One-Dimensional Convolutional Neural Networks with Infrared Spectroscopy for Classifying the Origin of Printing Paper

Sung-Wook Hwang,[a] Geungyong Park,[b] Jinho Kim,[b] Kwang-Ho Kang,[c,*] and Won-Hee Lee [b,*]

Herein, the challenge of accurately classifying the manufacturing origin of printing paper, including continent, country, and specific product, was addressed. One-dimensional convolutional neural network (1D CNN) models trained on infrared (IR) spectrum data acquired from printing paper samples were used for the task. The preprocessing of the IR spectra through a second-derivative transformation and the restriction of the spectral range to 1800 to 1200 cm^{-1} improved the classification performance of the model. The outcomes were highly promising. Models trained on second-derivative IR spectra in the 1800 to 1200-cm^{-1} range exhibited perfect classification for the manufacturing continent and country, with an impressive F1 score of 0.980 for product classification. Notably, the developed 1D CNN model outperformed traditional machine learning classifiers, such as support vector machines and feed-forward neural networks. In addition, the application of data point attribution enhanced the transparency of the decision-making process of the model, offering insights into the spectral patterns that affect classification. This study makes a considerable contribution to printing paper classification, with potential implications for accurate origin identification in various fields.

DOI: 10.15376/biores.19.1.1633-1651

Keywords: Classification; Convolutional neural network; Printing paper; Infrared spectroscopy; Data point attribution

Contact information: a: Human Resources Development Center for Big Data-Based Glocal Forest Science 4.0 Professionals, Kyungpook National University, 80 Daehak-ro, Buk-gu, Daegu 41566, Republic of Korea; b: Department of Wood Science and Technology, College of Agriculture and Life Sciences, Kyungpook National University, 80 Daehak-ro, Buk-gu, Daegu 41566, Republic of Korea; c: HP Printing Korea, 26 Yeonnaegaeul-ro, Sujeong-gu, Seongnam-si, Gyeonggi-do 13105, Republic of Korea;
** Corresponding authors: kwangho.kang@hp.com; leewh@knu.ac.kr*

INTRODUCTION

The rapid advancement of modern technology has been expected to considerably reduce paper consumption in many fields owing to increasing digitalization. However, in reality, the consumption of paper has been increasing owing to various complex factors such as the need for document backup, security concerns, packaging, technological disparities, and user preferences (Shah *et al*. 2023). The development of printing and output technologies further highlights the significance of paper. Printing technology has revolutionized document creation, image reproduction, and data storage, considerably affecting business, education, and research domains. The quality and characteristics of

printing paper are crucial components of these technologies, and fine analysis and distinction of these properties represent one of the vital challenges in this context.

Identifying the origin of paper products is essential for supporting global efforts to combat the illegal trade in timber products and promote sustainability (Australian Government 2012; Korea Legislation Research Institute 2020; Federal Register 2021; European Commission 2023). Furthermore, ensuring optimal printing quality for a particular printing device may require the identification of paper products that are suitable or unsuitable for the device. Paper origin identification is an advanced technology that offers practical value and innovations in various fields, including the prevention of document forgery and the development of new materials and manufacturing processes.

The combination of spectroscopy and multivariate analysis has proven to be a promising approach in classification problems involving various materials (Soriano-Disla *et al*. 2014; Chang *et al*. 2015; Hwang *et al*. 2016; Horikawa *et al*. 2019; Hwang *et al*. 2021). Infrared (IR) spectra are used to measure the IR emission of objects at specific wavelengths; these data can be used to discern the unique characteristics of paper (Stuart 2004; Trafela *et al*. 2007; Causin *et al*. 2010). Recent advancements in machine learning have further improved the predictive performance of models, making them more accurate and robust (Coppola *et al*. 2023; Hwang *et al*. 2023). Machine learning algorithms have already been used to process spectral data and identify the distinctive signatures of different types of paper through pattern recognition and feature extraction (Meza Ramirez *et al*. 2021). Recent research combines laser-induced breakdown spectroscopy (LIBS) and machine learning to achieve diverse goals. One aspect involves enhancing judicial expertise by analyzing ink marks in handwriting identification using LIBS and machine learning (Feng *et al.* 2023). Another facet addresses the misclassification of recyclable waste, employing LIBS and machine learning to create an effective online source tracing system (Chen *et al*. 2023). The system successfully identifies and categorizes smoke from waste paper incineration, demonstrating the possibility of tracing the source of waste paper.

Herein, a one-dimensional convolutional neural network (1D CNN) model using IR spectra was developed to accurately classify the manufacturing origin of printing paper, including the continent, country, and product. This model processes the spectral data of printing paper, learning patterns and features from the data. Furthermore, data point attribution analysis was used to understand how specific absorption bands in the IR spectrum contribute to the classification decisions of the model. This process enhanced the transparency of the decision-making process of the model and improved its interpretability. This article presents the results of machine learning-based classification of the manufacturing origin of printing paper, contributing to the existing body of research in this field.

EXPERIMENTAL

Printing Papers

Herein, 65 commercial products from 24 different manufacturers spanning 11 countries were used for printing paper classification (Table 1). Each product was categorized based on its country of production rather than the manufacturer's country of registration. The products in the sample exhibited considerable variation, with the majority

originating from China (28 products), whereas Finland was represented by only one product.

The majority of the selected products were typical office printing papers with a weight range of 70 to 90 grams per square meter (gsm). Some products exceeded 100 gsm and were intended for special documents, promotional materials, business cards, and other similar applications. Among the 28 products, those of Chinese origin were A4 or A3-sized printing papers with a weight of 70 to 80 gsm, manufactured by 12 companies. Among the 65 products tested, four were composed of recycled paper, whereas a unique product from the United States was produced from cotton. In addition, three original equipment manufacturer products with unverified manufacturer and country of origin information were included in the study. These items were incorporated into the analysis to predict their respective countries of origin.

Table 1. Number of Papers Analyzed and Country of Manufacture

No.	Country	Number of Analyzed Product
1	China	28
2	India	8
3	Indonesia	2
4	Korea	6 (2 recycled products included)
5	Thailand	2
6	Austria	3
7	Finland	1
8	Germany	2 (1 recycled product included)
9	Canada	2
10	USA	9 (1 recycled and 1 cotton paper products included)
11	Brazil	2

Dataset

IR spectra

The IR spectra of printing paper samples spanning the wavenumber range of 4000 to 400 cm^{-1} were acquired using attenuated total reflection infrared (ATR-IR) spectroscopy (ALPHA-P, Bruker Optics, Ettlingen, Germany). The spectral resolution was set to 4 cm^{-1}, and average spectra derived from 16 repeated scans were obtained. ATR-IR spectroscopy can analyze a wide range of samples, including liquids, solids, and powders. It requires minimal sample preparation, allowing for quick and direct analysis. For each printing paper product, the IR spectra from 5 samples were collected, resulting in a dataset comprising 325 spectra for the classification model.

Data preprocessing

The IR spectra were preprocessed using a Savitzky–Golay filter (Savitzky and Golay 1964). The original spectra were transformed into second-derivative spectra using a third polynomial with 21-point smoothing. This preprocessing was used to consistently adjust the baseline of the spectra and amplify peaks, thus emphasizing differences between the spectra (Hwang *et al*. 2016).

The IR spectra in the range of 4000 to 400 cm^{-1} comprise 2545 input variables, including zero-filled points. The IR spectra may contain information that is either noisy or

not useful for sample characterization. Excess input variables are a primary factor increasing the computational cost of the model. Therefore, herein, the IR data from two regions were used for model training: 4000 to 400 cm^{-1} (the entire range) and 1800 to 1200 cm^{-1} (the selected range). The selected range is suitable for paper characterization (Kim and Eom 2016), and it corresponds to 425 input variables.

Through data preprocessing and selection, four datasets were generated from the original IR spectra, including the entire range (Dataset A) and selected range (Dataset B) of the original IR spectra, as well as the entire range (Dataset C) and selected range (Dataset D) of the second-derivative spectra. These four datasets were then used to develop respective classification models through Euclidean (L2) norm-based vector normalization using Eq. 1. (Fig. 1),

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

where v is the vector (IR spectrum) to be normalized, v_i is ith element (data point) of vector v, and n is the number of vector elements.

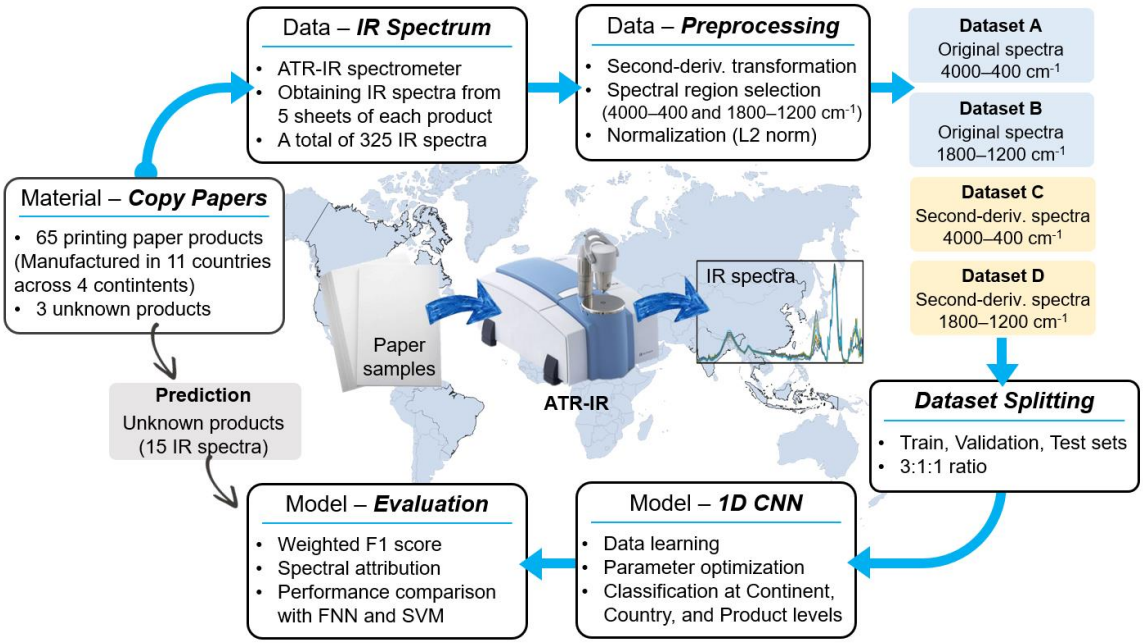

Fig. 1. Diagram for the classification of printing paper using 1D CNN

Dataset splitting

The datasets were split into training, validation, and test sets in a 3:1:1 ratio to build and evaluate the classification models. This ratio represented the minimum requirement for allocating data to each subset in product-level classification. The data were partitioned using stratified random sampling to maintain the specified split ratio for all classes.

Principal Component Analysis (PCA)

To analyze the IR spectral data of printing paper, PCA was conducted on four datasets. Through the PCA, the high-dimensional IR data were transformed into a new orthogonal coordinate system with six principal components (PCs). The transformed data

were subsequently visualized in a two-dimensional (2D) space to investigate the structure and patterns in the IR data for printing paper.

1D CNN Classification Model

1D CNN architecture

The CNNs are fundamental for deep learning; they are predominantly used in image processing, where they excel in feature extraction from 2D data to facilitate image recognition and classification. Similarly, 1D CNNs, operated based on the same technical principles, are used to extract features from 1D data for predictive purposes.

The architecture of the 1D CNN models tailored for the classification of printing paper in this study is illustrated in Fig. 2. The used 1D CNN networks comprise two convolutional layers and two fully connected layers, with each convolutional layer forming a module in conjunction with a max-pooling layer. These modules abstract and extract features from the input data, specifically from the IR spectrum, through data abstraction and down-sampling. Rectified linear unit (ReLU) was used as the activation function. The learned features from the convolutional modules are passed to a network composed of one flatten layer, two fully connected layers, and one softmax layer for training and performing prediction tasks using the input data.

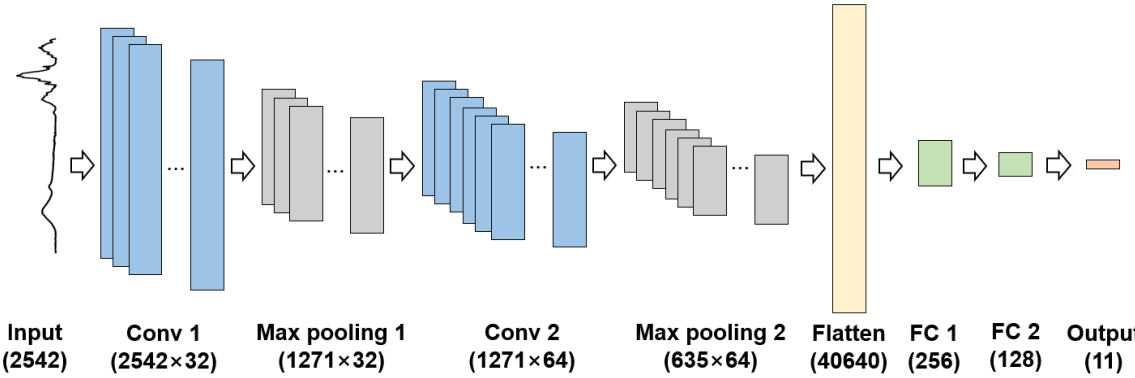

Fig. 2. Architecture of the 1D CNN model for printing paper classification. Numbers in parentheses indicate layer shapes. Notes: Conv, convolution layer; FC, fully connected layer

The details of the hyperparameters tested and their application within the network for establishing the 1D CNN model are shown in Table 2. These hyperparameters were optimized through loop-based testing. Each 1D CNN model for printing paper classification was trained for 700 epochs using categorical crossentropy as the loss function.

Evaluation metric

Printing paper products are inequally distributed across manufacturing countries; thus, the evaluations of the classification performance of models using accuracy may be biased because of oversampled classes. Consequently, in this study, the weighted F1 score was used for assessing the classification performance of the 1D CNN models. The F1 score, which is the harmonic mean of precision (Eq. 2) and recall (Eq. 3), is a commonly used performance metric in classification problems with class imbalance; it is defined in Eq. 3.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

where TP is the true positives, FP is the false positives, and FN is the false negatives.

$$\text{F1}_i = 2 \times \frac{\text{P}_i \times \text{R}_i}{\text{P}_i + \text{R}_i} \quad (4)$$

where F1_i, P_i, and R_i are the F1 score, precision, and recall for class i, respectively.

Table 2. Detailed Hyperparameters Used for Building the 1D CNN Model

Hyperparameter Configuration		
• kernel_size = [3, 5, 7] • filters = [16, 32, 64] • pool_size = [2, 4] • dense_units = [128, 256]	• dropout_rate = [0.5, 0.3] • learning_rate = [0.0001, 0.001, 0.01, 0.1] • optimizer = [SGD, Adam, RMSProp]	
Layer	Layer Shape	Hyperparameters
Conv_1	(n_features, filters)	kernel_size, filters
Max_pool_1	(n_features / pool_size, filters)	pool_size
Conv_2	(n_features / pool_size, filters × 2)	kernel_size, filters
Max_pool_2	((n_features / pool_size) / pool_size), filters × 2)	pool_size
Flatten	((n_features / pool_size) / pool_size) × (filters × 2)	-
Dense_1	dense_units	dense_units
Dropout_1	dense_unit	dropout_rate
Dense_2	dense_units / 2	dense_units / 2
Dropout_2	dense_units / 2	dropout_rate

Notes: kernel_size, the size of the convolutional kernel; filters: the number of filters applied in the convolutional layers; pool_size: the size of the pooling window in max pooling layers; dense_units: the number of nodes in the dense layers; dropout_rate: the rate at which dropout is applied in dropout layers; learning_rate: the learning rate used in the training; optimizer: the optimization algorithm chosen for training the model; SGD, stochastic gradient descent; Adam, adaptive moment estimation; RMSProp, root mean squared propagation; The values in square brackets represent the values of each hyperparameter used in building the model; Layer, the specific layer in the network architecture; Layer Shape, the shape or dimensions of the layer; Hyperparameters, the hyperparameters applied to each layer; Conv, convolution layer; Max_pool, maximum pooling layer; Dense, dense layer; n_features, the number of data points comprising the input data.

The weighted F1 score used for the assessment of 1D CNN model performance takes into account class imbalances by calculating the weights for each class (Eq. 5) and incorporates them into their respective F1 scores (Eq. 6). Through this process, the weighted F1 score assesses individual classes and the overall model performance even for imbalanced datasets.

$$w_i = \frac{N_i}{T_i} \quad (5)$$

where w_i is the weight of class i; N_i is the number of samples in class i; and T_i is the total number of samples.

$$\text{Weighted F1}_i = \sum_{i=1}^{N} \text{w}_i \times \text{F1}_i \quad (6)$$

Data point attribution

To assess the effect of individual data points within the given input IR data on the predictions of the 1D CNN models, the gradient-weighted activation mapping (Grad-CAM) method was used for data point attribution. Data point attribution is a fundamental tool for interpreting model predictions by tracing back the output of the model.

Data point attribution involves computing the gradient of the loss function to understand its sensitivity to each parameter and input data point. The gradient value indicates the extent to which a given data point affects the loss function, with a higher absolute gradient value signifying a substantial effect of that data point on the output of the model. The results of data point attribution were visualized alongside the IR spectra to quantitatively determine the importance of each data point and facilitate model interpretation.

Prediction of unknown products

Three products with unknown manufacturing information were used to predict their origins using the developed models. The PCA was performed on their IR spectra to analyze their relationships with existing data. Subsequently, they were input into the established 1D CNN models to calculate the prediction probabilities for each class (Fig. 1). When inputting the unknown products into the 1D CNN model, the IR spectra were preprocessed in the same way as those used in model construction.

Model Comparison

The classification performance of the constructed 1D CNN models was compared with those of conventional machine learning classifiers: feed-forward neural network (FNN) and support vector machine (SVM). They were trained on the same four datasets used to establish the 1D CNN models, thus constructing their respective classification models.

FNN

The FNN with a backpropagation algorithm was used as a benchmark method for the 1D CNN. When constructing the models, ReLU was adopted as the activation function and crossentropy was used as the loss function. Stochastic gradient descent and adaptive moment estimation (Adam) were used to optimize the loss function. The initial learning rate ranged from 0.0001 to 0.1, with a maximum of 1000 iterations. The FNN architecture was configured with either one or two hidden layers, each containing 12, 256, or 512 nodes. A grid search was conducted to determine the optimal parameters and network structure for the FNN models.

SVM

To facilitate performance comparison, SVM models were constructed using the radial basis function kernel (Vert *et al*. 2004), a technique that projects data into a high-dimensional space to determine hyperplanes. During model construction, the parameter cost, responsible for regulating the misclassification cost of the training data, was set in the range of 10^0 to 10^5. In addition, the parameter Gamma, which governs the Gaussian kernel used for nonlinear classification, was configured within the range of 10^{-1} to 10^{-6}. These parameters were optimized *via* a grid search.

RESULTS AND DISCUSSION

IR Spectral Characteristics of Printing Paper

The IR spectrum contains valuable information for capturing and interpreting the characteristics of paper. Figure 3 presents the IR spectra of select samples from the four datasets. In the original IR spectra (Fig. 3a and 3b), distinct peaks at 3600 to 3000 cm^{-1} were assigned to OH groups (Hofstetter *et al*. 2006), peaks at 2890 to 2780 cm^{-1} to CH stretching (Xiao *et al*. 2015), at 1647 to 1635 cm^{-1} to adsorbed water (Olsson and Salmén 2004), at 1430 to 1416 cm^{-1} to the CH_2 bending of crystalline cellulose (Schwanninger *et al*. 2004; Delmotte *et al*. 2008), and at 1200 to 900 cm^{-1} to the fingerprint peaks of cellulose (Garside and Wyeth 2003).

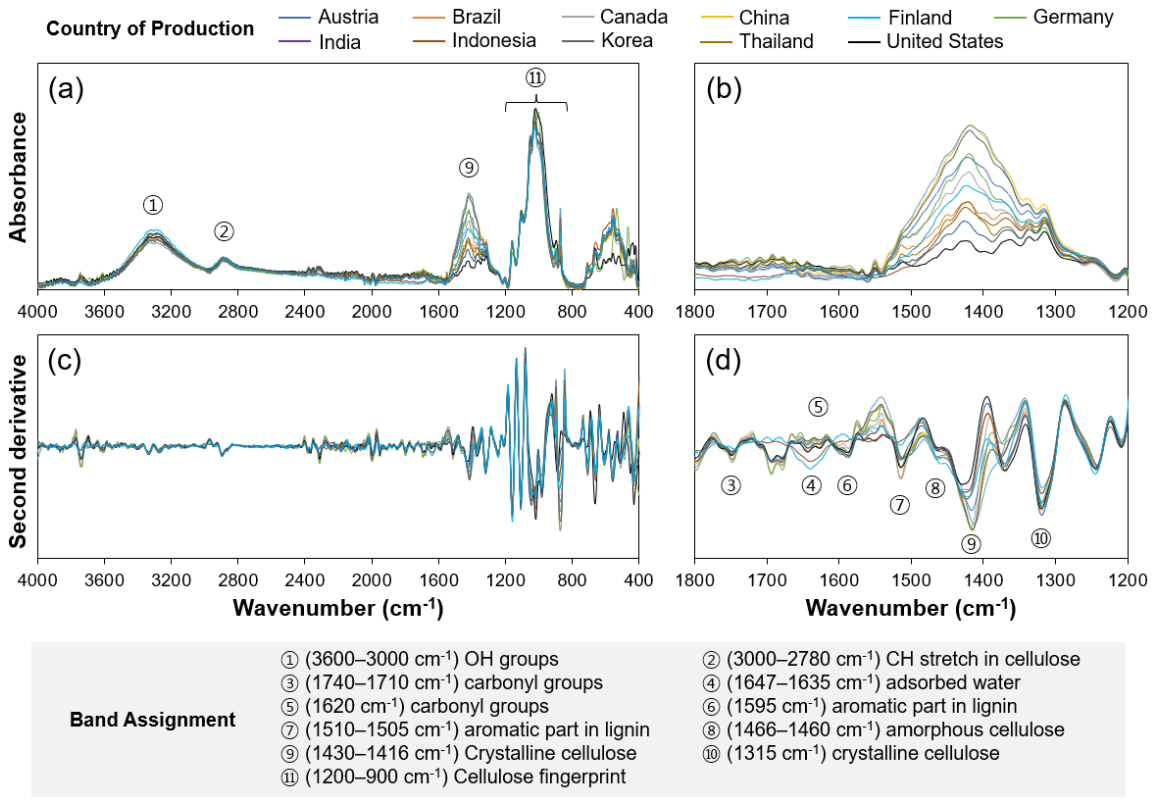

Fig. 3. Entire IR spectra of printing paper samples and the selected region: original (a, b) and second-derivative data (c, d)

In the second-derivative spectra (Fig. 3c and 3d), peaks were amplified, accentuating the distinctions between spectra. Moreover, the 1600 to 1200 cm^{-1} region, where multiple peaks in the original spectra overlap, is distinctly separated through the second-derivative transformation (Fig. 3d). In the second-derivative spectra, several absorption bands, apart from those prominently present in the original spectra, are enhanced. These include bands assigned to carbonyl groups at 1740 cm^{-1} (Schwanninger *et al*. 2004), aromatic parts in lignin at 1510 cm^{-1} (Pandey and Pitman 2003) and 1244 cm^{-1} (Delmotte *et al*. 2008), amorphous cellulose at 1466 to 1460 cm^{-1} (Hajji *et al*. 2016), and crystalline cellulose at 1315 cm^{-1} (Colom and Carrillo 2002).

PCA

PCA is a useful technique for extracting and analyzing patterns and structures in high-dimensional data. Figure 4 shows score plots depicting the first two PCs derived from the four IR datasets.

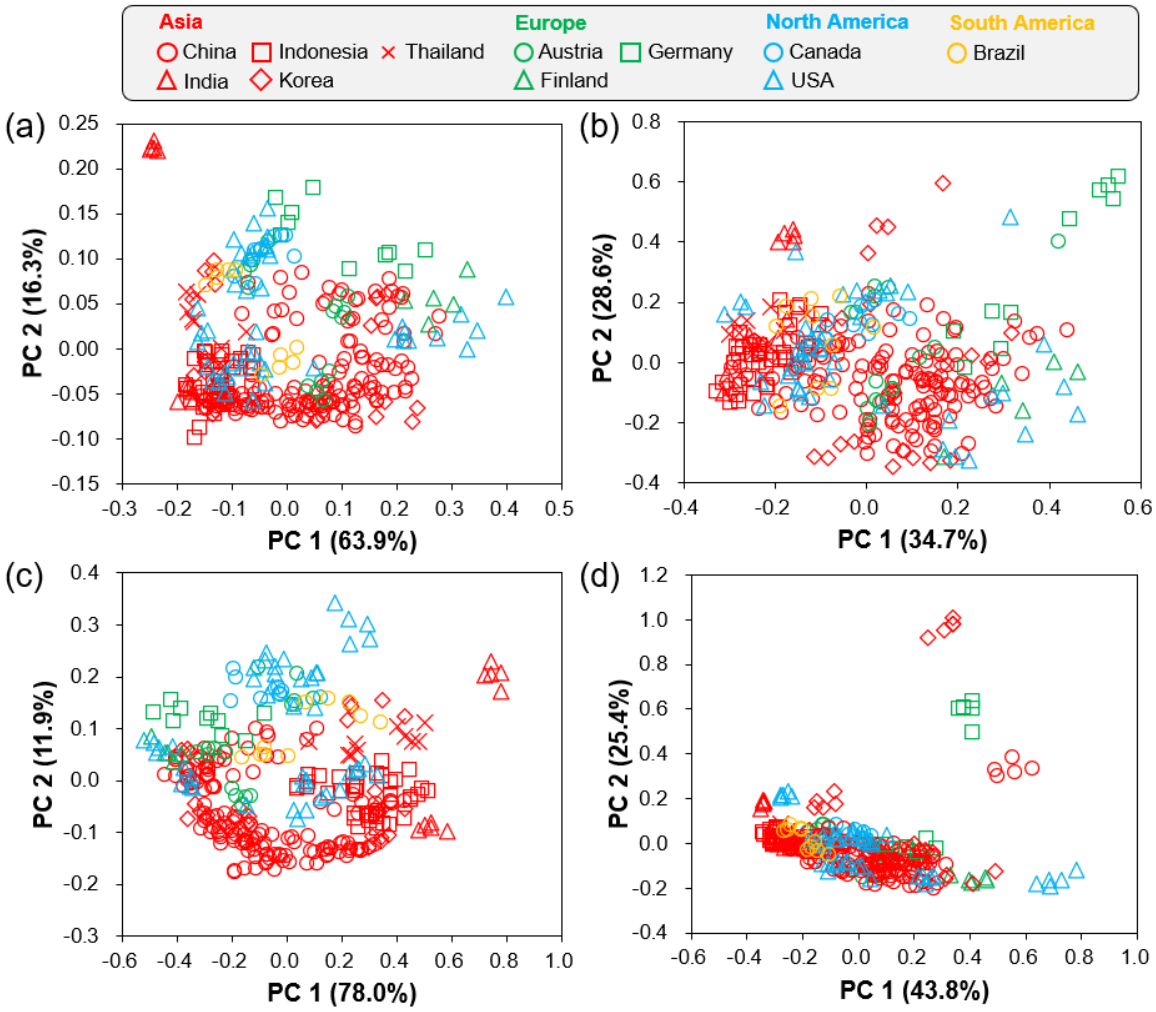

Fig. 4. PC score plots for the first two PCs in the 4000–400 cm^{-1} (a) and 1800–1200 cm^{-1} (b) regions of the original IR spectra, and score plots in the 4000–400 cm^{-1} (c) and 1800–1200 cm^{-1} (d) regions of second-derivative spectra

In all the score plots, data points from the majority of samples are mixed, forming a unified, large cluster, whereas some samples are grouped into smaller, distinct clusters. Notably, in contrast to other score plots (Fig. 4a, 4b, 4c), the smaller clusters are more prominently separated from the larger cluster in the score plots of second-derivative spectra in the 1800 to 1200 cm^{-1} region (Fig. 4d). These clusters correspond to products manufactured in Korea, Germany, China, and the United States, with each cluster comprising data points from the same product. Korean and German products have been identified as recycled paper. The second-derivative spectra of four products isolated from the large cluster in the PC score plot (Fig. 4d) and the loading values of the first two PCs are shown in Fig. 5. The absorbance bands at 1510 and 1595 cm^{-1} (Fig. 5b), assigned to the aromatic part of lignin, contributed to the positioning of Korean and German recycled products in the high-PC2 region of the score plot. They showed stronger peaks in those regions than in the control IR spectrum. These results suggest that unbleached pulp was possibly used in the manufacturing of the recycled products. The characteristics of the IR spectra and PC loading of the Chinese products, forming a distinct cluster, closely resemble those of the Korean and German recycled products. The isolated cluster of the United States products with high PC1 values exhibited a conspicuously strong peak at 1416 cm^{-1} (Fig. 5a), which was assigned to the crystalline cellulose. Furthermore, the substantial negative value in the PC1 loading for this region indicated that it was a distinctive feature of these products. Variations in the crystallinity of cellulose in printing paper samples were attributed to differences in cooking methods and conditions (Gümüşkaya *et al*. 2003). However, given that the printing paper is kraft pulp based, factors other than cooking methods likely contributed to this effect because the use of recycled pulp in the manufacturing of this product cannot be discounted (Sheikhi *et al*. 2010).

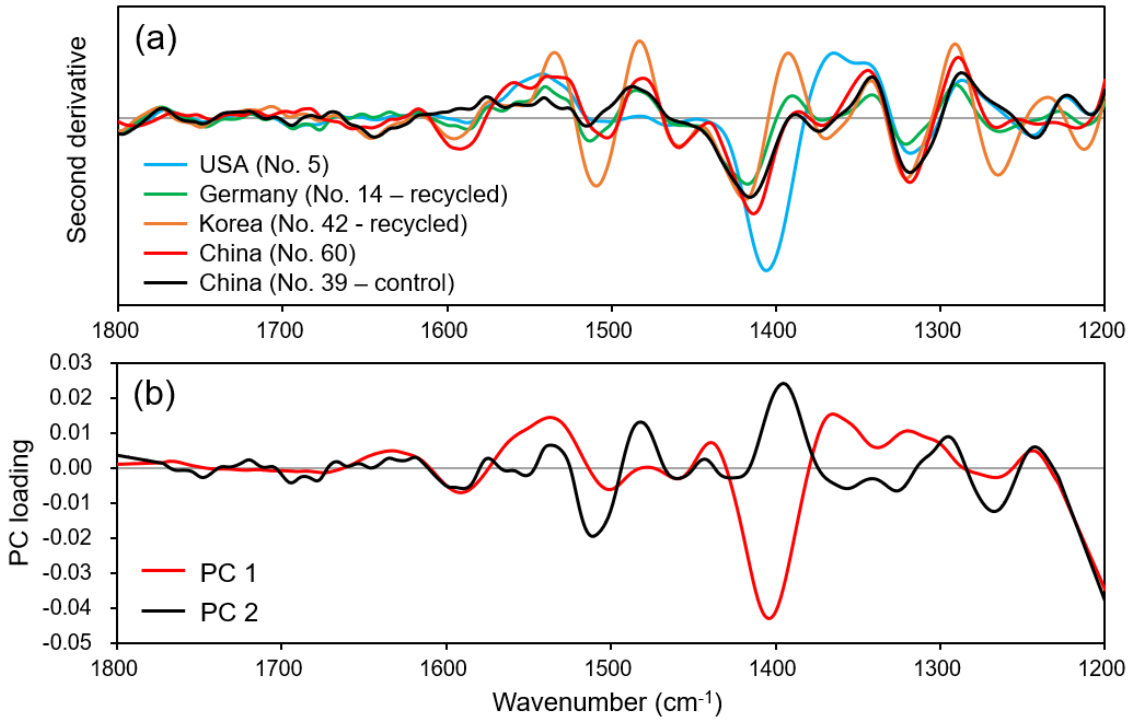

Fig. 5. Loadings for the first two PCs in the 1800 to 1200 cm^{-1} region of the second-derivative IR spectra. Numbers in parentheses indicate product number for printing paper samples.

Classification of Printing Papers

1D CNN models

The 1D CNN models trained on the IR spectra were constructed for the classification of printing paper samples. The learning curves of the models for classifying the country of manufacturing shown in Fig. 6 demonstrate a smaller loss difference between the training and validation curves in the selected range of 1800 to 1200 cm^{-1} (Fig. 6b), as opposed to the entire IR spectra (Fig. 6a). The variation between the training and validation curves can provide insights into model overfitting and generalization performance, with the smaller difference in the selected range suggesting that models trained on the IR data from this region are more likely to possess predictive capabilities for new data (Anzanello and Fogliatto 2011).

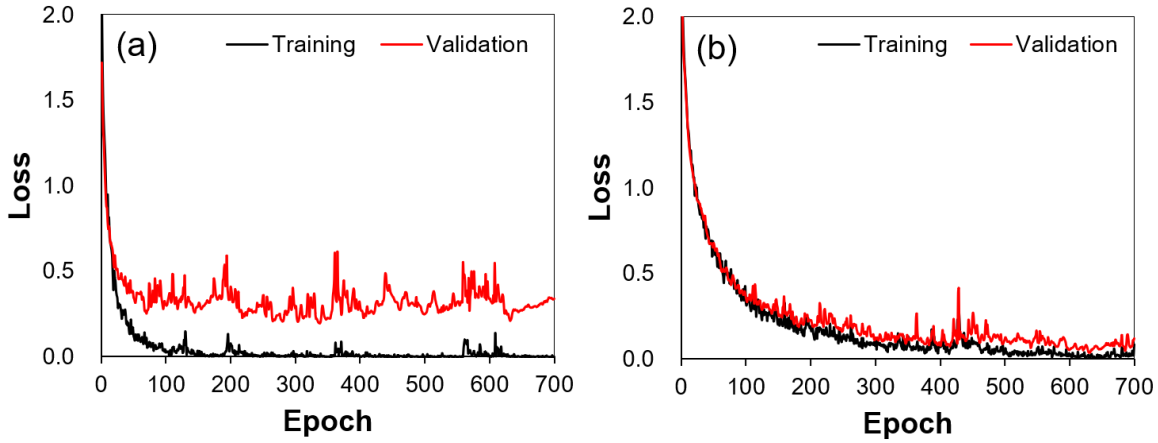

Fig. 6. Learning curves of the 1D CNNs for the training on IR spectral data and model validation for the classification of the country of manufacturing for printing paper samples. Learning curve for the entire IR data from 4000 to 400 cm^{-1} (a) and learning curve for selected data in the 1800 to 1200 cm^{-1} range (b)

Figure 7 presents the F1 scores for the 1D CNN models trained on the IR spectral data for the classification of printing paper samples. The hyperparameters applied to each model, determined through loop-based optimization, are detailed in Table 3. In the classification of the continent of manufacturing, all models exhibited F1 scores exceeding 0.967, confirming that printing paper samples share similar characteristics by continent. In the classification of the country of manufacturing, models trained on the IR data from the selected region of 1800 to 1200 cm^{-1} exhibited higher performance, with all classes perfectly classified regardless of spectral preprocessing.

The classification performance of the 1D CNN models at the product level was lower than that at other classification levels. Models trained on the original and second-derivative IR spectra encompassing the entire spectra (4000 to 400 cm^{-1}) showed the F1 scores of 0.788 and 0.818, respectively. These findings were attributed to the inherent challenges associated with product-level classification, which involves a considerable number of classes (65), where each class is represented by only 5 samples. This limitation prevented the models from adequately learning the distinctive features of each individual class. However, the use of spectral data from the selected region (1800 to 1200 cm^{-1}), notably improved the F1 scores, which reached 0.939 and 0.980, respectively. These results

suggest that the selected spectral region is well-suited for the characterization of printing paper samples.

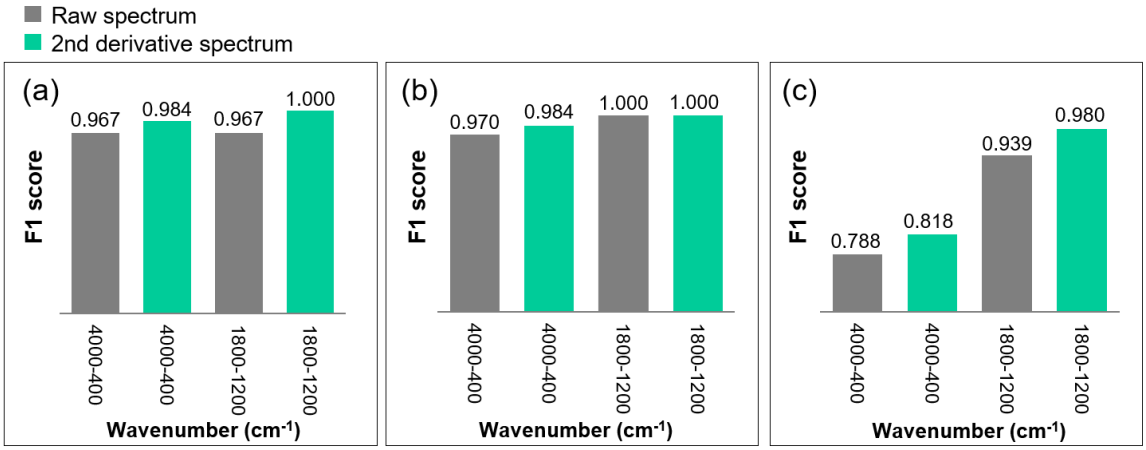

Fig. 7. Weighted F1 scores of the 1D CNN models for the classification of printing paper manufacturing continents (a), countries (b), and products (c)

Table 3. Performance of the 1D CNN Models in Printing Paper Classification and Their Optimal Hyperparameter Combinations

CLS Level	IR Spectrum		F1 Score	Hyperparameters						
	WN (cm^{-1})	Preproc.		Kernel Size	Filter	Pool Size	Dense Units	Dropout Rate	Learning Rate	Optimizer
Continent	4000–400	Original	0.967	3	16	2	128	0.5	0.0001	RMSProp
		Second deriv.	0.984	3	16	2	256	0.3	0.01	RMSProp
	1800–1200	Original	0.967	3	16	2	128	0.3	0.0001	RMSProp
		Second deriv.	1.000	3	16	2	128	0.5	0.0001	Adam
Country	4000–400	Original	0.970	7	32	4	128	0.3	0.0001	Adam
		Second deriv.	0.984	5	16	4	256	0.3	0.0001	RMSProp
	1800–1200	Original	1.000	7	16	2	128	0.5	0.001	RMSProp
		Second deriv.	1.000	3	16	2	128	0.5	0.0001	RMSProp
Product	4000–400	Original	0.788	3	64	2	256	0.5	0.0001	Adam
		Second deriv.	0.818	7	32	4	256	0.5	0.001	RMSProp
	1800–1200	Original	0.939	7	32	4	256	0.3	0.001	RMSProp
		Second deriv.	0.980	5	64	4	256	0.5	0.0001	RMSProp

Notes: CLS, classification; IR, infrared; WN, wavenumber; Preproc., preprocessing; Second deriv., second derivative; Adam, adaptive moment estimation; RMSProp, root mean squared propagation.

In the classification across all tested levels, the application of the second derivative and narrowing of the spectral region improved the classification performance of the 1D CNN models. These preprocessing techniques, particularly the narrowing of the spectral

region, improved the performance of the models in various classification problems for materials such as paper and wood. The effectiveness of these preprocessing techniques was reaffirmed in this study.

Misclassified class

In the classification of the continent of manufacturing, some samples from Europe were incorrectly classified as originating from Asia or North America. Regarding country classification, models trained on the original IR spectra misclassified certain samples from Brazil and the United States as originating from Austria and Finland, respectively. In the case of the second-derivative spectra, one sample from Korea was misclassified as originating from China. In the score plots (Fig. 4), misclassified samples are positioned close to the samples of the countries to which they were erroneously assigned, suggesting their spectral similarities.

Model Comparison

The classification performance of the 1D CNN models for printing paper classification was compared with that of the FNN and SVM models (Table 4). Across all tested classification levels and datasets, the 1D CNN models outperformed the reference models. Among the studied models, the SVM models exhibited the lowest classification performance. Although the FNN models showed performance close to that of the 1D CNN models in the continent-level classification, as the classification level became more granular, the performance gap between the two model types widened.

Table 4. Performance Comparison for the 1D CNN, FNN, and SVM Models in Printing Paper Classification

Classification Level	Wavenumber (cm^{-1})	Preprocessing	F1 Score		
			1D CNN	FNN	SVM
Continent	4000–400	Original	0.967	0.903	0.592
		Second derivative	0.984	0.923	0.592
	1800–1200	Original	0.967	0.949	0.934
		Second derivative	1.000	1.000	0.920
Country	4000–400	Original	0.970	0.837	0.268
		Second derivative	0.984	0.872	0.268
	1800–1200	Original	1.000	0.910	0.917
		Second derivative	1.000	0.985	0.888
Product	4000–400	Original	0.788	0.654	0.551
		Second derivative	0.818	0.616	0.507
	1800–1200	Original	0.939	0.813	0.788
		Second derivative	0.980	0.904	0.914

1D CNN, one-dimensional convolutional neural network; FNN, feed-forward neural network; SVM, support vector machine.

The second-derivative transformation and spectral region selection of the IR data also contributed to improve the classification performance of the FNN and SVM models. However, for some SVM models, the application of second-derivative preprocessing decreased F1 scores. This indicates that the optimal preprocessing technique varies among the models. These findings confirm the superior classification performance of the 1D CNN model for printing paper classification.

Prediction of Unknown Products

The established 1D CNN model was used to predict the country of manufacturing for products with undisclosed origins. Table 5 presents the predicted probabilities for these unknown products, derived from the softmax classifier within the 1D CNN. The model assigned an 80% probability to product 1 of Korean origin and predicted that unknown products 2 and 3 were of Chinese origin with probabilities of 69% and 45%, respectively.

Table 5. Predicted Probabilities for the Country-level Classification of Unknown Samples by the 1D CNN Model

Product	AUT	BRA	CAN	CHN	FIN	GER	IND	IDN	KOR	THA	USA
Unknown 1	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.80	0.00	0.00
Unknown 2	0.00	0.00	0.00	0.69	0.00	0.00	0.00	0.00	0.00	0.00	0.31
Unknown 3	0.00	0.01	0.03	0.45	0.00	0.00	0.00	0.27	0.01	0.02	0.21

Notes: AUT, Austria; BRA, Brazil; CAN, Canada; CHN, China; FIN, Finland; GER, Germany; IND, India; IDN, Indonesia; KOR, Korea; THA, Thailand; USA, United States; The 1D CNN model was trained with the second-derivative IR spectra in the 1800–1200 cm^{-1} region

Figure 8 shows unknown products on the PC score plot with other products used in model construction.

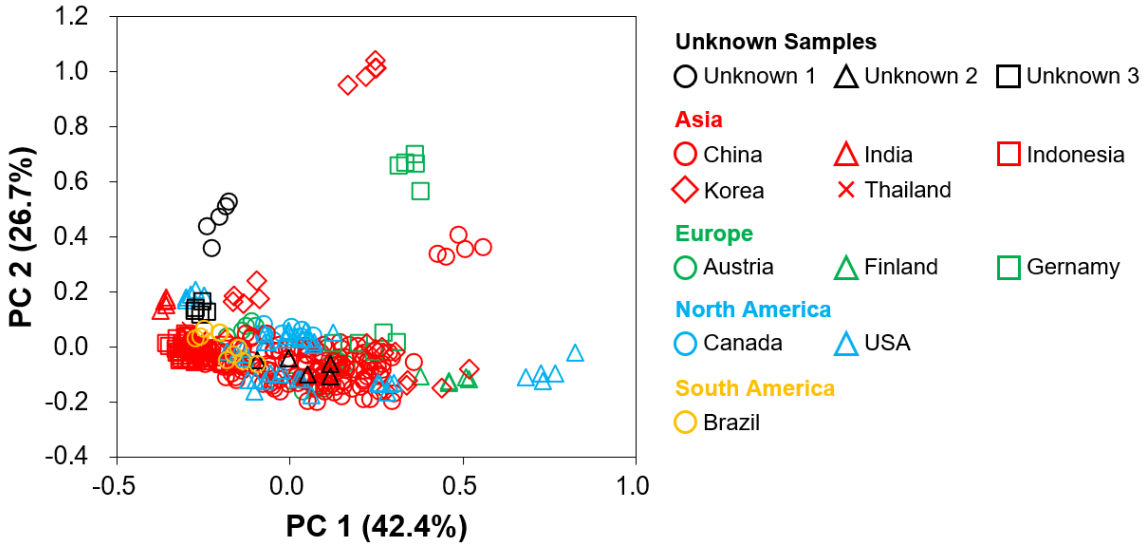

Fig. 8. PC score plot for two PCs in second-derivative IR spectra in the range of 4000 to 400 cm^{-1}

Unknown product 1, predicted to be of Korean origin, is positioned close to a specific Korean product on the score plot but forms an independent cluster. The presence of this cluster with a high PC2 value suggests the possibility that unbleached pulp may have been used in the manufacturing of this product.

Unknown product 2, predicted to be of Chinese origin, was positioned in a large cluster among the points of many other Chinese products. Although unknown product 3 was also predicted to be of Chinese origin, in the graph, it is close not only to Chinese products but also to products originating from the United States and India. Notably, the predicted probabilities for this product are relatively high for American and Indian origins, amounting to 21% and 27%, respectively (Table 6). Thus, the score plot provides interpretive support for the predictions generated by the 1D CNN model.

Data Point Attribution

Data point attributions were implemented using Grad-CAM from the 1D CNN models to analyze the differences in spectral regions contributing to the classification (Fig. 9). Figure 9a shows the second-derivative IR spectra of paper products from the United States and Korea in the range of 1800 to 1200 cm^{-1}, alongside their corresponding spectral point attributions.

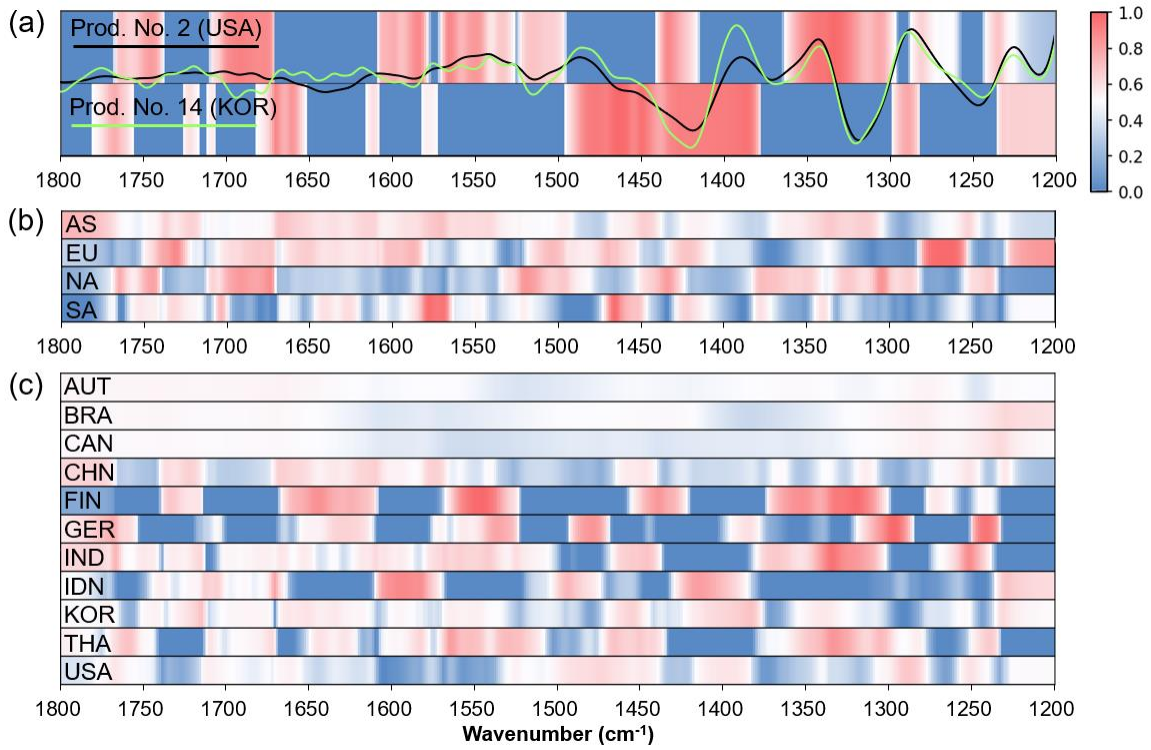

Fig. 9. Visualization of spectral point attribution for 1D CNN models trained with second-derivative IR spectra in the range of 1800 to 1200 cm^{-1} for classification at the product (a), manufacturing continent (b), and manufacturing country (c) levels. Notes: Prod., printing paper products; AS, Asia; EU, Europe; NA, North America; SA, South America; AUT, Austria; BRA, Brazil; CAN, Canada; CHN, China; FIN, Finland; GER, Germany; IND, India; IDN, Indonesia; KOR, Korea; THA, Thailand; USA, United States

In the classification of the American product, notable contributions were identified at 1685 and 1558 cm^{-1}, assigned to C=O stretching, and at 1430 to 1416 and 1315 cm^{-1}, ascribed to crystalline cellulose. Conversely, in the classification of the Korean product, a substantial contribution was observed in the 1375 to 1466 cm^{-1} region, assigned to CH_2 bending in crystalline and amorphous cellulose. These results highlight the variation in the spectral contributions to the classification of each product.

Figure 9b illustrates the spectral attribution for classification at the continent level, highlighting the variations in absorption bands contributing to classification across continents. For Asian products, relatively high contributions were observed in the broad region of 1500 to 1800 cm^{-1}, assigned to C=O stretching. Regarding European products, a notable contribution was observed near 1277 cm^{-1}, assigned to C–H deformation. Distinct regions were noted for North American products, with prominent color at 1740, 1685 to 1660, and 1315 cm^{-1}, assigned to C=O of carbonyl groups, C=O stretching, and crystalline cellulose, respectively. For South American products, high contributions are observed at 1560 and 1466 cm^{-1}, assigned to C=O and amorphous cellulose, respectively.

In the case of classification at the country level, the contributions of key IR absorption bands for paper exhibit diverse patterns among different countries (Fig. 9c). Furthermore, the spectral attribution in country-based classification differs from that in the continent classification. These findings emphasize that the IR absorption bands contributing to classification vary depending on the classification level, and they demonstrate that the 1D CNN model has been established by learning comprehensive information from critical spectral regions. Thus, spectral attribution analysis enhanced interpretability for printing paper classification, providing a comprehensive understanding of the impact of IR data on the classification process of the 1D CNN models.

CONCLUSIONS

1. One-dimensional convolutional neural network (CNN) models trained on the infrared (IR) spectral data of printing paper samples exhibited high performance in classifying the origin of printing paper, including the continent, country, and product name.
2. The preprocessing of the IR spectra through a second-derivative transformation improved the classification performance of 1D CNN models. In addition, narrowing the IR spectral data to the range of 1800 to 1200 cm^{-1} proved to be effective for enhancing the model performance.
3. The model trained with the second-derivative IR spectra in the range of 1800 to 1200 cm^{-1} achieved perfect classification for the continent and country of manufacturing, with an F1 score of 0.980 in product classification. The 1D CNN model outperformed the support vector machine (SVM) and feed-forward neural network (FNN) models trained on the same dataset.
4. Spectral point attribution using gradient-weighted activation mapping (Grad-CAM) demonstrated that the pattern of IR absorbance bands contributing to the decisions of the 1D CNN model varies depending on the classification level and provided insight into the classification decisions of the 1D CNN model.

ACKNOWLEDGEMENTS

This study was conducted with the support of the 'R&D Program for Forest Science Technology (Project No. FTIS-2019149C10-2323-0301)' provided by the Korea Forest Service (Korea Forestry Promotion Institute) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-RS-2023-00246356).

REFERENCES CITED

Anzanello, M. J., and Fogliatto, F. S. (2011). "Learning curve models and applications: Literature review and research directions," *International Journal of Industrial Ergonomics* 41(5), 573-583. DOI: 10.1016/j.ergon.2011.05.001

Australian Government. (2012). "Illegal Logging Prohibition Regulation 2012," Federal Register of Legislation, Australian Government, (https://www.legislation.gov.au/Details/F2022C00209), Accessed 30 Oct 2023.

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

Chang, Y. S., Yang, S. Y., Chung, H., Kang, K. Y., Choi, J. W., Choi, I. G., and Yeo, H. (2015). "Development of moisture content prediction model for Larix kaempferi sawdust using near infrared spectroscopy," *Journal of the Korean Wood Science and Technology* 43(3), 304-310. DOI: 10.5658/WOOD.2015.43.3.304

Chen, Z., Zhai, R., Cai, Y., Ye, Y., Sun, Z., and Liu, Y. (2023). "Online source tracing of waste paper by smoke based on laser-induced breakdown spectroscopy," *Journal of Laser Applications* 35, article 042031. DOI: 10.2351/7.0001226

Colom, X., and Carrillo, F. (2002). "Crystallinity changes in lyocell and viscose-type fibres by caustic treatment," *European Polymer Journal* 38(11), 2225-2230. DOI: 10.1016/S0014-3057(02)00132-5

Coppola, F., Frigau, L., Markelj, J., Malešič, J., Conversano, C., and Strlič, M. (2023). "Near-infrared spectroscopy and machine learning for accurate dating of historical books." *Journal of the American Chemical Society* 145, 12305-12314. DOI: 10.1021/jacs.3c02835

Delmotte, L., Ganne-Chedeville, C., Leban, J. M., Pizzi, A., and Pichelin, F. (2008). "CP-MAS 13C NMR and FT-IR investigation of the degradation reactions of polymer constituents in wood welding," *Polymer Degradation and Stability* 93(2), 406-412. DOI: 10.1016/j.polymdegradstab.2007.11.020

European Commission. (2023). "EU Timber Regulation," *European Commission*, (https://environment.ec.europa.eu/topics/forests/deforestation/illegal-logging/timber-regulation_en), Accessed 20 Oct 2023.

Federal Register. (2021). "Implementation of Revised Lacey Act Provisions," *The Daily Jounal of the United States Government, Federal Register*, (https://www.federalregister.gov/documents/2021/07/02/2021-14155/implementation-of-revised-lacey-act-provisions), Accessed 20 Oct 2023.

Feng, J., Wan, E., Han, B., Chen, Z., Liu, X., and Liu, Y. (2023). "Research on identification of ink marks based on machine learning and laser-induced breakdown

spectroscopy," *Journal of Laser Applications* 35, article 012017. DOI: 10.2351/7.0000895

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

Gümüskaya, E., Usta, M., and Kirci, H. (2003). "The effects of various pulping conditions on crystalline structure of cellulose in cotton linters," *Polymer Degradation and Stability* 81(3), 559-564. DOI: 10.1016/S0141-3910(03)00157-5

Hajji, L., Boukir, A., Assouik, J., Pessanha, S., Figueirinhas, J. L., and Carvalho, M. L. (2016). "Artificial aging paper to assess long-term effects of conservative treatment. Monitoring by infrared spectroscopy (ATR-FTIR), X-ray diffraction (XRD), and energy dispersive X-ray fluorescence (EDXRF)," *Microchemical Journal* 124, 646-656. DOI: 10.1016/j.microc.2015.10.015

Hofstetter, K., Hinterstoisser, B., and Salmén, L. (2006). "Moisture uptake in native cellulose–the roles of different hydrogen bonds: A dynamic FT-IR study using Deuterium exchange," *Cellulose* 13, 131-145. DOI: 10.1007/s10570-006-9055-2

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*," *Applied Biochemistry and Biotechnology* 188, 1066-1076. DOI: 10.1007/s12010-019-02965-8

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., Hwang, U. T., Jo, K., Lee, T., Park, J., Kim, J. C., Kwak, H. W., Choi, I. G., and Yeo, H. (2021). "NIR-chemometric approaches for evaluating carbonization characteristics of hydrothermally carbonized lignin," *Scientific Reports* 11(1), article 16979. DOI: 10.1038/s41598-021-96461-x

Hwang, S. W., Chung, H., Lee, T., Kim, J., Kim, Y., Kim, J. C., Kwak, H. W., Choi, I. G., and Yeo, H. (2023). "Feature importance measures from random forest regressor using near-infrared spectra for predicting carbonization characteristics of kraft lignin-derived hydrochar," *Journal of Wood Science* 69(1), 1. DOI: 10.1186/s10086-022-02073-y

Kim, K. J., and Eom, T. J. (2016). "Study on the aging characteristics of paper with principal component analysis," *Journal of Korea TAPPI* 48(6), 144-149. DOI: 10.7584/JKTAPPI.2016.12.48.6.144

Korea Legislation Research Institute. (2020) "Act on the sustainable use of timbers," *Korea Law Translation Center, Korea Legislation Research Institute,* (https://elaw.klri.re.kr/kor_service/lawView.do?hseq=53257&lang=ENG), Accessed 20 Oct 2023.

Meza Ramirez, C. A., Greenop, M., Ashton, L., and Rehman, I. U. (2021). "Applications of machine learning in spectroscopy," *Applied Spectroscopy Reviews* 56, 733-763. DOI: 10.1080/05704928.2020.1859525

Olsson, A. M., and Salmén, L. (2004). "The association of water to cellulose and hemicellulose in paper examined by FTIR spectroscopy," *Carbohydrate Research* 339(4), 813-818. DOI: 10.1016/j.carres.2004.01.005

Pandey, K. K., and Pitman, A. J. (2003). "FTIR studies of the changes in wood chemistry following decay by brown-rot and white-rot fungi," *International Biodeterioration & Biodegradation* 52(3), 151-160. DOI: 10.1016/S0964-8305(03)00052-0

Savitzky, A., and Golay, M. J. (1964). "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry* 36(8), 1627-1639.

Schwanninger, M. J. C. R., Rodrigues, J. C., Pereira, H., and Hinterstoisser, B. (2004). "Effects of short-time vibratory ball milling on the shape of FT-IR spectra of wood and cellulose," *Vibrational Spectroscopy* 36(1), 23-40. DOI: 10.1016/j.vibspec.2004.02.003

Shah, P., Choi, H. K., and Kwon, J. S. I. (2023). "Achieving optimal paper properties: A layered multiscale kMC and LSTM-ANN-based control approach for kraft pulping," *Processes* 11(3), 809. DOI: 10.3390/pr11030809

Sheikhi, P., Talaeipour, M., Hemasi, A. H., Eslam, H. K., and Gumuskaya, E. (2010). "Effect of drying and chemical treatment on bagasse soda pulp properties during recycling," *BioResources* 5(3), 1702-1716.

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., and McLaughlin, M. J. (2014). "The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties," *Applied Spectroscopy Reviews* 49(2), 139-186. DOI: 10.1080/05704928.2013.811081

Stuart, B. H. (2004). *Infrared Spectroscopy: Fundamentals and Applications*, John Wiley & Sons, Hoboken, NJ.

Trafela, T., Strlic, M., Kolar, J., Lichtblau, D. A., Anders, M., Mencigar, D. P., and Pihlar, B. (2007). "Nondestructive analysis and dating of historical paper based on IR spectroscopy and chemometric data evaluation," *Analytical Chemistry* 79(16), 6319-6323. DOI: 10.1021/ac070392t

Vert, J. P., Tsuda, K., and Schölkopf, B. (2004). "A primer on kernel methods," in: *Kernel Methods in Computational Biology*, B. Schölkopf, J. P. Vert, and K. Tsuda (eds.), MIT Press, Cambridge, MA.

Xiao, S., Gao, R., Lu, Y., Li, J., and Sun, Q. (2015). "Fabrication and characterization of nanofibrillated cellulose and its aerogels from natural pine needles," *Carbohydrate Polymers* 119, 202-209. DOI: 10.1016/j.carbpol.2014.11.041

Article submitted: October 23, 2023; Peer review completed: November 11, 2023; Revised version received and accepted: November 20, 2023; Published: January 24, 2024.
DOI: 10.15376/biores.19.1.1633-1651