, in the high dimensional space, the choice has to be made to select the hyperplane that provides the largest separation of the classes. As in our case we have six different classes corresponding to six different tea products, the traditional SVM binary classifier was combined with the error correcting output code (ECOC) ensemble method. The ECOC deals with the multiclass problem by creating several independent binary classification models.

The SVM along with ECOC was implemented in Matlab via the Statistics and Machine Learning Toolbox (R2016b). The coding design used for the ECOC-SVM model was one-vs-one, where a model was developed with one class being assigned positive and another class being assigned negative and all other classes were neglected. The algorithm exhausts all combinations of class pair assignments leading to  $k(k-1)/2$  models, where  $k$  is the number of classes to be considered. High dimensional mapping of the data was performed with a radial basis function (RBF) kernel function with a scale parameter of 2.5. The spectra from the images for six different classes were selected in a supervised way using the “roipoly” function in Matlab. The “roipoly” function provides the graphical interface to manually extract the information from the data cube. For each individual tea class, spectra were extracted from 200 pixels, which

were selected at random from the image collected, leading to a total of 1200 spectra for classification model development. For the development of a robust model, the model was cross-validated with a 10-fold cross-validation method. In 10-fold cross-validation, the calibration data is divided into ten equal parts. For making the model, 9 out of 10 parts were used and to cross-validate, the 10th part was used. This was then repeated ten times, and the average prediction accuracy was recorded. The whole process was performed with 100 iterations and the mean accuracy and standard deviation were recorded.

## 5.4 Results

### 5.4.1 Noisy and de-noised spectra from simulated images



*Figure 5.2: A single spectrum at pixel location (370, 135) extracted from the following image sets: (a). VNIR, (b). VNIR + Gaussian noise, and (c). VNIR + Mixed noise. 20 random wavelengths were used to simulate Gaussian and mixed noise images.*

Figure 5.2 presents the spectrum extracted from the raw reflectance (5.2a), GN added to reflectance (5.2b) and MN added to reflectance (5.2c) data from a spatial location at (370, 135) in the simulated hyperspectral image. The spectral noise was added at 20 different random wavelengths over the data cube and its effect can be seen in (5.2b,c) when compared to the raw reflectance profile in 5.2(a).

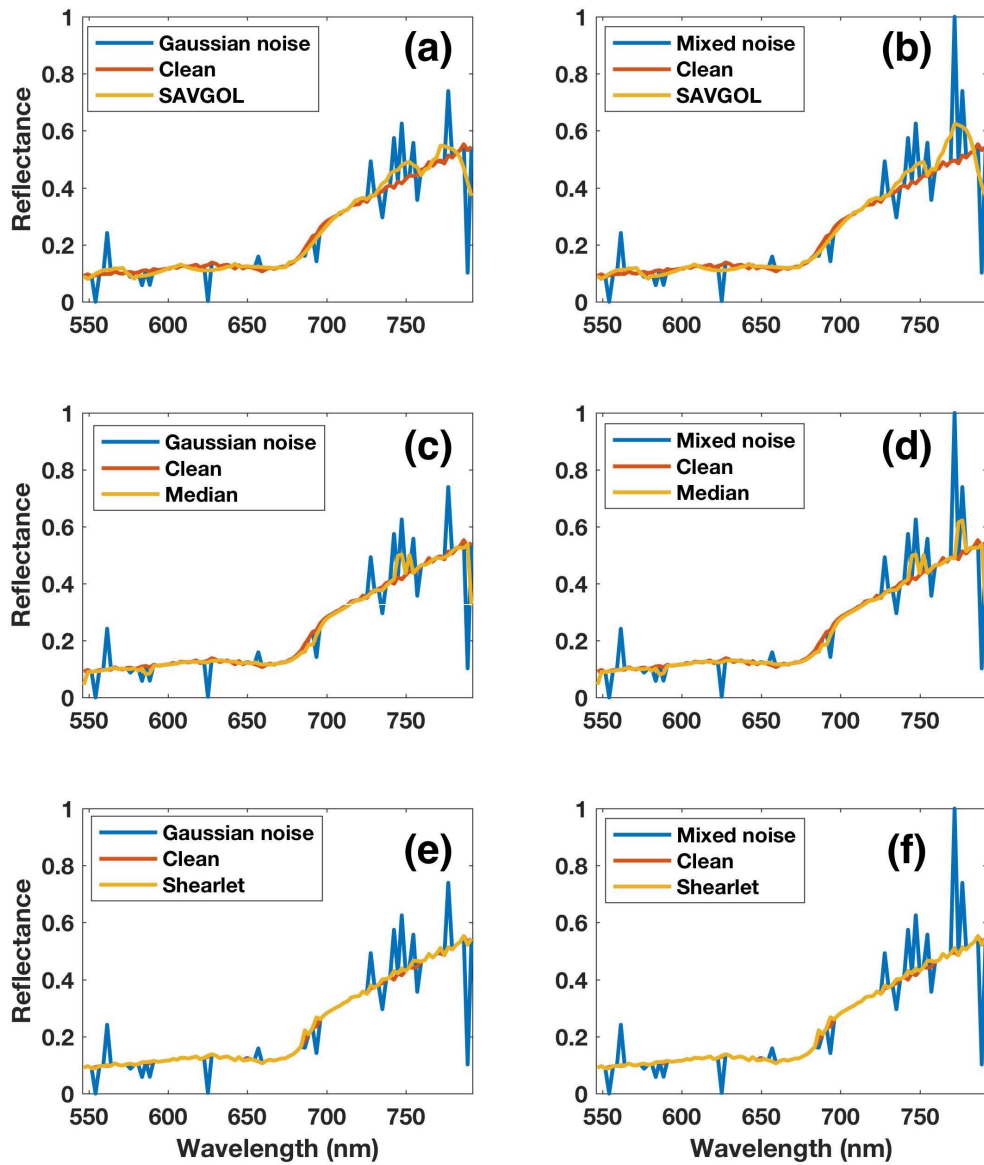


Figure 5.3: Noisy and de-noised spectra from pixel location (370, 135), the red line shows the clean spectrum, the blue line shows the noisy spectrum and the yellow line shows the spectrum after de-noising. (a). SAVGOL smoothing filter applied to the spectrum with mixed noise added, (c). Median filtering applied to spectrum with Gaussian noise added, (d). Median filtering applied to spectrum with mixed noise added, (e). de-noising with shearlet-based method for spectrum with Gaussian noise added, and (f). de-noising with shearlet-based method for spectrum with mixed noise added.



Figure 5.3 presents the effect of different de-noising methods applied to the spectra shown in Figure 5.2. Figure 5.3 (a,b) shows the results of SAVGOL smoothing, with a 15 point window and a second order polynomial, of the spectra with Gaussian and mixed noise added. Figure 5.3 (c,d) shows the results from application of median filtering with a 4 point window. Figure 5.3 (e,f) illustrates the result of utilising the shearlet-based de-noising methodology. The spectra in red in Figure 5.3 represent the clean reflectance profiles, while the solid blue and yellow lines represent the noisy and de-noised spectra, respectively.

In Figure 5.3, it can be seen that the shearlet-based de-noising method outperformed SAVGOL smoothing and median filtering. The reason for the poor performance of SAVGOL smoothing in the case of simulated noise can be understood as being due to the window size and the smoothing function used. Since SAVGOL smooths each spectrum by fitting a polynomial to a window of adjacent wavelengths, if noise contributes significantly to several of the wavelengths, then the polynomial fitting will be less effective. Median filtering works better than SAVGOL smoothing since even if there are several outlier intensity values in the spectrum, they will not have as much influence on the smoothed value. Nevertheless, there are peaks to be found in the spectra after median filtering, due to the presence of several peaks within the window used for the smoothing, resulting in the median value being influenced by this noise. Furthermore, the peaks resulting from median filtering now appear at new wavelengths. This is because median filtering is performed for each wavelength resulting in transferral of noise to wavelengths adjacent to the noisy wavelengths. However, in the shearlet-based methodology, the de-noising of the waveband is performed using the shearlet coefficients of the adjacent low GN image planes, therefore, the high-intensity noise wavelengths do not affect the spectral de-noising as in the case of SAVGOL smoothing and median filtering. This is because the high-intensity noise wavelengths have no influence when performing the de-noising since they are not considered in the calculation of the weighted average of shearlet coefficients.

In Figure 5.3 (e,f), it can be seen that some small differences can still be found in the

spectrum after de-noising. The reason for these small disturbances can be understood as resulting from the averaging of the shearlet coefficient, especially when the automatically selected GN image planes are distant, as averaging with the shearlet coefficients of these GN planes leads to small disturbances in the spectral profile. However, these disturbances are minute compared to the noise present in the spectrum after smoothing with SAVGOL or median filtering. To quantify these small disturbances and the spectral similarity, spectral correlation was used.

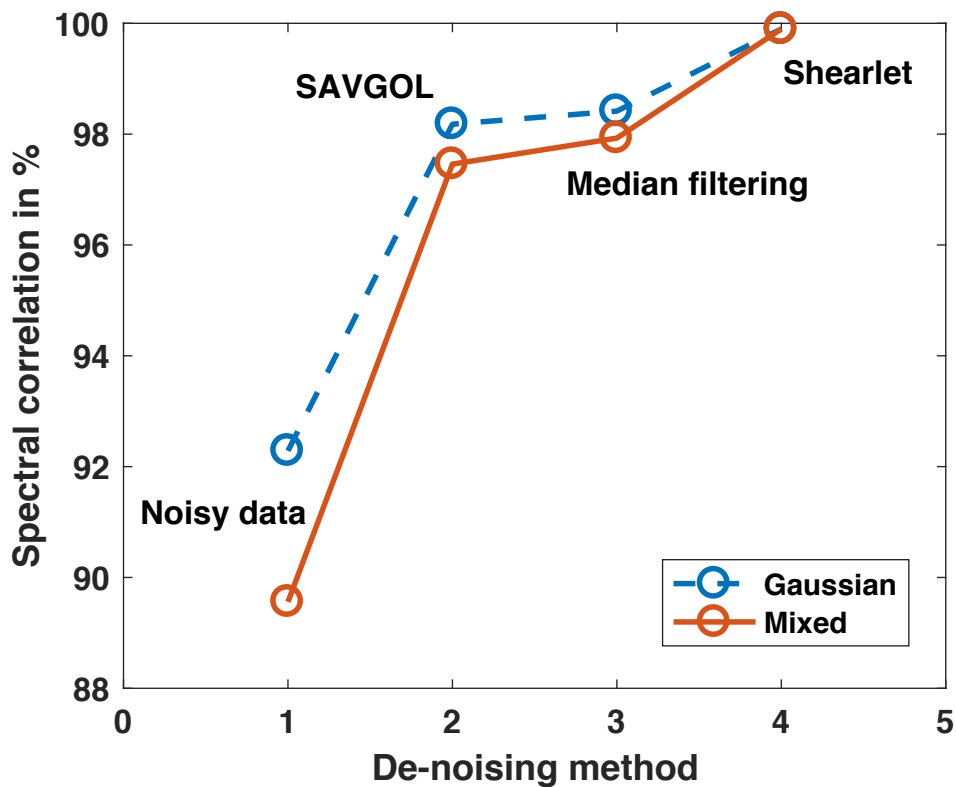


Figure 5.4: Correlation for the spectra extracted from the pixel at location (370, 135) from noisy and de-noised data with clean data. The blue line dashed shows the Gaussian noise and the red line shows the mixed noise.

Figure 5.4 presents the spectral correlation calculated between a single spectrum extracted at a pixel location of 370,135 from, on the one hand, the noisy, SAVGOL, median filtered and shearlet de-noised images, and, on the other hand, the clean image. A summary related to all the pixels correlation can be found in Annex 3 Figure A3.1.

Further in Annex 3, a summary of time required by different techniques is presented in Table A3.1. The blue dashed line and red solid line shows the GN and MN cases, respectively. The spectral correlation was calculated to quantify the similarity of the spectral profiles obtained from different de-noising techniques. It can be seen in Figure 5.4, that the spectral correlation for the spectrum with GN added was always high for all the de-noising techniques compared to the spectrum with MN added. The reason is that the MN noise is much more complicated, resulting in a higher number of noisy wavelengths in the spectrum. Furthermore, the new shearlet-based method gave the highest correlation of 99.9% followed by median filtering and then SAVGOL smoothing.

#### 5.4.2. Noisy and de-noised image planes from simulated images

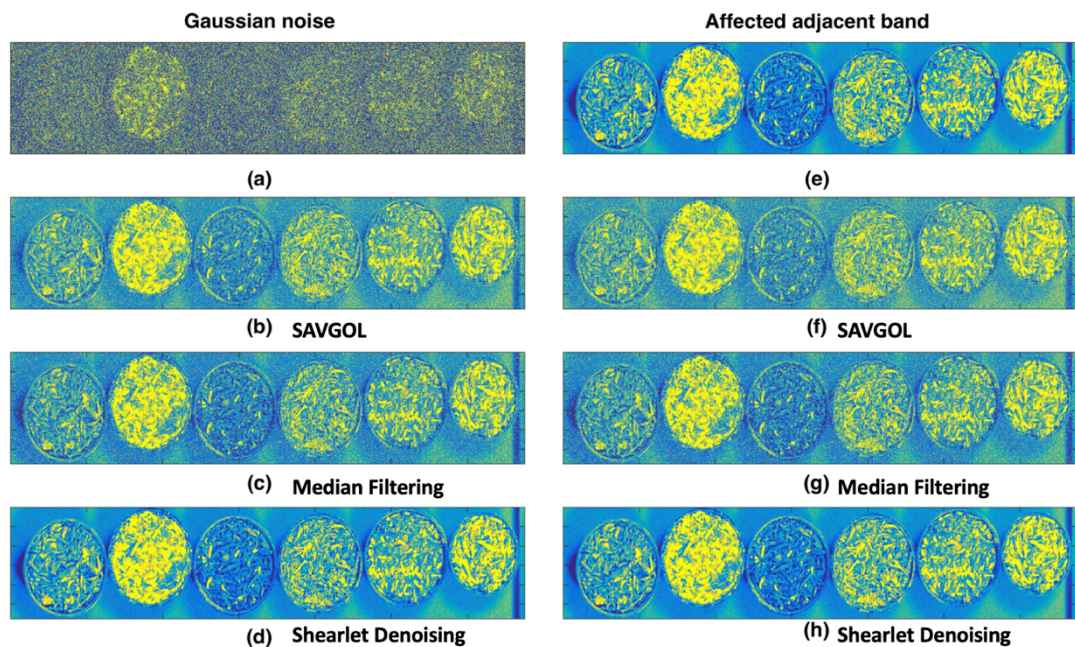


Figure 5.5: The image planes corresponding to six different tea samples (oolong, black, green, yellow, Pu-erh and white (from left to right)) and the simulated noisy and de-noised image planes for data with Gaussian noise added. (a). Reflectance image plane (588 nm), (b) SAVGOL smoothed image plane, (c) Median filtered image plane, (d) Shearlet de-noised image plane, (e). Clean adjacent reflectance image plane

(586 nm), (f). Adjacent image plane after SAVGOL smoothing, (g) Adjacent image plane after median filtering, and (h). Adjacent image plane after shearlet de-noising.

Figure 5.5 presents the image planes (at 588 and 586 nm) with GN added used in the simulation studies. The plane corresponding to 588 nm is the de-noised band whereas 586 nm represents the clean adjacent band affected by the de-noising methods. Figure 5.5 (a,e) presents the original reflectance image plane with Gaussian noise added, 5.5 (b,f) the SAVGOL smoothed data, 5.5 (c,g) the median filtered data and 5.5 (d,h) the image planes after application of the shearlet-based de-noising method. The six circular objects in the image represent six different commercial tea products, i.e. oolong, black, green, yellow, Pu-erh and white. The presence of the Gaussian noise can be seen as a ‘fog’ over the image plane (5.5 (a)). It can be seen clearly by visual inspection that the shearlet-based de-noising method outperformed both SAVGOL and median filtering to give a clearer image.

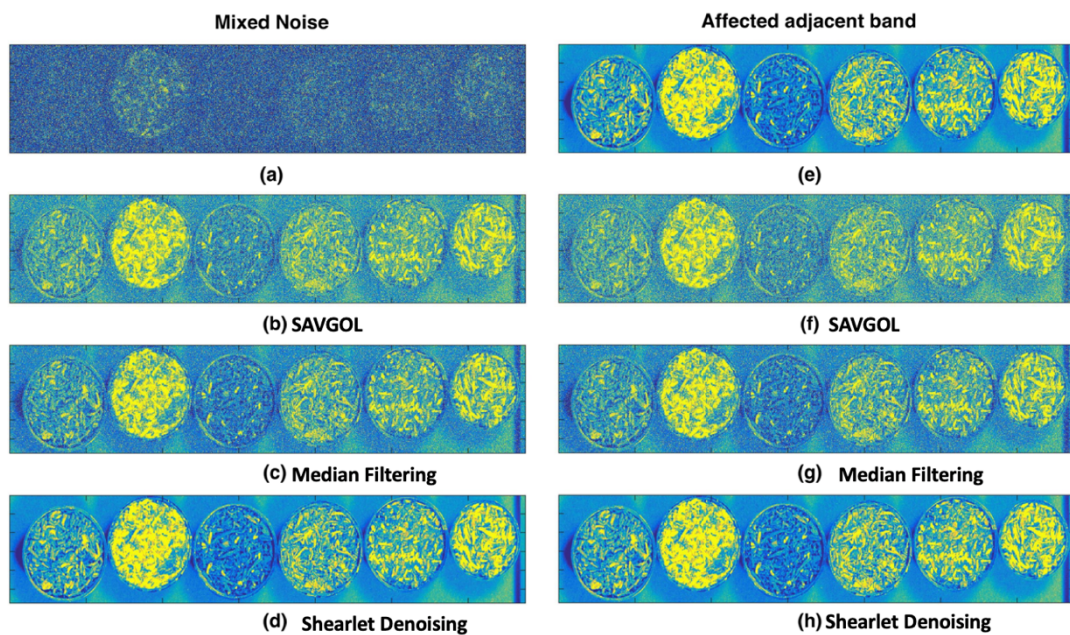


Figure 5.6: The image planes corresponding to six different tea samples (oolong, black, green, yellow, Pu-erh and white (from left to right)) and the simulated noisy and de-noised image planes for data with mixed noise added. (a). Reflectance image plane (588 nm), (b) SAVGOL smoothed image plane, (c). Median filtered image plane,

*(d) Shearlet de-noised image plane, (e). Clean adjacent reflectance image plane (586 nm), (f) Adjacent image plane after SAVGOL smoothing, (g). Adjacent image plane after median filtering, and (h). Adjacent image plane after shearlet de-noising.*

Figure 5.6 presents the image planes (for 588 and 586 nm) with MN added used in the simulation studies. Figure 5.6 (a,e) presents the reflectance image plane with MN added, 5.6 (b,f) the SAVGOL smoothed data, 5.6 (c,g) the median filtered data, and 5.6(d,h) the image planes after application of the shearlet-based de-noising method. The presence of mixed noise (5.6 (a)) can be seen as a ‘fog’ accompanied by some high-intensity (bright) pixels resulting from the spike noise. The shearlet-based de-noising method clearly outperformed SAVGOL and median filtering. However, median filtering seems to provide better results than SAVGOL smoothing for image planes containing GN or MN. This is because SAVGOL smoothing dilutes the noise of several consecutive wavelengths by fitting the polynomial whereas the median filter is calculated directly based on the intensities present inside the window resulting in a better de-noising. However, both the median filter and SAVGOL smoothing also affect the consecutive wavelengths by spreading the noise. The spreading is more important when the noise is present in consecutive wavelengths. In the case of SAVGOL smoothing, the noisy wavelengths dominate the shape of the polynomial and lead to spreading of the noise to the consecutive wavelengths. In the case of median filtering, a larger number of noisy wavelengths in a small window affects the calculation of the median. Also, median filtering does not use any measure to pre-identify the amount of noise. Therefore, the correction is performed for each wavelength, which results in a distribution of the noise to the adjacent wavelengths. The spreading of the noise can be seen in Figures 5.5 (f) and 5.6 (f), and 5.5 (g) and 5.6 (g) resulting from SAVGOL and median filtering, respectively, when compared to Figures 5.5(e) and 5.6(e), the clean image planes. However, this was not the case with the shearlet-based method as it does not affect any adjacent wavelengths because the method starts with identifying the wavelengths containing noise resulting in wavelength-specific de-noising.

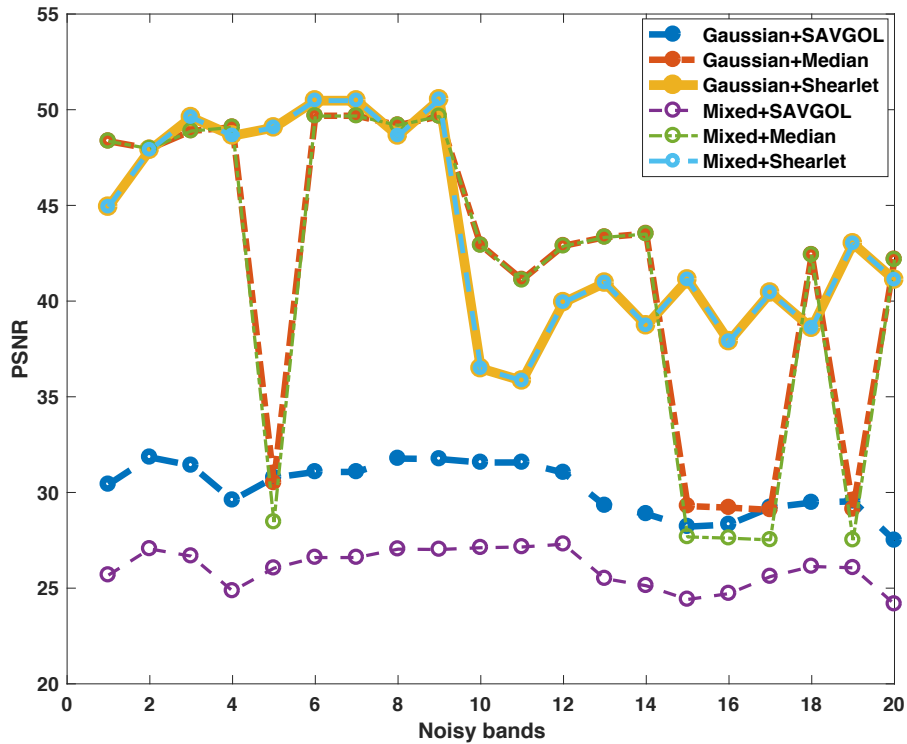


Figure 5.7: Peak signal-to-noise ratio (PSNR) for the 20 wavelengths with added noise: purple dashed (SAVGOL smoothed data with mixed noise), thick blue dashed (SAVGOL smoothed data with Gaussian noise), thick red dashed and dot (median filtered data with Gaussian noise), green dashed and dot (median filtered data with mixed noise), thick yellow (shearlet de-noised data with Gaussian noise) and dashed sky-blue (shearlet de-noised data with mixed noise).

The potential of the shearlet-based de-noising method is further quantified using the PSNR as presented in Figure 5.7. The PSNR represents the ratio of the maximum possible signal intensity to the corrupting noise present in the signal. Figure 5.7 shows the PSNR for the 20 randomly selected image planes used to simulate the noise. The PSNRs were estimated taking the raw reflectance image planes with no added noise (i.e. clean) as the reference for the de-noised image planes. It can be seen in Figure 5.7 that the shearlet-based de-noising method (in yellow and sky-blue) increased the PSNR and attained the highest levels for the majority of image planes where GN and

MN had been added. Median filtering increased the PSNR more than SAVGOL smoothing and was as effective as the new shearlet-based de-noising method for several wavelengths. The PSNR obtained from the shearlet-based de-noising methodology was the same for data containing GN (in yellow) and MN (in sky blue). The PSNR obtained for SAVGOL smoothed data with added noise was higher for GN (in dashed thick blue) compared to MN (in dashed purple). Similarly, the PSNR obtained with the median filter was higher for data with added GN (in thick dashed and dot red) at several wavelengths compared to MN (in dashed and dot green). The reason for this is that MN is more complex than GN and therefore with median filtering and SAVGOL smoothing, the PSNR increased more for the GN than for the MN case. The improved PSNR indicates that the signal contains more information compared to the noise and that data modelling based on the improved PSNR should be more successful. However, a higher PSNR does not guarantee successful modelling because the affected adjacent wavelengths resulting from the de-noising method is also a concern.

Table 5.2: SVM classification model accuracies (%) obtained using different de-noising methods with the simulated and real datasets (see Table 5.1). The value given is the mean  $\pm$  one standard deviation resulting from 100 iterations of a 10-fold cross validated model.

<b>Images type</b>	<b>Noisy data</b>	<b>Savitzky-Golay smoothing</b>	<b>Median filtering</b>	<b>Shearlet de-noising</b>
VNIR + Gaussian Noise (Simulated)	55.63 $\pm$ 0.42	57.18 $\pm$ 0.52	79.00 $\pm$ 0.48	87.37 $\pm$ 0.30
VNIR + Mixed Noise (simulated)	56.21 $\pm$ 0.41	49.77 $\pm$ 0.44	78.23 $\pm$ 0.47	87.36 $\pm$ 0.35
VNIR (Real)	42.9 $\pm$ 0.44	67.44 $\pm$ 0.55	59.67 $\pm$ 0.56	78.27 $\pm$ 0.55

The improvement in the classification performance of the SVM classifier after de-noising can be noted in Table 5.2. The shearlet-based de-noising method gave the highest accuracy of  $87.37 \pm 0.30\%$  and  $87.36 \pm 0.35\%$  for data cubes with added GN and MN, respectively. The median filter was second in terms of classification accuracy with comparable accuracy obtained for the data cubes with added GN ( $79.00 \pm 0.48\%$ ) and MN ( $78.23 \pm 0.47\%$ ). SAVGOL smoothing resulted in the lowest accuracy of  $57.18 \pm 0.52\%$  and  $49.77 \pm 0.44\%$ , for the data with GN and MN added, respectively. For the data containing MN and subjected to SAVGOL smoothing, the accuracy was even lower than for the data containing MN ( $56.21 \pm 0.41\%$ ). The poor performance of SAVGOL smoothing is due to the spreading out of the effect of the noise over several consecutive wavelengths by the fitting of the polynomial inside the



window. This reason can also explain the better performance of median filter compared to SAVGOL smoothing.

### 5.4.3 De-noised VNIR image set spectra

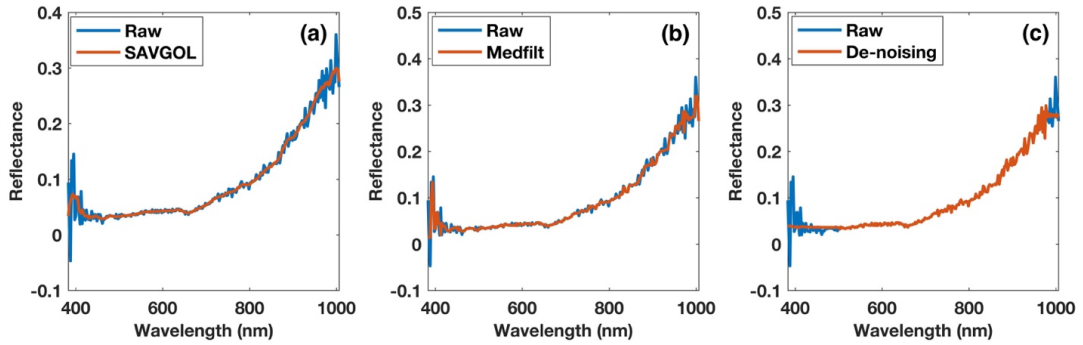


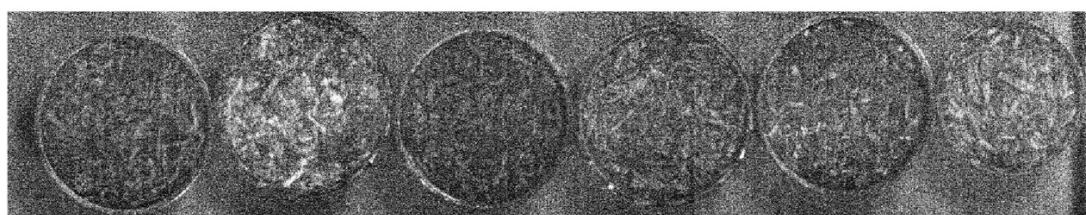
Figure 5.8: A single spectrum at pixel location (370,135) extracted from the VNIR hypercube, blue represents the raw reflectance signal, and red represents the de-noised spectra. (a). SAVGOL smoothing, (b). Median filtering, and (c). Shearlet-based de-noising method.

Figure 5.8 presents a complete spectrum (383-1000 nm) extracted from the raw reflectance and the de-noised VNIR hypercubes for the six different commercial tea products. Figure 5.8 (a) presents the spectrum after SAVGOL smoothing, Figure 5.8 (b) after median filtering and Figure 5.8 (c) shows the results of the shearlet-based de-noising method. The red line depicts the spectrum after the de-noising treatment, with the raw reflectance spectrum given in blue. It can be seen that the raw reflectance spectrum (blue) contains noise at different wavelengths over the complete range, especially at the beginning (383-500 nm) and end (900-1000 nm) of the spectrum. In Figure 5.8 (a) it can be seen that SAVGOL filter smooths the spectrum quite well from 500 to 900 nm. However, at the edges of the spectrum, SAVGOL smoothing does not work as well. In Figure 5.8 (b) it can be seen that median filtering also performed well between 500-900 nm, but was less effective than SAVGOL smoothing at the very noisy ends of the spectrum. In the case of the shearlet-based method (Figure 5.8 (c)), it can be seen that the spectrum around the edges (383-500 nm and 900-1000 nm) is smoother compared to both SAVGOL and median filtering. This is because the

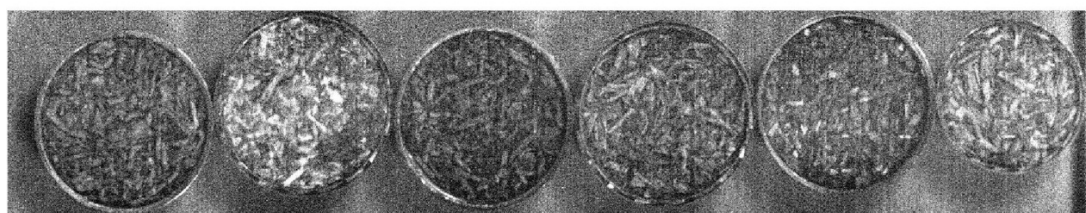
shearlet-based de-noising method utilises the shearlet coefficients of the neighbouring low Gaussian noise image planes to perform the weighted averaging. With the algorithm, a total of 59 wavelengths were identified as containing noise, and were de-noised automatically.

The VNIR (383-1000 nm) reflectance spectral profiles of food products contain different chemical information such as pigments, moisture content, and physical information such as particle size. Noise-free spectral features related to this information must be extracted to form the basis for the success of any classification modelling. In the case of tea products, the VNIR spectra contain information related to different chemical components. Some key wavelengths identified in a previous study were 485 nm corresponding to the total liquor colour, 522 – 625 nm to thearubigins constituent group TRS1, 688 and 732 nm to thearubigins, 706 nm to total polyphenols, 743 nm to liquor brightness and 745 nm to theaflavin [26].

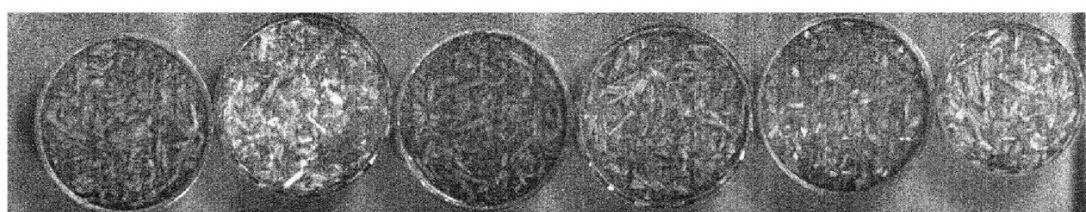
#### 5.4.4 De-noised VNIR image set image planes



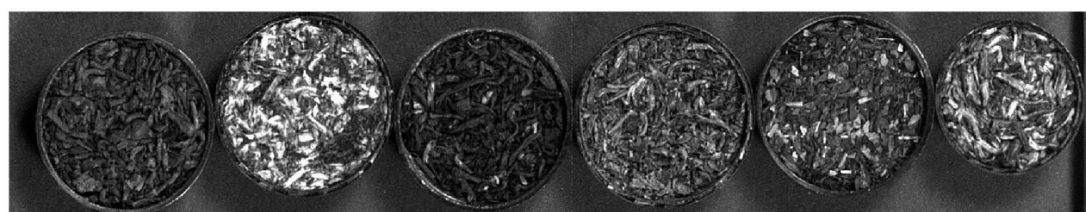
(a)



(b)



(c)



(d)

*Figure 5.9: Noisy and de-noised image planes (405 nm) extracted from real hyperspectral data from a VNIR camera. (a). VNIR noisy image plane, (b). SAVGOL de-noised image plane, (c). Median filter de-noised image plane, and (d) Shearlet de-noised image plane.*

Figure 5.9 presents the noisy and de-noised image planes for the real VNIR cube. Figure 5.9 (a) shows the raw noisy image planes from the VNIR hypercube corresponding to 405 nm, where intense noise is visible. The chosen waveband used

to display the image plane was automatically identified by the algorithm for de-noising. Figures 5.9 (b), (c) and (d) present the results of de-noising with SAVGOL smoothing, median filtering and the shearlet-based de-noising method, respectively. In Figure 5.9 (d), it can be seen that after de-noising the image planes reveal six different tea samples thus demonstrating the improved performance of the presented de-noising method over SAVGOL (Figure 5.9 (b)) and median filtering (Figure 5.9 (c)).

#### 5.4.5 SVM classification on VNIR image set

The results of classification (Table 5.2) showed an improvement in the classification accuracy after utilising the shearlet-based de-noising method compared to raw reflectance, SAVGOL smoothed and median filtered data. The reason for this is the high signal-to-noise ratio obtained after the de-noising of the image planes. The accuracy of the model was increased from  $42.90 \pm 0.44\%$  to  $78.27 \pm 0.55\%$  after de-noising with the shearlet-based methodology. For SAVGOL smoothing, the classification accuracy increased to  $67.44 \pm 0.55\%$  and for median filtering, it increased to  $59.67 \pm 0.56\%$ . In this case, SAVGOL smoothing outperformed median filtering because at the two ends of the spectra there are many consecutive noisy peaks that are considered in the calculation of the median value.

### 5.5 Conclusions

The data generated from HSI often contain noise. For an efficient data processing strategy, it is important to deal with the noise present in the data by either reducing or removing it. However, simply removing noise (wavelengths) can lead to loss of information from the dataset. Therefore, exploiting different data pre-processing techniques to reduce the noise in the datasets is always the better option as the information in the dataset is largely retained.

Commonly used methods such as SAVGOL smoothing and median filtering can deal with a small amount of noise and if the noise is not present in neighbouring wavelengths. However, when the noise level increases and the noise is present in consecutive wavelengths, SAVGOL smoothing and median filtering can result in

distorted spectral profiles leading to spreading of the noise to adjacent wavelengths. Furthermore, both SAVGOL smoothing and median filtering are performed for every waveband as no steps are present to determine the wavelengths that need to be de-noised automatically. This, results in over-smoothing of the spectral profiles. Another drawback is the need to determine the correct window size, which is often performed manually. Different window sizes might result in an improvement in the de-noising results, however, visual inspection is needed to select the optimum window size.

However, the presented shearlet-based technique deals with the noise in an intelligent, fully automatic way by first classing the wavelengths into non-noisy, low GN and MN, thus reducing the chances of over-smoothing of the wavelengths and spreading of noise to adjacent wavelengths. The method then deals with the spatial and spectral noise synergistically and also adapts to the type of noise present in the data. The de-noising of the low GN noise wavelengths is performed through retention of the sparsity of the shearlet coefficients, which is completely independent of other wavelengths. Finally, to de-noise the MN wavelengths, the method fuses the information from the shearlet coefficients of the neighbouring GN wavelengths. This study also demonstrated the potential of the shearlet based de-noising methodology as seen in the visual improvement of image planes, the increase in spectral correlation, the increase in PSNR for image planes and the improved classification accuracy of the multi-class SVM model, compared to SAVGOL smoothed and median filtered data. The shearlet-based methodology is a useful technique for automatic de-noising of close-range HSI data where the spectral domain exhibits broad signals (i.e., information from adjacent spectral bands is correlated). Hence, this methodology will provide new opportunities for the use of a wide variety of HSI techniques, e.g. NIR, UV, visible and fluorescence, for real-time decision making such as in a process environment.

## 5.6 References

1. A.A. Gowen, C.P. O'Donnell, P.J. Cullen, G. Downey, J.M. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, *Trends Food Sci. Technol.* 18 (2007) 590–598. doi:<https://doi.org/10.1016/j.tifs.2007.06.001>.
2. L.M. Kandpal, J. Tewari, N. Gopinathan, P. Boulas, B.-K. Cho, In-process control assay of pharmaceutical microtablets using hyperspectral imaging coupled with multivariate analysis, *Anal. Chem.* 88 (2016) 11055–11061. doi:[10.1021/acs.analchem.6b02969](https://doi.org/10.1021/acs.analchem.6b02969).
3. P.R. Wahl, I. Pucher, O. Scheibelhofer, M. Kerschhaggl, S. Sacher, J.G. Khinast, Continuous monitoring of API content, API distribution and crushing strength after tableting via near-infrared chemical imaging, *Int. J. Pharm.* 518 (2017) 130–137. doi:<https://doi.org/10.1016/j.ijpharm.2016.12.003>.
4. K.B. Ferreira, A.G.G. Oliveira, A.S. Gonçalves, J.A. Gomes, Evaluation of hyperspectral imaging visible/near infrared spectroscopy as a forensic tool for automotive paint distinction, *Forensic Chem.* 5 (2017) 46–52. doi:<https://doi.org/10.1016/j.forc.2017.06.001>.
5. F. Zapata, F.E. Ortega-Ojeda, C. García-Ruiz, Revealing the location of semen, vaginal fluid and urine in stained evidence through near infrared chemical imaging, *Talanta*. 166 (2017) 292–299. doi:<https://doi.org/10.1016/j.talanta.2017.01.086>.
6. P. Mishra, M.S.M. Asaari, A. Herrero-Langreo, S. Lohumi, B. Diezma, P. Scheunders, Close range hyperspectral imaging of plants: A review, *Biosyst. Eng.* 164 (2017) 49–67. doi:<https://doi.org/10.1016/j.biosystemseng.2017.09.009>.

7. M. Wahabzada, A.-K. Mahlein, C. Bauckhage, U. Steiner, E.-C. Oerke, K. Kersting, Plant phenotyping using probabilistic topic models: Uncovering the hyperspectral language of plants, *Sci. Rep.* 6 (2016) 22482. doi:10.1038/srep22482.
8. A.A. Gowen, Y. Feng, E. Gaston, V. Valdramidis, Recent applications of hyperspectral imaging in microbiology, *Talanta.* 137 (2015) 43–54. doi:https://doi.org/10.1016/j.talanta.2015.01.012.
9. H. Grahn, P. Geladi, Techniques and applications of hyperspectral image analysis, John Wiley & Sons, 2007.
10. J.M. Amigo, H. Babamoradi, S. Elcoroaristizabal, Hyperspectral image analysis. A tutorial, *Anal. Chim. Acta.* 896 (2015) 34–51. doi:https://doi.org/10.1016/j.aca.2015.09.030.
11. A. Karami, R. Heylen, P. Scheunders, Band-specific shearlet-based hyperspectral image noise reduction, *IEEE Trans. Geosci. Remote Sens.* 53 (2015) 5054–5066. doi:10.1109/TGRS.2015.2417098
12. M. Vidal, J.M. Amigo, Pre-processing of hyperspectral images. Essential steps before image analysis, *Chemom. Intell. Lab. Syst.* 117 (2012) 138–148. doi:https://doi.org/10.1016/j.chemolab.2012.05.009.
13. A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures., *Anal. Chem.* 36 (1964) 1627–1639. doi:10.1021/ac60214a047.
14. J. Steinier, Y. Termonia, J. Deltour, Smoothing and differentiation of data by simplified least square procedure, *Anal. Chem.* 44 (1972) 1906–1909. doi:10.1021/ac60319a045.

15. P. Mishra, A. Herrero-Langreo, P. Barreiro, J.M. Roger, B. Diezma, N. Gorretta, L. Lleó, Detection and quantification of peanut traces in wheat flour by near infrared hyperspectral imaging spectroscopy using principal-component analysis, *J. Near Infrared Spectrosc.* 23 (2015) 15–22. doi:10.1255/jnirs.1141.
16. P. Mishra, C.B.Y. Cordella, D.N. Rutledge, P. Barreiro, J.M. Roger, B. Diezma, Application of independent components analysis with the JADE algorithm and NIR hyperspectral imaging for revealing food adulteration, *J. Food Eng.* 168 (2016) 7–15.
17. D.L. Donoho, De-noising by soft-thresholding, *IEEE Trans. Inf. Theory.* 41 (1995) 613–627.
18. G. Chen, S.-E. Qian, Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage, *IEEE Trans. Geosci. Remote Sens.* 49 (2011) 973–980. doi:10.1109/TGRS.2010.2075937
19. A. Karami, M. Yazdi, A.Z. Asli, Noise reduction of hyperspectral images using kernel non-negative tucker decomposition, *IEEE J. Sel. Top. Signal Process.* 5 (2011) 487–493. doi:10.1109/JSTSP.2011.2132692
20. R. Heylen, D. Burazerovic, P. Scheunders, Fully constrained least squares spectral unmixing by simplex projection, *IEEE Trans. Geosci. Remote Sens.* 49 (2011) 4112–4122. doi : 10.1109/TGRS.2011.2155070
21. T. Hobza, L. Pardo, I. Vajda, Robust median estimator in logistic regression, *J. Stat. Plan. Inference.* 138 (2008) 3822–3840. doi:<https://doi.org/10.1016/j.jspi.2008.02.010>.



22. G. Easley, D. Labate, W.-Q. Lim, Sparse directional image representations using the discrete shearlet transform, *Appl. Comput. Harmon. Anal.* 25 (2008) 25–46. doi:<https://doi.org/10.1016/j.acha.2007.09.003>.
23. G. Kutyniok, K. Guo, D. Labate, Sparse multidimensional representations using anisotropic dilation and shear operators. *Wavelets und Splines (Athens, GA, 2005)*, G. Proceedings of the International Conference on the Interactions between Wavelets and Splines, Athens, GA, May 16-19, 2005.
24. W.-Q. Lim, The discrete shearlet transform: A new directional transform and compactly supported shearlet frames, *IEEE Trans. Image Process.* 19 (2010) 1166–1180. doi:[10.1109/TIP.2010.2041410](https://doi.org/10.1109/TIP.2010.2041410)
25. X. Liu, M. Tanaka, M. Okutomi, Single-image noise level estimation for blind denoising, *IEEE Trans. Image Process.* 22 (2013) 5226–5237. doi:[10.1109/TIP.2013.2283400](https://doi.org/10.1109/TIP.2013.2283400)
26. N. Panigrahi, C.S. Bhol, B.S. Das, Rapid assessment of black tea quality using diffuse reflectance spectroscopy, *J. Food Eng.* 190 (2016) 101–108. doi:<https://doi.org/10.1016/j.jfoodeng.2016.06.020>.

## Chapter 6 : Application of a spatio-spectral total variation method for automatic correction for different types of noise in close-range hyperspectral images

Puneet Mishra<sup>1</sup>, Hemant Kumar Aggarwal<sup>2</sup>, Alison Nordon<sup>1</sup>

<sup>1</sup>*WestCHEM, Department of Pure and Applied Chemistry and Centre for Process Analytics and Control Technology, University of Strathclyde, Glasgow, G1 1XL, United Kingdom*

<sup>2</sup>*Department of Electrical and Computer Engineering, University of Iowa, IA 52242, USA*

**Paper Published as:** To be submitted to Applied Spectroscopy or any chemometric journal

### **Contribution:**

- Performed HSI NIR measurements
- Conducted chemometrics analysis of data
- Wrote the first draft of article
- Corrected article based on comments of co-authors

### 6.1 Abstract

Close-range hyperspectral (HS) images may contain several types of noise such as Gaussian, salt-and-pepper, and dead striped lines. Noise can originate from a number of sources including detector sensitivity, and fluctuations in the power supply, dark current, external illumination, and experimental conditions. The noise in data should be removed or reduced before performing any data modelling. In the present work, an automatic, general de-noising method is presented for dealing with different types of noise in close-range HS images. The method utilises a spatio-spectral total variation (SSTV) model to simultaneously de-noise both the spatial and spectral dimensions of

HS images. Noisy near-infrared (NIR) HSI data were simulated by adding different types of noise (Gaussian, salt-and-pepper or striped, or a combination of all 3 (mixed)) to a HS image of 10 different cut-tear-curl (CTC) grade black-tea products. The performance of the de-noising method was then evaluated by comparison of the de-noised data with the original HS image. Image planes and spectra extracted from the noisy HS images showed a reduction in the level of noise after de-noising. Upon de-noising, the peak signal-to-noise ratio (PSNR) increased typically by a factor of 2 – 3, and the structure similarity index measure (SSIM) increased to approximately 1 and 0.9 for HSI data that had one type and mixed types of noise added, respectively. There was also a significant improvement in the classification accuracy of support vector machine (SVM) models for different CTC grade black tea products after de-noising of the HS images. As the method was able to reduce the level of noise in HS images for the 3 types of noise considered both individually and in combination, it can be considered to be a general de-noising method. Furthermore, the method can readily be automated and has wide applicability for the de-noising of data from different modalities of chemical imaging where spatial and spectral correlation is present.

*Keywords:* total variation, optimisation, split-Bregman, de-noising, black tea

## 6.2 Introduction

Hyperspectral imaging (HSI) is a non-destructive technique that combines traditional imaging with spectroscopy to support rapid analysis of spatially-resolved spectral properties of materials. HSI was initially developed for remote-sensing applications and was limited to the visible (VIS) and near-infrared (NIR) spectral regions. However, it is now well-developed in the close-range domain and has been used across a variety of spectral techniques such as VIS [1], NIR [2,3], mid-infrared (MIR) [4], Raman [5], fluorescence [6], terahertz [7] and x-ray [8]. Furthermore, applications of HSI can be found in different fields of research such as plant science [9], food science [10], archaeology [11], medicine [12], forensics [13] and is further expanding.

Unlike traditional imaging, HSI combines imaging with spectroscopy. HSI generates data by recording the spectral signature for each pixel of the image resulting in a 3D

data cube. The first two dimensions of the 3D cube represent the spatial information and the third dimension represents the spectral information [9]. The acquisition of HS images is not as simple as the single shot (snapshot) acquisition method that is possible with traditional RGB imaging. Typically, HS image acquisition can be performed via two different scanning methods: the point scan and the line scan. The point scan configuration involves utilising a single element detector with movement of either the camera or the sample to generate a spatial representation of the spectral properties of the sample. On the other hand, the line scan configuration involves a multi-element array detector to cover a larger spatial area, thus, resulting in faster acquisition speeds compared to that achievable with the point scan modality; this means that the line scan system is usually preferred over the point scan system. However, the multi-element array detector may bring stripe noise, which can affect both the spatial and spectral information, leading to poor data modelling [14,15]. Furthermore, noise can originate from a number of different sources including fluctuations in the power supply, dark current, external illumination and experimental conditions. Hence, close-range HS images can contain different types of noise such as Gaussian, salt-and-pepper, and horizontal or vertical dead striped lines [16]. All of these different types of noise either need to be removed or reduced in magnitude before performing any data modelling otherwise the noise can substantially reduce the model performance and lead to low-quality results.

In the chemometrics field, there are the different methods that can reduce noise such as application of a piecewise smoothing function or median filtering in the spatial and spectral domains of HSI. However, many of these methods are manual and require inputs such as the window size and specification of the smoothing function by the user. Further, the choice of window size and smoothing function is often decided by visual interpretation, which can be time consuming. An automatic wavelength-specific shearlet-based noise reduction method was proposed by Mishra et al., (2019) [17] to overcome many of the issues associated with smoothing and filtering methods. The method, which operates in a push-button automatic sense, utilises the shearlet coefficient of the image planes to simultaneously reduce the noise present in the spatial and spectral domains. The method was able to successfully de-noise NIR-HSI data and resulted in an improvement in the accuracy of classification models for six different

commercial tea products. However, the noise considered were limited to Gaussian and spike, and the method was not designed to consider the stripe noise that can be present in HSI data. In comparison, Xu et al., (2017) [14] presented a chemometric-based method to deal with stripe noise. However, their method was not designed to deal with other types of noise such as Gaussian and salt-and-pepper. Therefore, an automatic, general method that can be used to de-noise different types of noise, both individually and in combination (i.e. mixed noise), is needed.

The aim of this work is to apply an automatic de-noising technique that can reduce or remove the most common types of noise encountered in close-range HSI, i.e., Gaussian, salt-and-pepper noise, and horizontal or vertical dead striped lines, from the raw reflectance hypercube. The methodology utilises a spatio-spectral total variation (SSTV) model to simultaneously de-noise both the spatial and spectral dimensions of close-range HS images, and has previously been used to de-noise remote-sensing data [16]. Noisy NIR-HSI data were simulated by adding different types of noise (Gaussian, salt-and-pepper or striped, or a combination of all 3 (mixed)) to a HS image of 10 different CTC grade black tea products. The performance of the de-noising method was assessed by visual inspection and quantified by estimating the peak signal-to-noise ratio (PSNR), the structure similarity index measure (SSIM) and classification accuracies of support vector machine (SVM) models.

## 6.3 Material and methods

### 6.3.1 Samples and imaging measurements

Ten different cut-tear-curl (CTC) grade black tea products were sourced in loose leaf from Unilever, Colworth Park, United Kingdom. The CTC is a method of processing black teas in which the leaves are passed through a series of cylindrical rollers with hundreds of sharp teeth that crush, tear and curl the tea into small and hard pellets. Samples were provided in sealed packaging and were stored at ambient temperature until the day of the imaging experiment. The tea samples were transferred into ten black plastic circular containers (diameter = 3.3 cm and depth = 1.3 cm) for analysis. Images were acquired with a push-broom line scan NIR-HSI camera (Model name: RED EYE 1.7) from INNO-SPEC (Nurnberg, Germany) operating in the spectral

range of 950 – 1700 nm. The camera utilises an InGaAs sensor and generates a spatial map of  $320 \times 256$  pixels, with a pixel size of  $30 \times 30 \mu\text{m}^2$  and a spectral resolution of 3.2 nm. Two 50 W halogen light sources, positioned at  $45^\circ$  to each other, were used to illuminate the samples. The samples were presented automatically to the camera with an independent stage motor system (Zolix TSA 200 BF). An image was acquired, using an integration time of 300 ms, of the ten black tea samples placed adjacent to each other in their respective containers in the field of view of the camera. Each image comprised more than 3000 pixels (spectra) for the individual tea samples. Prior to acquisition of an image of the tea samples, a set of white (Spectralon diffuse reflectance standard) and dark references were recorded for radiometric calibration. Four noisy HS images were simulated by adding different types of noise (Gaussian, spike, striped lines, or a combination of all 3 (mixed)) to the HS image of the ten black tea samples. Gaussian noise was added to the HS image utilising the *awgn* function from Matlab (R2016b, Mathworks, Natick, USA) with a signal-to-noise ratio of 20. Salt-and-pepper noise was added utilising the *imnoise* function from Matlab with a noise density of 0.05. Striped noise was added by manually setting 15 columns and 15 rows to zero intensity in 15 randomly selected wavelengths; the columns and rows were selected at random.

### 6.3.2 Data pre-processing

The effects of variation in the illumination intensity, the detector sensitivity and transmission properties of the optics were corrected by radiometric calibration utilising dark and white reference images. The correction was performed for every pixel in the HS image according to equation (6.1):

$$I_R(i,j,k) = \frac{I_{raw}(i,j,k) - I_{dark}(i,j,k)}{I_{white}(i,j,k) - I_{dark}(i,j,k)} \quad (6.1)$$

where,  $I_R$  is the calibrated reflectance,  $I_{raw}$  is the raw intensity measured from the test sample,  $I_{dark}$  is the intensity of the dark response,  $I_{white}$  is the intensity of the uniform white reference, and  $i$  and  $j$  are spatial coordinates and  $k$  is the wavelength in the image.

### 6.3.3 Spatio-spectral total variation model

A HSI hypercube of size  $n \times p \times q$ , where  $n \times p$  represents the spatial dimensions

and  $q$  represents the number of wavelengths, can be depicted as a matrix  $X = [x_1, x_2, \dots, x_q]$  where  $x_i$  is a vector of size  $np \times 1$  formed from the vertical concatenation of the columns of each  $n \times p$  image plane. The model for Gaussian and sparse noise can be given by equation (6.2):

$$Y = X + S + G \quad (6.2)$$

where  $X$  is the image in the absence of noise,  $Y$  is the noisy image,  $S$  is sparse noise and  $G$  is Gaussian noise. The term sparse noise is used here to describe all those types of noise that are only present at certain pixels such as salt-and-pepper and dead stripes. This is a general model used to formulate the HSI de-noising problem [16, 18]. To deal with the spatial and spectral correlation present in HS images, the images can be described by piecewise smoothing functions and can be modelled with TV regularisation. The TV of a greyscale image  $x$  is given by equation (6.3):

$$TV(x) = \|D_h x\|_1 + \|D_v x\|_1 \quad (6.3)$$

where  $D_h$  and  $D_v$  are the horizontal and vertical 2D finite differencing operators, respectively and  $\|\cdot\|_1$  denotes the 1-norm. The hyperspectral TV (HTV) model for a HS image  $X$  with  $q$  wavelengths is given by equation (6.4):

$$HTV(X) = \sum_{i=1}^q TV(x_i) \quad (6.4)$$

However, the HTV model only accounts for the spatial correlation in an image. Therefore, to deal with the spectral correlation that is also present in a HS image, the spatio-spectral total variation (SSTV) model [16] is used as shown in equation (6.5):

$$SSTV(X) = \|D_h X D\|_1 + \|D_v X D\|_1 \dots (6.5)$$

where  $D$  is a 1D finite differencing operator applied on the spectral signature of each pixel. This model explores the spatial and spectral correlation simultaneously. The de-noising problem with SSTV is given by equation (6.6):

$$\min_{X, S} \|Y - X - S\|_F^2 + \lambda \|S\|_1 + \mu SSTV(X) \quad (6.6)$$

where  $\|\cdot\|_F$  is the Frobenius norm, and  $\lambda$  and  $\mu$  are regularisation parameters. This

results in a high-dimensional non-differentiable optimisation problem in terms of  $X$  and  $S$  as given by equation (6.7):

$$\min_{X,S} \|Y - X - S\|_F^2 + \lambda \|S\|_1 + \mu \|D_h X D\|_1 + \mu \|D_v X D\|_1 \quad (6.7)$$

Since  $X$  is not separable, the problem can be rewritten in a constrained form:

$$\min_{X,S} \|Y - X - S\|_F^2 + \lambda \|S\|_1 + \mu \|P\|_1 + \mu \|Q\|_1 \quad (6.8)$$

subject to  $P = D_h X D$  and  $Q = D_v X D$ . Equation (6.8) can be written as an unconstrained optimisation problem using a quadratic penalty function:

$$\begin{aligned} \min_{P,Q,X,S} \|Y - X - S\|_F^2 + \lambda \|S\|_1 + \mu \|P\|_1 + \mu \|Q\|_1 + \vartheta \|P - D_h X D\|_F^2 + \\ \vartheta \|Q - D_v X D\|_F^2 \end{aligned} \quad (6.9)$$

where  $\vartheta$  is the regularisation parameter. Equation (6.9) contains multiple regularisation terms, which can be solved using the split-Bregman approach [16, 19]. The Matlab code for the TV de-noising method can be accessed at [20]. To de-noise the HS image, the raw reflectance HS image was inserted into the algorithm in Matlab and the output was the de-noised image. The default parameters used for  $\lambda$ ,  $\mu$  and  $\vartheta$  were 0.1, 0.2 and 0.2, respectively.

#### 6.3.4 Quality parameters

The quality of the de-noised HS images was quantified using the peak signal-to-noise ratio (PSNR) and the structure similarity index measure (SSIM). The PSNR and SSIM values for de-noised images were calculated with respect to the corresponding image planes in the absence of noise. The PSNR calculation is given by equation (6.10):

$$PSNR = 10 \log_{10} \left( \frac{peakval^2}{MSE} \right) \quad (6.10)$$

where *peakval* is either specified by the user or selected from a range that is dependent on the image datatype (e.g. 255 for a uint8 image). *MSE* is the mean square error between the de-noised image plane and the corresponding image in the absence of noise.



The estimation of SSIM [20] is based on the computation of three terms, namely a luminance term ( $l$ ), a contrast term ( $c$ ) and a structural term ( $s$ ). The overall index is a multiplicative combination of the three terms as shown in equation (6.11):

$$SSIM(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma \quad (6.11)$$

where

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_2}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

and  $\mu_x$  and  $\mu_y$  are the local means, and  $\sigma_x$  and  $\sigma_y$  are the standard deviations of images  $x$  (reference image) and  $y$  (chosen image plane), respectively,  $\sigma_{xy}$  is the cross-covariance for images  $x$  and  $y$ ,  $\alpha$ ,  $\beta$  and  $\gamma$  are exponent terms, which were set to 1, and  $C_1 = (k_1L)^2$ ,  $C_2 = (k_2L)^2$  and  $C_3 = C_2/2$  where  $k_1 = 0.01$ ,  $k_2 = 0.03$  and  $L = 255$ .

### 6.3.5 Classification with support vector machines

Classification of the 10 black tea products was performed using multi-class error correcting output code (ECOC) models containing SVM binary learners, which used a one-versus-one coding design. The data were transformed using a quadratic kernel. For every black tea sample, spectra were extracted from 200 pixels, which were selected at random from the image collected, leading to 2000 pixels in total for calibration of the classification models. The models were cross-validated with a 10-fold cross-validation method. Furthermore, this whole calibration procedure was performed with 100 iterations, and the mean validation accuracy and standard deviation were recorded. The trained classifiers were later used to generate the classification maps for the black tea samples contained in the image, which comprised more than 3000 pixels per sample. The ECOC-SVM models were implemented in Matlab using the Statistics and Machine Learning Toolbox (R2016b).

## 6.4 Results

### 6.4.1 De-noising

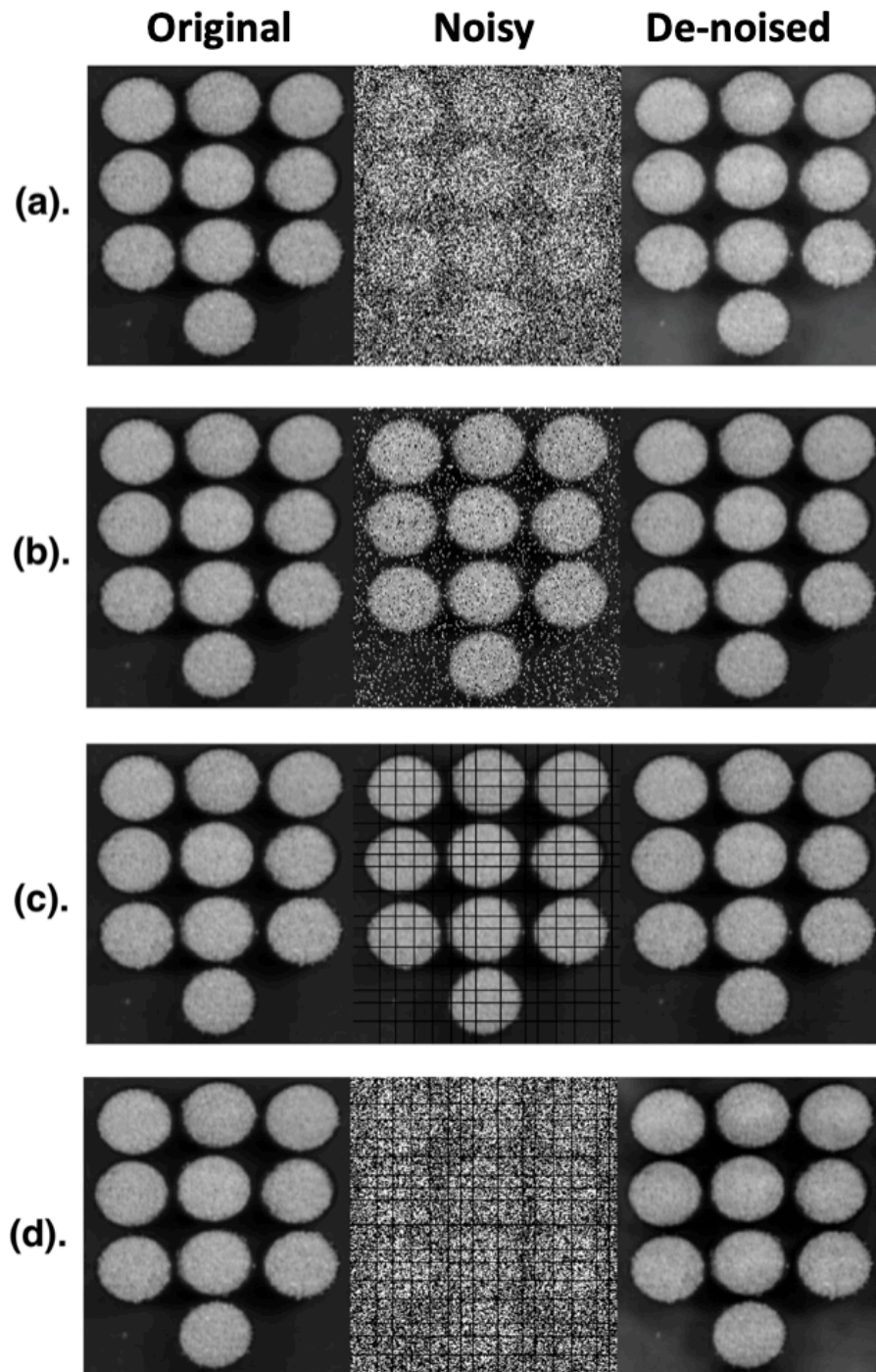
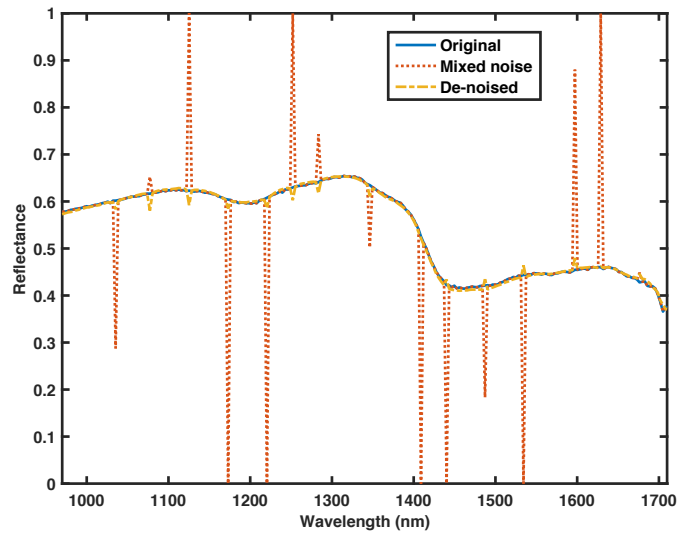


Figure 6.1: Greyscale images produced using the image plane at 1035 nm containing different types of noise: (a) Gaussian, (b) salt and pepper, (c) dead stripe, and (d) a combination of Gaussian, salt and pepper, and dead stripe noise. In each case, the

*original, noisy and de-noised images are given which can be understood as mentioned at the top in the figure.*

Figure 6.1 (a) to (d) show greyscale images produced using the image plane at 1035 nm (selected to allow visualisation of the data hypercubes) with no noise added (original), after the addition of different types of noise and the corresponding de-noised images. Four different noise cases were considered: Gaussian, salt-and-pepper, dead stripe lines, and a mixture of Gaussian, salt-and-pepper and dead stripe lines. It can be seen that the de-noising method was able to reduce the level of noise present in the data for all types of noise, with the appearance of the de-noised images comparable to that of the original image plane, i.e., with no noise added. Figure 6.2 presents a sample spectrum extracted from the original (blue), noisy (mixed noise; red) and de-noised (yellow) images from a pixel with x and y co-ordinates of 108 and 29, respectively. The noise was added by randomly selecting 15 different wavelengths, and their presence can be seen in the red trace in Figure 6.2 corresponding to a spectrum from the noisy image. After de-noising the spectrum was very similar to the original spectral profile, showing the potential of the de-noising method. However, the de-noised spectrum still contains some small peaks at the wavelengths where the noise was added; these peaks can be removed automatically using a filter operation.



*Figure 6.2: Sample spectra extracted from the original, mixed noise and de-noised images from the pixel with  $x$  and  $y$  co-ordinates of 108 and 29, respectively.*

### 6.4.2 PSNR and SSIM

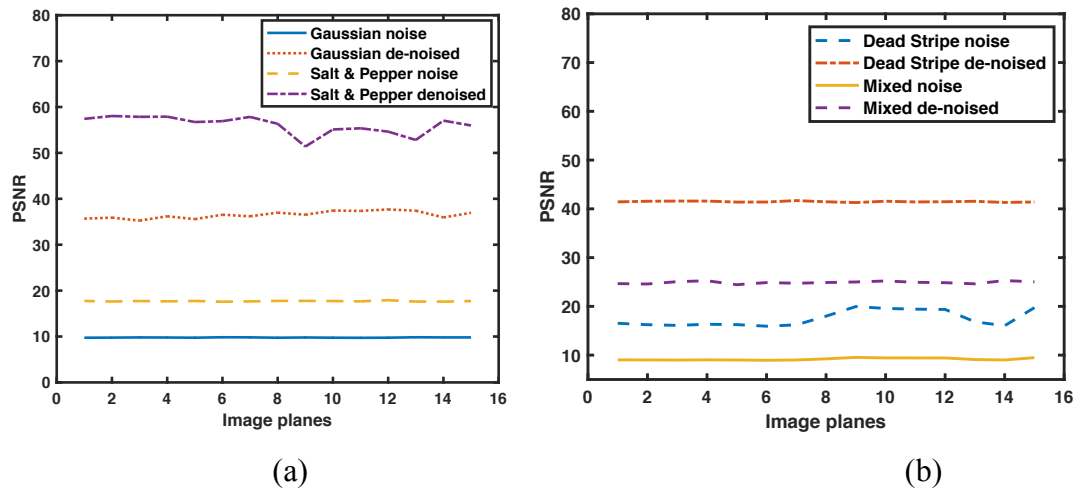


Figure 6.3: PSNR for noisy and de-noised images. (a) Gaussian and salt and pepper noise, and (b) dead stripe and mixed noise.

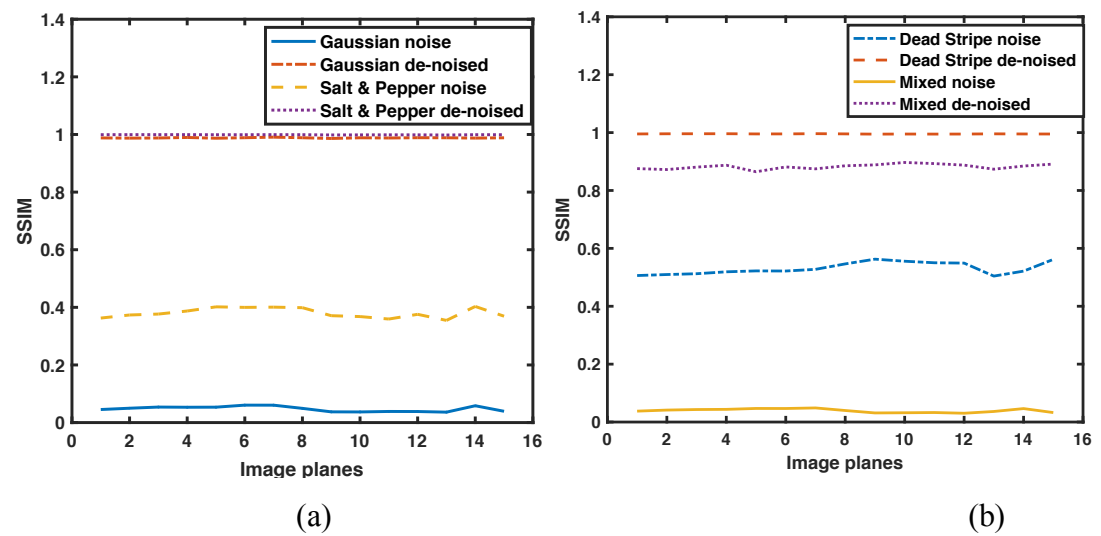


Figure 6.4: SSIM for noisy and de-noised images. (a) Gaussian and salt and pepper noise, and (b) dead stripe and mixed noise.

Figures 6.3 and 6.4 present the PSNR and SSIM, respectively, for the 15 image planes with noise added before and after de-noising with respect to the original hypercube (i.e. in the absence of noise). It can be seen in Figure 6.4 that the PSNR was lowest for mixed noise followed by Gaussian, salt-and-pepper and dead stripe lines. The image

to which mixed noise had been added had a lower PSNR as it contained more noise compared to the images containing one type of noise, i.e., Gaussian, salt-and-pepper and dead stripe lines. However, after de-noising the salt-and-pepper data resulted in the highest PSNR for all the 15 image planes followed by the dead stripe, Gaussian and mixed noise data. Similarly, for the SSIM, the mixed noise data had the lowest SSIM for all the 15 image planes followed by the Gaussian, salt-and-pepper, and striped noise. The SSIM increased after de-noising of images containing all types of noise considered; the highest SSIM was obtained for the de-noised images to which salt-and-pepper and dead stripe noise had been added followed by Gaussian and mixed noise.

### 6.4.3 Classification

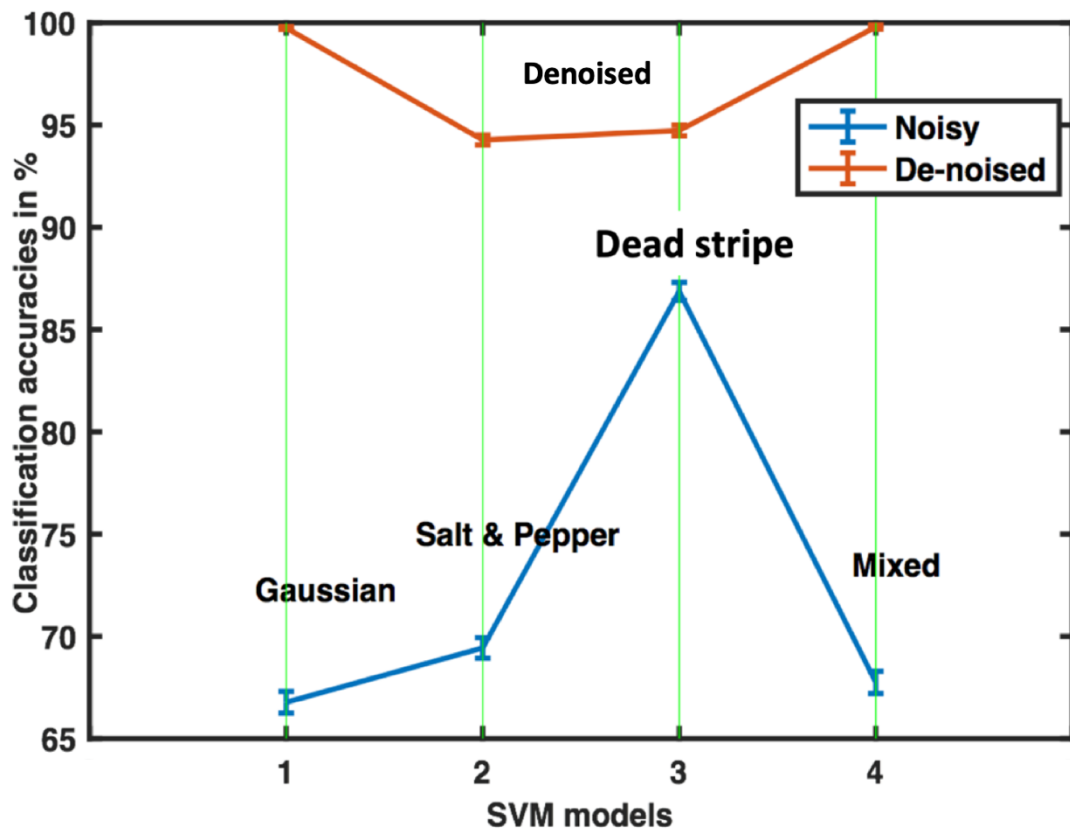


Figure 6.5: Classification accuracies of SVM models for noisy and de-noised data.

Figure 6.5 presents the classification accuracies of the SVM models developed for noisy and de-noised data for classifying ten black tea samples. A classification

accuracy of  $87.81 \pm 0.35$  % was obtained for the clean data alone without any pre-processing. In Figure 6.5, it can be seen that the SVM model accuracies were affected when constructed using images containing Gaussian, salt-and-pepper, and mixed noise resulting in model accuracies around 70%. However, the model accuracy was around 90% for a SVM model constructed from images containing dead stripe noise. A reason for high model accuracy of dead stripe noise could be due to less noise effected pixels considered during the calibration modelling. De-noising resulted in an improvement in the accuracies of all models from  $66.77 \pm 0.55\%$ ,  $69.43 \pm 0.50\%$ ,  $86.87 \pm 0.44\%$  and  $67.75 \pm 0.55\%$  to  $99.75 \pm 0.06$  %,  $94.27 \pm 0.24\%$ ,  $94.73 \pm 0.26$  % and  $99.80 \pm 0.09\%$  for Gaussian, salt-and-pepper, dead stripe and mixed noise data, respectively. The accuracies quoted are the mean  $\pm$  one standard deviation resulting from 100 iterations of a 10-fold cross validated model).

## 6.5 Conclusions

The data resulting from HSI contains different types of noises such as Gaussian, salt-and-pepper and dead striped lines. The noise affects both the spatial and spectral dimensions, leading to lower model accuracies for, for example, classification of samples. The removal of noise is important for the development of accurate and robust models. However, removing one particular type of noise is not the solution as other remaining noise also affects the data modelling. SSTV de-noising is a general method and is capable of reducing different kinds of noise that may be present in close-range HS images. The method explores the structure by utilising the 2D total variation in the spatial domain and the 1D spectral variation in the spectral domain. The results showed that the accuracies of the SVM models were significantly improved by de-noising. The de-noising method is fully automatic and has wide applicability for the de-noising of data from different modalities of chemical imaging where spatial and spectral correlation is present.

## 6.6 References

1. H. Lee, M.S. Kim, W.-H. Lee, B.-K. Cho, Determination of the total volatile basic nitrogen (TVB-N) content in pork meat using hyperspectral fluorescence imaging, *Sensors Actuators B Chem.* 259 (2018) 532–539. doi:<https://doi.org/10.1016/j.snb.2017.12.102>.
2. C. Wakholi, L.M. Kandpal, H. Lee, H. Bae, E. Park, M.S. Kim, C. Mo, W.-H. Lee, B.-K. Cho, Rapid assessment of corn seed viability using short wave infrared line-scan hyperspectral imaging and chemometrics, *Sensors Actuators B Chem.* 255 (2018) 498–507. doi:<https://doi.org/10.1016/j.snb.2017.08.036>.
3. S. Shrestha, M. Knapič, U. Žibrat, L.C. Deleuran, R. Gislum, Single seed near-infrared hyperspectral imaging in determining tomato (*Solanum lycopersicum* L.) seed quality in association with multivariate data analysis, *Sensors Actuators B Chem.* 237 (2016) 1027–1034. doi:<https://doi.org/10.1016/j.snb.2016.08.170>.
4. F. Rosi, C. Miliani, R. Braun, R. Harig, D. Sali, B.G. Brunetti, A. Sgamellotti, Noninvasive Analysis of Paintings by Mid-infrared Hyperspectral Imaging, *Angew. Chemie Int. Ed.* 52 (2013) 5258–5261. doi:[10.1002/anie.201209929](https://doi.org/10.1002/anie.201209929).
5. L.M. Kandpal, B.-K. Cho, J. Tewari, N. Gopinathan, Raman spectral imaging technique for API detection in pharmaceutical microtablets, *Sensors Actuators B Chem.* 260 (2018) 213–222. doi:<https://doi.org/10.1016/j.snb.2017.12.178>.
6. H. Lee, M.S. Kim, W.-H. Lee, B.-K. Cho, Determination of the total volatile basic nitrogen (TVB-N) content in pork meat using hyperspectral fluorescence imaging, *Sensors Actuators B Chem.* 259 (2018) 532–539. doi:<https://doi.org/10.1016/j.snb.2017.12.102>.



7. A. Doria, G.P. Gallerano, E. Giovenale, A. Casini, C. Cucci, M. Picollo, M. Poggesi, L. Stefani, K. Fukunaga, M. Tamassia, Vis-NIR Hyperspectral and Terahertz Imaging Investigations on a Fresco Painting on “Tavella” by Alessandro Gherardini, *J. Infrared, Millimeter, Terahertz Waves*. 38 (2017) 390–402. doi:10.1007/s10762-017-0357-2.
8. S. Pani, S.C. Saifuddin, F.I.M. Ferreira, N. Henthorn, P. Seller, P.J. Sellin, P. Stratmann, M.C. Veale, M.D. Wilson, R.J. Cernik, High Energy Resolution Hyperspectral X-Ray Imaging for Low-Dose Contrast-Enhanced Digital Mammography, *IEEE Trans. Med. Imaging*. 36 (2017) 1784–1795. doi:10.1109/TMI.2017.2706065.
9. P. Mishra, M.S.M. Asaari, A. Herrero-Langreo, S. Lohumi, B. Diezma, P. Scheunders, Close range hyperspectral imaging of plants: A review, *Biosyst. Eng.* 164 (2017) 49–67. doi:https://doi.org/10.1016/j.biosystemseng.2017.09.009.
10. Y. Liu, H. Pu, D.-W. Sun, Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications, *Trends Food Sci. Technol.* 69 (2017) 25–35. doi:https://doi.org/10.1016/j.tifs.2017.08.013.
11. M. Doneus, G. Verhoeven, C. Atzberger, M. Wess, M. Ruš, New ways to extract archaeological information from hyperspectral pixels, *J. Archaeol. Sci.* 52 (2014) 84–96. doi:https://doi.org/10.1016/j.jas.2014.08.023.
12. G. Lu, B. Fei, Medical hyperspectral imaging: a review., *J. Biomed. Opt.* 19 (2014) 10901. doi:10.1117/1.JBO.19.1.010901.

13. G.J. Edelman, E. Gaston, T.G. van Leeuwen, P.J. Cullen, M.C.G. Aalders, Hyperspectral imaging for non-contact analysis of forensic traces, *Forensic Sci. Int.* 223 (2012) 28–39. doi:<https://doi.org/10.1016/j.forsciint.2012.09.012>.
14. J.-L. Xu, D.-W. Sun, A.A. Gowen, Chemometric methods applied in the image plane to correct striping noise in hyperspectral chemical images of biomaterials, *J. Chemom.* (2017) e2986. doi:10.1002/cem.2986.
15. A.A. Gowen, M.R. Dorrepaal, Multivariate Chemical Image Fusion of Vibrational Spectroscopic Imaging Modalities, *Mol.* 21 (2016). doi:10.3390/molecules21070870.
16. H.K. Aggarwal, A. Majumdar, Hyperspectral Image Denoising Using Spatio-Spectral Total Variation, *IEEE Geosci. Remote Sens. Lett.* 13 (2016) 442–446. doi:10.1109/LGRS.2016.2518218.
17. P. Mishra, A. Karami, A. Nordon, D.N. Rutledge, J-M. Roger, Automatic denoising of close-range hyperspectral images with a wavelength-specific shearlet-based image noise reduction method, *Sensors and Actuators B: Chemical* 281 (2019) 1034-1044. doi.org/10.1016/j.snb.2018.11.034.
18. H. Zhang, W. He, L. Zhang, H. Shen, Q. Yuan, Hyperspectral Image Restoration Using Low-Rank Matrix Recovery, *IEEE Trans. Geosci. Remote Sens.* 52 (2014) 4729–4743. doi:10.1109/TGRS.2013.2284280.
19. T. Goldstein, S. Osher, The Split Bregman Method for L1-Regularized Problems, *SIAM J. Imaging Sci.* 2 (2009) 323-343. doi:10.1137/080725891.
20. H. K. Aggarwal, Mixed Noise Reduction, 2015. [Online]. Available: <http://www.mathworks.com/matlabcentral/fileexchange/49145-mixed-noise-reduction>, accessed 24<sup>th</sup> December 2018.

21. Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Trans. Image Process.* 13 (2004) 600–612. doi:10.1109/TIP.2003.819861.

## Chapter 7 : Fusing spectral and textural information in near-infrared hyperspectral imaging to improve green tea classification modelling

Puneet Mishra<sup>1\*</sup>, Alison Nordon<sup>1\*</sup>, Mohd Shahrime Mohd Asaari<sup>2</sup>, Guoping Lian<sup>3,4</sup>, Sally Redfern<sup>3</sup>

<sup>1</sup>*WestCHEM, Department of Pure and Applied Chemistry and Centre for Process Analytics and Control Technology, University of Strathclyde, Glasgow, G1 1XL, United Kingdom*

<sup>2</sup>*Vision Lab, Department of Physics, Campus Drie Eiken, University of Antwerp, Edegemsesteenweg 200-240, 2610, Antwerp, Belgium*

<sup>3</sup>*Unilever R&D Colworth, Colworth Science Park, Sharnbrook, Bedford, MK44 1LQ, United Kingdom*

<sup>4</sup>*Department of Chemical and Process Engineering, University of Surrey, Guildford GU2 7XH, United Kingdom*

**Paper Published as:** P. Mishra, A. Nordon, M.S. Mohd Asaari, G. Lian, S. Redfern, Fusing spectral and textural information in near-infrared hyperspectral imaging to improve green tea classification modelling, *J. Food Eng.* 249 (2019) 40–47. doi:<https://doi.org/10.1016/j.jfoodeng.2019.01.009>.

### **Contribution:**

- Performed HSI NIR measurements
- Conducted chemometrics analysis of data
- Wrote the first draft of article
- Corrected article based on comments of co-authors

## 7.1 Abstract

Hyperspectral imaging (HSI) can acquire data in two modes: imaging and spectroscopy, revealing the spatially-resolved spectral properties of materials. Traditional HSI processing in the close-range domain primarily focuses on the spectral information with minimal utilisation of the spatial information present in the data. The present work describes a methodology for utilising the spatial information present in HSI data to improve classification modelling over that achievable with spectral information alone. The methodology has been evaluated using near infrared (NIR) HSI data of sixteen green tea products from seven different countries. The methodology involves selecting and sharpening an image plane to enhance the textural details. The textural information is then extracted from the statistical properties of the grey level co-occurrence matrix (GLCM) of the sharpened image plane using a moving window operation. Finally, the textural properties are combined with the spectral information using one of the three different levels of data fusion, i.e. raw data level, feature level and decision level. Raw data-level fusion involved concatenating the spectral and textural data before performing the classification task. The feature-level fusion involved performing principal component analysis (PCA) on spectral and textural information and combining the PC scores obtained prior to performing classification. Decision-level fusion involved a majority voting scheme to enhance the final classification maps. All the classification tasks were performed using multi-class support vector machine (SVM) models. The results showed that combining the textural and spectral information during modelling resulted in improved classification of the sixteen green tea products compared to models built using spectral or textural information alone.

*Keywords:* chemical imaging; texture; support vector machine (SVM); grey level co-occurrence matrix (GLCM); data fusion; green tea.

## 7.2 Introduction

Computer vision and image processing have benefited from the exploration of spatially-resolved physical properties of materials in analytical chemistry [1]. The combination of imaging with spectroscopy, known as hyperspectral imaging (HSI), has complemented imaging by allowing simultaneous exploration of spatial and spectral properties of materials in a fast and non-destructive way. Although HSI was primarily developed for remote sensing [2], it is now a well-established technique in close-range laboratory settings [3, 4, 5, 6]. HSI has been used for the study of a wide range of food products such as wheat flour [7], olive oil [8], herbal tea [9], seeds [10], coffee [11], beans [12] and many more [13].

The information generated by HSI takes the form of hypercubes where the first two dimensions represent the spatial information of the imaged scene and the third dimension adds the spectral information to the pixels [12]. The extraction of meaningful information from the hypercube requires advanced pattern recognition and data modelling. Although, HSI data is rich in information, not all the information present is needed to perform the data modelling. The traditional HSI processing approach includes selection of the region of interest (ROI) over the image plane to extract the relevant spectra. The selected spectra are then used to perform different types of modelling such as data visualisation, regression, and classification. The models developed are used to predict the scores for each pixel to represent prediction or classification maps [14]. This modelling approach aids in visualising the spatial distribution of the predicted values or classes. However, the complementary information present in the spatial domain, e.g., texture, is not generally used in the construction of calibration models based on spectra [15]. In the predecessor of close-range HSI, i.e. remote sensing, the importance of information present in the spatial domain of HSI is well realised. In particular, utilising the spatial information to improve classification modelling is widely employed [16]. The spatial information can be used either pre or post-classification modelling to improve the classification accuracies and classification maps.

There are some extra benefits to the application of HSI in close-range settings, compared to the remote-sensing domain, which further motivates the use of spatial

information. One of the benefits is the high spatial resolution of the images, which reduces the number of mixed pixels in the imaged scene leading to improved image quality. The other is the artificial dark-field illumination used to enhance the contrast of regions where illumination interferes with the edges, scratches, imprints, slots and elevations over the imaging scene, leading to detailed information about the physical features of samples [17]. The spatial information that is primarily of interest in the case of close-range HSI is textural. Texture can be understood as a quantitative measure of the arrangement of intensities in a region [18]. Therefore, it is necessary to calculate texture from statistical analysis of an image plane. There are different ways of extracting textural information from an image plane. Estimating the grey level co-occurrence matrices (GLCMs) has gained widespread interest in the close-range HSI processing domain [19, 20, 21, 22, 23, 24, 25]. A reason for its popularity is that the statistical properties extracted from GLCMs can be used to represent, compare and classify texture. Since the GLCM-based texture calculation can only be performed on a monochromatic image, an image at a single wavelength is usually selected from the HS image and subjected to GLCM analysis [20, 24]. Furthermore, utilising textural information in conjunction with spectral information can be realised in a data fusion approach to combine the two types of information at three different levels, i.e., low, middle and high. The low-level data fusion of spectral and textural information utilises the spectral and textural data in raw form and performs concatenation of the data matrices before the data modelling. Mid-level fusion involves doing some feature transformation prior to performing the fusion such as utilising principal component analysis (PCA) to capture the most important variation in the feature vector and later concatenating the scores obtained for the corresponding features. High-level involves decision-level fusion where the output from different models is usually fused based on some decision criteria to enhance the final output such as classification maps.

The aim of this work is to present a methodology for fusing spectral and textural information to improve the modelling of near-infrared (NIR) HSI data. To demonstrate the potential of fusing textural and spectral information, the classification of sixteen green tea products from seven different countries was considered. High-quality green tea products are mainly characterised by the flavour that they impart, which involves

two primary sensory perceptions, i.e. taste and aroma. The distinct taste and aroma of any tea product are derived from its geographical origin as they are unique to the climate and soil conditions in which the plants were grown. Typically, discrimination of green tea products via sensory analysis is performed using an expert human panel. Sensory analysis involves assessment of tea products in leaf and/or extracted liquor form on the basis of appearance, colour, aroma and taste, along with the overall quality of the samples. However, distinguishing tea products based on sensory analysis is a time-consuming and expensive task as it requires an expert human panel. Furthermore, sensory analysis is subjective, and it can be inconsistent and unpredictable owing to physiological and psychological differences between tasters [26]. One more limitation is that the expert panel cannot be used as an on-line technique for grading of tea products [27]. In recent years, different analytical techniques have been explored for assessment of tea products of which HSI is one. NIR HSI, in comparison to visible HSI, provides access to the chemical information present in samples. NIR HSI has recently been used to discriminate between different types of tea products [28], although only the spectral information was used to build the classification models. However, leaf tea products also have a rich amount of textural detail present in their leaves; such textural information has previously been used to classify tea products [17, 29]. However, utilising texture alone is not a robust modelling solution as textural properties are affected by variations in illumination intensity [30]. Therefore, in this work, we utilise the textural information as supplementary information to enhance NIR spectroscopy-based classification of green tea products.

## 7.3 Material and methods

### 7.3.1 Samples

Sixteen green tea samples, differing in geographical origin, were sourced in loose-leaf form from Unilever R&D, Colworth Science Park, United Kingdom. All the samples were provided in sealed packaging and were stored at ambient temperature until analysis. All samples were green in colour and exhibited some textural differences owing to variations in the shape and size of the leaves. The sixteen samples originated from seven different countries: Argentina (one), South India (five), Sri Lanka (two),



China (two), Japan (two), Kenya (three) and Sumatra (one). Imaging experiments were performed by presenting the sample in a circular black plastic cap (diameter = 3.3 cm, depth = 1.3 cm). The sixteen tea samples were each analysed in a different cap to avoid any cross-contamination.

### 7.3.2 Hyperspectral imaging measurements

Imaging was performed with a push-broom line scan NIR HSI camera (*Model name: RED EYE 1.7*) from INNO-SPEC (Nurnberg, Germany). The camera has an InGaAs sensor and generates a spatial map of 320 x 256 pixels, and has pixel dimensions of 30 x 30  $\mu\text{m}^2$ . Images were acquired over the spectral range of 950 – 1765 nm with a spectral resolution of 3.2 nm. Two halogen light sources, each with a power of 50 W, were used to illuminate the samples. For image acquisition, the sixteen tea samples were placed on the translation stage, which was controlled via an independent stage motor system (Zolix TSA 200 BF). The speed of the translation stage, 2.5  $\text{mm s}^{-1}$ , was optimised using a checkerboard to avoid any distortion in the shape of the image arising from the overlapping of spectral and spatial information. The distance from the lens to the translation stage was 15 cm. Prior to acquisition of an image, a set of white (Spectralon diffuse reflectance standard) and dark references were recorded for radiometric calibration. Each image comprised more than 2000 pixels (spectra) per individual green tea sample and was acquired using an integration time of 300 ms.

### 7.3.3 Data analysis

#### 7.3.3.1 Image pre-processing

Variations in signal arising from illumination intensity, the detector sensitivity and the transmission properties of the optics were corrected by radiometric calibration utilising dark and white reference images. The correction was performed for every pixel in the HS image according to equation (7.1):

$$I_R(i,j,k) = \frac{I_{raw}(i,j,k) - I_{dark}(i,j,k)}{I_{white}(i,j,k) - I_{dark}(i,j,k)} \quad (7.1)$$

where,  $I_R$  is the calibrated reflectance,  $I_{raw}$  is the raw intensity measured from the test sample,  $I_{dark}$  is the intensity of the dark response,  $I_{white}$  is the intensity of the uniform

white reference, and  $i$  and  $j$  are spatial coordinates and  $k$  is the wavelength of the image. The spectral range of the hypercube was reduced from 950 – 1765 nm to 967.11 – 1700 nm to remove noise. A moving window Savitzky-Golay (SAVGOL) filter [31] (15-point width and second order polynomial) was applied to each pixel of the image to remove random noise, e.g. spikes, from spectra. Further, to reduce light scattering effects arising from inhomogeneity of the sample surface, the spectra were normalised using the standard normal variate (SNV) [32]. Smoothing and normalisation were performed using the *savgol* and *snv* functions, respectively, from PLS\_Toolbox (version 8.11, Eigenvector Research Inc., USA).

### 7.3.3.2 Texture estimation

#### 7.3.3.2.1 Selection of image plane

Textural analysis requires a single image plane to enable extraction of the GLCM properties. Since some spectral bands are noisy compared to others in HSI, the best image plane can be chosen on the basis of two different image quality parameters: the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM). The PSNR and SSIM were calculated with respect to the mean image plane (reference image), obtained from averaging the intensities of pixels along the spectral dimension. The PSNR can be calculated using equation (7.2):

$$PSNR = 10 \log_{10} \left( \frac{peakval^2}{MSE} \right) \quad (7.2)$$

where *peakval* is either specified by the user or selected from a range that is dependent on the image datatype (e.g. 255 for a uint8 image) and *MSE* is the mean square error between the chosen image plane and the reference image.

The SSIM [33] is based on the computation of three terms, namely the luminance term ( $l$ ), the contrast term ( $c$ ) and the structural term ( $s$ ). The overall index is a multiplicative combination of the three terms calculated by equation (7.3):

$$SSIM(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma \quad (7.3)$$

where

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_2}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

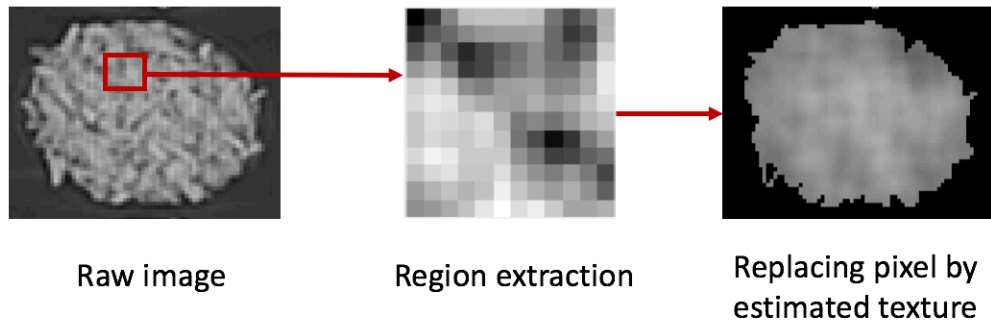
$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

and  $\mu_x$  and  $\mu_y$  are the local means, and  $\sigma_x$  and  $\sigma_y$  are the standard deviations of images  $x$  (reference image) and  $y$  (chosen image plane), respectively,  $\sigma_{xy}$  is the cross-covariance for images  $x$  and  $y$ ,  $\alpha$ ,  $\beta$  and  $\gamma$  are exponent terms, which were set to 1, and  $C_1 = (k_1L)^2$ ,  $C_2 = (k_2L)^2$  and  $C_3 = C_2/2$  where  $k_1 = 0.01$ ,  $k_2 = 0.03$  and  $L = 255$ . The best image plane was selected based on the maximum PSNR and SSIM.

#### 7.3.3.2.2 Sharpening of the image plane

The raw HS images obtained had soft edges owing to the limited focus and/or low spatial resolution of the camera resulting in low contrast between adjacent pixel intensities. Therefore, the image plane was sharpened to enhance the textural details. The enhanced textural details obtained with sharpening should result in more accurate calculation of the GLCM properties. Typically, the aim of sharpening is to increase the contrast along the edges where different colours meet. In the present work, the unsharp masking technique was used to perform image sharpening. This technique sharpens the image by first estimating a “blurred” negative image mask from the original image, which is then subtracted from the original image creating an image that is less blurry than the original [34]. Textural analysis was then performed on the sharpened image via estimation of the statistical properties of the GLCM.

#### 7.3.3.2.3 Estimating GLCM properties



*Figure 7.1: Schematic of the window operation performed for extracting textural features.*

The textural information of the image has variations in the greyscale as a function of spatial position. Different pixels in the image share spatial relationships in terms of greyscale intensities, which is spatial correlation. A common method to represent the relationship between greyscale pixels is via GLCMs [19, 20, 21, 22, 23, 24, 25]. The GLCM aims to describe the textural information present in the image by defining how often pairs of pixels with a specific value and spatial relationship occur in an image. The GLCM is a square matrix whose elements represent the probabilities of a pixel being at a distance from another pixel with a fixed spatial relationship. These values of the elements represent the conditional probabilities of all pairwise combinations of greyscale levels in the spatial window. Statistical measures can further be applied to these conditional probabilities to generate the textural properties. In the present work, twenty different statistical measures were estimated resulting in twenty different textural information maps. The twenty statistical properties considered were the correlation, autocorrelation, contrast, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, variance, sum average, sum variance, sum entropy, difference variance, difference entropy, two information measures of correlation, inverse moment difference, inverse difference normalised and inverse difference moment normalised. Further information on the use of statistical metrics for estimating textural properties can be found in [35, 36, 37]. In the present work, the GLCM estimation was performed utilising the *graycomatrix* command in Matlab (R2016b, Mathworks, USA). A square window with a size of  $11 \times 11$  pixels<sup>2</sup>, which was moved over the image plane (see Figure 7.1), was used for the GLCM estimation. The window

size was selected based on the number of pixels required to cover the largest tea leaves, and was an odd number to give equal coverage of the pixels around the centre pixel. In this process, the greyscale intensity of the centre pixel was replaced with the estimated textural property of the GLCM. To make the GLCM uniform around the exterior area of the sample, a patch mask was defined, which included replacing the individual pixel intensity values by their mean intensities. Textural analysis resulted in the calculation of 20 image planes corresponding to the 20 statistical metrics given above; all 20 textural image planes were used in subsequent analysis.

### 7.3.3.3 Feature transformation with PCA

In the present work, two PCA models were built to transform the spectral and textural information separately. The number of principal components was selected such that >99% of the variance in the data was retained. The PCA decomposition was performed in Matlab utilising the PLS\_Toolbox.

### 7.3.3.4 Data fusion scheme

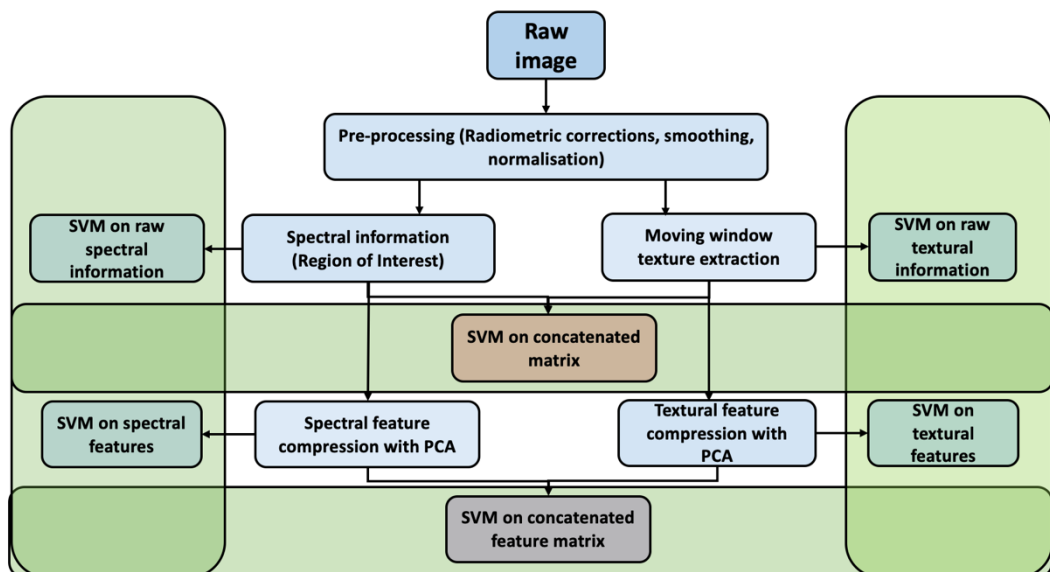


Figure 7.2: Schematic for raw data-level and feature-level fusion.

Once the 20 textural features were obtained from the data, the fusion of textural information with the spectral information was performed. The scheme for raw data-

level and feature-level fusion is depicted in Figure 7.2. Raw data-level fusion was performed by concatenating the texture with the spectral information. In the case of feature-level fusion, two separate PCA models were constructed to extract the relevant features from the spectral and textural cubes. The extracted features were then concatenated before performing the classification modelling. In the case of decision-level fusion, all the classification maps obtained from raw- and feature-level data fusion were used within a majority voting scheme and the final classification map was updated.

#### *7.3.3.5 Classification with support vector machines*

In the chemometrics domain, there are different methods to perform the classification of spectral features [38]. However, in the image processing domain the support vector machine (SVM) has gained popularity for the classification of fused spectral and textural information [39]. Classification of the 16 green tea products was performed using multi-class error correcting output code (ECOC) models containing SVM binary learners, using a one-versus-one coding design. High dimensional mapping of the data was performed using a quadratic kernel. For every green tea sample, spectra and/or textural information were extracted from 400 pixels, selected at random from the image, leading to 6400 pixels in total for the calibration of the classification models. The models were cross-validated with the 10-fold cross-validation method. This whole calibration procedure was performed with 100 iterations with the mean validation accuracy and standard deviation recorded. The trained classifiers were later used to generate the classification maps for the tea samples contained in the image, which comprised more than 2000 pixels per sample. The ECOC-SVM models were implemented in Matlab using the Statistics and Machine Learning Toolbox (R2016b).

## 7.4 Results

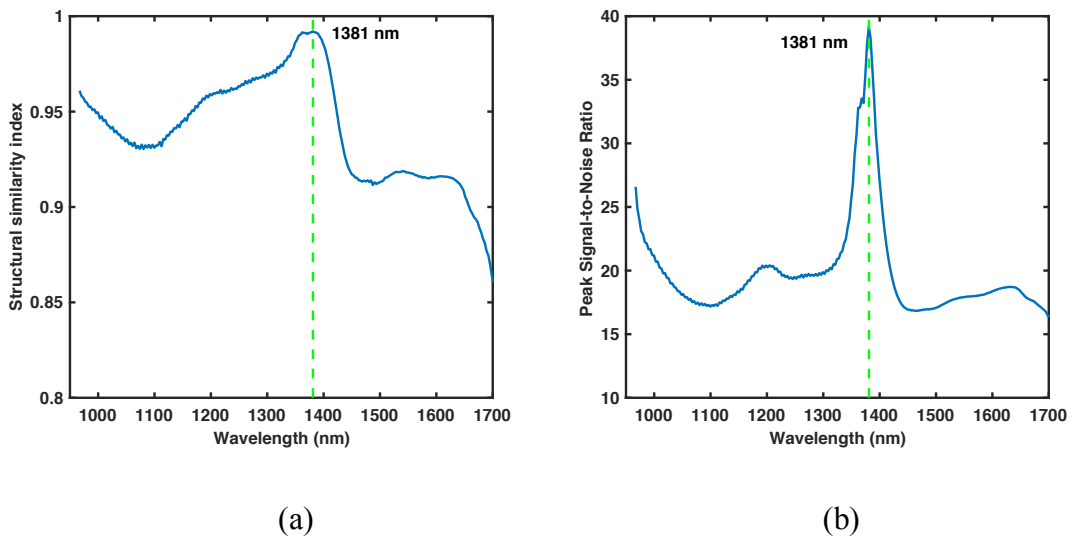
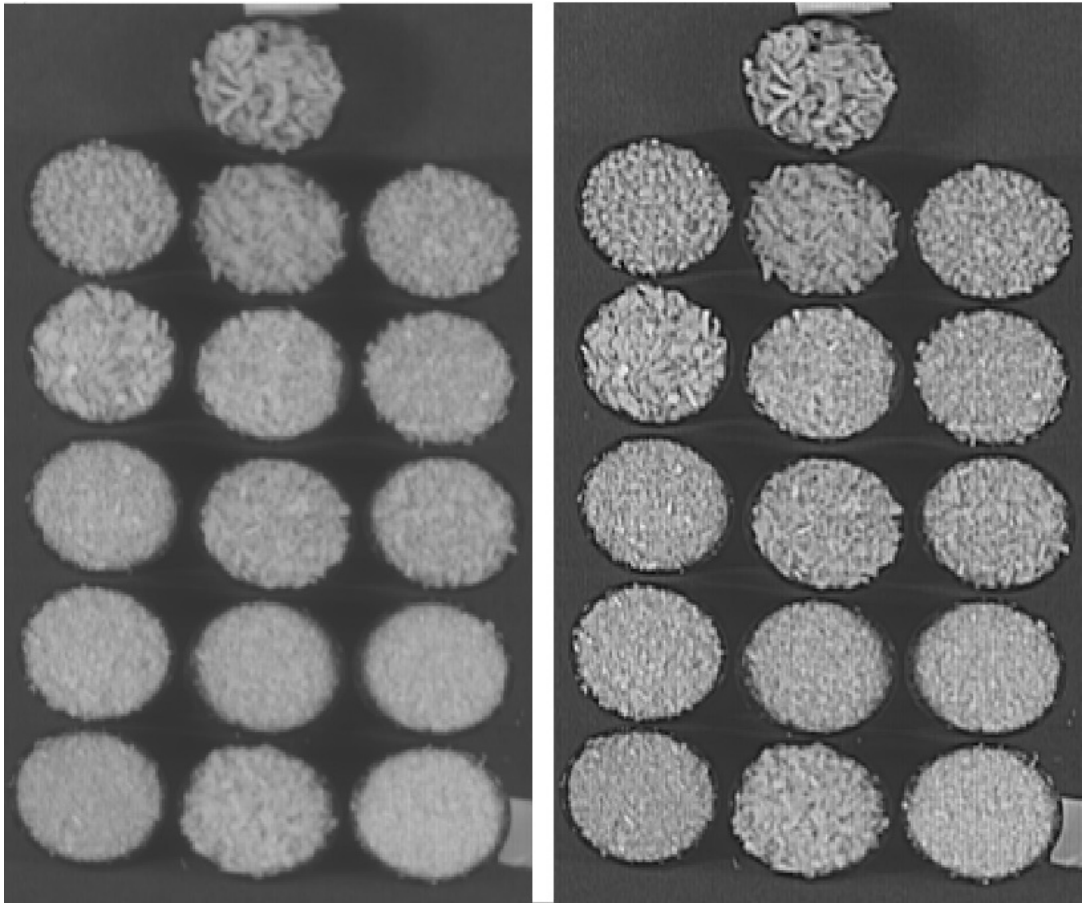


Figure 7.3: Criteria used for selection of the best image plane on which to perform sharpening and textural analysis: a) SSIM and b) PSNR for all image planes in the range 967.11 – 1700 nm.

Figure 7.3 presents the SSIM and PSNR obtained for each HSI image plane in the range 967.11 – 1700 nm. It can be seen in Figure 7.3(a), that the SSIM value was highest for the image plane at 1381 nm. The higher the SSIM value, the more similar the image of interest is to the reference image. For example, an SSIM value of one signifies that the image is exactly the same as the reference image, whereas, a SSIM value of zero indicates that there is no similarity between the image plane and the reference image. In Figure 7.3(b), it can be seen that the image plane at 1381 nm also has the highest PSNR value. A high PSNR value indicates that there is more information present (relative to the noise) in the image plane at 1381 nm compared to image planes at other wavelengths. The image plane corresponding to 1381 nm is presented in Figure 7.4(a). Figure 7.4(b) presents the same image plane after sharpening. It can be seen that before sharpening, the image plane is blurred, however, this is reduced after sharpening and the textural details are more evident.



(a)

(b)

*Figure 7.4: Greyscale images produced using the image plane at 1381 nm (a) without and (b) with sharpening.*



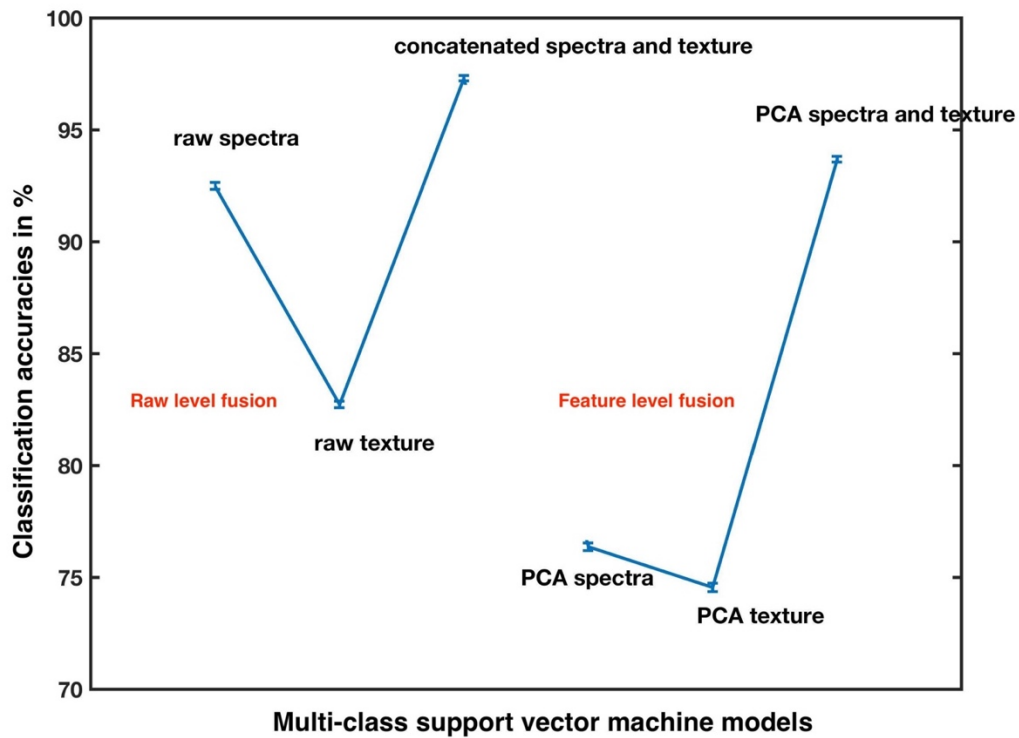


Figure 7.5: Mean classification accuracies (in percent) of the 16 green tea products obtained for the calibration samples (pixels) using models built with raw data and PCA features. In both cases, models were built using spectral information alone, textural information alone and fused spectral and textural information. The error bars denote  $\pm 1$  standard deviation ( $n = 100$ ).

Figure 7.5 presents the mean classification accuracies of the 16 green tea products obtained for the calibration samples (pixels) using multi-class SVM models developed with spectral and textural information. The accuracies are presented as the mean  $\pm$  one standard deviation for 100 iterations. Confusion matrices showing classification accuracies for individual classes obtained using raw data and feature-level SVM models are given in Figures A4.1 and A4.2, respectively, of Annex 4. It can be seen from Figure 7.5 that the models built with the spectral information alone were more accurate than those constructed using only textural information. Combining textural information with spectral information resulted in an improvement in the model accuracy. Improvements were observed for both raw data-level fusion as well as feature-level fusion. The model accuracy for fusion of data at the raw level was higher compared than that at the feature level. It could be that the features extracted using PCA contain less information than the raw data. The features were selected so as to

retain 99% of the variance in the data whereas the raw data retains all of the information. and therefore, this could account for the higher accuracy of the raw data models. Use of supervised feature selection algorithms such as partial least squares discriminant analysis (PLS-DA) could improve the performance of the feature-level models.

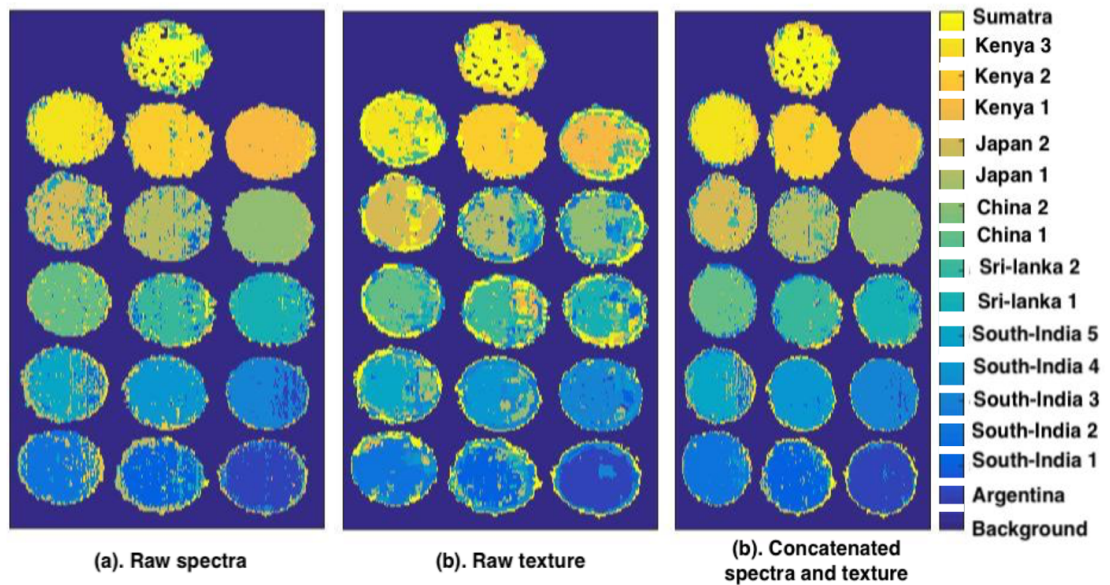


Figure 7.6: Classification maps for the 16 green tea products obtained from SVM modelling of (a). raw spectral information, (b). raw textural information, and (c). concatenated raw spectral and textural information.

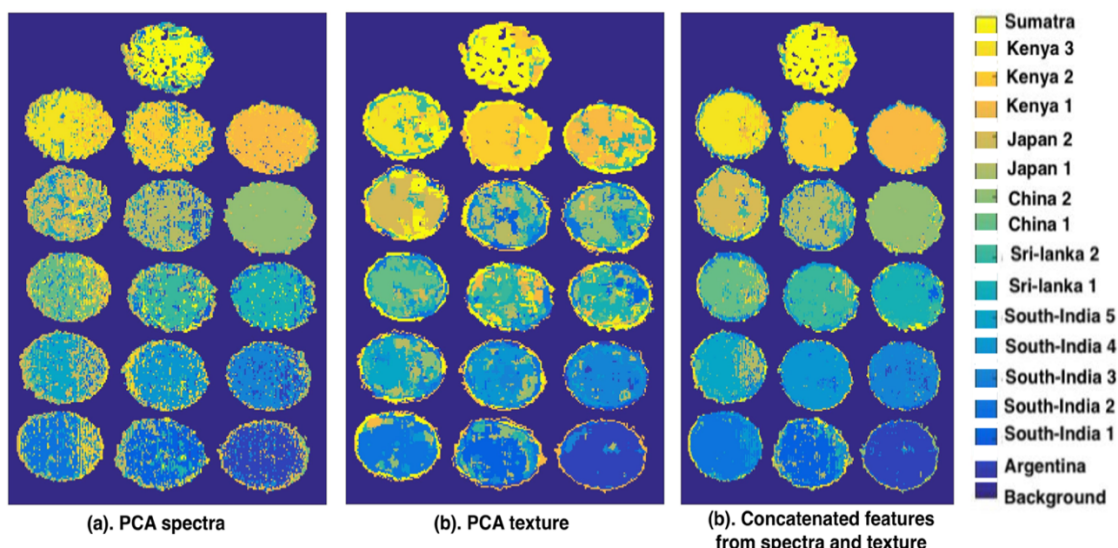


Figure 7.7: Classification maps for the 16 green tea products obtained from SVM modelling of (a). PCA features extracted from spectral information, (b). PCA features extracted from textural information, and (c). concatenated PCA features from spectral and textural information.

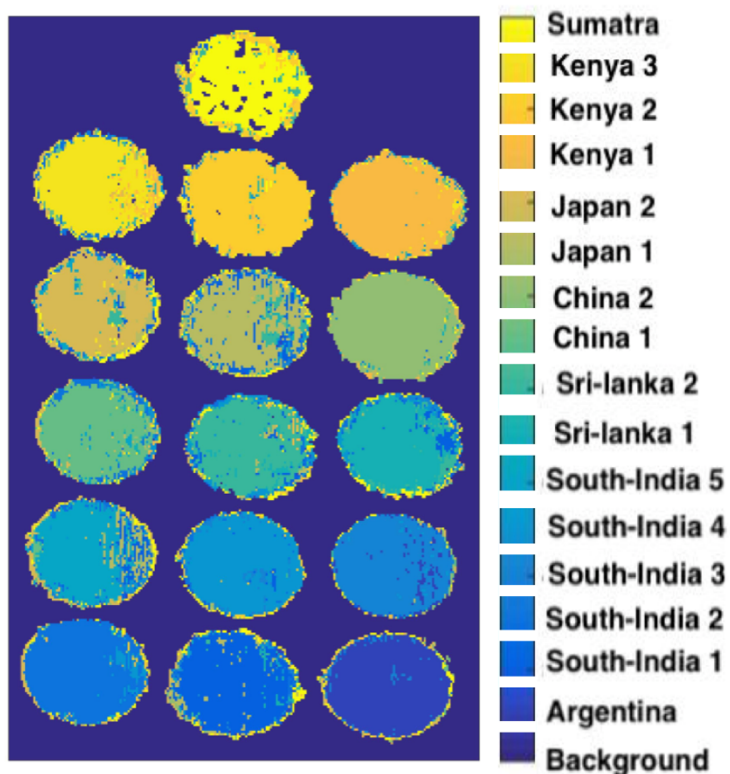


Figure 7.8: Classification maps for the 16 green tea products obtained from decision-level data fusion, using a majority voting scheme, of the six classification maps

*obtained from SVM modelling of spectral information, textural information, and spectral and textural information using raw data (Figure 7.6) and PCA features (Figure 7.7).*

Figure 7.6 and Figure 7.7 presents the classification maps for the 16 green tea products obtained from application of the raw data and feature-level SVM models, respectively, to the complete image. Every circular object in the classification maps is a different green tea sample, comprising more than 2000 pixels per sample, and the different colours reflect different classes. In Figure 7.6, the three classification maps were obtained from three different SVM models built using raw spectral data (Figure 7.6a), raw textural data (Figure 7.6b) and concatenated raw spectral and textural data (Figure 7.6c). Similarly, in Figure 7.7 the three classification maps were obtained from three different SVM models built using the scores obtained from PCA of spectral data (Figure 7.7a), the scores obtained from PCA of textural data (Figure 7.7b) and the concatenated scores obtained from separate PCA models of spectral and textural data (Figure 7.7c). Figure 7.8 provides the output of a majority voting scheme performed on all six classification maps, i.e., three from the raw data (Figure 7.6) and three from the extracted features (Figure 7.7). Majority voting was performed by assigning the pixel value to the class that occurred most frequently in all six classification maps. It can be seen from visual inspection of Figures 7.6, 7.7 and 7.8 that improved classification maps (i.e. an increase in the number of pixels inside the circular area belonging to the same class) were obtained for models built using fused spectral and textural information. This improvement can be quantified by calculating the percentage of correctly classified pixels as shown in Figure 7.9. It can be seen that the highest values were obtained for models built using raw data-level fusion (~84%), followed by decision-level fusion (~83%), with the least number of correctly classified pixels obtained using feature-level data fusion (~78%). Fusion of spectral and textural information at all levels (raw, feature and decision) gave improved model accuracies compared to spectral or textural information alone at the relevant level (i.e. raw or PCA features) leading to an improvement in the classification maps. These results are consistent with HSI studies of meat products [20, 21, 23, 24, 25] where improved

classification or property prediction was obtained with models built using both spectral and textural information.

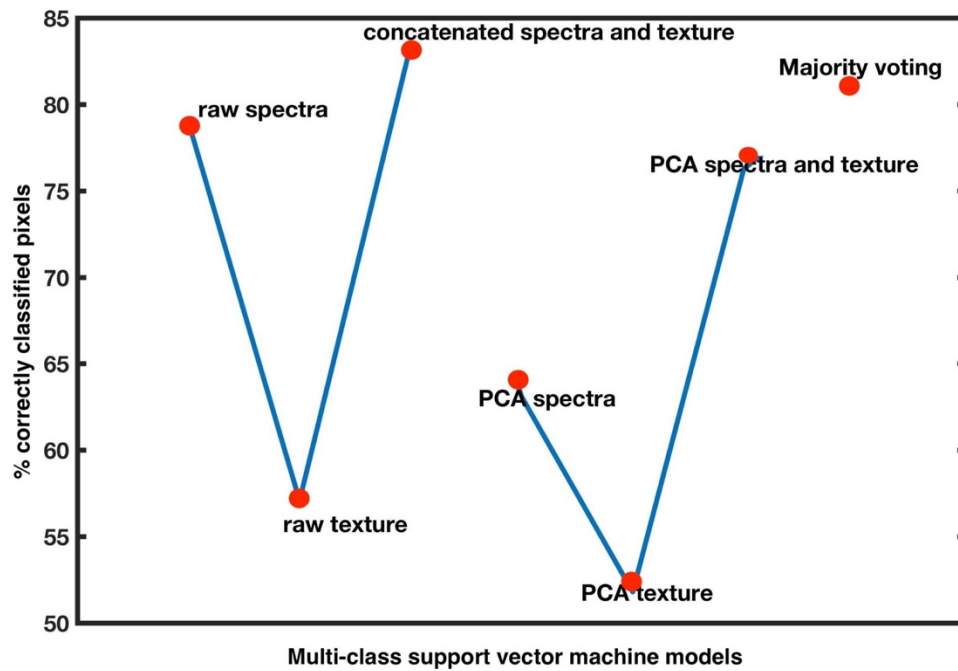


Figure 7.9: Percentage of pixels correctly identified in the classification maps for the 16 green tea products obtained using six different SVM models and decision-level fusion by majority voting.

### 7.5 Conclusions

The spectral and spatial domains of HSI generate complementary information, and synergistic processing of the information can lead to enhanced classification model accuracies and improved classification maps. The present work fused spectral and textural data at three different levels to demonstrate the usefulness of textural information in HSI for classification of green teas. The highest classification accuracy ( $97.30 \pm 0.12\%$  for the calibration samples) was obtained using the raw data-level fusion, which was superior to that obtained for feature-level data fusion. In this case, feature extraction resulted in information loss. However, use of supervised feature selection methods, such as PLS-DA, could improve the performance of the feature-level models. Decision-level fusion provided classification maps of comparable quality to those obtained using raw data-level fusion. In conclusion, the extracted

textural information is always complementary as it can support the development of enhanced understanding of the samples and further model improvement. However, it should be noted that the decision to use the textural information in data modelling has to be based on the samples imaged, as samples with high textural information can contribute positively to model improvement whereas model with no such textural details will merely increase the computation load. Therefore, the methodology developed will be useful in the assessment of a variety of food products (e.g., tea, spices, meat and fruit) where consideration of both spectral and textural information is required for, e.g., quality control and counterfeit detection.

## 7.6 References

1. L.F. Capitán-Vallvey, N. López-Ruiz, A. Martínez-Olmos, M.M. Erenas, A.J. Palma, Recent developments in computer vision-based analytical chemistry: A tutorial review, *Anal. Chim. Acta.* 899 (2015) 23–56. doi:<https://doi.org/10.1016/j.aca.2015.10.009>.
2. A.F. Goetz, G. Vane, J.E. Solomon, B.N. Rock, Imaging spectrometry for Earth remote sensing., *Science.* 228 (1985) 1147–1153. doi:[10.1126/science.228.4704.1147](https://doi.org/10.1126/science.228.4704.1147).
3. A.A. Gowen, C.P. O'Donnell, P.J. Cullen, G. Downey, J.M. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, *Trends Food Sci. Technol.* 18 (2007) 590–598. doi:<https://doi.org/10.1016/j.tifs.2007.06.001>.
4. G. Elmasry, M. Kamruzzaman, D.-W. Sun, P. Allen, Principles and Applications of Hyperspectral Imaging in Quality Evaluation of Agro-Food Products: A Review, *Crit. Rev. Food Sci. Nutr.* 52 (2012) 999–1023. doi:[10.1080/10408398.2010.543495](https://doi.org/10.1080/10408398.2010.543495).
5. Y.-Z. Feng, D.-W. Sun, Application of Hyperspectral Imaging in Food Safety Inspection and Control: A Review, *Crit. Rev. Food Sci. Nutr.* 52 (2012) 1039–1058. doi:[10.1080/10408398.2011.651542](https://doi.org/10.1080/10408398.2011.651542).
6. M. Manley, Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials, *Chem. Soc. Rev.* 43 (2014) 8200–8214. doi:[10.1039/C4CS00062E](https://doi.org/10.1039/C4CS00062E).
7. P. Mishra, C.B.Y. Cordella, D.N. Rutledge, P. Barreiro, J.M. Roger, B. Diezma, Application of independent components analysis with the JADE

- algorithm and NIR hyperspectral imaging for revealing food adulteration, *J. Food Eng.* 168 (2016) 7–15.
8. D.M. Martínez Gila, P. Cano Marchal, J. Gámez García, J. Gómez Ortega, On-line system based on hyperspectral information to estimate acidity, moisture and peroxides in olive oil samples, *Comput. Electron. Agric.* 116 (2015) 1–7. doi:<https://doi.org/10.1016/j.compag.2015.06.002>.
  9. M. Sandasi, W. Chen, I. Vermaak, A. Viljoen, Non-destructive quality assessment of herbal tea blends using hyperspectral imaging, *Phytochem. Lett.* 24 (2018) 94–101. doi:<https://doi.org/10.1016/j.phytol.2018.01.016>.
  10. C. Nansen, G. Zhao, N. Dakin, C. Zhao, S.R. Turner, Using hyperspectral imaging to determine germination of native Australian plant seeds, *J. Photochem. Photobiol. B Biol.* 145 (2015) 19–24. doi:<https://doi.org/10.1016/j.jphotobiol.2015.02.015>.
  11. N. Caporaso, M.B. Whitworth, M.S. Fowler, I.D. Fisk, Hyperspectral imaging for non-destructive prediction of fermentation index, polyphenol content and antioxidant activity in single cocoa beans, *Food Chem.* 258 (2018) 343–351. doi:<https://doi.org/10.1016/j.foodchem.2018.03.039>.
  12. K. Phuangsombut, T. Ma, T. Inagaki, S. Tsuchikawa, A. Terdwongworakul, Near-infrared hyperspectral imaging for classification of mung bean seeds, *Int. J. Food Prop.* 21 (2018) 799–807. doi:[10.1080/10942912.2018.1476378](https://doi.org/10.1080/10942912.2018.1476378).
  13. Y. Liu, H. Pu, D.-W. Sun, Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications, *Trends Food Sci. Technol.* 69 (2017) 25–35. doi:<https://doi.org/10.1016/j.tifs.2017.08.013>.



14. J.M. Amigo, H. Babamoradi, S. Elcoroaristizabal, Hyperspectral image analysis. A tutorial, *Anal. Chim. Acta.* 896 (2015) 34–51. doi:<https://doi.org/10.1016/j.aca.2015.09.030>.
15. A. de Juan, Hyperspectral image analysis. When space meets Chemistry, *J. Chemom.* 32 (2018) e2985–n/a. doi:10.1002/cem.2985.
16. L. Wang, C. Shi, C. Diao, W. Ji, D. Yin, A survey of methods incorporating spatial information in image classification and spectral unmixing, *Int. J. Remote Sens.* 37 (2016) 3870–3910. doi:10.1080/01431161.2016.1204032.
17. A. Laddi, S. Sharma, A. Kumar, P. Kapur, Classification of tea grains based upon image texture feature analysis under different illumination conditions, *J. Food Eng.* 115 (2013) 226–231. doi:<https://doi.org/10.1016/j.jfoodeng.2012.10.018>.
18. M. Tuceryan, A.K. Jain, Texture analysis, in: *Handb. Pattern Recognit. Comput. Vis.*, World Scientific, 1993: pp. 235–276.
19. J. Ma, D.-W. Sun, J.-H. Qu, H. Pu, Prediction of textural changes in grass carp fillets as affected by vacuum freeze drying using hyperspectral imaging based on integrated group wavelengths, *LWT - Food Sci. Technol.* 82 (2017) 377–385. doi:<https://doi.org/10.1016/j.lwt.2017.04.040>.
20. C. Garrido-Novell, A. Garrido-Varo, D. Pérez-Marín, J.E. Guerrero, Using spectral and textural data extracted from hyperspectral near infrared spectroscopy imaging to discriminate between processed pork, poultry and fish proteins, *Chemom. Intell. Lab. Syst.* 172 (2018) 90–99. doi:<https://doi.org/10.1016/j.chemolab.2017.11.011>.

21. D. Yang, D. He, A. Lu, D. Ren, J. Wang, Combination of spectral and textural information of hyperspectral imaging for the prediction of the moisture content and storage time of cooked beef, *Infrared Phys. Technol.* 83 (2017) 206–216. doi:<https://doi.org/10.1016/j.infrared.2017.05.005>.
22. S. Fan, B. Zhang, J. Li, C. Liu, W. Huang, X. Tian, Prediction of soluble solids content of apple using the combination of spectra and textural features of hyperspectral reflectance imaging data, *Postharvest Biol. Technol.* 121 (2016) 51–61. doi:<https://doi.org/10.1016/j.postharvbio.2016.07.007>.
23. D. Liu, H. Pu, D.-W. Sun, L. Wang, X.-A. Zeng, Combination of spectra and texture data of hyperspectral imaging for prediction of pH in salted meat, *Food Chem.* 160 (2014) 330–337. doi:<https://doi.org/10.1016/j.foodchem.2014.03.096>.
24. H. Pu, D.-W. Sun, J. Ma, J.-H. Cheng, Classification of fresh and frozen-thawed pork muscles using visible and near infrared hyperspectral imaging and textural analysis, *Meat Sci.* 99 (2015) 81–88. doi:<https://doi.org/10.1016/j.meatsci.2014.09.001>.
25. W. Cheng, D.-W. Sun, H. Pu, Y. Liu, Integration of spectral and textural data for enhancing hyperspectral prediction of K value in pork meat, *LWT - Food Sci. Technol.* 72 (2016) 322–329. doi:<https://doi.org/10.1016/j.lwt.2016.05.003>.
26. D. Huo, Y. Wu, M. Yang, H. Fa, X. Luo, C. Hou, Discrimination of Chinese green tea according to varieties and grade levels using artificial nose and tongue based on colorimetric sensor arrays, *Food Chem.* 145 (2014) 639–645. doi:<https://doi.org/10.1016/j.foodchem.2013.07.142>.

27. W. He, X. Hu, L. Zhao, X. Liao, Y. Zhang, M. Zhang, J. Wu, Evaluation of Chinese tea by the electronic tongue: Correlation with sensory properties and classification according to geographical origin and grade level, *Food Res. Int.* 42 (2009) 1462–1467. doi:<https://doi.org/10.1016/j.foodres.2009.08.008>.
28. P. Mishra, A. Nordon, J. Tschannerl, G. Lian, S. Redfern, S. Marshall, Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products, *J. Food Eng.* 238 (2018) 70-77. doi:<https://doi.org/10.1016/j.jfoodeng.2018.06.015>.
29. Z. Tang, Y. Su, M.J. Er, F. Qi, L. Zhang, J. Zhou, A local binary pattern based texture descriptors for classification of tea leaves, *Neurocomputing.* 168 (2015) 1011–1023. doi:<https://doi.org/10.1016/j.neucom.2015.05.024>.
30. N.R. Sarkar, Machine vision for quality control in the food industry, *Instrum. Methods Qual. Assur. Foods.* (1991) 167–187.
31. A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures., *Anal. Chem.* 36 (1964) 1627–1639. doi:[10.1021/ac60214a047](https://doi.org/10.1021/ac60214a047).
32. R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra, *Appl. Spectrosc.* 43 (1989) 772–777. doi:[10.1366/0003702894202201](https://doi.org/10.1366/0003702894202201).
33. Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Trans. Image Process.* 13 (2004) 600–612. doi:[10.1109/TIP.2003.819861](https://doi.org/10.1109/TIP.2003.819861).

34. A. Polesel, G. Ramponi, V.J. Mathews, Image enhancement via adaptive unsharp masking, *IEEE Trans. Image Process.* 9 (2000) 505–510. doi:10.1109/83.826787.
35. D.A. Clausi, An analysis of co-occurrence texture statistics as a function of grey level quantization, *Can. J. Remote Sens.* 28 (2002) 45–62. doi:10.5589/m02-004.
36. L.K. Soh, C. Tsatsoulis, Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices, *IEEE Trans. Geosci. Remote Sens.* 37 (1999) 780–795. doi:10.1109/36.752194.
37. R.M. Haralick, K. Shanmugam, I. Dinstein, Textural Features for Image Classification, *IEEE Trans. Syst. Man. Cybern. SMC-3* (1973) 610–621. doi:10.1109/TSMC.1973.4309314.
38. F. Marini, Classification methods in chemometrics, *Curr. Anal. Chem.* 6 (2010) 72–79.
39. R. Seifi Majdar, H. Ghassemian, A probabilistic SVM approach for hyperspectral image classification using spectral and texture features, *Int. J. Remote Sens.* 38 (2017) 4265–4284. doi:10.1080/01431161.2017.1317941.

## Chapter 8 : Conclusions and future work

### 8.1 Conclusions

Tea product analysis is of key importance to different industries working in the field of tea processing, procurement, packaging and marketing for product quality control. At present, tea analysis is largely limited to the ‘tea taster’ which can be categorised as slow, costly and subjectively biased. Further, the number of samples to be analysed by tea tasters is limited and therefore cannot be used as an on-line tool for an automated high-throughput scenario. Further, in an industrial scenario there is a need for rapid fast and non-destructive sensors which can support this assessment and to some extent can replace the different high-end techniques such as HPLC, GCMS which are used to perform this task. The NIR-HSI in particular can support in non-destructive real time prediction of different chemical constituents such as caffeine, polyphenols, Gallic acid, epigallocatechin, theobromine, Theaflavin, moisture, amino acids and Lignin. However, the main motivation is to have a technique which can provide rapid and cost-effective analysis of tea products as this can support on-line analysis of tea products and processes. Keeping this as the motivation, the 1st chapter of the thesis explored the literature for the possibility of utilising NIRS and HSI by identifying some most recent applications of NIRS for analysing tea products. Learnings from the 1st chapter include the understanding of the assessment of tea products based on both physical quality and chemical assessment, which define the taste profile for the tea. The literature survey showed an increased interest in the use of rapid non-destructive sensors such as NIR point spectroscopy for chemical assessment of tea processes and products. For physical assessment, there are increasing applications of RGB imaging for assessment of shape, size and texture of tea products to explain the physical quality. However, the literature reviewed showed that a combination of imaging and spectroscopy, i.e., HSI is still emerging for analysis of tea products and processes. Further, not only for tea but in general in recent years, HSI has emerged as a potential tool for the rapid non-contact assessment of food products. There is also growing interest in utilising HSI for the on-line monitoring of food products and processes. However, there are different challenges related to utilising HSI as an on-line monitoring tool and these are mainly related to the handling and processing of the data

generated by HSI. These challenges need to be dealt with also in the case for utilising HSI for assessment of tea product and processes. Further, the implementation of HSI for a high-throughput application can be understood in two ways, the first is as a continuous monitoring tool which generate a continuous stream of data and the second is the implementation to HSI for discrete image acquisition for example monitoring agricultural plants every day for phenotyping related applications. The methodologies presented in the work are more adaptable to the discrete HSI data compared to online monitoring tasks.

The main focus of the research after the learnings from chapter 1 was to utilise NIR HSI for non-destructive assessment of tea products and to deal with different challenges associated with processing and handling of the data. Chapter 2 and 3 dealt mainly with analysing tea products, where the research showed that utilising NIR for classification of different tea products is possible and HSI can be moved forward as a potential tool for on-line assessment of tea products. Particularly, in chapter 2 experiments relating to the classification of commercial tea products were carried out. NIRS data is multivariate and therefore visualising its structure directly is not possible. In chapter 2, different linear and non-linear neighbourhood-based data visualisation techniques were compared to traditional PCA, showing an improvement in performance for visualising NIR data. The better performance of neighbourhood methods was based on the fact the when there are multiple classes in the data with small and high variances, then visualising all the classes together inside a single PCA will try to separate the classes with maximum variance where the small variance classes will be forced to lie next to each other. In such a case, the method which can consider the neighbourhood information can benefit by improving the data visualisation by separating even the low variance classes. Learnings from chapter 2 led the way to chapter 3, where more complex samples, comprising only green tea products, were analysed. In chapter 3, the task was more complex because all the green tea products have very similar tea profiles. However, SVM modelling supported the classification of green tea products based on individual products, and countries and continents of origin. It is to be kept in mind, that the samples analysed in the analysis was a subset of samples provided by Unilever, Colworth Park, UK. A study performed

on such a small set cannot be directly translated to broader samples and for that purpose there is a need to perform such a study on a large number of samples considering global variability. Further, in chapter 3, different wavelength selection methods (sequential and filter based) were used to identify the subset of wavelengths that were important for maintaining the classification accuracy while reducing the number of variables required to perform the modelling.

After developing the learning on the potential of NIR HSI for tea products analysis in chapter 2 and chapter 3, the next stage was to consider implementation of HSI in a real-world industrial scenario for tea products analysis. The challenges associated with implementation mainly relate to data processing and handling. Chapters 4 – 7 consider these challenges and deal with them one after another based on the complexity of the challenges. The first challenge was related to the size of the HS image and their storage for the long term. Typically, HSI cameras record the images in binary format and therefore the size of the images is directly proportional to the spatial and spectral dimension of the images. In chapter 4 to deal with it, a method for HS image compression based on the decomposition of spatial and spectral information is presented. The method utilises the 2D-DWT for preconditioning of the HS image and later utilising the variance-based method to perform the decomposition of the HS image into a set of matrices. The methods can help to store individual images by storing them in the form of decomposed information, which can be reconstructed later to perform the analysis. Such a method is of high use where a large number of individual HS images are generated and the storage of them is needed for future recall or reference and most recently for developing deep learning-based models.

The second major challenge related to the implementation of HSI for automated continuous application was to deal with the pre-processing of raw hyperspectral cubes. HSI cameras due to various internal and external factors sometimes provide a measurement which is accompanied by noise. To deal with such types of noises, in chapter 5, a method for automatically pre-processing the HS images is presented. The automatic method benefited over the traditional methods used in the chemometrics domain for pre-processing of images such as SAVGOL and median filtering. Further,

the automatic method does not require the user to check each and every image for noise and decide on how to correct the images and can support the on-line implementation of HSI. However, the method presented in chapter 5 can only deal with Gaussian and spiked noise, and a combination of the two. The method cannot be used to de-noise dead striped lines. The need to find a de-noising method that can be used with a wide variety of different types of noise, including dead striped lines, laid the foundation for chapter 6. In chapter 6, a method based on the total variation model is presented. The results showed that the total variation-based method can deal also with the dead stripped lines and can be used as a general method for automatic denoising. Chapter 5 also shows that to perform the classification of 6 different tea products used in the chapter 2, the VNIR system can provide a classification accuracy of 87 %. The VNIR system are sometimes a preferred choice over the NIR systems because of the cheap systems available for imaging in VNIR spectral range. However, the NIR range has the benefit of accessing the chemical information present in the samples as the NIR spectra reflects overtones and combinations of the fundamental vibration of chemical bonds.

The final challenge related to utilising HSI and in particular for tea analysis was extracting and utilising the physical information about the samples and to use it for improved modelling. Tea samples of different grades and geographical origin possess different physical properties in terms of shape and size. Such tea samples when viewed in bulk with imaging exhibit different textures which is distinct for a particular sample. Distinct textural information from individual tea samples can be used to support the classification models which are solely based on NIR information. Texture extraction from HS images laid the foundation for chapter 7, where a methodology based on GLCM is presented for extracting textural information. The methodology showed that the extracted texture from the HSI can be combined with the NIR information via data fusion, which can finally lead to improved classification accuracies.

In summary, the study concludes that HSI can be a potential tool for non-destructive assessment of tea product and processes. Further, to automate the HSI for an industrial scenario study presented automated pre-processing, compression and texture fusion



methodology which in this study were used with the HSI images of tea products can be extended/modified based on application areas. A complete list of all the outputs generated in this study can be referred to in Annex 5.

## 8.2 Suggestions for future work

The work presented on the assessment of tea products is in its very initial stage. There are a large number of opportunities related to developing NIR HSI as a tool for assessment of tea products. Of particular interest is to work directly in conjunction with the industries to deal directly with the real case scenario. The models and methods related to tea product analysis generated in this work were limited to the number of samples and cannot be directly translated to an industry scenario. Implementing HSI for tea analysis into an industrial scenario will require training new models based in new samples, however, the similar schemes presented in this work can be followed. The implementation for HSI for an industrial conveyor belt scenario will also include examining the thickness of tea leaves and the penetration depth of the radiation. Also, it will include exploring strategies for removing the background material if it effect the imaged signal for the tea product.

In the present work, HSI was used as a tool to perform classification of tea products, however, in future work HSI can be used as a tool for performing quantification of chemical constituents by use of image regression methods. The non-destructive chemical quantification can especially support the monitoring of the tea fermentation process, which is a solid-state fermentation process involving enzymatic oxidation of chopped fresh tea leaves. In that case, predicting the concentration of key analytes can help to monitor and optimise the process to obtain the best quality tea products. In that case, the classification models developed in this work will not be valid but new regression models based on new samples can be trained.

In future work, there are also opportunities to link the sensory properties of the tea products judged by the 'Tea Taster' with that of the NIR data through HSI. In that case, new regression models involving modelling between NIR and tea taste scores as reference are needed to be developed. Most commonly, the industries deal mainly with the three tea products i.e. green, black and oolong tea products as also covered in this

work. However, the work related to HSI can be extended to different tea products which come from the processing of fresh leaves of *camellia sinensis* plants.

In future, there are opportunities to perform system integration utilising HSI as a complete solution for an automated industrial use case scenario for tea products assessment. The integrated system can have its own software system for automatic acquisition, handling and processing of data based on the methodologies presented in this work.

Apart from the application based on the tea, the different data processing methodologies presented in this work for dealing with the HS data can have the broad range of applications where the HSI can be implemented as a continuous monitoring tool. Finally, there is an increasing interest from tea industry side requiring miniature sensors which can be used directly in the tea auctions before purchasing the tea products to access the quality of tea products in real-time for making effective decisions regarding the purchase. In such cases, the miniature NIR sensor such as microNIR from Viavi Solutions, DLP NIR from Texas Instrument can be explored. Further, there is an open door for application of portable miniature handheld HSI cameras if they appear in the market in the coming years like the specim IQ which work in VNIR range.

Related to the data processing methodologies presented in the work, the future work can be related to developing a methodology which combines the de-noising and compression step. Such a methodology will make the two-step process to a single step which can lead to time-saving while performing continuous measurements. Furthermore, for a real-case industrial scenario integrating these methodologies to the pipeline can also be foreseen as a future direction of work.

In conclusion, utilising the HSI as a continuous monitoring tool is in its initial stage and many future opportunities exist for making it a day to day usable technology in an industrial case scenario.

## Annex 1

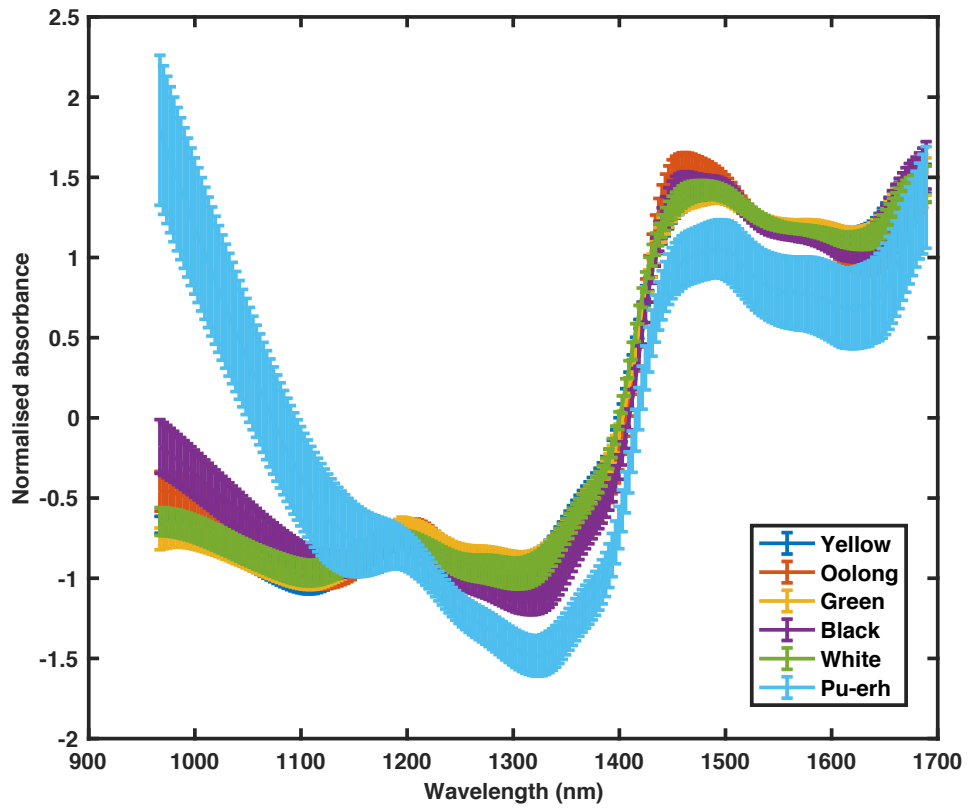
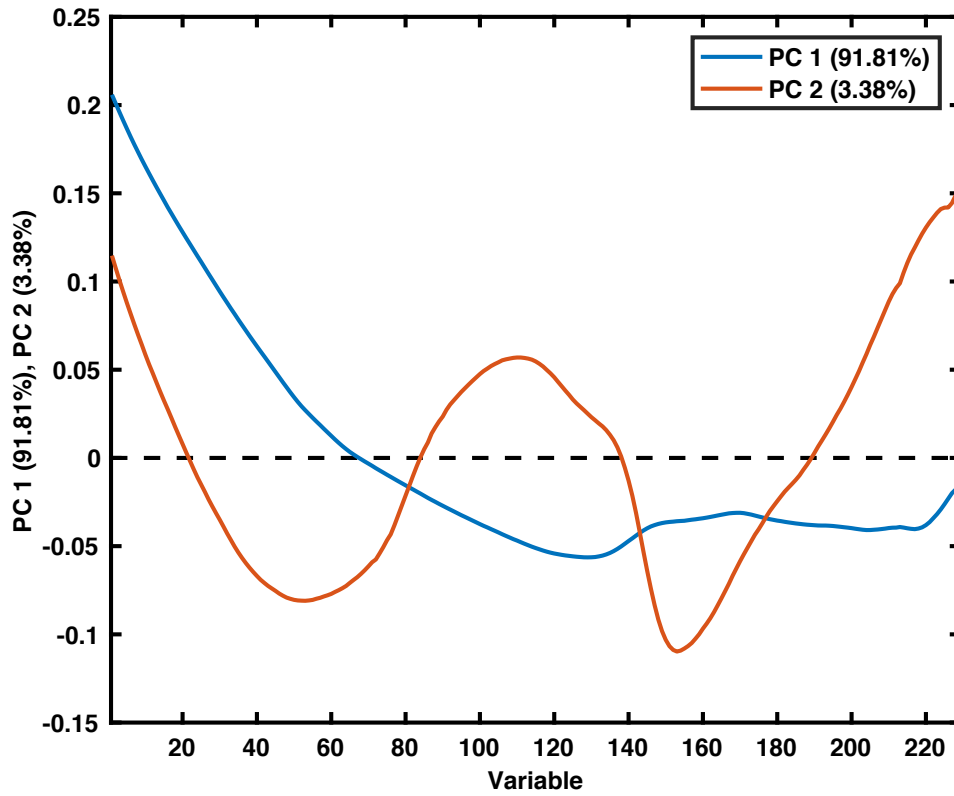


Figure A1.1: Standard deviation for spectra of individual tea products.



*Figure A1.2: PC1 and PC2 for the PCA performed an compared with other data visualisation tools.*

## Annex 2

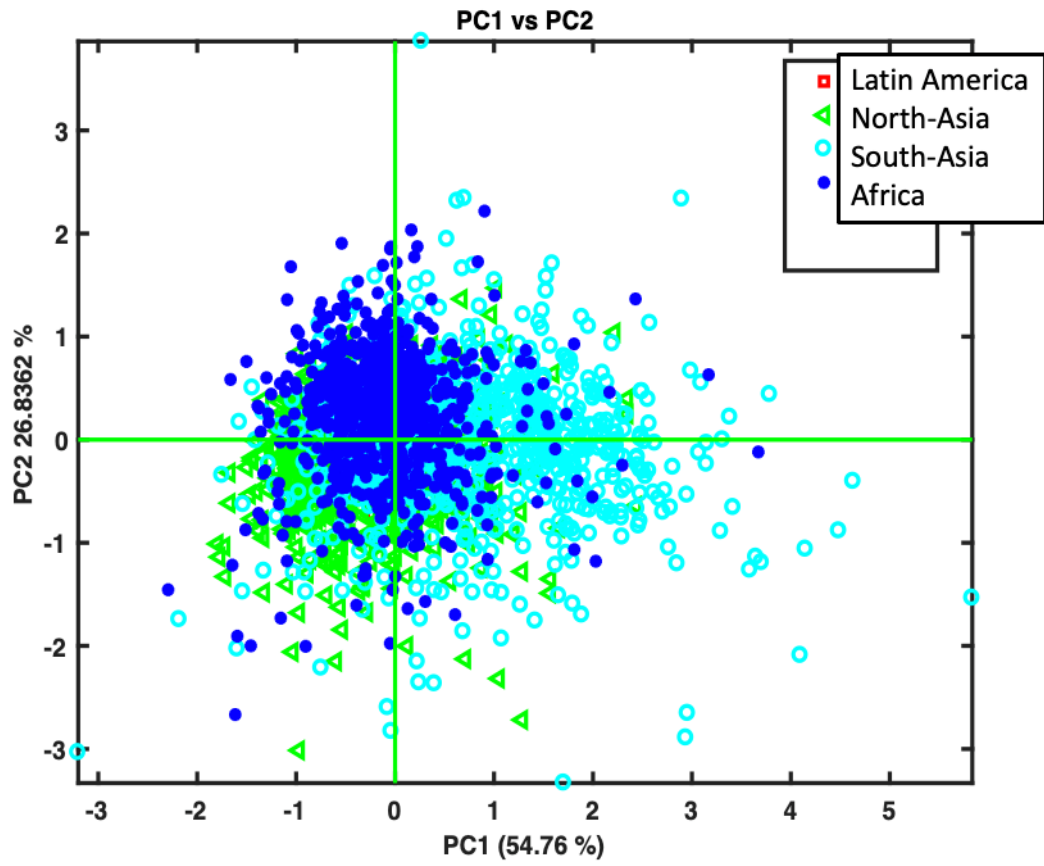


Figure A2.1: Principal component scores (PC1 v PC2) for visualising data based on continent of origin

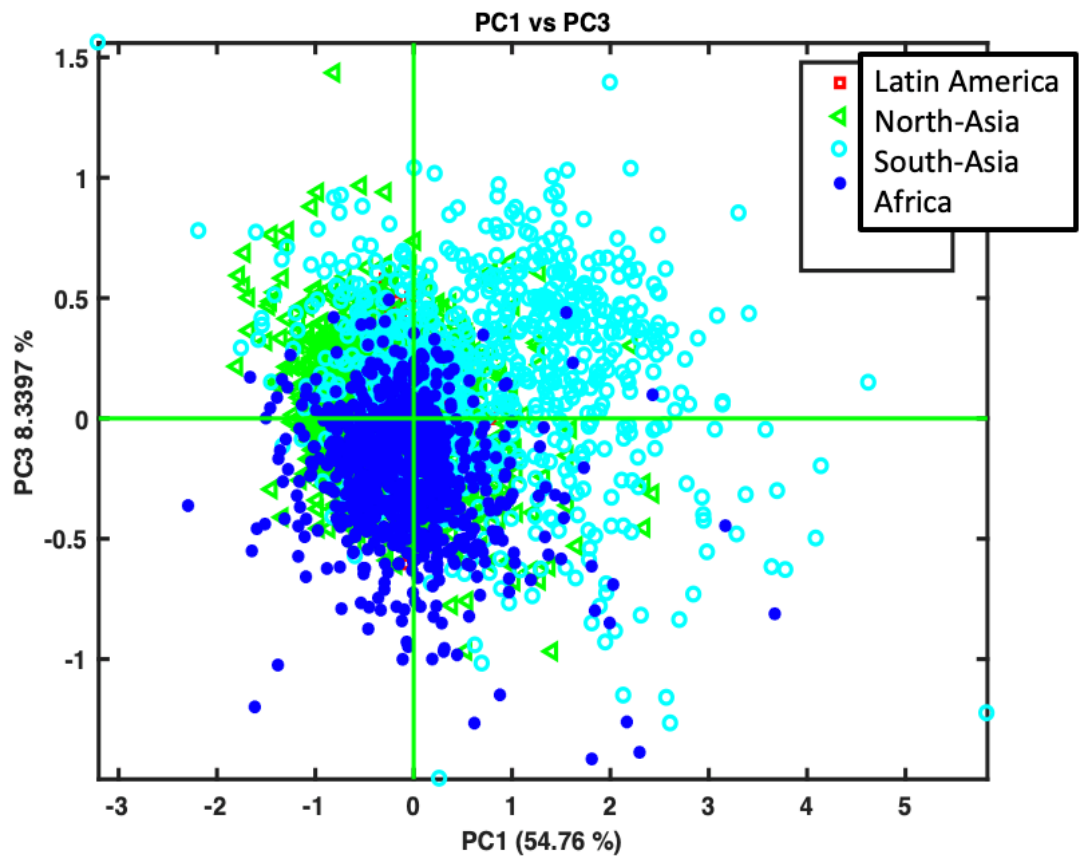


Figure A2.2: Principal component scores (PC1 v PC3) for visualising data based on continent of origin.

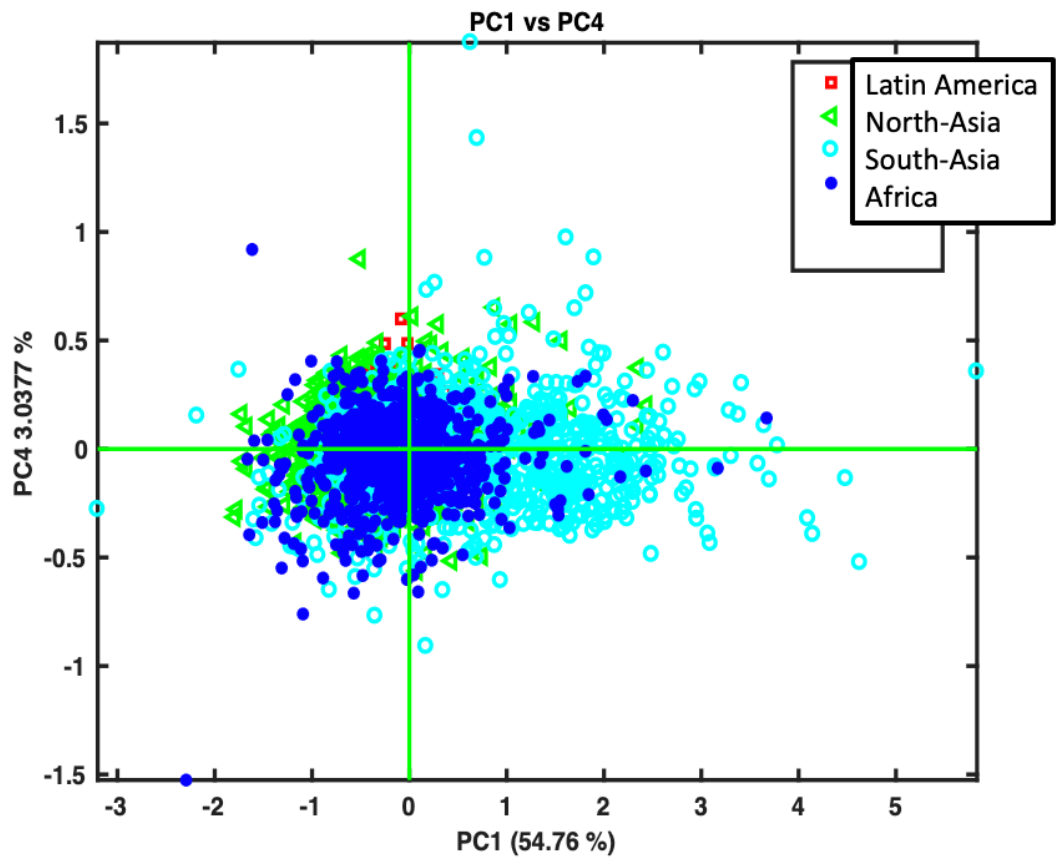


Figure A2.3: Principal component scores (PC1 v PC4) for visualising data based on continent of origin.

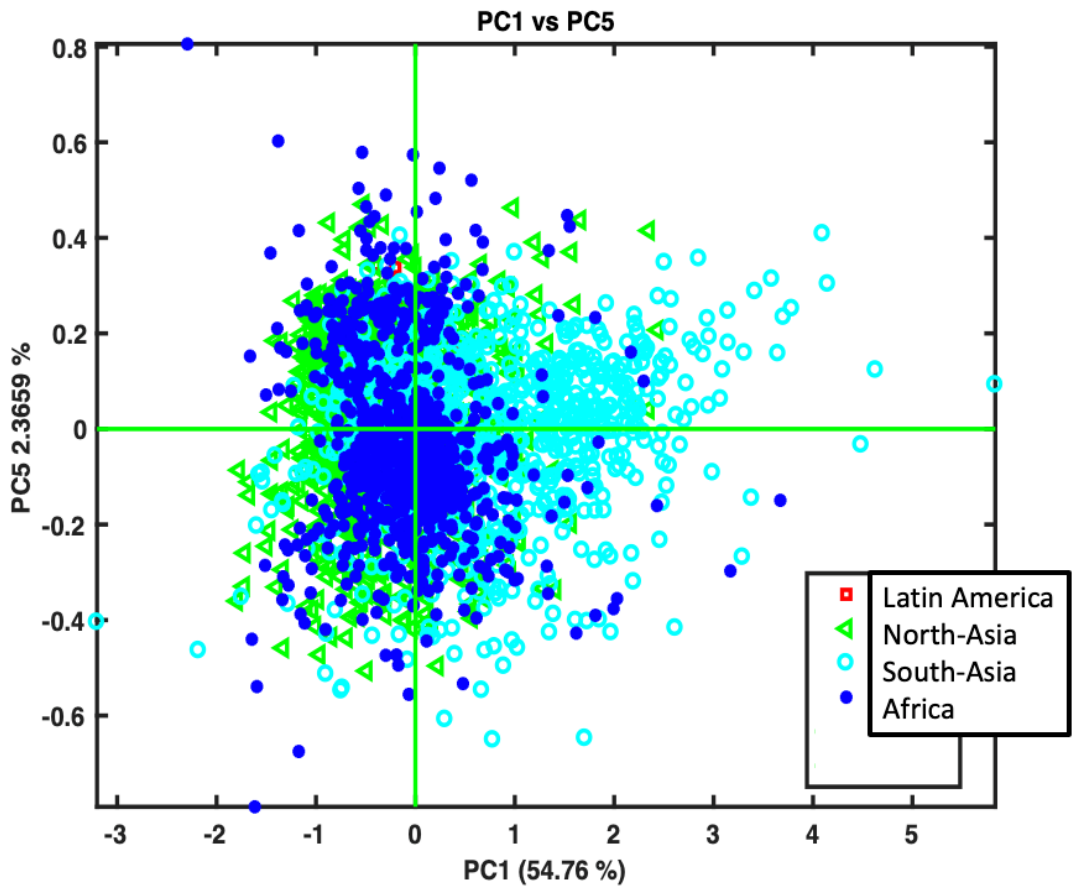


Figure A2.4: Principal component scores (PC1 v PC5) for visualising data based on continent of origin.



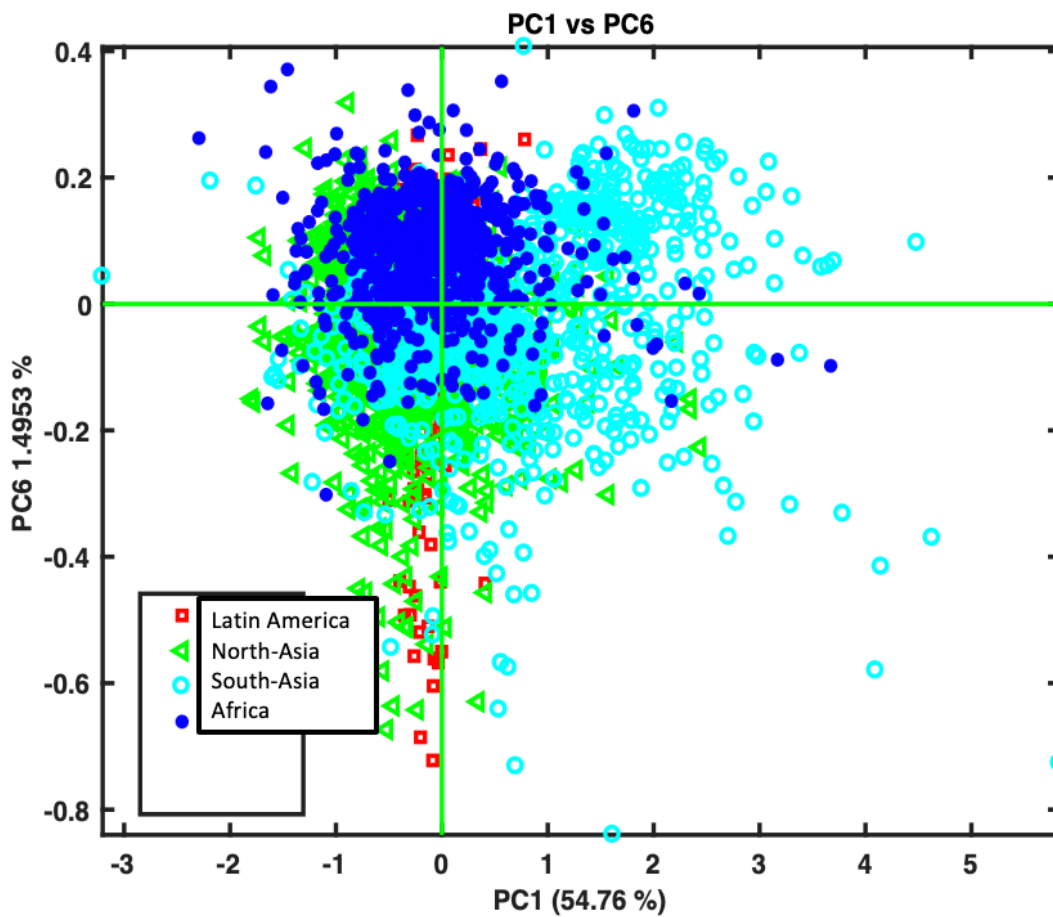


Figure A2.5: Principal component scores (PC1 v PC6) for visualising data based on continent of origin.

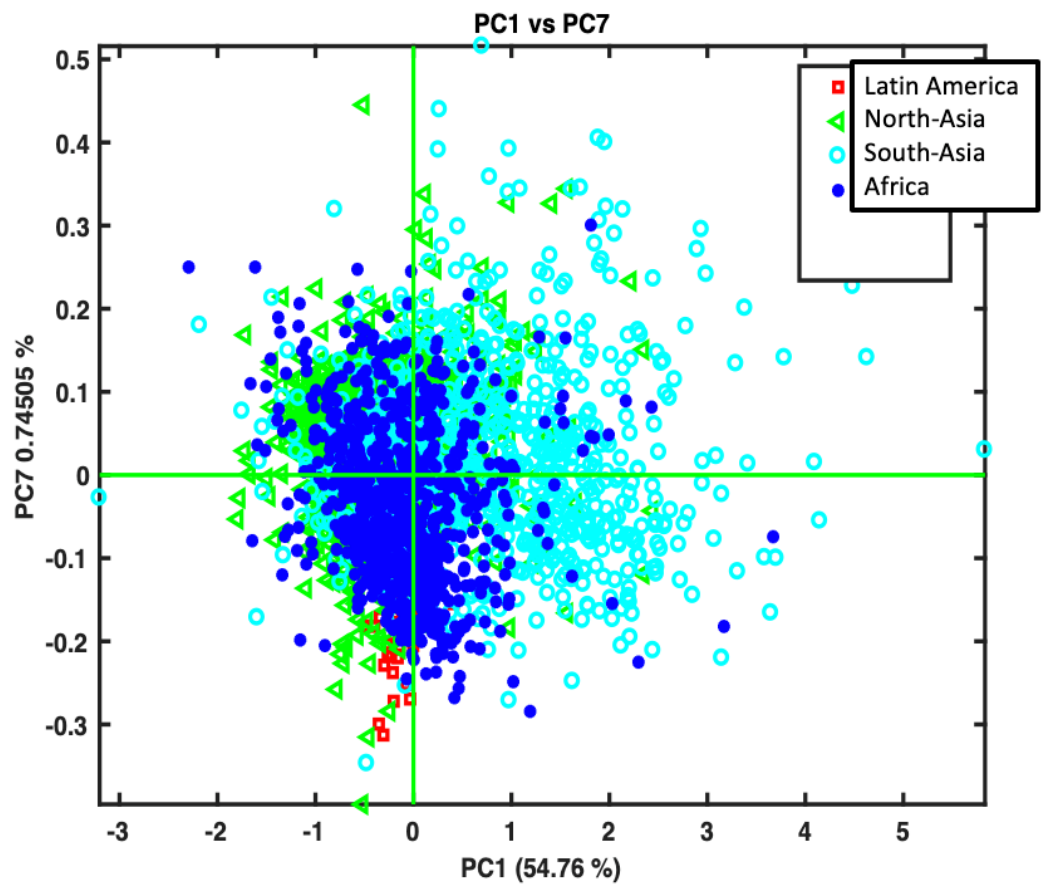


Figure A2.6: Principal component scores (PC1 v PC7) for visualising data based on continent of origin.

Confusion matrix plot for SVM models:

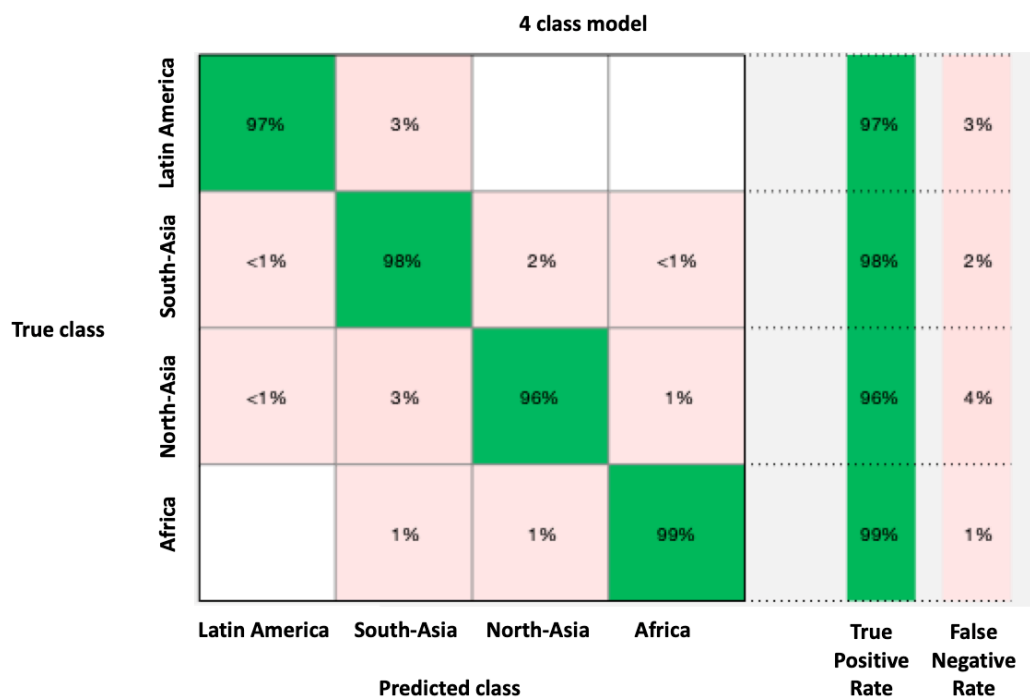


Figure A2.7: Confusion matrix for SVM model developed for classification of tea samples on the basis of continent of origin (4 classes).

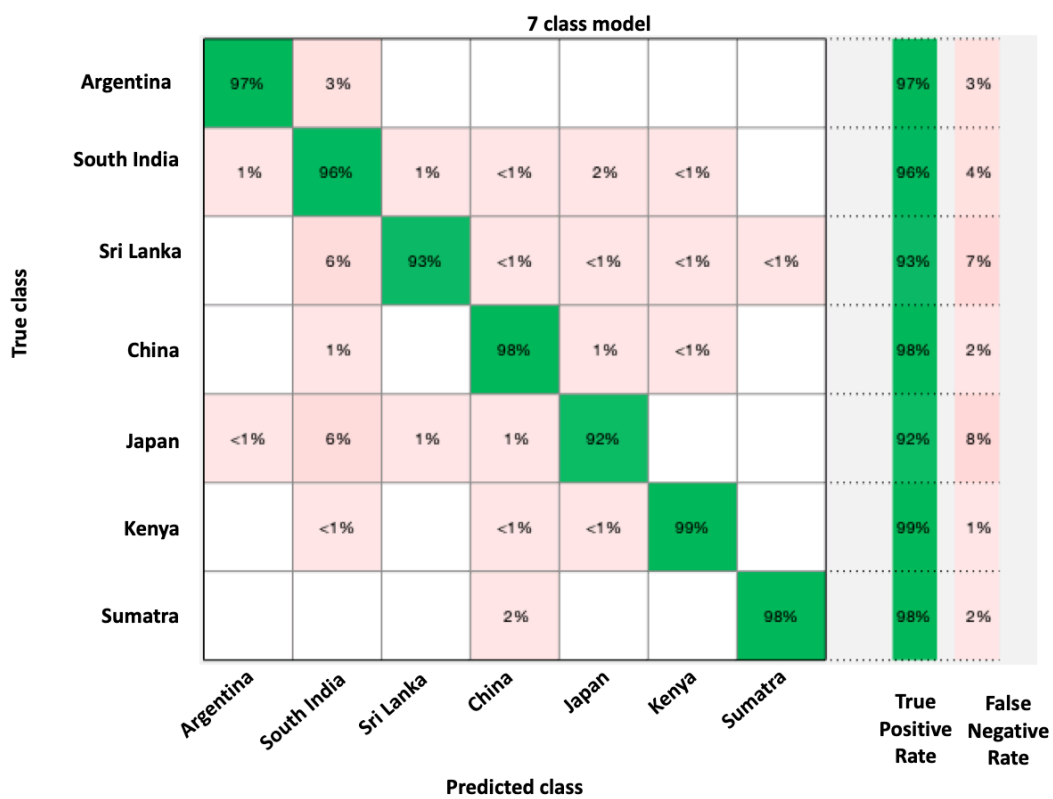


Figure A2.8: Confusion matrix for SVM model developed for classification of tea samples on the basis of country of origin (7 classes).

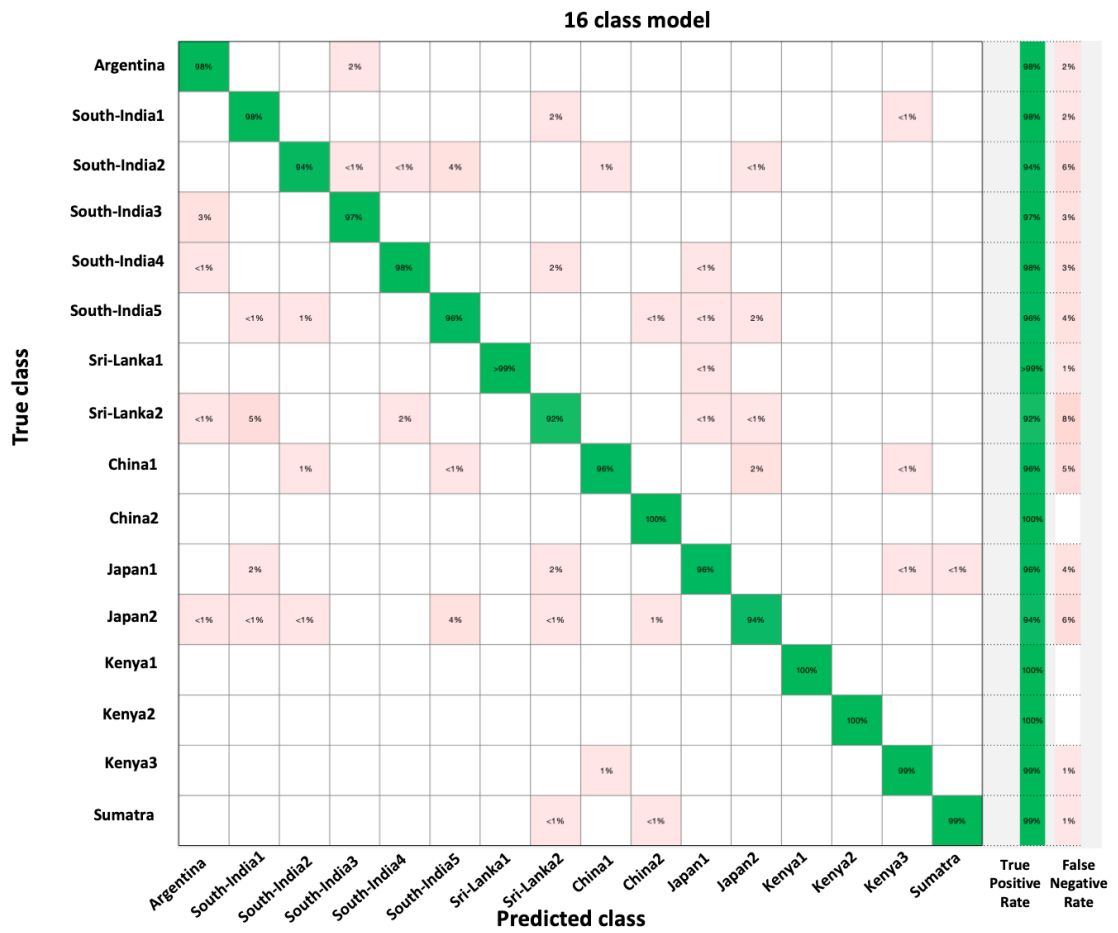
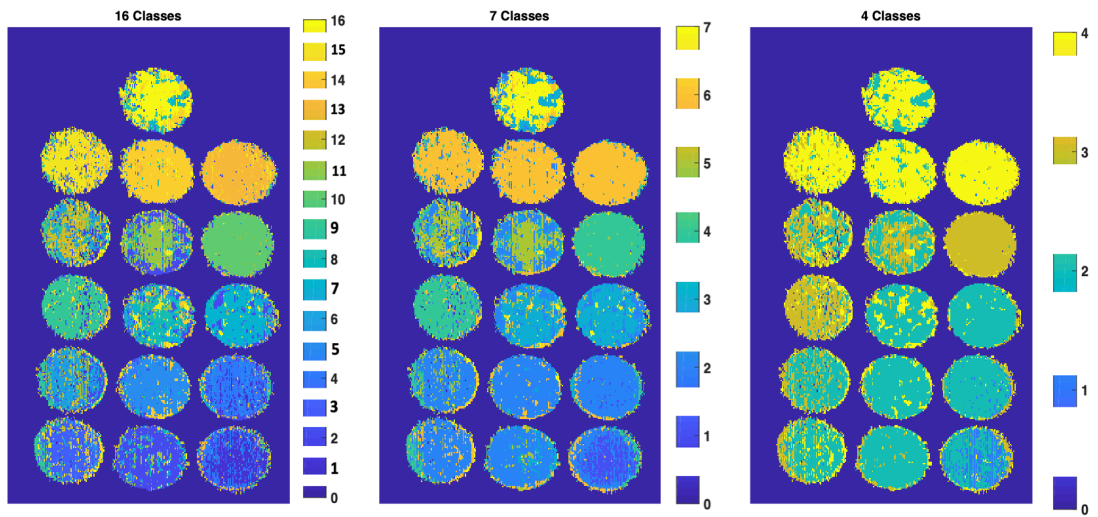


Figure A2.9: Confusion matrix for SVM model developed for classification of tea samples on the basis of product (16 classes).



*Figure A2.10: Principal component scores (PC1 v PC2) for visualising data based on continent of origin*

### Annex 3

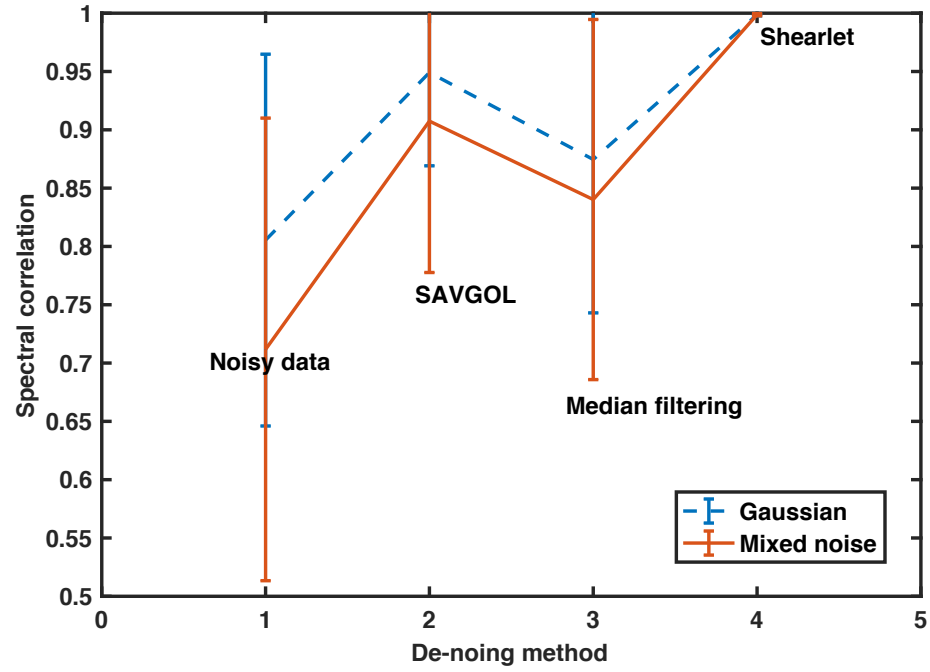


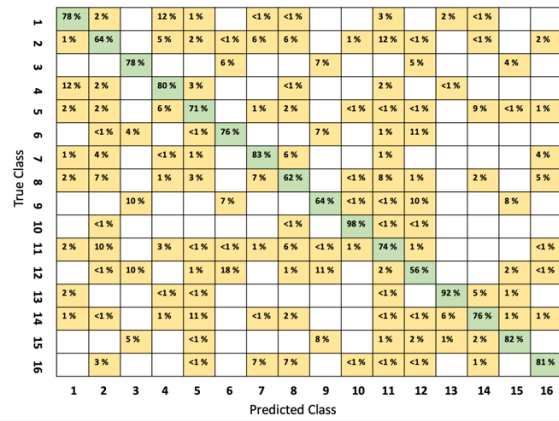
Figure A3.1: Principal component scores (PC1 v PC2) for visualising data based on continent of origin

*Table A3.1: Time spent in running each algorithm for denoising.*

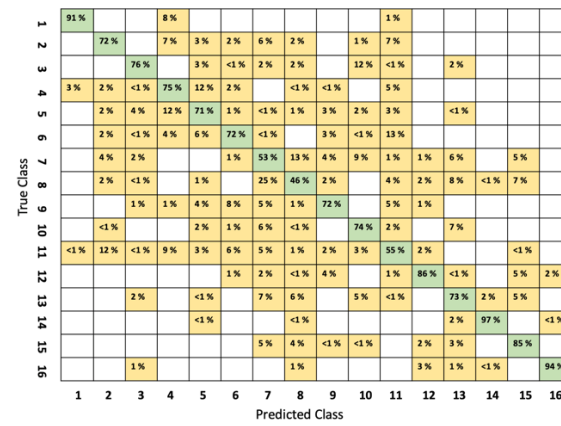
Method	Gaussian noise (Savgol)	Gaussian noise (Median filtering)	Gaussian noise (Shearlet)	Mixed noise (Savgol)	Mixed noise (Median filtering)	Mixed noise (Shearlet)
Time required (Seconds)	0.08	0.16	23.43	0.20	0.33	23.35



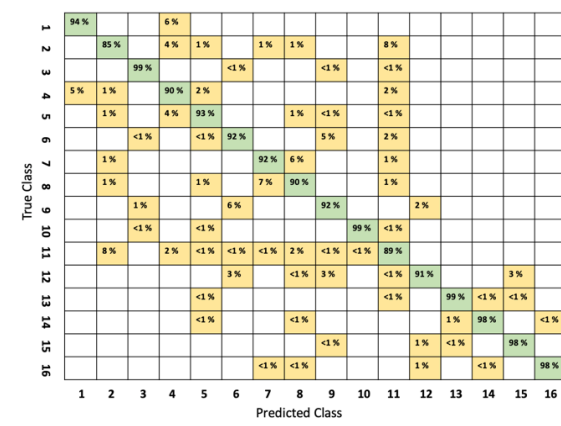




(a)



(b)



(c)

Figure A4.2: Confusion matrix for (a) Classification based on PCA features from NIR spectral data, (b) Classification based on PCA features from textural information and (c) Classification based on concatenated PCA features from NIR spectral and textural data.

## Annex 5

### List of publications and external presentations

#### Published journal articles

P. Mishra, A. Nordon, M.S. Mohd Asaari, G. Lian, S. Redfern, Fusing spectral and textural information in near-infrared hyperspectral imaging to improve green tea classification modelling, *J. Food Eng.* 249 (2019) 40–47. doi:<https://doi.org/10.1016/j.jfoodeng.2019.01.009>.

P. Mishra, A. Karami, A. Nordon, D.N. Rutledge, J.-M. Roger, Automatic de-noising of close-range hyperspectral images with a wavelength-specific shearlet-based image noise reduction method, *Sensors Actuators B Chem.* 281 (2019) 1034–1044. doi:<https://doi.org/10.1016/j.snb.2018.11.034>.

P. Mishra, A. Nordon, J. Tschannerl, G. Lian, S. Redfern, S. Marshall, Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products, *J. Food Eng.* 238 (2018) 70-77. doi:<https://doi.org/10.1016/j.jfoodeng.2018.06.015>.

#### Presentations

##### *Oral*

**(Invited)** P. Mishra, A. Karami, A. Nordon, D.N. Rutledge, J.-M. Roger, **SciX 2018**, Atlanta, Georgia, USA (October 2018), De-noising, compression and data fusion: new approaches for automatically managing and processing hyperspectral imaging.

P. Mishra, A. Karami, A. Nordon, D.N. Rutledge, J.-M. Roger, **Chemometrics in Analytical Chemistry (CAC)-2018**, Halifax, Nova Scotia, Canada (June 2018) |Automatic pre-processing of hyperspectral images with a wavelength specific shearlet-based image noise reduction method.

P. Mishra, A. Nordon, M.S. Mohd Asaari, G. Lian, S. Redfern, **Spring SciX**, Glasgow, United Kingdom, April 2018, Data fusion for hyperspectral imaging.

P. Mishra, A. Nordon, M.S. Mohd Asaari, G. Lian, S. Redfern, **APACT-2018**, Newcastle, United Kingdom (April 2018), Fusion of textural and spectral information of hyperspectral images for improving green tea classification.

P. Mishra, A. Karami, A. Nordon, D.N. Rutledge, J.-M. Roger, **CPACT research day**, University of Strathclyde (October 2018), De-noising, compression and data fusion: new approaches for automatically managing and processing hyperspectral imaging.

P. Mishra, A. Nordon, J. Morris, **CPACT research day**, GlaxoSmithKline, Stevenage, United Kingdom (March 2018), Spectral data fusion or multiblock data analysis: A summary of methods available.

**(Invited)** P. Mishra, A. Nordon, Tensor Meeting, University of Edinburgh, United Kingdom (March, 2017), t-distributed Stochastic Neighbour Embedding (t-SNE) for visualising high dimension hyperspectral data.

### ***Poster***

**(Invited)** P. Mishra, A. Nordon, **EFSA-2018**, Parma, Italy, In-situ spectral sensing to support evidence-based risk assessment for food safety, traceability and authenticity.

P. Mishra, A. Nordon, J. Morris, **Scandinavian Symposium on Chemometrics**, Naantali, Finland, 19-22 June 2017, Visualizing hyperspectral data with linear and non-linear dimensionality reduction methods

P. Mishra, A. Nordon, J. Tschannerl, G. Lian, S. Redfern, S. Marshall, J. Morris, **International conference on near-infrared spectroscopy (ICNIRS) 2017**, Copenhagen, Denmark, Near infrared hyperspectral imaging for classification of tea products.

P. Mishra, A. Nordon, H. Ferrero, B. Sarup, G. Sin, **EUROPACT 2017**, Potsdam, Germany, held during 10-12 May 2017, Nondestructive analysis of vegetable oils using spectroscopic measurements.

## **Award and achievements**

- CPACT's Malcolm McIvor award 2018, Glasgow, United Kingdom
- Accepted and funded proposal for European Researchers Night 2018, Brussels, Belgium
- Early stage researcher award from European Food Safety Authority (EFSA) 2018, Parma, Italy
- Accepted and funded proposal for participation in Falling Walls Lab 2017, Brussels, Belgium

## **Secondments and industrial trainings**

**4-15 September 2017**

**Unilever, Sharnbrook, United Kingdom**

- Introduction to tea products analysis
- Methodologies for tea products analysis
- Visit to green houses
- Introduction to tea processing

**4 June-19 October 2018**

**Unilever, Sharnbrook, United Kingdom**

- Tea products analysis with NIR sensors
- Chemometric modelling for predicting chemical constituents

**5-16 November 2018**

**Bayer AG, Leverkusen, Germany**

- Analysis of hyperspectral images from a continuous line for assessment of drought studies in plants
- Exploring automatic hyperspectral imaging facilities at Bayer AG

**14 January – 15 February 2019**

**Bayer AG, Leverkusen, Germany**

- Analysis hyperspectral images from a continuous line for assessment of drought studies in plants from automatic phenotyping platforms
- Exploring automatic hyperspectral imaging facilities at Bayer Crop science in Frankfurt, Germany
- Developing pipeline for hyperspectral images processing