Classification Analysis of Copy Papers Using Infrared Spectroscopy and Machine Learning Modeling

Yong-Ju Lee,[a] Tai-Ju Lee,[b,*] and Hyoung Jin Kim [a,*]

The evaluation and classification of chemical properties in different copy-paper products could significantly help address document forgery. This study analyzes the feasibility of utilizing infrared spectroscopy in conjunction with machine learning algorithms for classifying copy-paper products. A dataset comprising 140 infrared spectra of copy-paper samples was collected. The classification models employed in this study include partial least squares-discriminant analysis, support vector machine, and K-nearest neighbors. The key findings indicate that a classification model based on the use of attenuated-total-reflection infrared spectroscopy demonstrated good performance, highlighting its potential as a valuable tool in accurately classifying paper products and ensuring assisting in solving criminal cases involving document forgery.

DOI: 10.15376/biores.19.1.160-182

Keywords: Attenuated-total-reflection infrared spectroscopy (ATR-IR); Partial least squares-discriminant analysis (PLS-DA); Support vector machine (SVM); K-nearest neighbor (KNN); Machine learning; Document forgery; Forensic document analysis

*Contact information: a: Department of Forest Products and Biotechnology, Kookmin University, 77 Jeongneung-ro, Seongbuk-gu, Seoul 02707 Republic of Korea; b: National Institute of Forest Science, Department of Forest Products and Industry, Division of Forest Industrial Materials, 02455, Seoul, Republic of Korea; *Corresponding authors: leetj@korea.kr; hyjikim@kookmin.ac.kr*

INTRODUCTION

Various types of copy papers are produced and sold worldwide, finding extensive usage in institutions such as schools, offices, and printing companies. Each copy paper possesses distinct properties, including basis weight, whiteness, and composition. These attributes are influenced by factors, such as the type and ratio of pulp; additives, such as fillers and sizing agents; and variations in manufacturing processes. Surface sizing agents, such as starch, are commonly employed in copy-paper manufacturing to enhance surface strength and printing suitability (Moutinho *et al.* 2011).

Despite the advances in information technology that have led to the concept of a paperless society, many tasks are still conducted using paper documents. Paper remains a crucial medium for recording various daily tasks and activities, including taking notes, jotting down memos, and formalizing important contracts (Ganzerla *et al.* 2009; Lee *et al.* 2023).

Meanwhile, owing to recent technological advances in office automation devices, document forgery has become more accessible, consequently underscoring the growing significance of paper document examination in fields such as forensic investigation and evidence analysis (Lee *et al.* 2023). In South Korea, a substantial number of document forgery cases have been documented. In 2020, the country recorded 2,217 cases of

document forgery related to official documents and 7,604 cases concerning private documents. Over a three-year period starting in 2018, the statistics indicate that nearly ten thousand criminal cases of document forgery are reported annually (Choi *et al.* 2018a; 2018b). Hence, the development of robust forensic capabilities for detecting forgery becomes paramount (Kim *et al.* 2016; Lee *et al.* 2023).

Various classification methods for paper have been reported, primarily relying on chemical analysis. The X-ray diffraction (XRD) method has been employed to investigate the inorganic substances used as paper fillers (Foner and Adan 1983; Causin *et al.* 2010). Spence *et al.* (2000) demonstrated that document paper could be identified by utilizing ICP-MS to trace the elements within the paper. Ganzerla *et al.* (2009) conducted a comprehensive study on the characteristics of ancient documents produced in Palazzo Ducale, Italy, employing various analytical tools, including scanning electron microscopy-energy dispersive X-ray spectrometry (SEM–EDX), high-performance liquid chromatography mass spectrometry (HPLC-MS/MS), and pyrolysis–gas chromatography and mass spectrometry (Py-GC/MS). Choi *et al.* (2018a; 2018b) utilized SEM-EDX to analyze fillers in 188 types of copy paper and further determined the mix ratios of fiber components through dissociation tests.

In recent years, the utilization of infrared spectroscopy has gained prominence for identifying paper types (Kher *et al.* 2001, 2005; Kumar *et al.* 2017; Kang *et al.* 2021; Kim *et al.* 2022). Infrared spectroscopy (IR) serves as a fundamental tool for investigating paper structure and pulp chemistry (Workman 1999; Pan and Nguyen 2007). The paper industry has been utilizing IR for non-destructive process control and rapid determination of specific parameters, including identification of paper (Hodges *et al.* 2006; Ganzerla *et al.* 2009; Jang *et al.* 2020; Seo *et al.* 2023). Kher *et al.* (2005) reported a successful attempt to identify six types of paper by analyzing mid-infrared wavelengths (MIR, 2500 to 4000 cm^{-1}) using an FT-IR spectrometer.

Additionally, IR spectroscopy combined with multivariate statistical methods has proven to be an effective approach for distinguishing similar paper products, modeling systematic data variances, and presenting data in a concise manner (Marcelo *et al.* 2014). Multivariate statistical methods offer several classification models, including principal component analysis (PCA), partial least squares-discriminant analysis (PLS-DA), support vector machine (SVM), and K-nearest neighbor (KNN) (Agarwal *et al.* 2021). PCA is an unsupervised model that uncovers relationships between samples and analytical data, resulting in distinct groupings of variables and samples. This grouping can be visualized as a reduction from a multidimensional space to a two-dimensional representation. Kim *et al.* (2016) utilized IR spectroscopy and PCA to effectively distinguish traditional paper products originating from Korea, China, and Japan. Similarly, Kang *et al.* (2021) and Kim *et al.* (2022) demonstrated the efficacy of IR and PCA in classifying paper products according to their respective continents or countries of origin.

In contrast, PLS-DA, SVM, and KNN are supervised models that employ labeled datasets to train the models for classifying new samples based on known classes. These models enable the prediction of a sample's class by leveraging its spectral characteristics and previously labeled data (Singh *et al.* 2023). Jang *et al.* (2020) utilized supervised models, including PLS-DA, SVM, and random forest, to predict the types of traditional Korean paper with varying raw materials. Hwang *et al.* (2023) also reported the feasibility of discriminating manufacturing origins with artificial neural networks (ANN) and infrared spectroscopy. Canals *et al.* (2008), Ruiz *et al.* (2011), and Xia *et al.* (2023) developed

algorithms for classifying a wide range of paper types, such as base, coated, printed, recycled, and hygiene papers using infrared spectra data.

Despite the considerable amount of research on the utilization of multivariate statistical methods in conjunction with IR spectroscopy for classifying diverse paper products, a significant research gap exists pertaining to the classification of copy papers within the same grade, employing IR spectroscopy combined with machine learning algorithms, such as PLS-DA, SVM, and KNN.

The main objective of this study was to examine the feasibility of employing IR in conjunction with machine learning algorithms for identifying copy paper products as same as printing paper and document paper. Therefore, a dataset comprising 140 IR spectra of copy-paper samples was collected. In addition, this study utilized PLS-DA, SVM, and KNN as the classification models. Moreover, the effectiveness of these three classification models were compared in terms of efficiency and being expeditious approaches for analyzing the constituent materials in copy paper.

EXPERIMENTAL

Materials

The samples were conditioned for more than 48 h at a temperature of 23 °C ± 1 °C and a relative humidity of 50% ± 2%, according to ISO 187 (1990). Table 1 provides information regarding paper products, manufacturers, as well as the physical and optical properties examined in this study.

Various non-destructive methods for paper analysis and identification are currently in use. These methods encompass the comparison of physical characteristics such as basis weight, thickness, apparent density, surface roughness, brightness, opacity, and whiteness (Lee *et al.* 2023). Table 2 provides details about the equipment and standards utilized for the evaluation of the physical and optical properties.

Analysis of the Inorganic Filler Content

Fillers can significantly influence the discrimination procedure for paper samples (Causin *et al.* 2010; Choi *et al.* 2018). The content and types of fillers in each paper product vary due to differences in processing parameters, manufacturing conditions, and the formulation of additives (Causin *et al.* 2010).

The analysis of inorganic filler content, including clay ($CaSiO_3$) and calcium carbonate ($CaCO_3$) and titanium dioxide (TiO_2), was performed using an ash content analyzer (Emtec, Germany). This measurement method is based on a combination of X-ray fluorescence analysis and the X-ray transmission method (Hu *et al.* 2020).

SEM-EDX

SEM-EDX allowed the production of images at high magnification and the determination of major elemental components in the samples. JSM 7401F (JEOL Ltd., Japan) was used for imaging and energy dispersive X-ray spectroscopy (EDX, X-Max, Oxford instruments, UK) for elemental analysis. The acceleration voltage of the electron beam was set at 10 kV for the imaging, and 15 kV for EDX, respectively.

Table 1. Physical and Optical Properties of Samples

Copy Paper Models and Manufacturers		
Code	Product	Manufacturer
A	Office ultra white	HP printing Co.
B	Multipurpose 20	HP printing Co.
C	G.R COPY	Daehan paper Co.
D	Milk	Hankuk paper Co.
E	Multipurpose ultra white	HP printing Co.
F	Premium inkjet and laser	Hammermill Co.
G	EQ plus premium	Arim paper Co.
H	EJK-SUPA4100	Elecom Co.
I	SW-101	Canon Co.
J	KJ-P19A4-250	Kokuyo Co.
K	Copy paper	Sustainable Earth Co.
L	Earth Choice	Domtar Co.
M	Great White 100	Hammermill Co.
N	Great White 30	Hammermill Co.

Physical Properties				
Code	Basis weight (g/m^2)	Thickness (μm)	Apparent density (g/m^3)	Bendtsen (mL/min)
A	76.3 ± 0.4	1.72 ± 0.01	0.84 ± 0.03	199.7 ± 30.7
B	78.9 ± 0.5	1.78 ± 0.01	0.84 ± 0.02	229.4 ± 31.4
C	74.5 ± 1.2	1.68 ± 0.03	0.76 ± 0.01	205.3 ± 14.8
D	75.5 ± 0.4	1.70 ± 0.01	0.78 ± 0.01	261.5 ± 23.4
E	75.8 ± 0.5	1.71 ± 0.01	0.82 ± 0.03	213.1 ± 54.9
F	90.0 ± 1.1	2.02 ± 0.02	0.85 ± 0.01	229.9 ± 54.9
G	86.7 ± 0.4	1.95 ± 0.01	0.86 ± 0.01	95.8 ± 10.1
H	88.4 ± 1.1	1.99 ± 0.02	0.84 ± 0.03	82.3 ± 8.2
I	73.2 ± 0.5	1.65 ± 0.01	0.89 ± 0.01	88.6 ± 6.3
J	78.7 ± 0.5	1.77 ± 0.01	0.82 ± 0.01	251.0 ± 23.6
K	75.5 ± 0.4	1.70 ± 0.01	0.85 ± 0.02	311.9 ± 44.7
L	76.1 ± 0.2	1.71 ± 0.01	0.85 ± 0.01	178.8 ± 19.7
M	75.9 ± 0.7	1.71 ± 0.02	0.80 ± 0.01	160.5 ± 20.5
N	74.7 ± 0.7	1.67 ± 0.02	0.79 ± 0.01	168.2 ± 13.4

Optical Properties			
Code	Brightness (R457, D65)	Whiteness (CIE, D65)	Opacity (D65)
A	103.4 ± 0.3	149.6 ± 0.9	91.3 ± 0.5
B	106.4 ± 0.3	156.6 ± 0.8	91.5 ± 0.4
C	96.7 ± 0.5	136.5 ± 1.0	94.2 ± 0.5
D	100.9 ± 0.4	147.4 ± 1.1	95.3 ± 0.1
E	106.1 ± 0.5	155.3 ± 0.8	91.2 ± 1.0
F	108.4 ± 0.2	153.0 ± 0.6	94.8 ± 0.3
G	108.0 ± 0.1	160.9 ± 0.2	93.9 ± 0.1
H	84.9 ± 0.3	82.9 ± 1.0	95.0 ± 0.2
I	104.5 ± 0.4	149.6 ± 0.7	93.0 ± 0.2
J	100.1 ± 0.1	141.2 ± 0.2	93.9 ± 0.1
K	103.3 ± 0.1	155.8 ± 0.4	90.4 ± 0.4
L	102.1 ± 0.3	138.9 ± 0.7	91.4 ± 0.4
M	100.4 ± 0.4	135.3 ± 2.1	90.4 ± 0.4
N	99.8 ± 0.8	138.0 ± 4.3	90.2 ± 0.3

Table 2. The information for Analysis of Physical and Optical Properties

Properties	Measurement	Equipment	Standard
Physical	Basis weight	Precisa, XT 220A, Switzerland	ISO 536
	Thickness	L&W thickness tester, Sweden	ISO 534
	Apparent density	-	
	Surface roughness	L&W Bendtsen tester, Sweden	ISO 8791-2
Optical	Brightness (R457, D65)	L&W Elrepho, Sweden	ISO 2471
	Whiteness (CIE, D65)		
	Opacity (D65)		

ATR-IR Analysis and Data Preprocessing

The ATR-IR spectra of the copy paper samples were determined using an ATR-IR (Bruker Optics, Germany). Every spectrum was recorded in the range of 4,000 to 400 cm^{-1} with a 4 cm^{-1} resolution, 32 scans, and the air absorbance was recorded as a reference standard. To eliminate undesired scatter effects, such as baseline shift and nonlinearity, as many spectra preprocessing steps as possible were applied, such as the selection of a spectral range of 1,800–800 cm^{-1}.

The total crystallinity index (TCI) (Nelson and O'Connor 1964) was calculated to derive the peak intensities from the spectral data. In addition, the chemical properties resulting from the raw materials were evaluated along with additives used in the production of copy papers. The TCI was computed by dividing the value at 1,372 cm^{-1}, corresponding to C-H bending, by the peak intensity at 2,900 cm^{-1}, corresponding to C-H and CH_2 stretching, as shown in Eq. 1.

$$TCI\ (\text{total crystallinity index}) = \frac{1{,}367\ cm^{-1}}{2{,}893\ cm^{-1}} \quad (1)$$

Additionally, the hydrogen bond intensity (HBI), which represents the bonding index of hydroxyl groups within cellulose, was calculated based on the peak intensity of 1,336 cm^{-1}, corresponding to C–OH, and 3,336 cm^{-1}, indicative of intermolecular hydrogen-bonding, as shown in Eq. 2 (Široký *et al.* 2010).

$$HBI\ (\text{hydrogen bond intensity}) = \frac{3{,}330\ cm^{-1}}{1{,}334\ cm^{-1}} \quad (2)$$

Proposed Approach

Machine learning process

The machine learning modeling for identification of copy paper is visualized in Fig. 1. First, the spectral pre-processing steps needed to identify an indispensable part of spectral data (Lasch 2012). This process involves, among others, outlier rejection, normalization, filtering, detrending, transformation, folding, and feature selection. In this study, the fifth-order polynomial Savitzky–Golay second-derivative was employed (Savitzky and Golay 1964). Next, the training dataset and test dataset were separated from the input dataset. The training data are employed to train the model, the development set evaluates various versions of the proposed model during development, and the test set confirms the answers to the primary research questions, such as the product names of copy papers. In this study, the input IR spectra dataset was divided into a training set and a test set in a 7:3 ratio. The stratified sampling method was applied to split the training and test sets to prevent sample selection bias resulting from the division process (Hens and Tiwari

2012; Ye *et al.* 2013). Then, the extracted IR spectra datasets were learned using PLS-DA, SVM, and KNN.

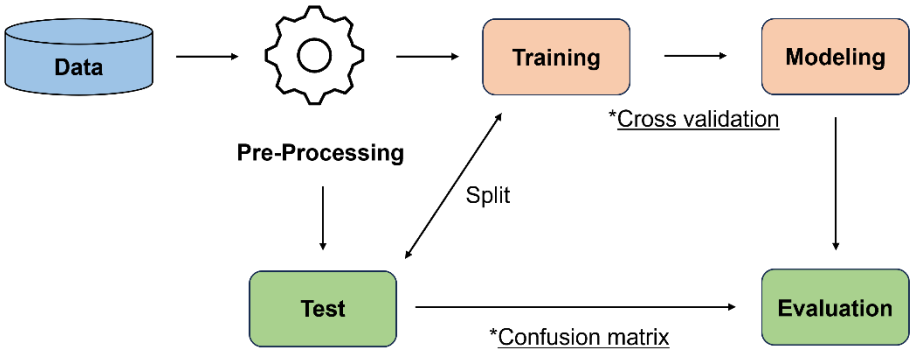

Fig. 1. The process of machine learning modeling for classification of copy paper

Machine learning model development

Three classification algorithms, namely PLS-DA, SVM, and KNN, were utilized for classification prediction. The complete data processing and classification procedures were performed using the R-statistical software (R Core Team, version 4.3.0, Auckland, New Zealand).

PLS-DA is widely recognized as one of the prominent classification techniques in the field of chemometrics. In addition, extensive research has been conducted and documented on PLS-DA and its properties (Indahl *et al.* 2007). The PLS-DA model can be expressed as a regression equation: Y = XB, where X represents an $n \times p$ matrix representing n samples, with each sample characterized by a vector of p feature values. Matrix Y is an $n \times k$ matrix comprising information about the class memberships of the samples, with k denoting the number of classes. The individual element, $y_{i,j}$, follows the structure described in Eq. 3.

$$y_{i,j} = \begin{cases} 1, if\ sample\ i\ belongs\ to\ class\ j \\ 0, otherwise \end{cases}, \quad (3)$$

where i and j represent the sample and class numbers, respectively, with i ranging from 1 to n and j ranging from 1 to k. The binary Y matrix exhibits a structured format, in which each row sums up to unity.

After estimating regression matrix B using the PLS2 algorithm, the prediction for a new set of samples can be performed as follows: $Y_{test} = X_{test}B$. However, the predicted values in the Y_{test} matrix are continuous numbers, requiring a conversion to class memberships. In this study, the class membership of each unknown sample is assigned based on the column index with the largest absolute value in the corresponding row of the Y_{test} matrix (Chevallier *et al.* 2006). Figure 2 shows the visualization of PLS-DA model.

The SVM is a nonparametric classifier that constructs a hyperplane to maximize the margin between classes. The hyperplane is built based on the training observations closest to different classes (Samanta *et al.* 2003). These training observations, known as support vectors, play a crucial role in constructing the separating hyperplane.

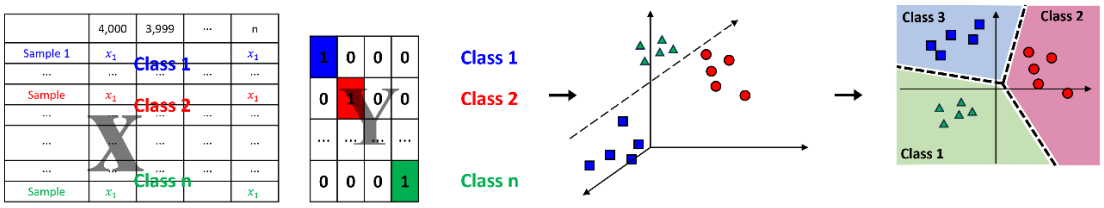

Fig. 2. Visualization of PLS-DA model

By using SVM, an optimal hyperplane must be determined that not only separates the classes but also maximizes the margin, representing the distance between the hyperplane and nearest data points from each class (Mancini *et al*. 2019). By maximizing the margin, SVM improves the generalization and robustness of the classifier when classifying new, unseen data points. SVM identifies the support vectors with the most significant influence on determining the position of the hyperplane. These support vectors are essential in defining the decision boundary between classes and contribute to the overall performance of the SVM classifier (Chauchard *et al.* 2008). In this study, a radial basis function (RBF) kernel was used to determine the hyperplane by projecting data onto a high dimensional feature space (Vert *et al.* 2004). The cost and gamma parameters of the RBF kernel SVM were optimized using grid searches with a logarithmic grid from 2^{-7} to 2^{7} for cost and from 10^{-5} to 10^{5} for gamma. Cost and gamma are parameters that control the cost of misclassification of the data and a Gaussian kernel for nonlinear classification, respectively. Figure 3 shows the nonlinear classification process of SVM model.

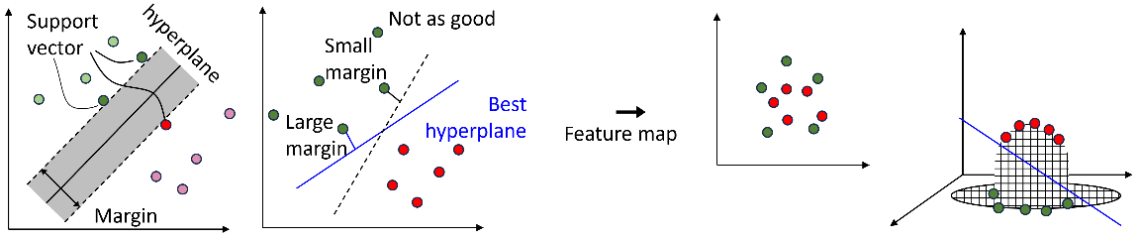

Fig. 3. Visualization of SVM model

The KNN procedure was employed to classify an unknown sample based on its proximity to previously categorized samples, similar to Mahalanobis' generalized distance technique (Tsuchikawa *et al.* 2003). Specifically, the predicted class of an unknown sample was determined by considering the classes of its KNNs. Analogous to polling, each of the *k*-closest training-set samples contribute one vote for its respective class. Subsequently, the unknown sample is assigned to the class receiving the highest number of votes. Therefore, the selection of an appropriate *k* value, representing the number of neighbors participating in the voting process, is crucial (Guo *et al.* 2003). The KNN procedure offers a flexible and intuitive approach to the classification, relying on the proximity of samples and the majority vote principle to assign classes to unknown samples. The selection of an appropriate *k* value must be considered carefully to achieve accurate and reliable classification results (Kabir *et al.* 2003). In this study, the number of nearest neighbors (*k*) was set to odd numbers in the range of 1 to 11, and the optimal k was determined using a grid search. Figure 4 shows the classification process of KNN model.

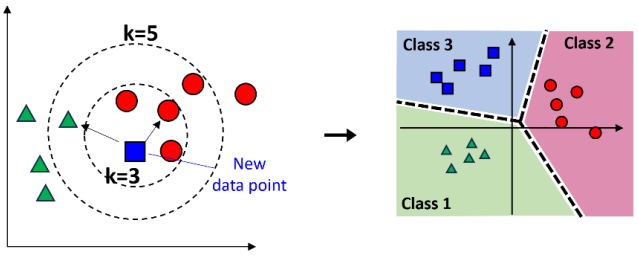

Fig. 4. Visualization of KNN model

Cross-validation

All the employed classification methods utilized cross-validation, a technique used to estimate the generalization error by utilizing holdout data (Xia *et al.* 2019). Among the available cross-validation techniques, leave-one-out cross-validation (LOOCV) is the most commonly used. LOOCV iteratively excludes a single data point from the training set, while utilizing the remaining data for model training, as depicted in Fig. 5. This process is repeated for each data point in the dataset. LOOCV is particularly advantageous for small datasets, as it minimally affects the available training data.

In this study, the classification models were evaluated using LOOCV to determine the predicted accuracies on the training datasets, that were calculated as the average of each operation.

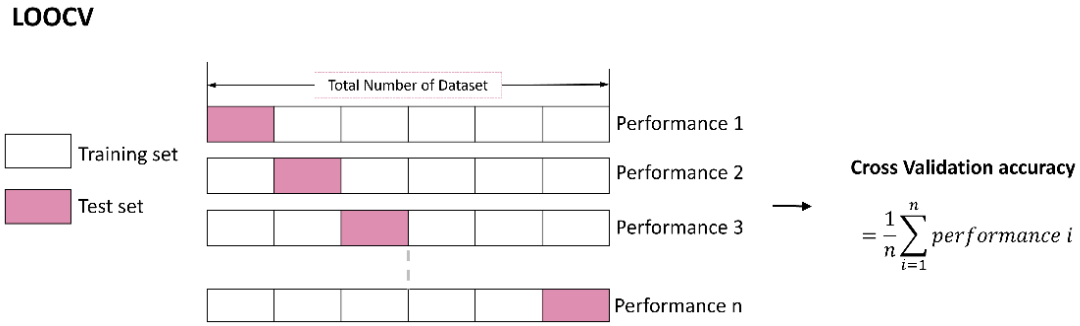

Fig. 5. The mechanism of leave one out cross validation (LOOCV)

Confusion matrix

A confusion matrix is widely used for evaluating the performance of classification algorithms (DeVries *et al.* 2003; Ruuska *et al.* 2018; Xu *et al.* 2020). It provides valuable information about the actual and predicted classifications made by the algorithm and is applicable to both two-class (binary) and multiclass classification problems (Xu *et al.* 2020). Figure 6 illustrates confusion matrix for a two-class classifier and a multiclass classifier.

In classification tasks, the accuracy of classifying observations into positive and negative categories must be assessed. When observations belonging to the positive class are correctly classified, they are referred to as true positives (TP), while correctly classified observations belonging to the negative class are termed true negatives (TN). Furthermore, instances of positive classes incorrectly classified as negative are referred to as false negatives (FNs), and instances of negative classes incorrectly classified as positive are termed as false positives (FPs).

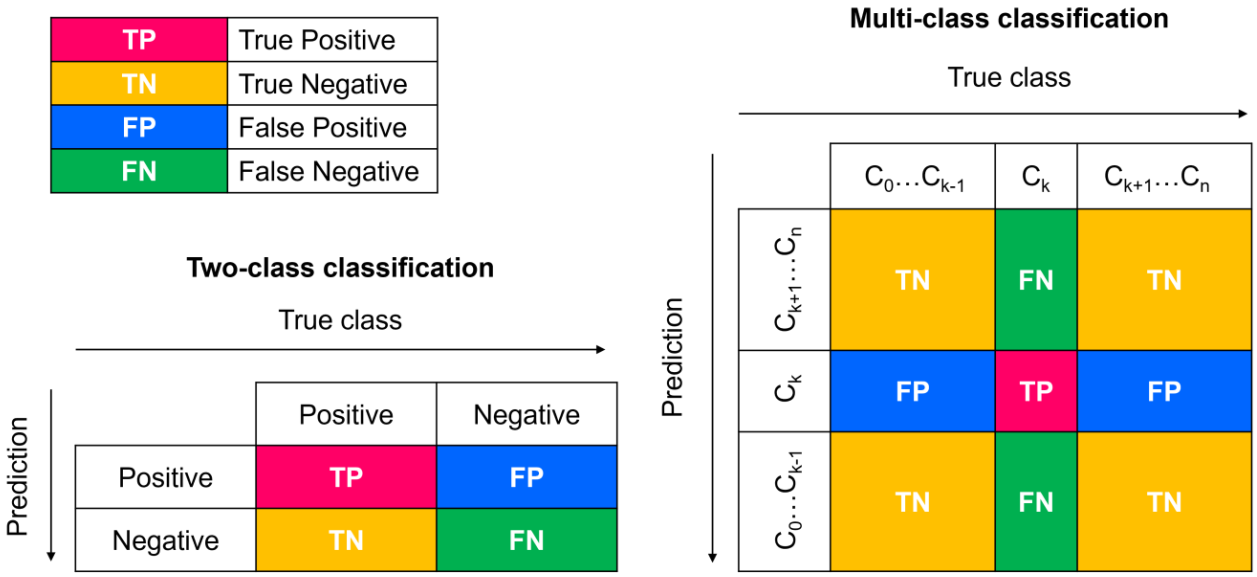

Fig. 6. Visualization of the confusion matrix for two-class and multiclass classifications with N classes

From these values, various performance indicators can be calculated to evaluate the classifier's ability to detect the target class (Piras *et al.* 2018). The commonly used indicators include accuracy, sensitivity, and specificity. Accuracy, sensitivity, and specificity are calculated as (TP + TN)/(TP + TN + FP + FN), TP/(TP + FN), and TN/(TN + FP), respectively.

RESULTS AND DISCUSSION

Inorganic Filler Content

Table 3 summarizes the inorganic filler content in copy paper samples. Titanium dioxide was not detected in any of the copy papers. From Table 1, employing analysis of variance (ANOVA) in R software and subsequent grouping, it was determined that there were no significant differences in the physical and optical properties of some copy papers at a 95% confidence level. This suggests that the evaluation of physical and optical method was not able to provide distinguishing features for classification and identification of copy paper product. Such differentiation became even more challenging when the manufacturers were the same.

On the other hand, the results in Table 3 suggest that the inorganic filler content has the feasibility of being a distinguishable feature of each copy paper compared to the physical and optical methods. For example, in the comparison between copy paper M and N, detecting the clay contents in M made it possible to find different features of both samples.

Nevertheless, the results showed that it was still hard to identify the copy papers A and E, which have similar $CaCO_3$ and total ash contents.

Table 3. Summary of Inorganic Filler Content

Code	Clay (%)	$CaCO_3$ (%)	Miscellaneous (%)	Total Ash Content (%)
A	-	15.3 ± 0.7	0.8 ± 0.01	16.1 ± 0.7
B	-	14.0 ± 0.4	0.8 ± 0.11	14.8 ± 0.4
C	1.3 ± 0.2	9.5 ± 1.1	-	10.9 ± 1.0
D	-	19.6 ± 0.4	0.1 ± 0.10	19.7 ± 0.4
E	-	15.0 ± 1.6	0.9 ± 0.20	15.9 ± 1.5
F	-	23.2 ± 0.8	1.5 ± 0.05	24.7 ± 0.8
G	-	21.1 ± 0.4	0.1 ± 0.10	21.2 ± 0.4
H	33.8 ± 0.7	8.5 ± 0.1	0.2 ± 0.00	42.4 ± 0.7
I	-	11.8 ± 0.3	0.2 ± 0.15	11.9 ± 0.3
J	0.3 ± 0.3	25.1 ± 0.6	-	25.4 ± 0.6
K	0.6 ± 0.3	11.2 ± 0.4	0.2 ± 0.10	12.0 ± 0.5
L	0.4 ± 0.2	15.4 ± 1.0	0.8 ± 0.11	16.6 ± 1.0
M	1.4 ± 0.3	12.7 ± 1.0	0.0 ± 0.10	14.2 ± 1.1
N	-	15.2 ± 0.8	0.1 ± 0.15	15.3 ± 0.6

SEM-EDX

The observation of the samples with SEM made it possible to study the fibers and fillers at high magnification. Figures 7 and 8 show the SEM and EDX elemental mapping images respectively. The copy paper H in Fig. 7(c) differed from the others: it was a machine finished coated paper but especially visible were some typical pigments of clay (Choi *et al.* 2018). The EDX mapping image also shows that the surface of copy paper H was coated with coating pigments.

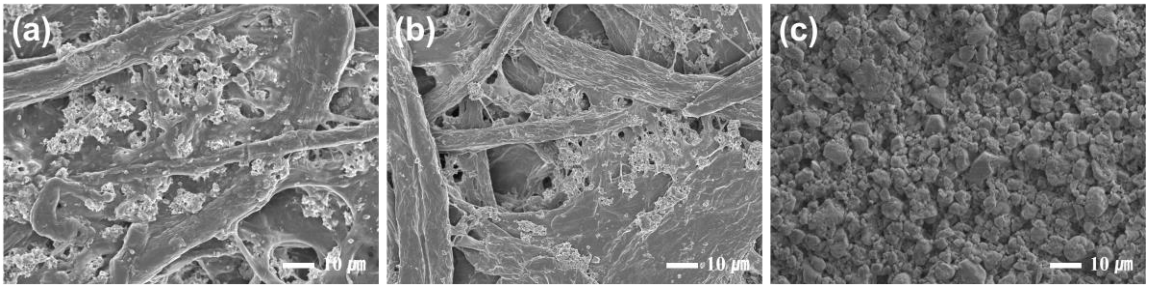

Fig. 7. SEM images of copy papers (a: Sample A, b: Sample E, c: Sample H).

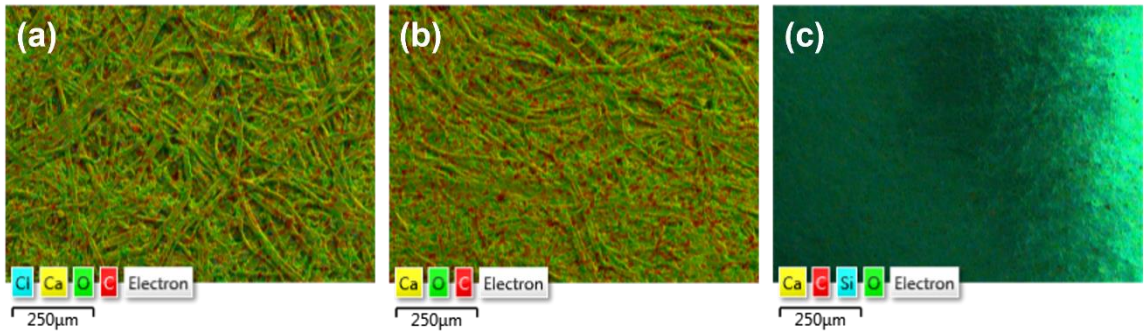

Fig. 8. EDX elemental mapping images of copy papers (a: Sample A, b: Sample E, c: Sample H).

However, excluding sample H, other samples also exhibited similar characteristics, making it challenging to demonstrate distinct differences between the samples. The samples A and E exhibited similar morphology, including fiber shapes and filler types, as shown in Fig. 7. The images showed small crystals of precipitated calcium carbonate (PCC) incorporated into the sizing of the sheets (Ganzerl *et al.* 2009). Figure 9 shows the EDX spectra of copy papers. The elemental analysis made with the probe EDX detected the amounts of calcium carbonate in all samples. The presence of Cl in copy paper A is probably a remnant of the process of bleaching, as shown in Fig. 9(a). Chlorine has been used since the late 18th century to produce printing paper (Seo *et al.* 2023).

The results demonstrated that copy papers which have similar filler types and contents were hard to classify.

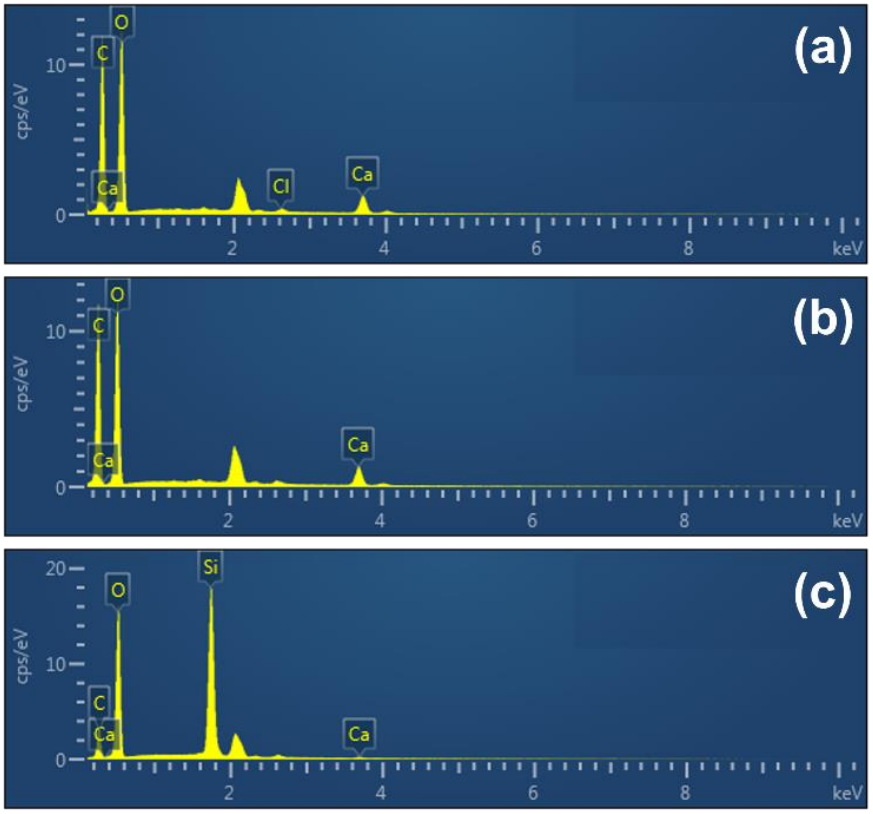

Fig. 9. EDX spectra of copy papers (a: Sample A, b: Sample E, c: Sample H)

ATR-IR Data

The spectra data were collected in the range of 4,000 to 400 cm^{-1}. However, as the key peaks associated with cellulose, hemicellulose, lignin, and moisture, which are major components of paper, are typically found in the range of 1,800 to 800 cm^{-1}, only the spectral region within 1,800 to 800 cm^{-1} was extracted and utilized for this study.

By focusing on this specific spectral range, the characteristic peaks relevant to the identification and analysis of cellulose, hemicellulose, lignin, and moisture in paper samples could be evaluated. This selective extraction of the spectral data allowed for a more precise and efficient analysis of paper composition and its constituents (Kim and Eom 2016).

Figure 10(a) shows the raw IR spectra, and Fig. 10(b) illustrates the second-derivative spectra. The data preprocessing using the Savitzky-Golay filter serves to make the baseline of spectra consistently adjusted and enhance the peaks, thus emphasizing differences between samples (Hwang *et al.* 2016). In the raw IR spectra shown in Fig. 10(a), several differences were revealed in the absorption bands at 1647 to 1635 cm^{-1} (water), 1422 cm^{-1} (CH_2 bending), 1337 cm^{-1} (amorphous cellulose), and 1200 to 900 cm^{-1} (cellulose fingerprint) for each copy paper (Garside and Wyeth 2003; Polovka *et al.* 2006; Ciolacu *et al.* 2010; Castro *et al.* 2011). However, in the second derivative spectra presented in Fig. 10(b), additional absorption peaks were observed at 1730, 1700, 1680, 1661, 1644, 1620, 1547, and 1510 to 1505 cm^{-1}.

The 1730 cm^{-1} peak was attributed to the oxidation of cellulose (Sistach *et al.* 1998). Differences at 1700 cm^{-1} may indicate the use of rosin in the sizing of paper products (Ganzerla *et al.* 2009). The peak at 1661 cm^{-1} suggests the presence of carbonyl groups along the cellulose chain, potentially keto or aldehyde groups linked to the phenyl ring of lignin (Proniewicz *et al.* 2001; 2002). The absorption peak at 1510 to 1505 cm^{-1} is characteristic of lignin. The peaks at 1644 and 1547 cm^{-1} (Amin I and II) may suggest the use of gelatine as glue (Calvini and Gorassini 2002; Gracia 2001). The vibrations of stretching of C=C aromatic ring at 1510 cm^{-1} and the peak at 1680 cm^{-1}, given by oxidation of keto or aldehydic groups linked to the ring (Ganzerla *et al.* 2009). The spectra at 1620 cm^{-1} are indicative of calcium carbonate and gypsum (Gracia 2001).

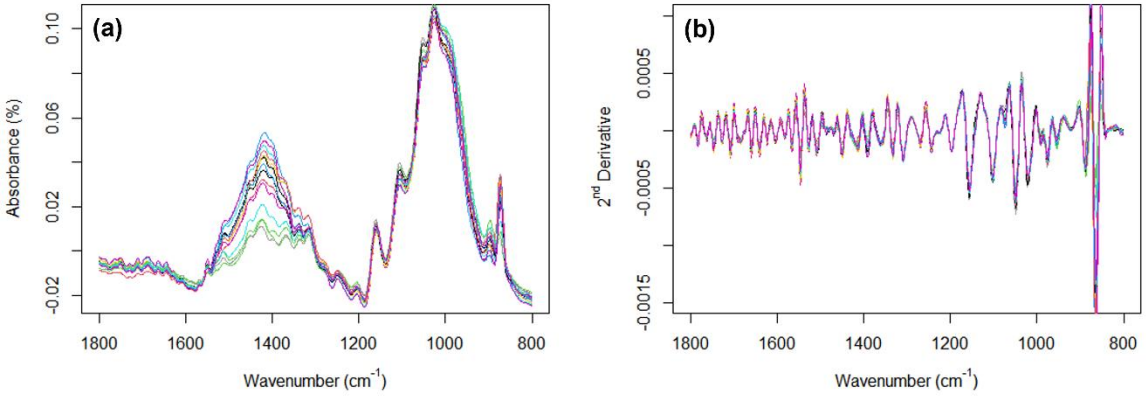

Fig. 10. Preprocessing of IR spectra with Savitzky–Golay algorithms: (a) Raw IR spectra and (b) 2nd derivative spectra

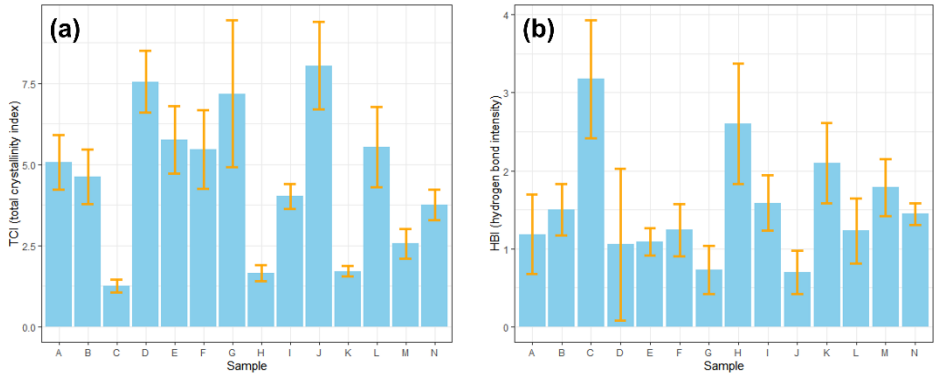

Fig. 11. (a) TCI and (b) HBI of copy papers.

Before the machine learning modeling with IR spectrum datasets, TCI and HBI were determined. Figure 11 shows the TCI and HBI of copy papers. As shown in Fig. 11, the results of TCI and HBI indicate an opposite trend called the "mirror effect" (Široký *et al.* 2010).

If the papers contain sizing agents or inorganic fillers on their surface, they may demonstrate a relatively higher crystallinity index (Kang *et al.* 2021; Kim *et al.* 2022). The TCI in Fig. 11(a) increased relatively with the higher content of $CaCO_3$, as shown in Table 3. Additionally, recycled fiber refers to pulp that has undergone multiple reuse cycles. The entire recycling process undergoes extensive physical and chemical treatments, resulting in significant amounts of cutting/decomposition of fibers. These phenomena primarily occur within the amorphous region of cellulose, which is distinctive from the crystalline region. Consequently, recycled fiber may exhibit a relatively higher crystallinity index (Kang *et al.* 2021).

The HBI serves as an indicator of the presence of hydroxyl groups (−OH), which play a significant role in the bonding between cellulose fibers. A high HBI value suggests excellent hydrogen-bonding ability with other fibers. When the paper has a high content of inorganic filler, the average distance between the wood fibers increases, and the chance of developing hydrogen bonding between wood fibers decreases (Han *et al.* 2020). The HBI corresponds to the results in Table 3 and aligns with the previously mentioned "mirror effect". However, in the case of copy paper H, which had a high clay content, it was deemed inconsistent with these findings.

Considering the TCI and HBI, there are still limitations in distinguishing the copy paper samples, especially copy paper A and E. Therefore, relying solely on the information provided by TCI and HBI to accurately specify and classify the manufacturing process and chemical properties of copy-paper products may be deemed unreasonable.

The PLS-DA Model

Figure 12 shows the PLS-DA score plots based on ATR-IR spectra. Table 4 shows the classification performance of PLS-DA. In the 4,000 to 400 cm^{-1} range, as illustrated in Fig. 12(a), when conducting PLS-DA on raw IR spectra without any preprocessing, the clusters exhibit severe aggregation. Consequently, the classification-model performance is adversely affected, resulting in low accuracy, as shown in Table 4. Furthermore, the utilization of second-derivative spectra in the range of 4,000 to 400 cm^{-1} had no significant effect on the classification performance of the PLS-DA model. However, the spectra in the range of 1,800 to 800 cm^{-1}, the clusters in Fig. 12(c) and Fig. 12(d) demonstrate a higher level of dispersion than those in Fig. 12(a). Specifically, the use of the second-derivative spectra in the range of 1,800 to 800 cm^{-1}, indicating the best performance on both the training and test sets, could prove the most effective approach.

Table 4. Classification Performance of PLS-DA models with IR Spectra

Wavenumber (cm^{-1})	Preprocessing	Accuracy	
		Train	Test
4000-400	Raw	0.622	0.738
	Second-derivative	0.653	0.691
1800-800	Raw	0.612	0.762
	Second-derivative	0.653	0.762

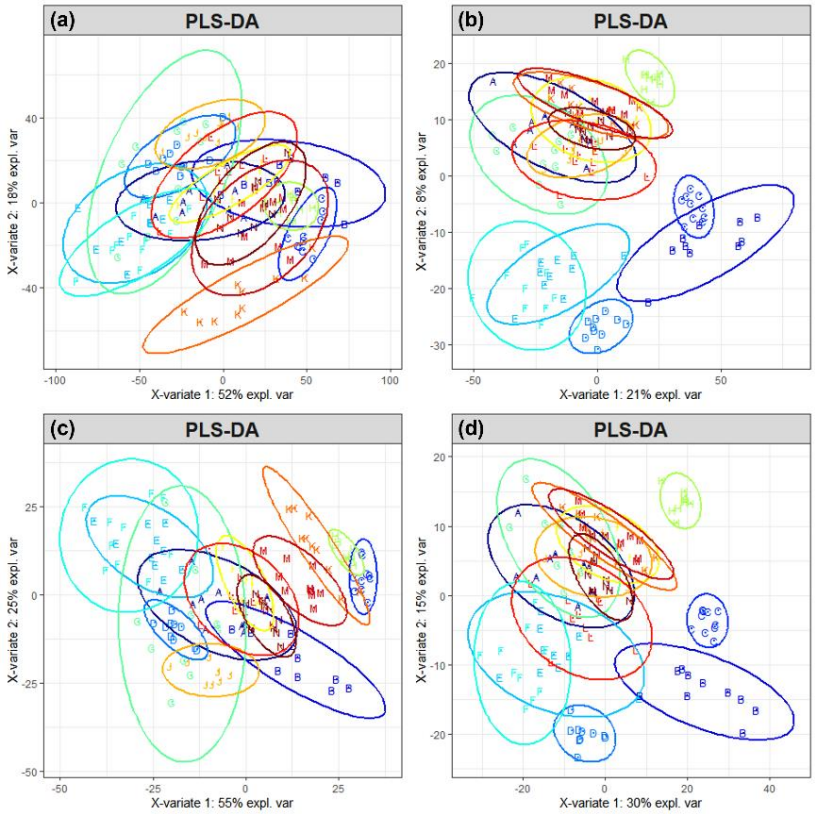

Fig. 12. PLS-DA score plot based on (a) the spectra in the range of 4,000 to 400 cm^{-1}, (b) second-derivative spectra in the range of 4,000 to 400 cm^{-1}, (c) the spectra in the range of 1,800 to 800 cm^{-1}, and (d) second-derivative spectra in the range of 1,800 to 800 cm^{-1}).

Figure 13(a) presents the range of the second-derivative spectra of 1,800 to 800 cm^{-1}, and Fig. 13(b) shows the principal component loading values within the same range.

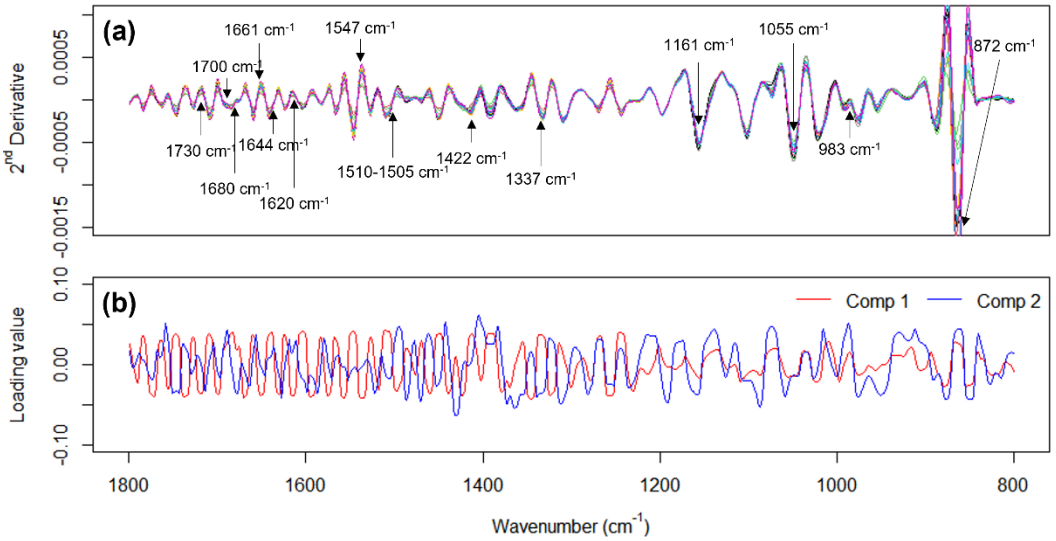

Fig. 13. Loading values of principal components 1 and 2 within 1,800 to 800 cm^{-1} for PLS-DA score plot in Fig. 12(d) (a: the second-derivative spectra of 1,800 to 800 cm^{-1} b: the principal component loading values).

These figures provide compelling evidence for identifying the significant spectra, which is essential for classification analysis. A key approach to achieving this is by comparing the principal component (PC) loading values obtained from the PLS analysis.

The comparison between the second-derivative spectra and PLS loading values revealed that cellulose, lignin, and additives, such as sizing agent, fillers and glue, were significant contributors to cluster formation. Significantly, in the range of 1,800 to 1500 cm^{-1} spectral data influenced scores of PC 1. Also, the range of 1,200 to 800 cm^{-1} spectral data influenced scores of PC 2. Therefore, the observed spread of copy papers towards the right side along PC 1 in Fig. 12(d) suggests distinctions from other paper types, potentially related to variations in inorganic fillers, sizing agents, pulp blend proportions, or differences in residual components like lignin and adhesives such as PAM, starch, *etc.* Figure 14 shows the second-derivative spectra extraction corresponding to PC 1.

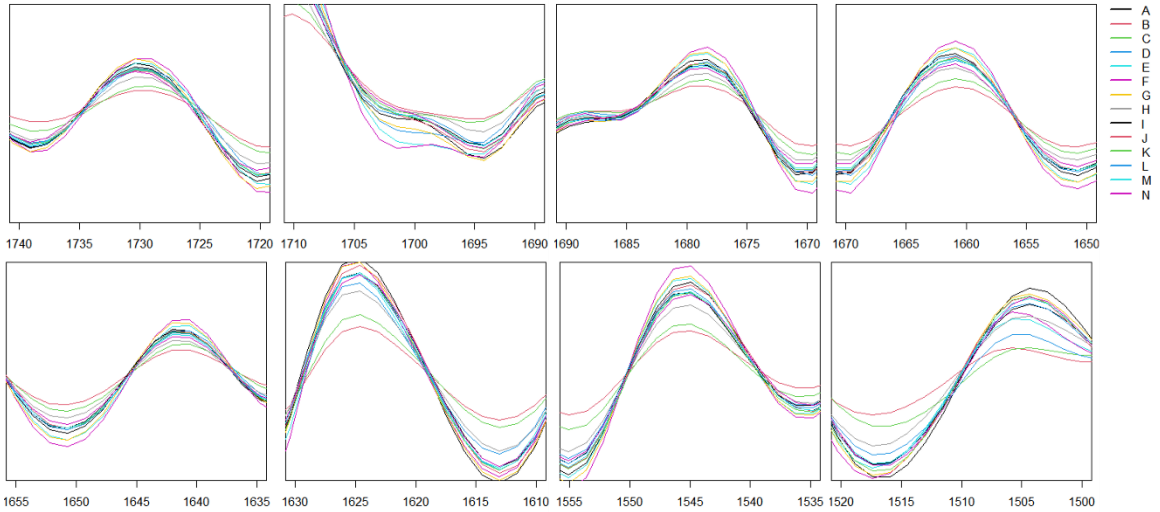

Fig. 14. second-derivative spectra extraction corresponding to PC 1

In principle component 2, numerous peaks were observed in the carbohydrate region, and the clusters of copy papers were arranged vertically (Fackler *et al.* 2011; Široký *et al.* 2010). The formation of clusters based on carbohydrate peaks suggested the influence of cellulose and hemicellulose. The difference peaks around the cellulose fingerprint at 1161 (C_1-O-C_4 antisymmetric stretching), 1055 (C_3-OH stretching), 983 (C_6-O_6 stretching minor), and 872 (C_2-H bending) cm^{-1} are assigned to the carbohydrates in cellulose and hemicellulose (Široký *et al.* 2010; Fackler *et al.* 2011; Horikawa *et al.* 2019; Kang *et al.* 2021; Kim *et al.* 2022).

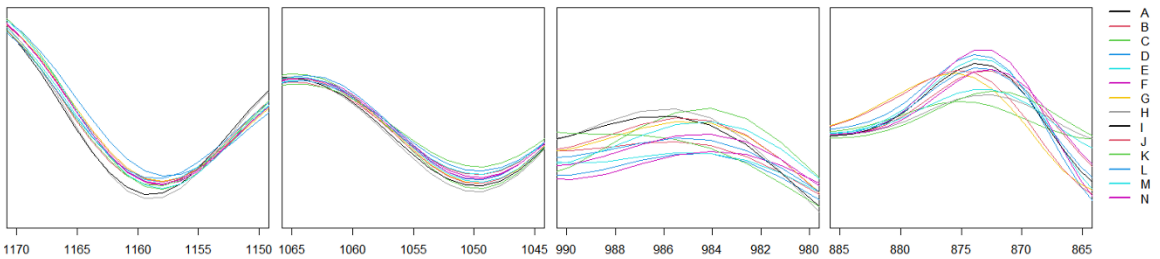

Fig. 15. second-derivative spectra extraction corresponding to PC 2

In other words, variations in cellulose and hemicellulose, resulting from different manufacturing processes or materials, influenced cluster formation. Figure 15 shows the second-derivative spectra extraction corresponding to PC 2.

Consequently, the classification and analysis of copy-paper samples were achieved by detecting chemical properties associated with the bonding of carbon and hydrogen as well as hydroxyl groups in carbohydrates, along with the presence of small amounts of constituents, such as lignin and chemical agent in the pulp, utilizing IR spectroscopy and machine learning modeling.

Classification-Performance Assessment

The performance of the three-classification model, *i.e.*, PLS-DA, SVM, and KNN, was assessed using evaluation measures, such as accuracy, sensitivity, and specificity, derived from a confusion matrix. For all the modeling, the second-derivative spectra in the range of 1,800 to 800 cm^{-1} dataset were used as confirmed in PLS-DA modeling. Table 5 shows the classification accuracy on the training and test datasets. And Table 6 presents the classification performance by accuracy, sensitivity, and specificity on test datasets.

The SVM model obtained the highest values than the other classification models; accuracy (1.00), sensitivity (1.00) and specificity (1.00). PLS-DA and SVM were fundamentally linear models; however, the SVM had the capability to conduct nonlinear classification by projecting data into a high-dimensional feature space using the kernel trick (Vert *et al.* 2004). The KNN is a straightforward algorithm that classifies data points based on their proximity to other data points in the feature space, without a prior data learning process. Considering Occam's razor theory, which suggests that if the errors on the training set are similar, a simpler model is more likely to have lower generalization error. KNN can be considered an effective tool for classification analysis of copy paper (Domingos 1998), accuracy on the training datasets (0.796), and the test datasets (0.929).

However, a specific model cannot completely replace the effectiveness of each model. Unlike PLS-DA, SVM and KNN do not provide logical explanations for classification rules. When selecting a model for predictive modeling through machine learning, a comprehensive review is required, considering factors such as data size, computational cost, and overfitting. Particularly in ATR-IR analysis, highlights the significance of selecting appropriate equipment and model for classification purposes (Xia *et al.* 2023). Figure 16 shows the summary of confusion matrix from three models. Table 6 summarizes the results of the three models, and the detailed assignment information of the test set is presented in Fig. 16. Figure 16 provides the information on misclassified cases for each classification model. Significantly, the SVM model with the second-derivative spectra in the range of 1,800 to 800 cm^{-1} dataset showed the correct prediction rate of 100%, as evident from Fig. 16(b).

Table 5. Classification Accuracy on Training and Test Datasets

Model	Accuracy	
	Train	Test
PLS-DA	0.653	0.762
SVM	1.000	1.000
KNN	0.796	0.929

Table 6. Comparison of the Classification Performance by Accuracy, Sensitivity, and Specificity on Test Dataset

Code	PLS-DA			SVM			KNN		
	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.	Acc.	Sen.	Spe.
A	0.76	0.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00
B		1.00	1.00		1.00	1.00		1.00	1.00
C		1.00	1.00		1.00	1.00		1.00	1.00
D		1.00	1.00		1.00	1.00		1.00	1.00
E		0.00	1.00		1.00	1.00		1.00	0.97
F		1.00	0.92		1.00	1.00		0.67	1.00
G		0.67	1.00		1.00	1.00		0.67	1.00
H		1.00	1.00		1.00	1.00		1.00	1.00
I		1.00	0.97		1.00	1.00		1.00	1.00
J		1.00	0.97		1.00	1.00		1.00	0.97
K		0.33	0.97		1.00	1.00		1.00	1.00
L		1.00	0.95		1.00	1.00		1.00	1.00
M		0.67	0.95		1.00	1.00		0.67	1.00
N		1.00	0.97		1.00	1.00		1.00	0.97

*Acc. = Accuracy; Sen. = Sensitivity; Spe. = Specificity

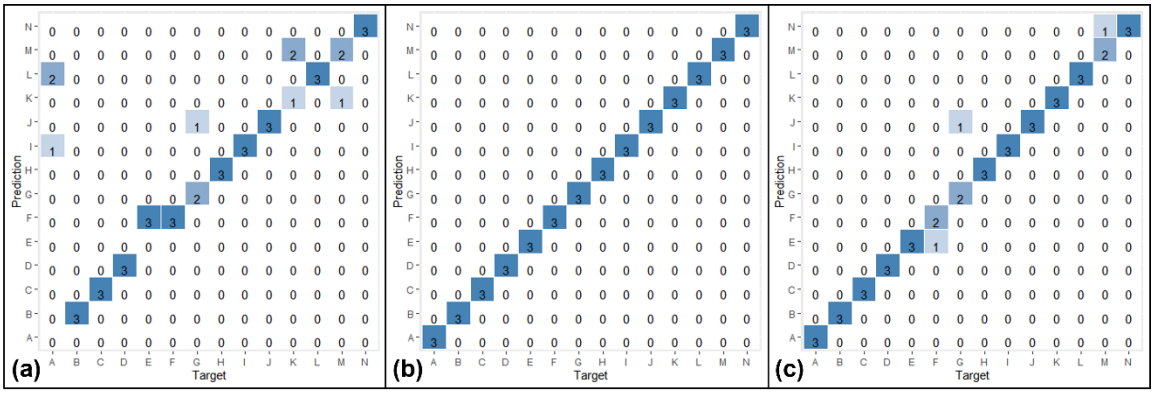

Fig. 16. Summary of the confusion matrices of (a) PLS-DA, (b) SVM, and (c) KNN

CONCLUSIONS

This study explored the feasibility of classifying copy papers based on product or manufacturer. The inorganic filler content, SEM-EDX, and ATR-IR analysis were performed. The analysis of inorganic filler content using ash content analyzer allowed for partial identification of copy paper products by different filler levels among them, although this was insufficient for same manufacturer such as Sample A and E. SEM-EDX analysis detected the distinguishable feature for Sample H based on specific pigment. However, this method also didn't provide adequate discrimination for comprehensive analysis.

To implement this approach, the ATR-IR analysis was employed, and classification models PLS-DA, SVM, and KNN, were developed for evaluation purposes. The key findings and conclusions were as follows:

When analyzing copy papers through IR spectral data, the polymeric characteristics (TCI) and chemical characteristics (HBI) of copy paper provided valuable information for identification. However, it is worth noting that relying solely on these characteristics may present limitations in classifying copy papers.

Preprocessing techniques, specifically the fifth-order polynomial Savitzky–Golay second-derivative, enhanced the clarity of significant spectra, resulting in more effective classification analysis. Notably, the utilization of the second-derivative spectra within the range of 1,800 to 800 cm^{-1} was the most effective for developing classification models.

The comparative analysis of the second-derivative spectra and PLS loading values revealed the substantial contributions of fillers, chemical agents, lignin and carbohydrates, including cellulose and hemicellulose, to cluster formation.

The results demonstrated that the most effective classification model using ATR-IR was SVM, which exhibited excellent performance with 100% accuracy, sensitivity, and specificity. The SVM model utilized in this research could be a valuable tool for classifying and detecting paper products.

ACKNOWLEDGMENTS

This study was carried out with the support of the R&D Program for Forest Science Technology (Project No. FTIS 2019150B10-2323-0301), which was provided by the Korea Forest Service (Korea Forestry Promotion Institute). Also, this work was supported by the National Institute of Forest Science (grant No. FP0400-2023-01).

REFERENCES CITED

Agarwal, A., Sharma, P., Alshehri, M., Mohamed, A. A., and Alfarraj, O. (2021). "Classification model for accuracy and intrusion detection using machine learning approach," *PeerJ. Comput. Sci.* 7, article e437. DOI: 10.7717/peerj-cs.437

Calvini, P., and Gorassini, A. (2002). "FTIR–deconvolution spectra of paper documents," *Restaurator* 23(1), 48-66. DOI: 10.1515/REST.2002.48

Canals, T., Riba, J., Cantero, R., Cansino, J., Domingo, D., and Iturriaga, H. J. T. (2008). "Characterization of paper finishes by use of infrared spectroscopy in combination with canonical variate analysis," *Talanta* 77(2), 751-757. DOI: 10.1016/j.talanta.2008.07.059

Castro, K., Princi, E., Proietti, N., Manso, M., Capitani, D., Vicini, S., Madariaga, J. M., and De Carvalho, M. L. (2011). "Assessment of the weathering effects on cellulose based materials through a multianalytical approach," *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 269(12), 1401-1410. DOI: 10.1016/j.nimb.2011.03.027

Causin, V., Marega, C., Marigo, A., Casamassima, R., Peluso, G., and Ripani, L. (2010). "Forensic differentiation of paper by X-ray diffraction and infrared spectroscopy," *Forensic Science International* 197(1-3), 70-74. DOI: 10.1016/j.forsciint.2009.12.056

Chauchard, F., Svensson, J., Axelsson, J., Andersson-Engels, S., and Roussel, S. (2008). "Localization of embedded inclusions using detection of fluorescence: Feasibility study based on simulation data, LS-SVM modeling and EPO pre-processing," *Chemometr. Intell. Lab. Syst.* 91(1), 34-42. DOI: 10.1016/j.chemolab.2007.08.008

Chevallier, S., Bertrand, D., Kohler, A., and Courcoux, P. (2006). "Application of PLS-DA in multivariate image analysis," *J. Chemom.* 20(5), 221-229. DOI: 10.1002/cem.994

Choi, K. H., Lee, J. H., and Ryu, J. Y. (2018a). "Analysis of modern document paper for identification of counterfeit document (I): Filler composition," *Journal of Korea TAPPI* 50(3), 28-35. DOI: 10.7584/JKTAPPI.2018.06.50.3.28

Choi, K. H., Lee, J. H., and Ryu, J. Y. (2018b). "Analysis of modern document paper for identification of counterfeit document (II): Fiber identification," *Journal of Korea TAPPI* 50(4), 15-24. DOI: 10.7584/JKTAPPI.2018.08.50.4.15

Ciolacu, D., Kovac, J., and Kokol, V. (2010). "The effect of the cellulose-binding domain from Clostridium cellulovorans on the supramolecular structure of cellulose fibers," *Carbohydrate Research* 345(5), 621-630. DOI: 10.1016/j.carres.2009.12.023

DeVries, T. J., Von Keyserlingk, M. A. G., Weary, D. M., and Beauchemin, K. A. (2003). "Technical note: Validation of a system for monitoring feeding behavior of dairy cows," *JDS Commun.* 86(11), 3571-3574. DOI: 10.3168/jds.S0022-0302(03)73962-9

Domingos, P. (1998). "Occam's two razors: The sharp and the blunt," *KDD*, pp. 37-43.

Fackler, K., Stevanic, J. S., Ters, T., Hinterstoisser, B., Schwanninger, M., and Salmén, L. (2011). "FT-IR imaging microscopy to localise and characterise simultaneous and selective white-rot decay within spruce wood cells," *Holzforschung* 65(3), 411-120. DOI: 10.1515/hf.2011.048

Foner, H. A., and Adan, N. (1983). "The characterization of papers by X-ray diffraction (XRD): measurement of cellulose crystallinity and determination of mineral composition," *Journal of the Forensic Science Society* 23(4), 313-321. DOI: 10.1016/S0015-7368(83)72269-3

Ganzerla, R., Gambaro, A., Cappelletto, E., Fantin, M., Montalbani, S., and Orlandi, M. (2009). "Characterization of selected paper documents from the archives of Palazzo Ducale (Venice), Italy using various analytical techniques," *Microchemical Journal* 91(1), 70-77. DOI: 10.1016/j.microc.2008.08.003

Garside, P., and Wyeth, P. (2003). "Identification of cellulosic fibres by FTIR spectroscopy-thread and single fibre analysis by attenuated total reflectance," *Studies in Conservation* 48(4), 269-275. DOI: 10.1179/sic.2003.48.4.269

Gracia, I. A. (2001). "Applicazioni della spettrofotometria IR allo studio dei beni culturali," *II prato*.

Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). "KNN model-based approach in classification," in: *On The Move to Meaningful Internet Systems 2003*: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003, Springer, pp. 986-996. DOI: 10.1007/978-3-540-39964-3_62

Han, J. S., Jung, S. Y., Kang, D. S., and Seo, Y. B. (2020). "Development of flexible calcium carbonate for papermaking filler," *ACS Sustainable Chemistry & Engineering* 8(24), 8994-9001. DOI: 10.1021/acssuschemeng.0c01593

Hens, A. B., and Tiwari, M. K. (2012). "Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method," *Expert Systems with Applications* 39(8), 6774-6781. DOI: 10.1016/j.eswa.2011.12.057

Hodges, R., Cullinan, H., and Krishnagopalan, G. A. J. T. J. (2006). "Recent advances in the commercialization of NIR (near-infrared) based liquor analyzers in the pulping and recovery area," *TAPPI J. 5*(11), 3.

Horikawa, Y., Hirano, S., Mihashi, A., Kobayashi, Y., Zhai, S., and Sugiyama, J. (2019). "Prediction of lignin contents from infrared spectroscopy: Chemical digestion and lignin/biomass ratios of *Cryptomeria japonica*," *Appl. Biochem. Biotechnol.* 188(4), 1066-1076. DOI: 10.1007/s12010-019-02965-8

Hu, L., Lu, X., and Ma, J. (2020). "Research review on devices and methods for rapid measurement of paper ash," *BioResources* 15(1), 2096-2110. DOI: 10.15376/biores.15.1.2096-2110

Hwang, S. W., Horikawa, Y., Lee, W. H., and Sugiyama, J. (2016). "Identification of *Pinus* species related to historic architecture in Korea using NIR chemometric approaches," *Journal of Wood Science* 62(2), 156-167. DOI: 10.1007/s10086-016-1540-0

Hwang, S. W., Park, G. Y., Kim, J. H., and Jung, M. J. (2023). "Predictive modeling of traditional Korean paper characteristics using machine learning approaches (Part 1): Discriminating manufacturing origins with artificial neural networks and infrared spectroscopy," *Journal of Korea Technical Association of the Pulp and Paper Industry* 55(4), 57-69. DOI: 10.7584/JKTAPPI.2023.8.55.4.57

Indahl, U. G., Martens, H., and Naes, T. (2007). "From dummy regression to prior probabilities in PLS-DA," *J. Chemom.* 21(12), 529-536. DOI: 10.1002/cem.1061

Jang, K. J., Heo, T. Y., and Jeong, S. H. (2020). "Classification option for Korean traditional paper based on type of raw materials, using near-infrared spectroscopy and multivariate statistical methods," *BioResources* 15(4), 9045-9058. DOI: 10.15376/biores.15.4.9045-9058

Kabir, M. F., Schmoldt, D. L., Araman, P. A., Schafer, M. E., and Lee, S. M. (2003). "Classifying defects in pallet stringers by ultrasonic scanning," *Wood Fiber Sci.*, 341-350.

Kang, K. H., Kim, J. H., and Kim, K. J. (2021). "Analysis of classification characteristics of Americas copy paper using the infrared spectroscopy and principal component analysis," *JKDISS* 32(6), 1195-1204. DOI: 10.7465/jkdi.2021.32.6.1195

Kher, A., Mulholland, M., Reedy, B., and Maynard, P. (2001). "Classification of document papers by infrared spectroscopy and multivariate statistical techniques," *Appl. Spectrosc.* 55(9), 1192-1198. DOI: 10.1366/0003702011953199

Kher, A., Stewart, S., and Mulholland, M. (2005). "Forensic classification of paper with infrared spectroscopy and principal components analysis," *J. Near Infrared Spectrosc.* 13(4), 225-229. DOI: 10.1255/jnirs.540

Kim, K. J., and Eom, T. J. (2016). "Classification of papers using IR and NIR spectra and principal component analysis," *Journal of Korea TAPPI* 48(1), 34-42. DOI: 10.7584/ktappi.2016.48.1.034

Kim, J. H. (2016). "The characteristics of counterfeit crime and countermeasures: The Korean case," *New Trend of Criminal Law* 52, 37-79.

Kim, J. H., Kang, K. H., and Kim, K. J. (2022). "A study on the classification of european copy papers with infrared spectroscopy and principal component analysis," *Journal of Korea TAPPI* 54(4), 34-41. DOI: 10.7584/JKTAPPI.2022.08.54.4.34

Kumar, R., Kumar, V., and Sharma, V. (2017). "Fourier transform infrared spectroscopy and chemometrics for the characterization and discrimination of writing/photocopier paper types: Application in forensic document examinations," *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* 170, 19-28. DOI: 10.1016/j.saa.2016.06.042

Lasch, P. (2012). "Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging," *Chemometrics and Intelligent Laboratory Systems* 117, 100-114. DOI: 10.1016/j.chemolab.2012.03.011

Lee, J., Kim, H., Yook, S., and Kang, T. Y. (2023). "Identification of document paper using hybrid feature extraction," *Journal of Forensic Sciences* 68(5), 1808-1815. DOI: 10.1111/1556-4029.15330

Mancini, M., Taavitsainen, V. M., and Toscano, G. (2019). "Comparison of three different classification methods performance for the determination of biofuel quality by means of NIR spectroscopy," *J. Chemom.* 33(7), e3145. DOI: 10.1002/cem.3145

Marcelo, M. C. A., Martins, C. A., Pozebon, D., and Ferrão, M. F. (2014). "Methods of multivariate analysis of NIR reflectance spectra for classification of yerba mate," *Anal. Methods* 6(19), 7621-7627. DOI: 10.1039/C4AY01350F

Moutinho, I. M. T., Ferreira, P. J. T., and Figueiredo, M. L. (2011). "Paper surface chemistry as a tool to improve inkjet printing quality," *BioResources* 6(4), 4259-4270. DOI: 10.15376/biores.6.4.4259-4270

Nelson, M. L., and O'Connor, R. T. (1964). "Relation of certain infrared bands to cellulose crystallinity and crystal lattice type. Part II. A new infrared ratio for estimation of crystallinity in celluloses I and II," *J. Appl. Polymer Sci.* 8(3), 1325-1341. DOI: 10.1002/app.1964.070080323

Pan, J., and Nguyen, K. L. (2007). "Development of the photoacoustic rapid-scan FT-IR-based method for measurement of ink concentration on printed paper," *Anal. Chem.* 79(6), 2259-2265. DOI: 10.1021/ac061732y

Piras, P., Sheridan, R., Sherer, E. C., Schafer, W., Welch, C. J., and Roussel, C. (2018). "Modeling and predicting chiral stationary phase enantioselectivity: An efficient random forest classifier using an optimally balanced training dataset and an aggregation strategy," *J. Sep. Sci.* 41(6), 1365-1375. DOI: 10.1002/jssc.201701334

Polovka, M., Polovková, J., Vizárová, K., Kirschnerová, S., Bieliková, L., and Vrška, M. (2006). "The application of FTIR spectroscopy on characterization of paper samples, modified by Bookkeeper process," *Vibrational Spectroscopy* 41(1), 112-117. DOI: 10.1016/j.vibspec.2006.01.010

Proniewicz, L. M., Paluszkiewicz, C., Wesełucha-Birczyńska, A., Barański, A., and Dutka, D. (2002). "FT-IR and FT-Raman study of hydrothermally degraded groundwood containing paper," *Journal of Molecular Structure* 614(1-3), 345-353. DOI: 10.1016/S0022-2860(02)00275-2

Proniewicz, L. M., Paluszkiewicz, C., Wesełucha-Birczyńska, A., Majcherczyk, H., Barański, A., and Konieczna, A. (2001). "FT-IR and FT-Raman study of hydrothermally degradated cellulose," *Journal of Molecular Structure 596*(1-3), 163-169. DOI: 10.1016/S0022-2860(01)00706-2

Ruiz, J. R. R., Canals, T., and Gomez, R. C. (2011). "Comparative study of multivariate methods to identify paper finishes using infrared spectroscopy," *IEEE Trans. Instrum. Meas.* 61(4), 1029-1036.

Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., and Mononen, J. (2018). "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behav. Processes* 148, 56-62. DOI: 10.1016/j.beproc.2018.01.004

Samanta, B. I. S. W. A. J. I. T., Al-Balushi, K. R., and Al-Araimi, S. A. (2003). "Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection," *Eng. Appl. Artif. Intell.* 16(7-8), 657-665. DOI: 10.1016/j.engappai.2003.09.006

Savitzky, A., and Golay, M. J. (1964). "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.* 36(8), 1627-1639. DOI: 10.1021/ac60214a047

Seo, J. H., Oh, Y. J., Hwang, K., Gwon, J., Ahn, B. J., Kim, K. J., and Lee, T. J. (2023). "Enhanced bleachability of chemi-thermomechanical pulp by hydrogen peroxide bleaching in ethanol-water media," *BioResources* 18(1), 1731-1741. DOI: 10.15376/biores.18.1.1731-1741

Singh, V. K., Tripathi, D. K., Deguchi, Y., and Wang, Z. (eds.). (2023). *Laser Induced Breakdown Spectroscopy (LIBS): Concepts, Instrumentation, Data Analysis and Applications*, 2 volume set," John Wiley and Sons.

Široký, J., Blackburn, R. S., Bechtold, T., Taylor, J., and White, P. (2010). "Attenuated total reflectance fourier-transform infrared spectroscopy analysis of crystallinity changes in lyocell following continuous treatment with sodium hydroxide," *Cellulose* 17(1), 103-115. DOI: 10.1007/s10570-009-9378-x

Sistach, M. C., Ferrer, N., and Romero, M. T. (1998). "Fourier transform infrared spectroscopy applied to the analysis of ancient manuscripts," *Restaurator* 19(4), 173-186. DOI: 10.1515/rest.1998.19.4.173

Spence, L. D., Baker, A. T., and Byrne, J. P. (2000). "Characterization of document paper using elemental compositions determined by inductively coupled plasma mass spectrometry," *Journal of Analytical Atomic Spectrometry* 15(7), 813-819. DOI: 10.1039/B001411G

Tsuchikawa, S., Yamato, K., and Inoue, K. (2003). "Discriminant analysis of wood-based materials using near-infrared spectroscopy," *J. Wood Sci.* 49(3), 275-280. DOI: 10.1007/s10086-002-0471-0

Vert, J. P., Tsuda, K., and Schölkopf, B. (2004). "A primer on kernel methods." *Kernel Methods in Computational Biology*. 47, 35-70.

Workman Jr, J. J. (1999). "Review of process and non-invasive near-infrared and infrared spectroscopy: 1993-1999," *Appl. Spectrosc. Rev.* 34(1-2), 1-89. DOI: 10.1081/ASR-100100839

Xia, J., Min, S., and Li, J. (2023). "Rapid analysis the type of customs paper using micro-NIR spectrometers and machine learning algorithms," *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* 290, article 122272. DOI: 10.1016/j.saa.2022.122272

Xia, J., Zhang, J., Zhao, Y., Huang, Y., Xiong, Y., and Min, S. (2019). "Fourier transform infrared spectroscopy and chemometrics for the discrimination of paper relic types," *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* 219, 8-14. DOI: 10.1016/j.saa.2018.09.059

Xu, J., Zhang, Y., and Miao, D. (2020). "Three-way confusion matrix for classification: A measure driven view," *Inf. Sci.* 507, 772-794. DOI: 10.1016/j.ins.2019.06.064

Ye, Y., Wu, Q., Huang, J. Z., Ng, M. K., and Li, X. (2013). "Stratified sampling for feature subspace selection in random forests for high dimensional data," *Pattern Recognition* 46(3), 769-787. DOI: 10.1016/j.patcog.2012.09.005

Article submitted: August 1, 2023; Peer review completed: October 14, 2023; Revised version received and accepted: October 31, 2023; Published: November 10, 2023.
DOI: 10.15376/biores.19.1.160-182