

Applying Fourier-Transform Infrared Spectroscopy and Self-Organizing Maps for Forensic Classification of White-Copy Papers

Loong Chuen Lee^{#1,*}, Choong-Yeun Liong^{#2}, Abdul Aziz Jemain^{#3}

[#]School of Mathematical Sciences, FST, Universiti Kebangsaan Malaysia, Bangi, 43600, Malaysia

E-mail: ¹lc_lee@ukm.edu.my; ²lg@ukm.edu.my; ³azizj@ukm.edu.my

*Forensic Science Program, FSK, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300, Kuala Lumpur, Malaysia

Abstract— White-copy A4 paper is an important kind of substrate for preparation of most formal as well as informal documents. It often encountered as questioned document in cases such as falsification, embezzlement or forgery. By comparing the questioned piece, (e.g. of a contract) against the rest deemed authentic, forgery indicator could be derived from inconsistent chemical compositions. However, classification and even differentiation of white-copy paper have been difficult due to highly similar physical properties and chemical composition. Self-organizing map (SOM) has been proven useful in many published works as a good tool for clustering and classification of samples, especially when involving high-dimensional data. In this preliminary paper, we explore the feasibility of SOM in classifying white-copy paper for forensic purposes. A total of 150 infrared spectra were collected from three varieties of white paper using Attenuated Total Reflectance Fourier-transform infrared (ATR-FTIR) spectroscopy. Each IR spectrum composed of over thousands of wavenumbers (i.e. input variables) and resembles chemical fingerprint for the sample. Comparative performance between raw wavenumbers and its reduced form (i.e. principal components, PCs) in SOM modeling also conducted. Results showed that SOM built with PCs is much efficient than built with raw wavenumbers, with classification accuracy of over 90% is obtained with external validation test. This study shows that SOM coupled with ATR-FTIR spectroscopy could be a potential non-destructive approach for forensic paper analysis.

Keywords— self-organizing map; forensic paper analysis; IR spectrum; classification

I. INTRODUCTION

Forensic document examinations (FDEs) is composed of several sub-fields, i.e. ink analysis, handwriting examinations, that all concentrate on a piece of evidence known as questioned document. The goal is to authenticate the questioned document in hands so to find evidence to include or exclude the possibility of forgery [1]. For instance, paper analysis is a part of the routine analysis that could be conducted to seek for indicators of forgery, substitution, or falsification [2]. If there is no identified suspect, classification and identification tasks could be attempted on the piece of evidence to guide investigators to a narrower scope of investigative direction; otherwise, differentiation task can be attempted which involves a direct comparison between known and questioned samples. Fig. 1 illustrates the divisions of FDEs and the purpose of analysis.

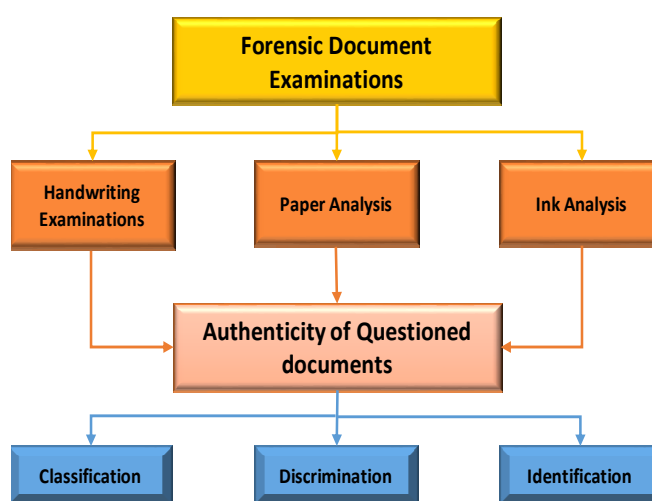


Fig. 1 Divisions of forensic document examinations and purpose of analysis

II. MATERIAL AND METHOD

A paper is usually composed of cellulose fibre derived from tree and some chemical additives to modify or improve the quality of final products, e.g. white copy paper or newsprint paper. Cellulose often selected as the main structural component in the paper because of its low price and availability. Chemically, it is organic compounds composed of carbon, hydrogen and oxygen atoms. In addition to that, hemicellulose and lignin also present together with raw cellulose and held together by lignin which the “natures glue” for the fibres. These are primarily natural components present in paper [3]. In current practice, differentiation of white A4 copy paper is seldom successful due to high similarity of composition between different paper manufacturers, except based on inorganic additives profiles.

Different kinds of instrumental technique have been applied to acquire organic or inorganic profiles of paper samples. Causin et al. [4] used diffuse reflectance UV-Vis-NIR spectroscopy to study optical brightener in the paper. X-ray based techniques are the favoured approaches to collect inorganic fingerprinting of paper samples. X-ray fluorescence (XRF) spectrometry is a well-established technique for elemental analysis at the micro and trace levels. XRF can be applied to generate an elemental profile of paper. On the other hand, X-ray diffraction (XRD) spectrometry depends on the crystalline structure of cellulose matrix and other trace elements to create analytical signals [5].

According to one recent review paper on forensic document examinations, spectroscopic techniques are replacing separation techniques and become the most popular analytical approaches in FDEs [6]. Attenuated Total Reflectance Fourier-Transform Infrared (ATR-FTIR) spectroscopy is especially favoured due to its versatility and non-destructiveness. It is a fast and cost-effective technique, as it does not require a sample preparation step [7].

The drawbacks of this technique are that the obtained infrared (IR) spectrum is often complicated and shows low signal intensity. To overcome that shortcoming, various multivariate statistical techniques could be applied to help pre-processing and interpreting high dimensional data [8-10]. Principal component analysis (PCA) and Linear Discriminant Analysis (LDA) being the most popular techniques when dealing with IR spectral-like data. Both techniques are powerful and supported with a vast number of published works [11]-[14].

Self-organizing map (SOM) is a part of pattern recognition methods that has grown in interest over the last years. Contrast to PCA or LDA, it is a non-linear approach that stems from artificial neural networks techniques [15]. On top of that, SOM can handle high-dimensional data without requiring the IR data to be reduced first. In essence, SOM can be a powerful visualization technique and classifier algorithm. It has been demonstrated to be a powerful statistical tool for classification task in various application fields [15]-[18].

In this preliminary paper, we study the feasibility of coupling advanced pattern recognition methods, i.e. SOM, with ATR-FTIR spectroscopy to differentiate three highly similar white-copy papers. ATR-FTIR spectroscopy has been applied in this study to collect organic profiles of paper in a non-invasive and non-destructive way.

A. Dataset

The dataset used in this study is composed of 150 IR spectra from three different varieties of white-copy papers as described in Table 1. For each paper types, at least 15 sheets are analysed from a single ream. Both surfaces of each single sheet are analysed using ATR-FTIR spectroscopy. The final dataset composed of 150 IR spectra that each described by 2701 wavenumbers (i.e. input variables). The same dataset has been studied previously using different statistical analysis strategies [19]-[21].

TABLE I
DETAILS OF SAMPLES USED IN THIS STUDY

ID	White	
	Manufacturer	Number of IR spectrum
IY	IK Yellow	29 x 2 = 58
OP	One Paper	29 x 2 = 58
SP	Save Pack	17 x 2 = 34
	Total	150

B. Self-Organizing Maps (SOMs)

SOM is one of the most widely used Artificial Neural Network (ANN) algorithms. It is mainly used to identify interesting patterns that it places similar data (i.e. in the input space) on the nearby map and present them as a clustering diagram [22]. By doing so, it helps to discover distribution of data and thus elucidating classes in data. Interested readers are directed to [23]-[24].

In this study, we aim to reveal hidden spatial distribution of the three varieties of paper using SOM. Although SOM is introduced as an unsupervised learning algorithm, in this study, we picked a supervised version of SOM in order to enable calculation of classification accuracy that will act as a figure of merit for selecting the best model.

C. Principal Component Analysis (PCA)

PCA is a classical dimensionality reduction (DR) method in statistical data analysis, i.e. reconstruct manifest variables into reduced hyperspace. PCA computes a set of orthogonal vectors, called the principal components (PCs) which explain the original data with maximum variance. These PCs represent a series of least squares fit the data, with each component being orthogonal to all the previous ones [25].

In this study, the over thousands of wavenumbers were reduced to few numbers of PCs via PCA. In this way, data dimension is reduced; noise could also be potentially excluded in the subsequent analysis.

D. Sampling Algorithms

Six different sampling algorithm, i.e. Honigs, Kennard-Stone, naes, Duplex, Select, Puchewein and are selected to split the IR dataset into six pairs of training and test sets according to pre-defined ratio. Working principles of each sampling algorithm will be described in preceding paragraphs.

Honigs algorithm is initially proposed to select training sample set in near-infrared diffuse-reflectance analysis, in which the most spectrally unique samples (i.e. maximum

absorbance) will be identified and selected. By incorporating sample with the larger variations in sample composition, the resulted predictive model will be more robust [26].

The Kennard-Stone algorithm performs training samples selection based on a uniform distribution over the predictor space. Data points are distributed according to Euclidean distance then the farthest pair of points will be assigned to the calibration set and removed from the list of points [27].

Contrast to KS, Naes algorithm incorporating K-means sampling into the selection of calibration samples from large multivariate datasets [28].

The Puchwein algorithm involves an iterative process in the elimination of similar samples as defined by the Mahalanobis distance. As such, there is not possible to the pre-defined number of samples [29].

The SELECT algorithm is another sampling algorithm working on the iterative procedure in which the observations having the highest number of neighbours within a given minimum distance (i.e. standardized Mahalanobis distance) is selected and its neighbours are discarded [30].

DUPLEX algorithm is similar to the Kennard-stone algorithm but allows to select both calibration and validation points that are independent [31].

E. Statistical Analysis

Fig. 2 summarizes the statistical analysis steps involved in this study. The dataset composed of 150 IR spectra from three varieties of white copy paper. Each spectrum is characterized by different absorbance value over 2701 wavenumbers.

For SOM modeling purposes, different types of input variables are derived from the raw IR spectra. Three IR spectral regions, i.e. MIR 1, MIR 2 and MIR 3, are derived from the raw spectra. MIR 1 resembles the global model whereas the other two regions, i.e. MIR 2 and 3, have been identified to be informative regions in previous papers [20-21]. For each IR region, its raw form (i.e. wn) and reduced form (i.e. PCs), respectively, is to be presented as input variables for SOM modeling. With that, a total of six SOM models have been built.

On the other hand, the 150 IR spectra are repeatedly split into a pair of training and test set, using six different sampling algorithms. For each pair of the dataset, six different prediction accuracy are derived from the six SOM models being constructed using the six different kinds of input variables. Selection of best model is based on external validation as it is known to be more accurate than internal validation (e.g. leave-one-out cross validation).

All statistical analysis was performed using the R software environment for statistical computing and graphics (R Core Team, Austria). The R package Kohonen offers supervised version of SOMs, i.e. xvf function [32]. The power of SOM is provided with exchange conditions that there are quite a number of parameters needed to be optimized before reaching a decision on the best model. For that, we have applied train function from caret package which sets up a grid of tuning parameters for SOM, fits each model and calculate a resampling-based performance measure [33].

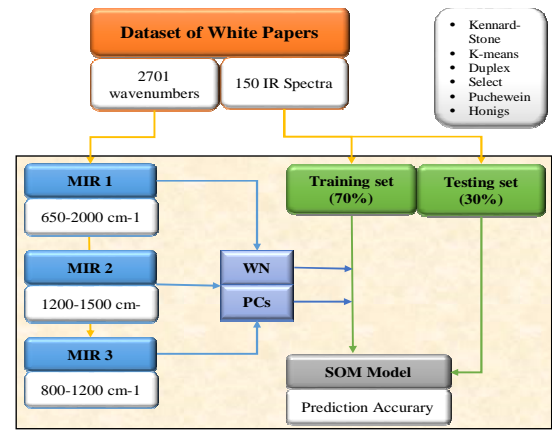


Fig. 2 Flow of statistical methods applied to the IR dataset of white copy papers as implemented in this study

III. RESULTS AND DISCUSSION

Fig. 3 illustrates the averaged IR spectra of the three paper varieties. It clearly seems visual inspection does not allow any form of clear discrimination between the three paper types. The most discriminating region being 1200 – 1500 cm-1 (i.e. MIR 2), whereas region between 800 – 1200 cm-1 (i.e. MIR 3) appeared to be an uninformative region (refer to Fig. 2). The three regions are to be used as input variables for SOM modeling.

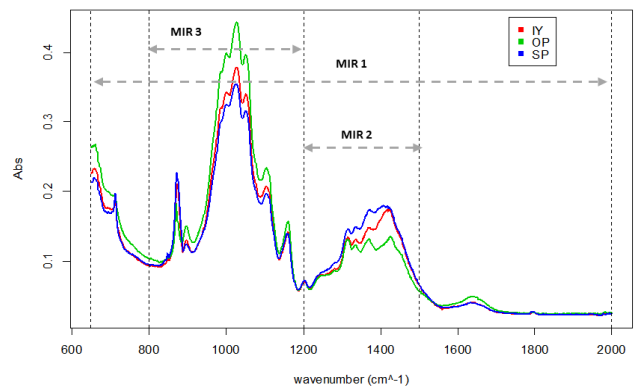


Fig. 3 Averaged IR spectra of the three paper types studied in this paper

This preliminary study aims to investigate benefits of using SOM in forensic classification problems using an example from FDE. SOM is a type of ANN that is expected to be good in handling high-dimensional data. However, based on the previous investigation on the same dataset [20-21], we decided to construct another set of input variables, i.e. Principal Components (PCs). In the following sections, performances of SOM models are discussed according to the three IR spectral regions, based on (a) type of input variables, i.e. raw and reduced form (PCs), and (b) type of column-wise manipulations (CWMs), i.e. mean centering (MC), variance scaling (VS) and autoscaling (AS). The consistency of performances of SOM models across the six different test sets is used to derive robustness of the model.

A. MIR 1

Fig. 4 shows the external validation results obtained with MIR 1, i.e. full spectral regions. Performances of raw wavenumbers and its reduced form (i.e. PCs) are respectively shown by panel (a) and (b), presented according to the form of data, i.e. raw untreated IR, mean-centered IR (MC), variance scaled IR (VS) and autoscaled IR (AS).

With raw wavenumbers as input variables, mean-centered IR gave similar performances like untreated raw IR spectra that both are showing slightly higher accuracy than the respective autoscaled or variance scaled IR spectra. Similar trend of performances also observed by SOM models constructed using PCs as input variables (Fig. 4 (b)).

By comparing the mean accuracy of these SOM models, PCs derived from untreated IR spectra showed the highest rank with mean accuracy 80%, slightly higher than 78% obtained with raw wavenumbers of untreated IR spectra.

B. MIR 2

MIR 2 composed of the primary informative regions for discriminating the three paper varieties. Contrast to MIR 1, a number of variables is much lesser, reduced from 2701 to 601, only preserved around 22% of raw variables. Based on Fig. 5, we noticed that all SOM models, using either raw wavenumbers or PCs, are approaching nearly similar external accuracy like those constructed with MIR 1 (Fig. 4), even though included smaller IR spectral regions. One more interesting features of these SOM models is all of them are showing quite good agreement among the six test sets contrast with previous models (Fig. 4).

Both batches of SOM models, built with either wavenumbers or PCs, are giving similar accuracy, between 74-75%. This is definitely relatively lower than the ones built with MIR 1 which achieved 80% accuracy. However, it was to consider robustness as one of the model selection criteria, we think MIR 2 is slightly superior to MIR 1.

C. MIR 3

MIR 3 representing an uninformative IR spectral region because differences observed among the three paper varieties most likely caused by elevated baseline which does not stem from inherent compositional variations of samples (Fig. 3). As illustrated by Fig. 6, most SOM models hardly approaching accuracy near to 80%, but located around 60-70%. Again, similar inconsistent accuracy rate across the six test sets are observed with SOM models constructed with MIR 3. Again, using PCs as input variables do not seem to outperform its raw form, i.e. wavenumbers.

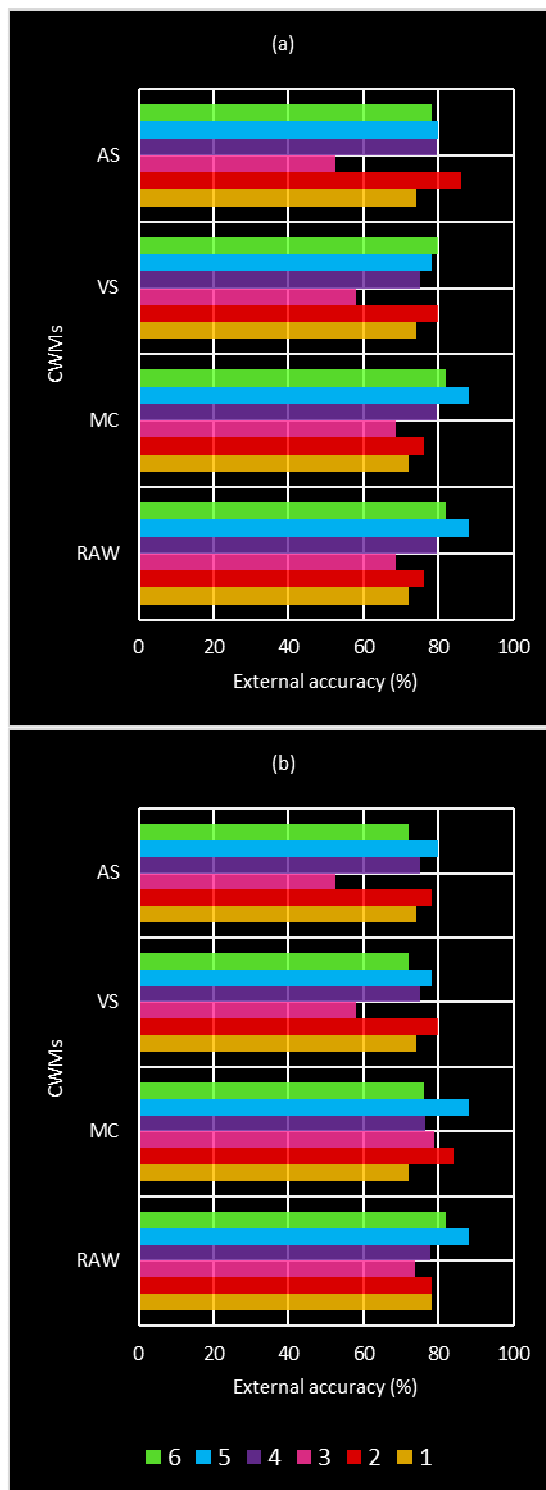


Fig. 4 External validation accuracy obtained with raw IR spectra (RAW) and its treated form, i.e. mean-centering (MC), variance scaling (VS) and autoscaling (AS), using (a) raw wavenumbers and (b) PCs, from MIR 1. The six test sets are respectively created with (1) Kennard-Stone, (2) Naes, (3) Select, (4) Puchwein, (5) Honigs and (6) DUPLEX

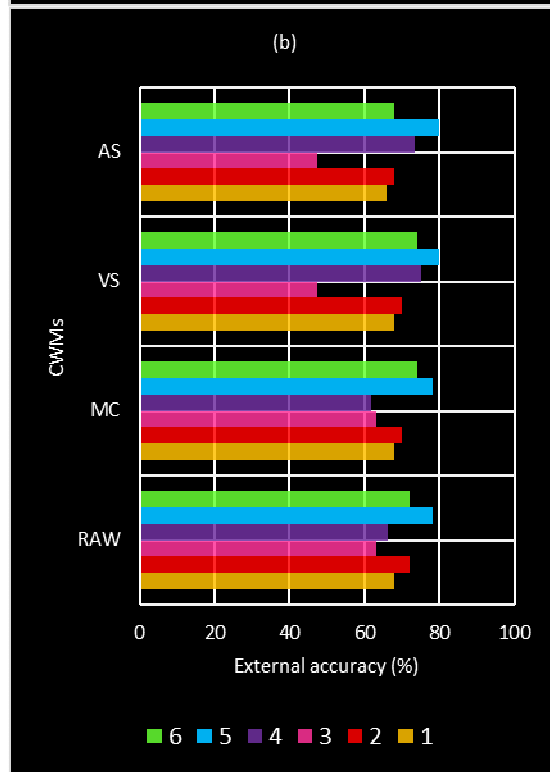
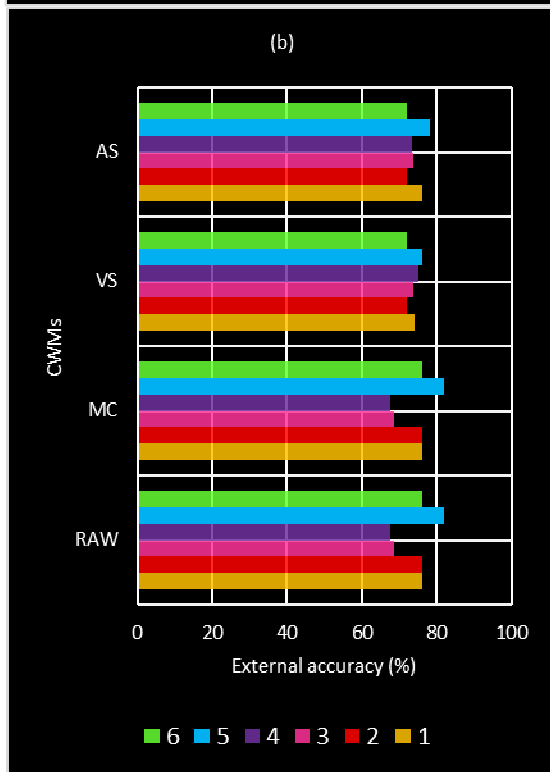
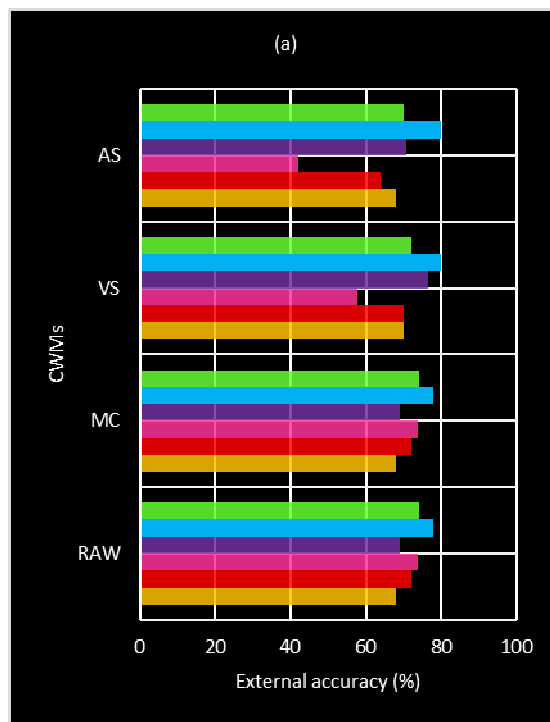
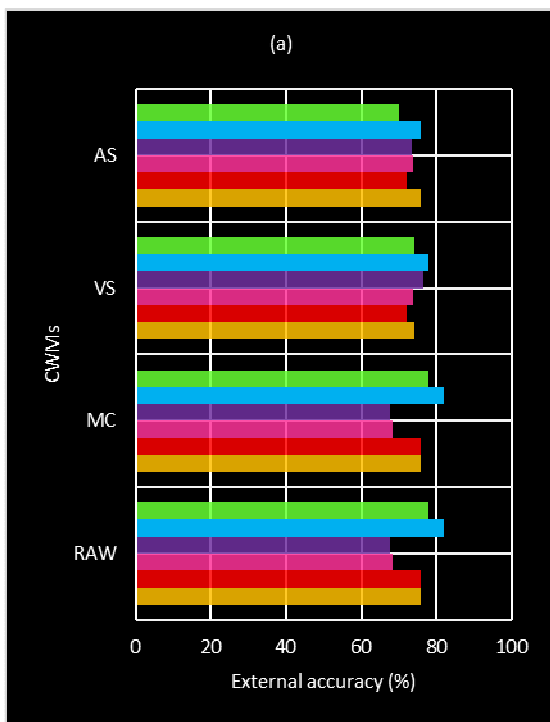


Fig. 5 External validation accuracy obtained with raw IR spectra (RAW) and its treated form, i.e. mean-centering (MC), variance scaling (VS) and autoscaling (AS), using (a) raw wavenumbers and (b) PCs, from MIR 2. The six test sets are respectively created with (1) Kennard-Stone, (2) Naes, (3) Select, (4) Puchwein, (5) Honigs and (6) DUPLEX

Fig. 6 External validation accuracy obtained with raw IR spectra (RAW) and its treated form, i.e. mean-centering (MC), variance scaling (VS) and autoscaling (AS), using (a) raw wavenumbers and (b) PCs, from MIR 3. The six test sets are respectively created with (1) Kennard-Stone, (2) Naes, (3) Select, (4) Puchwein, (5) Honigs and (6) DUPLEX

D. Comparative Analysis

On overall, MIR 1, 2 and 3 produced modelling accuracy ranged between 85-70%, with the former showed the highest mean accuracy at 80%. The distinguished difference between the three batches of SOM models are consistency accuracy over the six test sets with MIR 2 is the only one gave most robust models despite not able to produce the highest accuracy.

Based on a previous study [21], we noticed that PC 3 and 5 calculated from MIR 1 are the most discriminating variables for the three paper varieties. Therefore, we also built a series of SOM models using PC 3 and 5 constructed from untreated, mean-centered, variance scaled and autoscaled MIR 1 (Fig. 7). PCs derived from mean-centered IR spectra showed lowest accuracy contrast to PCs calculated from autoscaled IR spectra which showed over 90% accuracy. With that, we decided to use only PC 3 and 5 to build the final SOM model for data at hand.

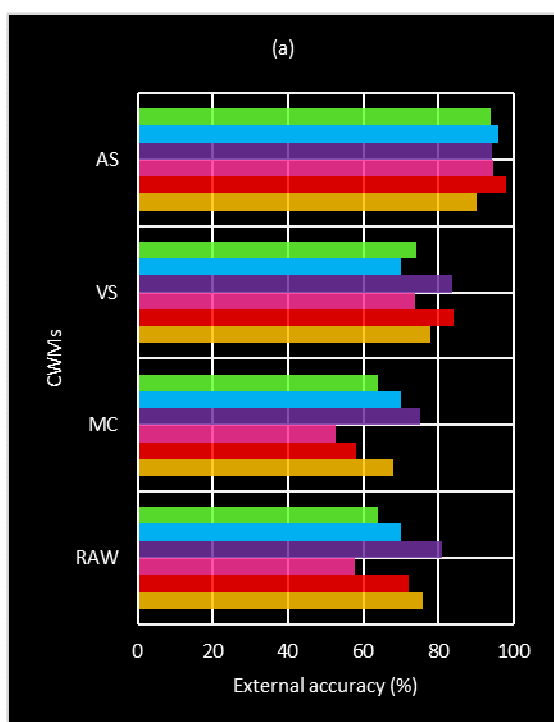


Fig. 7 External validation accuracy obtained with PC 3 and 5 calculated from raw, mean-centered, variance scaled and autoscaled MIR 1. The six test sets are respectively created with (1) Kennard-Stone, (2) Naes, (3) Select, (4) Puchwein, (5) Honigs and (6) DUPLEX.

E. SOM Model

In this study, a 3 x 4 map (i.e. in hexagonal network topology) is selected to map the 100 IR spectra of papers (i.e. the training set), and the weight given to the X map in the calculation of Tanimoto distance for updating Y is set at 0.5. This leads to the plots in Fig. 4. The network has captured most of the variability in the data as the three classes of paper are visible on the map with only two samples being misplaced. The plot indicates that all units composed of homogeneous samples except two units. IY showed the highest diversity in the patterns and is mapped with seven units.

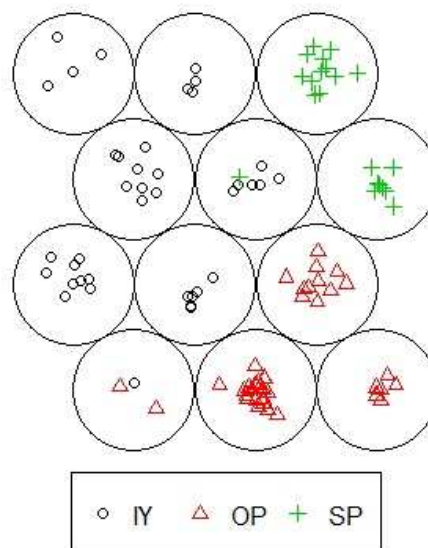


Fig. 8 Separations of 100 paper samples on a 3 x 4 topological regions

IV. CONCLUSION

In conclusions, the benefits of using SOM coupling with ATR-FTIR spectroscopy were demonstrated. Although it does not achieve 100% classification accuracy, major benefits of SOM is its ability to handle many complex data and is relevant to most the forensic science research that are targeting large databases in the future. Another advantage of SOMs that is relevant to forensics applications is that they can be updated easily because forensic analysis often accounted new data during investigations.

ACKNOWLEDGMENT

The authors would like to thank SAC Narenasagaran A/L Thangaveloo, Supt. Shaikh Abdul Adzis bin Shaikh Abdullah and Mr. Rizal of the Makmal Maktab PDRM, Cheras for their valuable help in acquiring the IR data. Financial support from the KPTM (grant no. FRGS/2/2013/ST06/UKM/02/1) is gratefully acknowledged.

REFERENCES

- [1] D. Ellen, *The Scientific Examination of Questioned Documents: Methods and Techniques*, London, UK: CRC Press, 1997.
- [2] J. S. Kelly and B. S. Lindblom, *Scientific Examination of Questioned Documents*, Boca Raton: CRC Press, 2006.
- [3] F. M. Udristoiu, I. G. Tanase, A. A. Bunaciu and H. Y. Aboul-Enenods, "Paper analysis: non-destructive and destructive analytical methods," *Appl. Spectrosc. Rev.*, vol. 47, pp. 550-570, 2012.
- [4] V. Causin, R. Casamassima, G. Marruncheddu, G. Lenzoni, G. Peluso and L. Ripani, "The discrimination potential of diffuse reflectance ultraviolet-visible-near infrared spectrophotometry for the forensic analysis of paper," *Forensic Sci. Int.* vol. 216, pp. 163-167, 2012.
- [5] A. van Es, J. de Koeijer and G. van der Peijl, "Discrimination of document paper by XRF, LA-ICP-MS and IRMS using multivariate statistical techniques" *Sci. Justice*, vol. 49, pp. 120-126, 2009.
- [6] M. Calcerrada and C. Garcia-Ruiz, "Review on the analysis of questioned documents," *Anal. Chimica Acta*, vol. 853, pp.143-166, 2015.

- [7] C. K. Muro, K. C. Doty, J. Bueno, L. Halamkova and I. K. Lednev, "Vibrational spectroscopy: recent developments to revolutionize forensic science," *Anal. Chem.*, vol. 87, pp. 306-327, 2015.
- [8] A. van Es, J. de Koeijer and G. van der Peijl, "Discrimination of document paper by XRF, LA-ICP-MS and IRMS using multivariate statistical techniques," *Sci. Justice*, vol. 49, pp.120-126, 2009.
- [9] L. D. Spence, A. T. Baker and J. P. Byrne, "Characterization of document paper using elemental compositions determined by coupled plasma mass spectrometry," *J. Anal. Atom. Spectro.*, vol. 15, pp. 813-819, 2000.
- [10] A. Kher, S. Stewart, and M. Mulholland, "Forensic classification of paper with infrared spectroscopy and principal component analysis," *J. Near Infrared Spectrosc.*, vol. 13, pp. 225-229, 2005.
- [11] A. Racz, K. Heberger and M. Fodor, "Quantitative determination and classification of energy drinks using near-infrared spectroscopy," *Anal. Bioanal. Chem.* vol. 408, pp. 6403-6411, 2016.
- [12] K. Porker, M. Zerner and D. Cozzolino, "Classification and Authentication of Barley (*Hordeum vulgare*) malt varieties: Combining attenuated total reflectance mid-infrared spectroscopy with chemometrics," *Food Anal. Methods*. In press
- [13] C. A. Lima, V. P. Goulart, L. Correa, D. M. Zzell, "Using Fourier transform infrared spectroscopy to evaluate biological effects induced by photodynamic therapy," *Lasers in Surgery and Medicine*. vol. 48, pp. 538-545, 2016.
- [14] G. Theophilou, K. M. G. Lima, P. L. Martin-Hirsh, H. F. Stringfellow and F. L. Martin, "ATR-FTIR spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue: classifying subtypes of human cancer," *Analyst*, vol. 141, pp. 585-594, 2016.
- [15] A. Yan, X. Nie, K. Wang and M. Wang, "Classification of Aurora kinase inhibitors by self-organizing map (SOM) and support vector machine (SVM)," *European Journal of Medicinal Chemistry*, vol. 61, pp. 73-83, 2013.
- [16] W. N. S. Mat-Desa, D. Ismail and N. N. Daeid, "Classification and source determination of medium petroleum distillates by Chemometric and Artificial neural networks: A self-organizing feature approach," *Anal. Chem.* vol. 83, pp. 7745-7754, 2011.
- [17] K. S. Mattingly, C. A. Ramseyer, J. J. Rosen, T. L. Mote and R. Muthyala, "Increasing water vapor transport to the Greenland Ice Sheet revealed using self-organizing maps," *Geophysical Research Letters*, vol. 43, pp. 9250-9258, 2016.
- [18] S-G. Lee and E-Y Cha, "Style classification and visualization of art painting's genre using self-organizing maps," *Hum. Cent. Comput. Inf. Sci.* vol. 6, pp. 1-11, 2016.
- [19] L. C. Lee, C-Y Liong, K. Osman, and A. A. Jemain, "Forensic differentiation of paper by ATR-FTIR spectroscopy technique and PLS-DA," in *AIP Conference Proceedings*, vol. 1750, 2016, p. 060016.
- [20] C-Y Liong, L. C. Lee, K. Osman and A. A. Jemain, "Genetic algorithm for wavenumber selection in forensic differentiation of paper by LDA," in *AIP Conference Proceedings*, vol. 1750, 2016, p. 060017.
- [21] L. C. Lee, C-Y Liong, K. Osman and A. A. Jemain, "Forensic classification of white paper based on ATR-FTIR spectrum and Multivariate statistical technique," in *7th Asian Forensic Sciences Network*, 2015, pp. 93.
- [22] T. Kohonen, "Self-Organized Formation of Topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59-69, 1982.
- [23] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52-65, 2013.
- [24] S-L Shieh and I-En Liao, "A new approach for data clustering and visualization using self-organizing maps," *Expert Syst. Appl.* vol. 39, pp. 11924-11933, 2012.
- [25] R. Bro, A. K. Smilde, "Principal component analysis," *Anal. Methods*, vol. 6, pp. 2812-2831, 2014.
- [26] D. E. Honigs, G. M. Hieftje, H. L. Mark and T. B. Hirschfeld, "Unique-Sample Selection via Near-Infrared Spectral Subtraction," *Anal. Chem.*, vol. 57, pp. 2299-2303, 1985.
- [27] R. W. Kennard and L. A. Stone, "Computer aided design of experiments," *Technometrics*, vol. 11, pp.137-148, 1969.
- [28] T. Naes, "The design of calibration in near infra-red reflectance analysis by clustering," *J. Chemo.* vol. 1, pp. 121-134, 1987.
- [29] G. Puchwein, "Selection of calibration samples for near-infrared spectrometry by factor analysis of spectra," *Anal. Chem.* vol. 60, pp. 569-573, 1988.
- [30] J. S. Shenk and M. O. Westerhaus, "Population definition, sample selection and calibration procedures for Near Infrared Reflectance Spectroscopy," *Crop. Science.* vol. 31, pp. 469-474, 1991.
- [31] R. D. Snee, "Validation of regression models: methods and examples," *Technometrics*, vol. 19, pp. 415-428, 1977.
- [32] Wehrens, R. (2015, Sept 04). Package 'kohonen' [Online]. Available: <https://cran.r-project.org/web/packages/kohonen/kohonen.pdf>
- [33] Kuhn, M. (2016, Apr 11). Package 'caret' [Online]. Available: <https://cran.r-project.org/web/packages/caret/caret.pdf>
- [34] Kuhn, M. (2016, Apr 11). Package 'caret' [Online]. Available: <https://cran.r-project.org/web/packages/caret/caret.pdf>
- [35] Stevens, A., Ramirez-Lopez, L (2015, Feb 20). Package 'prospectr' [Online]. Available: <https://cran.r-project.org/web/packages/prospectr/prospectr.pdf>