Analyst

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2025, **150**, 1785

Raman spectroscopy and machine learning for forensic document examination

Yong Ju Lee, [a] Chang Woo Jeong,[b] Hong Taek Kim,[c] Tai-Ju Lee[a] and Hyoung Jin Kim*[a]

Received 11th December 2024,
Accepted 18th March 2025

DOI: 10.1039/d4an01529k

rsc.li/analyst

Forensics relies on the differentiation and classification of document papers, particularly in cases involving document forgery and fraud. In this study, document papers are classified by integrating Raman spectroscopy with machine learning models, namely, random forest (RF), support vector machines (SVMs), and feed-forward neural networks (FNNs). Among the machine learning models, the RF model effectively calculated the feature importance and identified the critical spectral region contributing to classification, enhancing the transparency and interpretability of the result. Spectral preprocessing with the first derivative significantly improved the classification performance. The spectral range 200–1650 cm^{-1} was identified as a highly informative region for differentiation, reducing the number of input variables from 756 to 360 while enhancing the model accuracy. The FNN model outperformed the RF and SVM models, with an F1 score of 0.968. The results underscore the potential of combining Raman spectroscopy with machine learning for forensic document examination, offering an interpretable, computationally efficient, and robust approach for paper classification.

1. Introduction

Forensic investigations require the identification and differentiation of document papers, particularly in cases of document forgery related to crime scenes and taxation. Forensic-paper analysis plays a pivotal role in the detection of forgeries such as falsified dates on real estate contracts, helping to establish authorship, verify authenticity, and detect discrepancies in comparisons of seized paper samples.[1–8] Document forgery often involves the altering or replacement of pages without the knowledge of all involved parties. In such cases, the characteristics of the paper of the altered page can provide crucial evidence to support or refute a forgery claim.

Conventional methods such as fiber identification, filler composition analysis, and fluorescence analysis have been widely used in document examination.[9] However, these methods often require large sample sizes or destructive testing, which limit their applicability. Advanced non-destructive analytical techniques, including X-ray diffraction,[10,11] elemental analysis,[1–3] infrared spectroscopy,[6,12,13] image analysis,[5] and pyrolysis gas chromatography,[14] have expanded the toolbox of forensic document examination. Despite their effectiveness, these methods generate large datasets that are error-prone and time-consuming when processed manually.

These limitations have been overcome by various chemometric approaches for handling complex datasets.[15] Lee *et al.*[16] demonstrated the forensic value of integrating chemometric techniques with spectroscopic data. The classification and regression tree (CART) method, which combines attenuated total reflectance-Fourier transform infrared spectroscopy with principal component analysis (PCA), distinguished white copy paper with a prediction accuracy of nearly 90%. Similarly, diffuse reflectance ultraviolet–visible–near infrared spectroscopy combined with PCA discriminated among writing, office, and photocopy papers with an accuracy of up to 99.7%.[17]

Raman spectroscopy is a vibrational spectroscopic technique that analyzes molecular interactions through scattering rather than absorption. In this respect, Raman spectroscopy differs from infrared (IR) spectroscopy. While Raman spectroscopy primarily detects the vibrations of homonuclear bonds such as C=C and S–S, IR spectroscopy is more sensitive to polar functional groups such as C=O and C–O–C.[18,19] In chemometric approaches for forensic document examination, IR spectroscopy has been extensively studied while Raman spectroscopy remains underutilized. Few studies have integrated Raman spectroscopy with machine learning techniques for forensic applications.[6,12,16,20–22]

[a]*Department of Forest Products and Biotechnology, Kookmin University, 77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Republic of Korea. E-mail: hyjikim@kookmin.ac.kr*
[b]*Graduate School of Scientific Criminal Investigation, Chungnam National University, Daejeon 34134, Korea*
[c]*Department of Electrical Engineering, Korea University, Seongbuk-gu, Seoul, Republic of Korea*

The present study tests the abilities of different machine learning models—random forest (RF), support vector machine (SVM), and artificial neural networks (ANN)—in the classification of document-paper manufacturers from Raman spectral data. Among these models, the RF model can calculate the feature importance, identifying the critical spectral regions contributing to classification and enhancing the transparency and interpretability of the results. By incorporating feature importance into the model development process, this study achieves robust classification performance while minimizing the computational costs. The findings demonstrate that Raman spectroscopy combined with machine learning effectively analyzes forensic documents, offering an efficient and interpretable approach for paper classification.

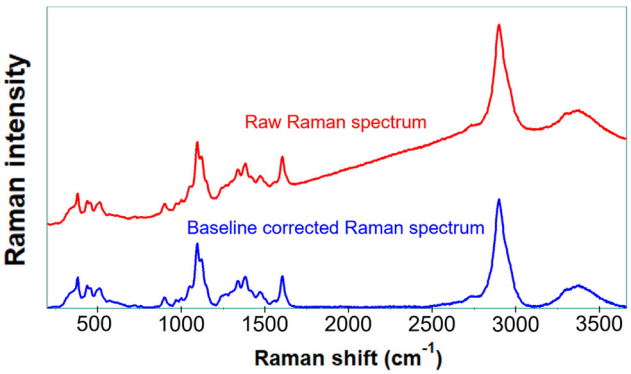

Fig. 1 Raw Raman spectrum and baseline corrected Raman spectrum of KOR1.

2. Experimental

2.1. Document paper

The methods were tasked with classifying 10 commercial document papers from 10 different products in five countries (see Table 1). Each product was selected based on its country of production and market share ranking. All samples were typical white-colored office papers with a grammage of 80 g m^{-2}.

2.2. Dataset

2.2.1. Raman spectra. The Raman spectrum of each paper was obtained on a confocal micro-Raman spectrometer (LabRAM Soleil, HORIBA France SAS, France) using a 532 nm, 5.2 mW laser as the excitation source. The Raman-scattered light was detected using a charge-coupled device detector (Syncerity, HORIBA Instruments Incorporated, USA) with a 600 g mm^{-1} grating configuration. All Raman spectra were collected within a 2 s acquisition time with 5 accumulations. The spectral range was 200–3600 cm^{-1} and the spatial resolution was less than 1 μm. Samples were focused under a microscope with a 100× objective lens. A total of 100 Raman spectra were acquired by measuring 10 spectra per sample from 10 different paper samples. Finally, the Raman spectra of 200–3000 cm^{-1} were used to construct classification models.

2.2.2. Data preprocessing. First, the Raman spectra were preprocessed to remove cosmic ray effects and baseline drift and the spectral quality was enhanced by denoising preprocessing.[21–23] The baseline was corrected with polynomial fitting.[24] Fig. 1 shows the raw Raman spectrum and baseline-corrected Raman spectrum of KOR1. Then, the Raman spectra were preprocessed using a Savitzky–Golay filter.[25] The original spectra were converted into first-derivative spectra using a third-degree polynomial with 13-point smoothing. Finally, the Euclidean (L2) norm vector of the preprocessed Raman spectrum was calculated as:

$$\text{Normalized vector} = \frac{v}{\sqrt{\sum_{i=1}^{n} |v_i|^2}}, \quad (1)$$

where v is the vector (Raman spectrum) to be normalized, v_i is the i^{th} element (*i.e.*, data point) of the vector, and n is the total number of elements in the vector. The dataset of normalized original and first derivative Raman spectra was used to construct classification models.

2.2.3. Dataset splitting. The Raman spectral dataset was divided into training and test sets at a ratio of 7 : 3. The training and test sets were used for model construction and validation, respectively. The ratio of each data class was preserved with a stratified sampling method. Threefold cross-validation was also performed to avoid overfitting and enhance the predictive performance of the models.

Table 1 Paper samples collected and analyzed in the present study

No.	Sample	Country	Manufacturer	Grammage (g m^{-2})
1	KOR1	Korea	A	80
2	KOR2			
3	KOR3			
4	KOR4			
5	IDN1	Indonesia	B	
6	IDN2		C	
7	CHN1	China	D	
8	CHN2		E	
9	THA	Thailand	F	
10	BRA	Brazil	G	

2.3. PCA

The underlying structure of the Raman spectral data was determined through PCA. In this step, the high-dimensional Raman spectra were projected onto a new orthogonal coordinate system represented by ten principal components (PCs). By visualizing the data in two-dimensional space, PCA enables the exploration of inherent patterns within the dataset, enhancing the interpretability of the data.

2.4. Random forest

The RF classifier[26] is an ensemble learning method that effectively mitigates premature convergence. Ensemble learning combines the outputs of multiple models, enhancing the prediction accuracy beyond the capabilities of individual classi-

fiers. RF can also analyze the importance of features and is commonly applied in pattern recognition tasks. In this study, the fundamental components of the RF model were decision trees (DTs).[27]

To increase diversity among the DTs, the RF algorithm employs random subsampling, which prevents the simultaneous use of all input variables and fosters the development of independent trees. The subsampling method of RF is bootstrap sampling, in which data points are randomly selected with replacement from the training dataset. The DTs are trained on approximately two-thirds of the data, known as in-bag samples; the remaining data, referred to as out-of-bag (OOB) samples, are reserved for validating the performance of the tree models.[28]

The probability of a data point being excluded from a set of m samples during random sampling with replacement is $(m - 1)/m$. When this process is repeated m times, the likelihood of a sample being excluded from all iterations converges to approximately 36.8%, as expressed in eqn (2):

$$\mathrm{OOB} = \lim_{m\to\infty}\left(1 - \frac{1}{m}\right)^{m} = \mathrm{e}^{-1} \approx 0.3678. \quad (2)$$

The RF model was developed by training multiple DTs on in-bag samples. The predictions of all trees were averaged to obtain the final classification of new data. Following the CART approach,[27] all DTs in the RF model were independently constructed without pruning. This study trialed different input variables (n_feature) of tree generation square root (sqrt), binary logarithm (log 2), and one-third (1/3) of the total spectral points, and different numbers of trees (n_tree) (10 to 500). The values of n_feature and n_tree were optimized by minimizing the OOB errors through a grid search approach.[29]

2.5. Feature-importance measure

The significance of the spectral variables was evaluated through the mean decrease impurity (MDI)[30] method, which identifies the Raman shifts dominating the document-paper classification. The process by which tree-based models assess the impact of input variables on the classification outcome is called feature-importance assessment or variable importance assessment.[13,29] The feature importance $I(n_j)$ is determined as:

$$I(n_j) = w_j C_j - w_{L_j} C_{L_j} - w_{R_j} C_{R_j}, \quad (3)$$

where n_j is the parent node, L_j and R_j are the left and right child nodes branched from n_j, respectively, w_j is the node weight, which equals the number of samples, and C_j is the impurity of n_j. The importance of variable i in a DT is computed as

$$I(f_i) = \frac{\sum_j I(n_j)}{\sum_{k\in n_{\mathrm{all}}} I(n_k)}, \quad (4)$$

where $I(f_i)$ represents the importance of variable i within the DT model. In an RF model, the importance scores from all DTs in the ensemble are aggregated to give the overall variable importance. Before aggregation, the importance score of each variable is normalized using eqn (5) to ensure consistency across the ensemble:

$$\mathrm{norm}I(f_i) = \frac{I(f_i)}{\sum_{j\in f_{\mathrm{all}}} I(f_j)}. \quad (5)$$

Subsequently, the final importance of the variable in the RF model is averaged over all DTs as follows:

$$I(\mathrm{RF}_i) = \frac{\sum_{j\in t_{\mathrm{all}}} \mathrm{norm}I(f_{i_j})}{T}, \quad (6)$$

where t denotes the DT model, $\mathrm{norm}I(f_{i_j})$ is the normalized importance of variable i in the RF model, and T is the total number of DTs.

2.6. Model comparison

The classification capabilities of the developed RF models were compared with those of two traditional machine learning algorithms: feed-forward neural networks (FNN) and SVMs. All models were trained on the same dataset, enabling direct comparison of their classification performances.

2.6.1. FNN. The ANN classifier was developed on an FNN architecture employing the backpropagation algorithm. The activation function was the rectified linear unit and the loss function was the cross-entropy. The loss function was optimized through stochastic gradient descent (SGD) and adaptive-moment estimation (Adam). The learning rate ranged from 0.0001 to 0.1 with a maximum of 1000 iterations. The FNN architecture included one or two hidden layers, each with 16, 32, 64, 128, 256, or 512 nodes. The hyperparameters were optimally configured by fine-tuning using a grid search method.

2.6.2. SVM. The SVM models were constructed with a radial basis function kernel.[31] The SVM algorithm seeks the optimal hyperplane that maximizes the margin between different data classes in a high-dimensional space. The performance of SVM is governed by two primary hyperparameters: cost and gamma. The cost parameter balances the margin maximization with the reduction of misclassification errors in the training data, whereas the gamma parameter determines the flexibility of the Gaussian kernel, thus affecting the model's ability to handle nonlinear relationships. In this study, the cost and gamma values were ranged from 2^{-5} to 2^{5} and from 10^{-5} to 10^{5}, respectively, and the hyperparameter combination was optimized through a grid search method.

2.7. Evaluation metric

In classification tasks, correctly identified observations in the positive and negative classes are referred to as true positives (TP) and true negatives (TN), respectively, positive-class observations that are misclassified as negative are labeled as false negatives (FN), and negative-class observations incorrectly classified as positive are identified as false positives (FP).[32]

The F1-score is a key performance metric that effectively balances the precision–recall tradeoff.[33] Precision measures the proportion of correctly identified positive cases among all predicted positive cases, and recall quantifies the model's ability to identify positive cases among all actual positive cases. Calculated as the harmonic mean of precision and recall, the F1-score more robustly determines the classification performance than accuracy, which may not adequately reflect the model's ability to handle FP and FN. The precision, recall, and F1-score metrics are, respectively, calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \tag{8}$$

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{9}$$

All data processing and classification modeling were conducted using R statistical software (R Core Team, ver. 4.4.1, Auckland, New Zealand).

3. Results and discussion

3.1. Raman spectral characteristics of the document papers

Fig. 2 presents the Raman spectra of the document paper samples. The prominent peaks at 280 and 1084 cm^{-1} correspond to lattice vibrations associated with translations/vibrations of the $(CO_3)^{2-}$ group and the symmetric stretching mode of the carbonate ion, both originating from calcium carbonate ($CaCO_3$).[34] These peaks are attributed to inorganic fillers, which are commonly incorporated during the papermaking process to reduce production costs and improve the optical properties of the paper.[35]

Additional peaks at 380 and 436 cm^{-1} correspond to torsional and flexural vibrations of the pyran ring and to bending and expansion vibrations of the CCO framework within the pyran ring, respectively.[36,37] The peaks at 508 and 1117 cm^{-1} are attributed to the C–O–C glycosidic linkages in cellulose,[38,39] and those at 1337 and 1380 cm^{-1} are associated with HCC, HCO, and HOC bending and with CH and CH_2 stretching in the carbohydrate components (cellulose and hemicellulose).[40,41]

The peak at 1602 cm^{-1} corresponds to aromatic ring stretching of lignin, while the peak at 1660 cm^{-1} is attributed to ring-conjugated C–C stretching of coniferyl alcohol and C=O stretching of coniferyl aldehyde, both occurring in lignin.[22,42] Finally, the peak at 2895 cm^{-1} represents CH and CH_2 stretching vibrations in cellulose.[40]

3.2. PCA of the Raman spectra

Fig. 3 presents the PCA score plots of the first two principal components (PCs) obtained from the original Raman spectra and the first derivative spectra. In the score plot of the original spectra (Fig. 3a), the data points form large clusters. In contrast, in the score plot of the first derivative spectra, some data points are separated from the main cluster, resulting in the

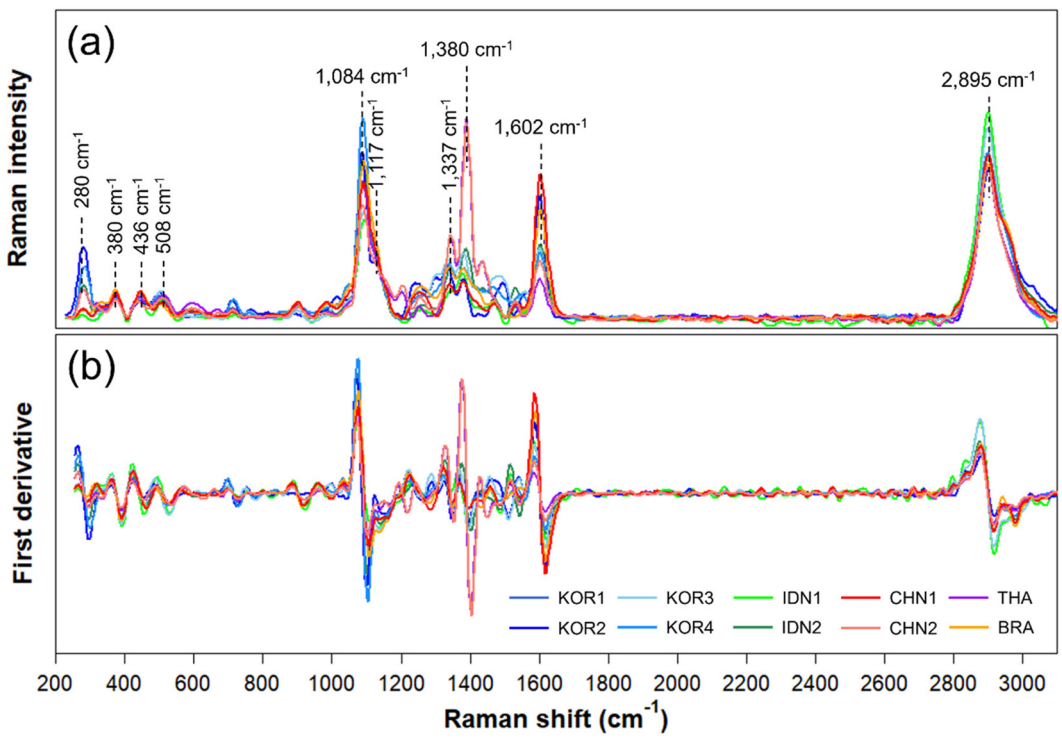

Fig. 2 Original (a) and first derivative (b) Raman spectra of the document paper samples.

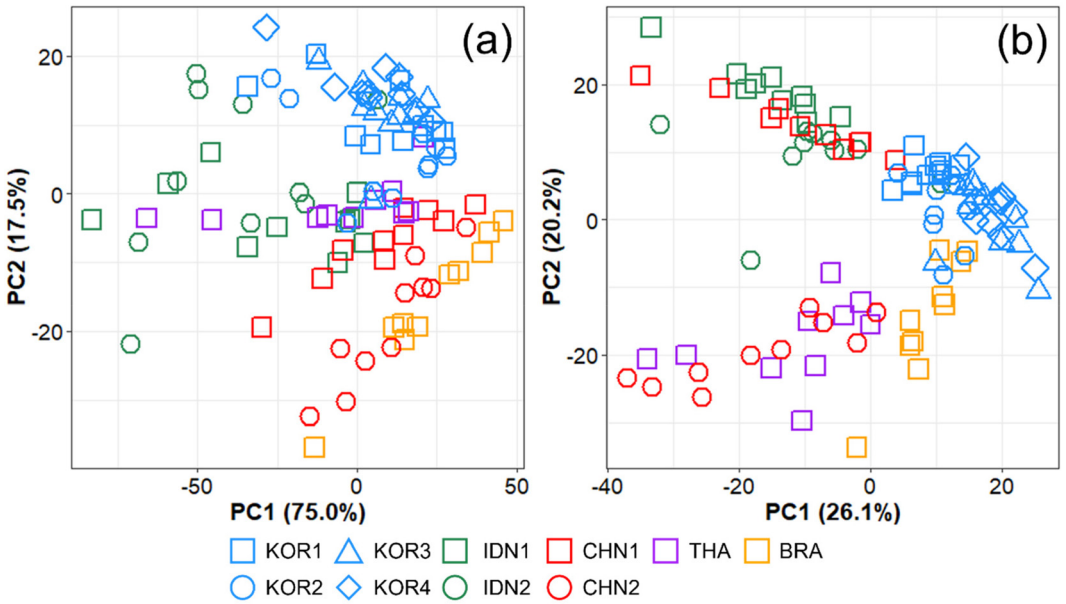

Fig. 3 PCA score plots of the original Raman spectra (a) and first derivative spectra (b). Percentages in parentheses are the scores of the explained variance of each PC.

formation of more distinct clusters. These results indicate that spectral preprocessing using the first derivative is a promising approach for enhancing the differentiation of document papers using Raman spectra.

Fig. 4a shows the Raman spectra of the KOR4, IDN2 and THA products and Fig. 4b presents the PC1 and PC2 loadings of the first derivative Raman spectra of the document paper. The peaks at 280 cm^{-1} and 1084 cm^{-1} (Fig. 4b) correspond to $CaCO_3$ from inorganic fillers, which explains their positioning along the PC1 axis, reflecting differences between KOR4 and THA. The variation in ash content, in itself, serves as evidence supporting these distinctions.[6,11] The peak at 1380 cm^{-1} was attributed to HCC, HCO, and HOC bending, as well as CH and CH_2 stretching in the carbohydrate components (cellulose and hemicellulose), which partially explains the differences in positioning along the PC2 axis as shown in Fig. 3b. The residual

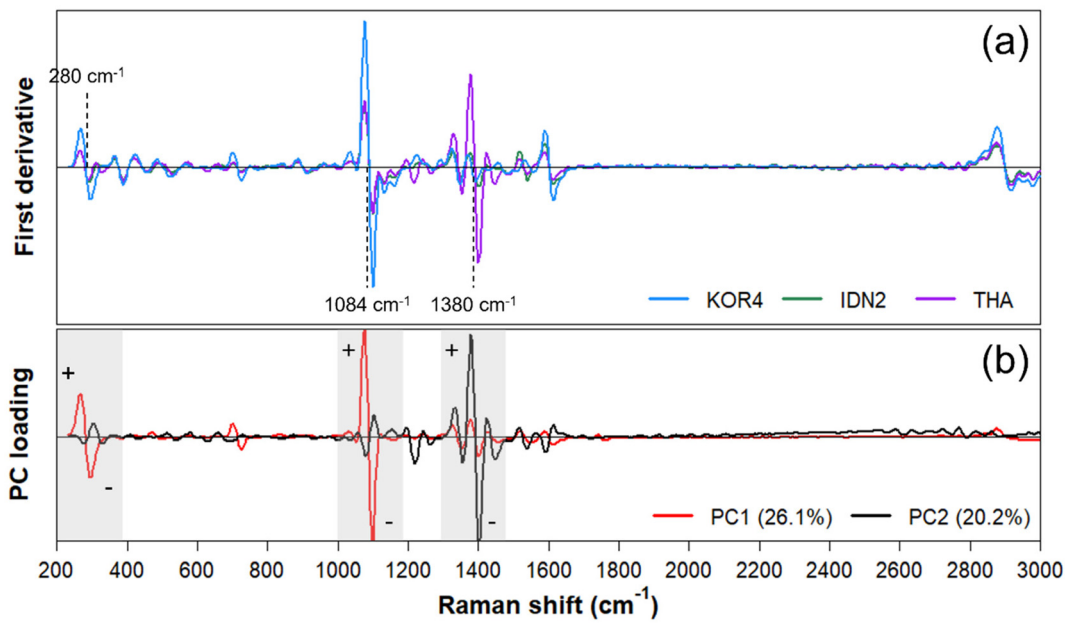

Fig. 4 Raman spectra of the KOR4, IND2 and THA samples (a) and loadings of the first two PCs of the first derivative Raman spectra of the document paper (b).

carbohydrates were influenced by the alkali charge, temperature, and processing time during cooking and bleaching, resulting in variations in the xylan and glucomannan yield of the final wood pulp—*i.e.*, the raw material for document paper.[43] The crystallinity of cellulose pulp is also affected by these processes. The cellulose crystallinity of printing paper depends on the cooking methods and processing conditions.[12,44] As document papers are kraft pulp-based, their cellulose crystallinity is likely influenced by additional factors. For instance, recycled pulp may have been used in the manufacturing process,[12,45] and the products may plausibly contain bleached chemo-thermomechanical pulp (BCTMP), a high-yield pulp that retains water-soluble components, particularly acetylated galactoglucomannan.[13] Recycled pulp and BCTMP, commonly employed as cost-saving measures in paper manufacturing, further impact the composition and properties of the final product.

3.3. Random forest models for classification of document papers

Fig. 5 shows the OOB error changes as each classification tree is added to the RF during training for document-paper classification. The original and first derivative Raman spectra exhibit obvious different trends. For original spectra, increasing the number of trees initially reduced the OOB errors in both cases but the first derivative spectra notably accelerated the reduction from that of the raw spectra. The OOB errors were minimized during the training process and the first derivative spectra overwhelmingly improved the performance from that of the original spectra. The optimized hyperparameters and classification performances (F1 scores) of the RF models are presented in Table 2.

Table 2 compares the performances of the RF models trained on the original and first derivative Raman spectra across various hyperparameter settings. After training on the original spectra, the OOB errors remained relatively high (0.474–0.500), indicating limited predictive accuracy of the RF methods. The test F1 scores ranged between 0.664 and 0.711, reflecting low classification performance.

Table 2 Classification performance of RF models on document papers

Raman spectra	Hyperparameters		OOB error	F1 score	
	n_feature	n_tree		Train	Test
Original spectra	sqrt	123	0.474	1.000	0.711
	log 2	284	0.500	1.000	0.669
	1/3	380	0.497	1.000	0.664
First derivative spectra	sqrt	86	0.271	1.000	0.843
	log 2	48	0.285	1.000	0.838
	1/3	30	0.243	1.000	0.875

The OOB errors were obviously lower (0.243–0.285) after training on the first derivative spectra, suggesting an enhanced predictive reliability of the RF models. Furthermore, the test F1 scores were markedly improved to 0.838–0.875, affirming that the spectral preprocessing with the first derivative enhances the classification performance of the RF models.

Among the tested hyperparameter settings, the "1/3" configuration for the first derivative spectra minimized the OOB error (0.243) and maximized the F1 score (0.875). Therefore, "1/3" was deemed the optimal configuration for this dataset. As highlighted by these findings, preprocessing steps such as spectral derivative decisively improve the accuracy and generalizability of machine learning models on Raman spectral data, emphasizing the necessity of proper spectral preprocessing for robust and reliable analytical applications.[22,23,46–48]

3.4. Feature-importance measures of the Raman spectra

The spectral regions contributing to document-paper classification by the RF models were identified through an MDI-based feature-importance analysis of the Raman spectral data. The results are visualized in Fig. 6.

Fig. 6a presents the feature importance results for the KOR4, IDN2, and THA products based on the first derivative

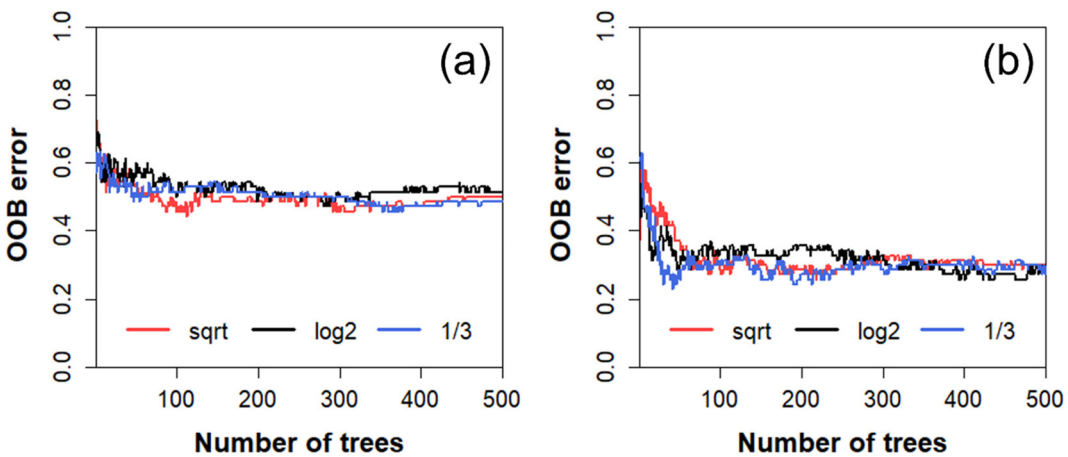

Fig. 5 Changes in out-of-bag (OOB) error rates with increasing number of classification trees in the classification of document papers (a: original Raman spectra and b: first derivative Raman spectra).

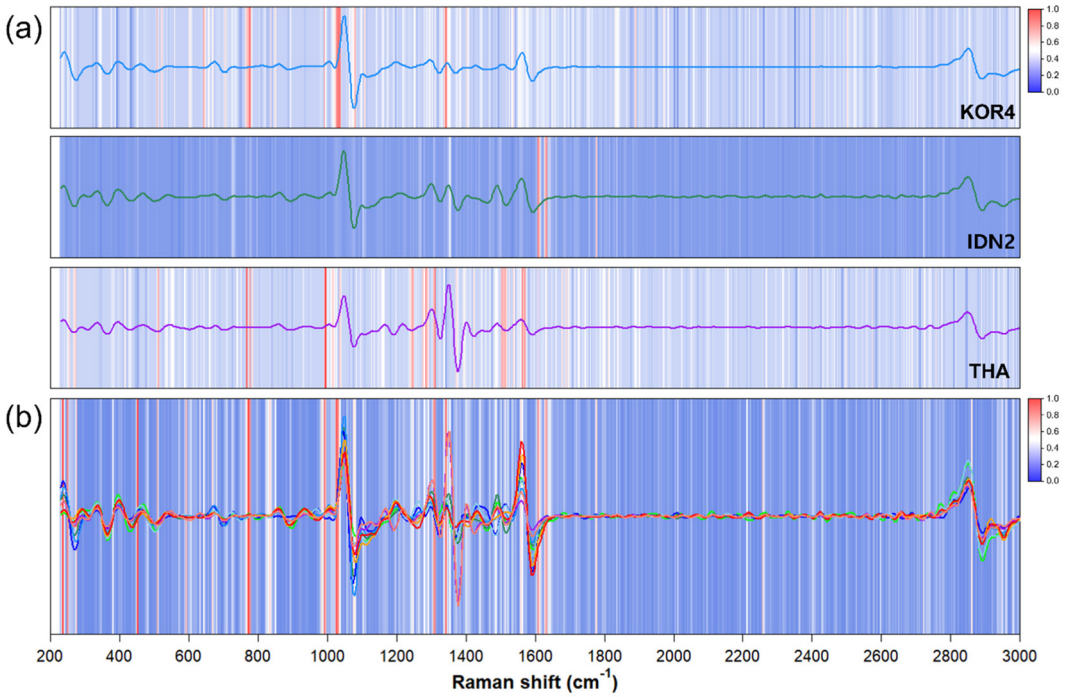

Fig. 6 Raman spectral-feature importance analysis of the RF model when classifying document paper samples. Spectral feature importance of the KOR4, IDN2, and THA products (a) and entire data classes (b). Red indicates higher importance, while blue indicates lower importance.

spectra. In Fig. 6, the red-colored regions indicate high contributions to the differentiation of data classes. The classification of KOR4 was notably influenced by lignin at 798 cm^{-1},[49] lattice vibrations associated with calcium carbonate at 1084 cm^{-1}, and HCC, HCO, and HOC bending, as well as CH and CH_2 stretching in carbohydrate components (cellulose and hemicellulose) at 1380 cm^{-1}. For IDN2, the region at 1602 cm^{-1} was identified as a significant contributor, corresponding to the aromatic ring stretching of lignin. The THA product exhibited a similar pattern. It is well known that a substantial amount of lignin must be removed during the chemical pulping and bleaching processes. However, in East Asian countries such as Korea and China, manufacturers rely on imported wood pulp as a raw material. Due to this dependence, they often select BCTMP as a cost-effective alternative.[13] BCTMP is produced by mechanically refining wood chips with a small amount of sodium sulfite, which facilitates sulfonation (Fig. 7). The sulfonation process removes resin components from the wood under mildly alkaline conditions while also providing a slight brightening effect.[50] Even after bleaching, a substantial amount of lignin residues remains, which affects the quality of the paper such as brightness and yellowness.

Fig. 6b illustrates the spectral feature importance derived from the entire dataset. In addition to the previously discussed features, 280 cm^{-1} and 436 cm^{-1} were identified as highly important Raman spectral variables, corresponding to the presence of calcium carbonate and the pyran ring, respectively.

Overall, the decision-making process of the RF model was primarily influenced by the presence of calcium carbonate and the type of wood pulp used in paper production, making these factors key discriminators of paper products. Moreover, the 200–1650 cm^{-1} spectral range in Raman spectroscopy appears to

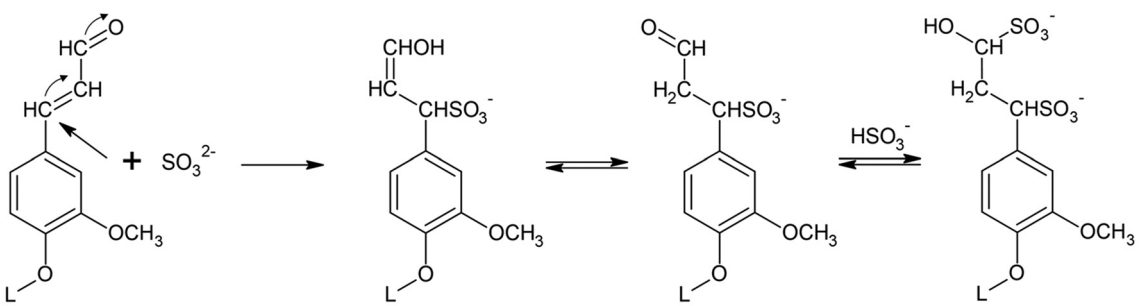

Fig. 7 Reaction of a lignin structure with sodium sulfite.

be a crucial region for tracing and classifying document papers, and it will be utilized as a selective variable for modeling.

This analysis highlights the effectiveness of spectral feature selection in improving classification performance by emphasizing the contributions of specific fillers and wood pulp types to the distinct spectral characteristics of document papers. Additionally, it underscores the advantage of reducing computational costs when constructing machine learning models.[13,29,51]

3.5. Variable selection and model comparison

This subsection compares the classification performances of the SVM, FNN, and RF models within the significant spectral region (200–1650 cm^{-1}) and the entire range (200–3000 cm^{-1}) of the Raman spectra of the document-paper samples. The validation of the selected region (200–1650 cm^{-1}) was performed with first derivative spectra. The performances of all models across both spectral ranges are compared in Table 3. The classification performances of all models were improved after restricting the spectral range to 200–1650 cm^{-1}, highlighting the significance of this region for differentiating document papers.

The F1 score of the SVM model improved from 0.732 to 0.935 after transitioning from the entire spectral range to the selected range, demonstrating that excluding the irrelevant variables enhances the robustness of the model. Similarly, the F1 score of the FNN model increased from 0.901 to 0.968 narrowing the spectral region from 200–3000 to 200–1650 cm^{-1}, emphasizing that the selected spectral range also increased the computational efficiency. The RF model also demonstrated an improvement in performance, with an increase in the F1 score from 0.875 to 0.903. Notably, the selected range reduced the number of input variables from 756 (in the 200–3600 cm^{-1} range) to 360 (in the 200–1650 cm^{-1} range). The reduced number of input variables not only enhances the robustness of the model by focusing on the most relevant spectral features but also significantly reduces computational costs.[12,13] Therefore, the classification models effectively utilize the critical spectral features within the 200–1650 cm^{-1} range, which correspond to calcium carbonate, cellulose, and lignin. FNN, which achieved the highest F1 score (0.965), appears to be a promising tool. However, considering Occam's razor,[52] which suggests that when accuracy is similar, the simplest model should be preferred, SVM can also serve as an effective alternative to FNN. Nevertheless, both FNN and SVM lack transparency in their decision-making processes, which limits the ability to interpret or justify their predictions. For this reason, the authors suggest that each classification model offers distinct advantages, and no single model can be considered a complete replacement for another.

As clarified by the above results, narrowing the spectral range to the most relevant region enhances the model's robustness and reduces the computational complexity. This finding underscores the potential of feature selection in developing efficient and scalable models for document-paper classification in practical applications. Identifying the relevant range (200–1650 cm^{-1}) is a focused, computationally efficient approach for analyzing document papers. Therefore, our work can make valuable contributions to forensic investigations and material classification tasks.

Table 3 Performance comparison of the SVM, FNN, and RF models in document-paper classification

Model	Spectral range (cm^{-1})	Hyperparameters	F1 score
SVM	200–3000	gamma = 10^{-3}, C = 2^3	0.732
	200–1650	gamma = 10^{-4}, C = 2^5	0.935
FNN	200–3000	hl_size = (32), lr = 0.001, optimizer = Adam	0.901
	200–1650	hl_size = (32), lr = 0.1, optimizer = SGD	0.968
RF	200–3000	n_feature = 1/3, n_tree = 30	0.875
	200–1650	n_feature = log, n_tree = 144	0.903

hl_size, hidden layer size; lr, learning rate; Adam, adaptive-moment estimation; SGD, stochastic gradient descent.

4. Conclusions

This study demonstrated the potential of integrating Raman spectroscopy with machine learning for forensic document examination. The RF model computed the feature importance, thereby enhancing the interpretability, and identified the 200–1650 cm^{-1} spectral range as the most informative region for classification. Within this narrowed range, the number of input variables was reduced from 756 to 360, largely lowering the computational complexity while improving the robustness and accuracy.

Spectral preprocessing with the first derivative boosted the classification performance of all models, but most obviously benefited the FNN model. The FNN model outperformed the RF and SVM classifiers, achieving the highest F1 score of 0.968. These findings highlight the superior accuracy and computational efficiency of the variable selection based on feature importance measures with first derivative Raman spectra, confirming the suitable choice for forensic document examination. This work advances the use of Raman spectroscopy and machine learning in forensic science, offering a scalable, interpretable, and efficient solution for document-paper classification in real-world scenarios.

However, this study has certain limitations. Contamination or aging can significantly alter the Raman spectral characteristics of paper, potentially reducing the applicability of the proposed approach. Methods that mitigate the spectral distortion caused by contamination or degradation will be incorporated in future work. In addition, the research scope will be expanded to larger datasets and a broader range of paper products. Advanced methods such as deep learning are expected to further enhance the classification performance and ensure scalability of the proposed framework to diverse forensic applications.

Author contributions

Conceptualization, Y. J. L.; methodology, Y. J. L. and C. W. J.; formal analysis, Y. J. L. and H.T.K.; investigation, Y. J. L. and

C. W. J.; data curation, Y. J. L. and H.T.K.; writing – original draft preparation, Y. J. L. and C. W. J.; writing – review and editing, Y. J. L. and T. J. L.; supervision, T. J. L. and H. J. K.; funding acquisition, H. J. K.

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors would like to acknowledge the financial support provided by the Ministry of Science and ICT (Information and Communication Technology) of the Korean government and the National Research Foundation of Korea (Grant No. RS-2023-00301889).

References

1 K. Jones, S. Benson and C. J. Roux, *Forensic Sci. Int.*, 2016, **262**, 97–107.
2 K. Jones, S. Benson and C. J. Roux, *Forensic Sci. Int.*, 2013, **231**, 364–374.
3 K. Jones, S. Benson and C. J. Roux, *Forensic Sci. Int.*, 2013, **233**, 355–364.
4 C. E. Berger, *Forensic Sci. Int.*, 2009, **192**, 1–6.
5 J. Lee, H. Kim, S. Yook and T. Y. Kang, *J. Forensic Sci.*, 2023, **68**, 1808–1815.
6 Y. J. Lee, T. J. Lee and H. J. Kim, *BioResources*, 2024, **19**(1), 160–182.
7 J. A. Green, *J. Forensic Sci.*, 2012, **57**, 1003–1007.
8 Y. J. Lee, C. W. Jeong and H. J. Kim, *BioResources*, 2024, **19**(4), 7591–7605.
9 J. Grant, *J. Forensic Sci. Soc.*, 1973, **13**, 91–95.
10 H. A. Foner and N. Adan, *J. Forensic Sci. Soc.*, 1983, **23**, 313–321.
11 V. Causin, C. Marega, A. Marigo, R. Casamassima, G. Peluso and L. Ripani, *Forensic Sci. Int.*, 2010, **197**, 70–74.
12 S.-W. Hwang, G. Park, J. Kim, K.-H. Kang and W.-H. Lee, *BioResources*, 2024, **19**, 1633–1651.
13 Y. J. Lee, S. W. Kweon, C. W. Jeong and H. J. Kim, *Spectrochim. Acta, Part A*, 2024, **327**, 125299.
14 H. Ebara, A. Kondo and S. Nishida, *Rep. Natl. Res. Inst. Police Sci.*, 1982, **2**(35), 88–98.
15 R. Kumar and V. C. Sharma, *Trends Anal. Chem.*, 2018, **105**, 191–201.
16 L. C. Lee, *J. Anal. Chem.*, 2021, **76**, 95–101.
17 R. Kumar, V. Kumar and V. Sharma, *Appl. Spectrosc.*, 2015, **69**, 714–720.
18 D. N. Sathyanarayana, *Vibrational spectroscopy: theory and applications*, New Age International, 2015.
19 V. Sharma, R. Chopra, N. Verma, P. K. Mishra and R. Cieśla, *TrAC, Trends Anal. Chem.*, 2024, **180**, 117989.
20 J. Xia, X. Du, W. Xu, Y. Wei, Y. Xiong and S. Min, *Spectrochim. Acta, Part A*, 2021, **248**, 119290.
21 H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, M. J. Walsh, M. R. McAinsh, N. Stone and F. L. Martin, *Nat. Protoc.*, 2016, **11**, 664–687.
22 W. Gao, T. Shu, Y. Guan, S. Ling, S. Liu and L. Zhou, *Carbohydr. Polym.*, 2022, **277**, 118793.
23 W. Gao, L. Zhou, S. Liu, Y. Guan, H. Gao and B. Hui, *Bioresour. Technol.*, 2022, **348**, 126812.
24 C. A. Lieber and A. Mahadevan-Jansen, *Appl. Spectrosc.*, 2003, **57**, 1363–1367.
25 A. Savitzky and M. J. E. Golay, *Anal. Chem.*, 1964, **36**, 1627–1639.
26 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
27 L. Breiman, *Classification and regression trees*, Routledge, 2017.
28 L. Breiman, *Manual on Setting Up, Using, and Understanding Random Forests*, Statistics Department, University of California, Berkeley, 2002.
29 S.-W. Hwang, H. Chung, T. Lee, J. Kim, Y. Kim, J.-C. Kim, H. W. Kwak, I.-G. Choi and H. Yeo, *J. Wood Sci.*, 2023, **69**, 1.
30 G. Louppe, L. Wehenkel, A. Sutera and P. Geurts, *Adv. Neural Inf. Process. Syst.*, 2013, **26**.
31 J.-P. Vert, K. Tsuda and B. Schölkopf, *Kernel Methods in Computational Biology*, 2004, **47**, 35–70.
32 M. Sokolova and G. Lapalme, *Inf. Process. Manage.*, 2009, **45**, 427–437.
33 M. Sokolova, N. Japkowicz and S. Szpakowicz, in *Proc. Australas. Joint Conf. Artif. Intell.*, Springer, Berlin, Heidelberg, 2006, pp. 1015–1021.
34 L. Borromeo, N. Egeland, M. W. Minde, U. Zimmermann, S. Andò, M. V. Madland and R. I. Korsnes, *Minerals*, 2018, **8**, 221.
35 J. S. Han, S. Y. Jung, D. S. Kang and Y. B. Seo, *ACS Sustainable Chem. Eng.*, 2020, **8**, 8994–9001.
36 M. Szymańska-Chargot, J. Cybulska and A. Zdunek, *Sensors*, 2011, **11**, 5543–5560.
37 U. P. Agarwal and S. A. Ralph, *Appl. Spectrosc.*, 1997, **51**, 1648–1655.
38 Q. Li and S. Renneckar, *Biomacromolecules*, 2011, **12**, 650–659.
39 K. Schenzel and S. Fischer, *Cellulose*, 2001, **8**, 49–57.
40 J. H. Wiley and R. H. Atalla, *Carbohydr. Res.*, 1987, **160**, 113–129.
41 R. S. Dassanayake, N. Dissanayake, J. S. Fierro, N. Abidi, E. L. Quitevis, K. Boggavarappu and V. D. Thalangamaarachchige, *Appl. Spectrosc. Rev.*, 2023, **58**, 180–205.
42 U. P. Agarwal, in *Advances in Lignocellulosics Characterization*, ed. D. S. Argyropoulos, TAPPI Press, Atlanta, 1999, pp. 201–225.

43 R. Aurell and N. Hartler, *Sven. Papperstidn.*, 1965, **68**, 97–102.
44 E. Gümüskaya, M. Usta and H. Kirci, *Polym. Degrad. Stab.*, 2003, **81**, 559–564.
45 P. Sheikhi, M. Talaeipour, A. H. Hemasi and H. K. Eslam, *BioResources*, 2010, **5**, 1702–1716.
46 W. Gao, Y. Guan, H. Huang, S. Liu, S. Ling and L. Zhou, *Cellulose*, 2024, **31**, 7697–7711.
47 W. Gao, Q. Jiang, Y. Guan, H. Huang, S. Liu, S. Ling and L. Zhou, *Int. J. Biol. Macromol.*, 2024, **269**, 132147.
48 W. Gao, T. Shu, Q. Liu, S. Ling, Y. Guan, S. Liu and L. Zhou, *ACS Omega*, 2021, **6**, 8578–8587.
49 P. Bock, P. Nousiainen, T. Elder, M. Blaukopf, H. Amer, R. Zirbs, A. Potthast and N. Gierlinger, *J. Raman Spectrosc.*, 2020, **51**, 422–431.
50 *Pulping Chemistry and Technology*, ed. M. Ek, G. Gellerstedt and G. Henriksson, Walter de Gruyter, Berlin, 2009, vol. 2.
51 Y. J. Lee, S. Y. Won, S. B. Park and H.-J. Kim, *Heritage Sci.*, 2024, **12**, 373.
52 P. Domingos, in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 1998, pp. 37–43.