Petroleum Science 22 (2025) 151–162

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

Effect of preprocessing on performances of machine learning-based mineral composition analysis on gas hydrate sediments, Ulleung Basin, East Sea

Hongkeun Jin [a, b, 1], Ju Young Park [a, f, 1], Sun Young Park [c], Byeong-Kook Son [c], Baehyun Min [d, e], Kyungbook Lee [a, f, *]

[a] Department of Geoenvironmental Sciences, Kongju National University, Gongju-si, Chungcheongnam-Do, 32588, Republic of Korea
[b] Tunnel & Underground Facility Division, ESCO Consultant & Engineers Company, Simin-Daero, Dongan-Gu, Anyang-Si, Gyeonggi-Do, 14057, Republic of Korea
[c] Petroleum and Marine Research Division, Korea Institute of Geoscience and Mineral Resources, 124, Gwahak-ro, Yuseong-gu, Daejeon, 34132, Republic of Korea
[d] Department of Climate and Energy Systems Engineering, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Republic of Korea
[e] Center for Climate/Environment Change Prediction Research, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Republic of Korea
[f] Yellow Sea Institute of Geoenvironmental Sciences, Gongju-si, Chungcheongnam-Do, 32588, Republic of Korea

ARTICLE INFO

Article history:
Received 27 September 2023
Received in revised form
30 October 2024
Accepted 18 November 2024
Available online 20 November 2024

Edited by Teng Zhu

Keywords:
Sample-based preprocessing
X-ray diffraction (XRD)
Machine learning
Mineral composition
Gas hydrate (GH)
Ulleung basin

ABSTRACT

Gas hydrate (GH) is an unconventional resource estimated at 1000–120,000 trillion m^3 worldwide. Research on GH is ongoing to determine its geological and flow characteristics for commercial production. After two large-scale drilling expeditions to study the GH-bearing zone in the Ulleung Basin, the mineral composition of 488 sediment samples was analyzed using X-ray diffraction (XRD). Because the analysis is costly and dependent on experts, a machine learning model was developed to predict the mineral composition using XRD intensity profiles as input data. However, the model's performance was limited because of improper preprocessing of the intensity profile. Because preprocessing was applied to each feature, the intensity trend was not preserved even though this factor is the most important when analyzing mineral composition. In this study, the profile was preprocessed for each sample using min-max scaling because relative intensity is critical for mineral analysis. For 49 test data among the 488 data, the convolutional neural network (CNN) model improved the average absolute error and coefficient of determination by 41% and 46%, respectively, than those of CNN model with feature-based preprocessing. This study confirms that combining preprocessing for each sample with CNN is the most efficient approach for analyzing XRD data. The developed model can be used for the compositional analysis of sediment samples from the Ulleung Basin and the Korea Plateau. In addition, the overall procedure can be applied to any XRD data of sediments worldwide.

1. Introduction

Gas hydrate (GH) is an ice-like solid compound formed by combining natural gas with water under a low temperature (<300 K) and high pressure (> 0.6 MPa) (Shahnazar and Hasan, 2014). Unlike conventional oil or natural gas, GH exists in relatively shallow (100–1100 m) sedimentary layers (Li et al., 2021). Research on GH as a next-generation alternative energy source has been conducted worldwide because its estimated reserves are approximately 1000–120,000 trillion m^3, and the amount of carbon dioxide emitted during combustion is 0.7 times that of gasoline (Kvenvolden et al., 1993; Ji et al., 2022).

GH is widely found in seafloor sediments. These sediments are composed of several minerals, which not only determine the characteristics of the sediments but also provide information on the local environment and origin of the sediments (McLennan et al.,

* Corresponding author.
E-mail address: kblee@kongju.ac.kr (K. Lee).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.petsci.2024.11.012

1993; Zhang et al., 2019; Wang et al., 2020). For example, determining the presence or absence of GH in reservoirs is possible by using sediment analysis like total organic carbon, X-ray fluorescence and X-ray diffraction (XRD).

When the particle size of a mineral is relatively large (e.g., 5 mm or more), its composition can be determined by modal analysis with the naked eye or a microscope. However, the composition and content of fine-grained particle sediments can only be determined using XRD analysis. XRD measures the intensity of diffracted X-rays by injecting them into powdered samples of 10 μm or less. Because each mineral has a unique intensity according to its crystal structure, it is easy to figure out the composition which composed of few minerals. In the case of natural sediment samples, the intensity profile is complicated because of various minerals. This difference increases the time required for analysis and the reliance on experts for interpretation when analyzing hundreds of samples.

Due to such complexity and the long analysis time, the development of an XRD composition analysis model incorporating machine learning has progressed. The machine learning model reduces analysis time by providing an initial solution for composition analysis to experts or allows non-experts to roughly understand the mineral composition of the sediment.

Domínguez-Olmedo et al. (2020) grouped minerals with the same properties to reduce ambiguity in the optical identification of opaque minerals based on a decision tree. Zeng et al. (2021) proposed a mineral identification method using the Mohs hardness scale and mineral images as input data to improve the traditional mineral identification method that relies on the analyst's experience. A deep convolutional neural network (CNN) has been used for image feature extraction. Okada et al. (2020) proposed an automatic mineral identification system to identify the type of mineral before processing it from the ore by combining high-resolution spectral data and a CNN. This nondestructive approach determines the type and crystal structure of the minerals contained in the rock. Kim et al. (2020) proposed a deep neural network model that predicts the weight ratio of minerals from logging data (i.e., neutron porosity, bulk density, sonic, and gamma ray logs) and core data (XRD analysis results) in the Canning Basin. Preprocessing was performed using principal component analysis to augment insufficient logging and mineral composition data. Dong et al. (2022) proposed DeepXRD, a deep learning algorithm that predicts XRD spectra by considering only the mineral composition. A study was conducted to build a machine learning model that predicts the mineral composition of a mixture by generating approximately 1.78 million theoretical intensity profiles of that mixture (Lee et al., 2020). In addition, research was conducted to generate XRD intensity profiles with various crystallographic characteristics for machine learning (Schuetzke et al., 2021). Lee et al. (2020) developed a compositional analysis model for a relatively simple mixture of 3–4 minerals using theoretically generated experimental data rather than natural rock samples.

In the Ulleung Basin (UB), two large-scale drilling expeditions were conducted: in September 2007 (UBGH-1) to determine the GH potential and in July 2010 (UBGH-2) to evaluate the GH resource volume (Ryu et al., 2013). The application of machine learning to approximately 500 actual sediment samples obtained from the UB was conducted by Park et al. (2022). Although the sample was composed of 12 minerals, including amorphous minerals with complex intensity profiles, confirming the possibility of automating composition analysis using machine learning was meaningful.

However, the model developed by Park et al. (2022) has limited accuracy for samples with abnormal mineral compositions. This limitation is caused by the general preprocessing method in machine learning and the normalization of each feature. Moreover, it does not consider the characteristics of the XRD experimental data. Owing to the nature of the XRD experimental data, even if the mineral composition is similar, there is a slight difference in the incident angle at which the intensity peak appears. Moreover, the scale differs depending on the experiment's conditions. In interpreting mineral composition, the relative ratio of the intensity profile to the overall trend is much more crucial than the exact angle and absolute size of the intensity peak. Because of these characteristics, if a preprocessing method is applied to each feature, the pattern of the intensity profile is damaged, making it difficult to interpret its mineral composition.

The purpose of this study was to improve the precision of a composition analysis model by applying a reasonable preprocessing method to XRD experimental data using domain knowledge. In the proposed method, min-max normalization is applied to each sample, not each feature. Therefore, the absolute scale was calibrated between 0 and 1, and the XRD profile shape was preserved. In addition, a suitable machine learning algorithm with newly preprocessed data was proposed by comparing several machine learning algorithms.

Section 2 presents the available XRD experimental data for the UB and the proposed preprocessing method. Section 3 conducts a prediction performance analysis based on a machine learning algorithm and a data preprocessing method.

2. Methodology

2.1. XRD experiment data

In total, 488 core samples were obtained from three boreholes (1-4, 1-9, and 1-10B) of UBGH-1 and five boreholes (2-1-1, 2-2-2, 2-5, 2-6, and 2-10) of UBGH-2 (Fig. 1). Table 1 lists the number and depths of the samples.

The 2θ values of the XRD intensity profile of the five boreholes (1-4, 1-9, 1-10B, 2-1-1, and 2-6) ranged from 3.01° to 64.99° in 0.02° increments (3100 in total) during the experiment (Park et al., 2022). The range of 2θ values of the XRD intensity profiles of the three boreholes (2-2-2, 2-5, and 2-10) ranged from 3.005° to 64.995° with an interval of 0.01° (6200 in total). Therefore, 6200 intensities were matched to 3100 intensities using linear interpolation.

XRD was performed using a PHILPS X'Pert MPD XRD instrument at the Korea Institute of Geoscience and Mineral Resources. The intensity profile was measured in counts per second (cps). The detailed XRD process is described by Park et al. (2022). The intensity profiles obtained from each sample were analyzed using the Rietveld method in SIROQUANT™ software (Rietveld, 1969; Taylor, 1991; Taylor and Matulis, 1994; Sietronics, 1996; Park et al., 2022).

In the SIROQUANT™ software, the agreement between the experimental (original) and the software-calculated profiles is represented by the chi-square (x^2). A x^2 value is used to determine whether a significant relationship exists between the experimental and calculated profiles (Ainane et al., 2021). A x^2 value close to 1 indicates reasonable match, but values below 3 are considered to have acceptable reliability (Sietronics, 1996). The mineral composition analysis data used in this study were obtained when the x^2 value was below 3.

Fig. 2 shows the mineral compositions of the 488 samples. The top and bottom boundaries of the box indicate the 3rd and 1st quartiles of the total data, respectively, and the orange line indicates the median. The open circles signs indicate what is outside the 1.5 × interquartile range from the quartile boundaries.

Table 2 lists the basic statistical values of the 488 datasets. Opal-A was identified as the main mineral, with an average composition of 30.1% and a wide distribution from 0% to 73.7%. Quartz had the second highest average composition at 18.8%, and some data showed high compositions from 30% to 55.9% (Fig. 2). Although

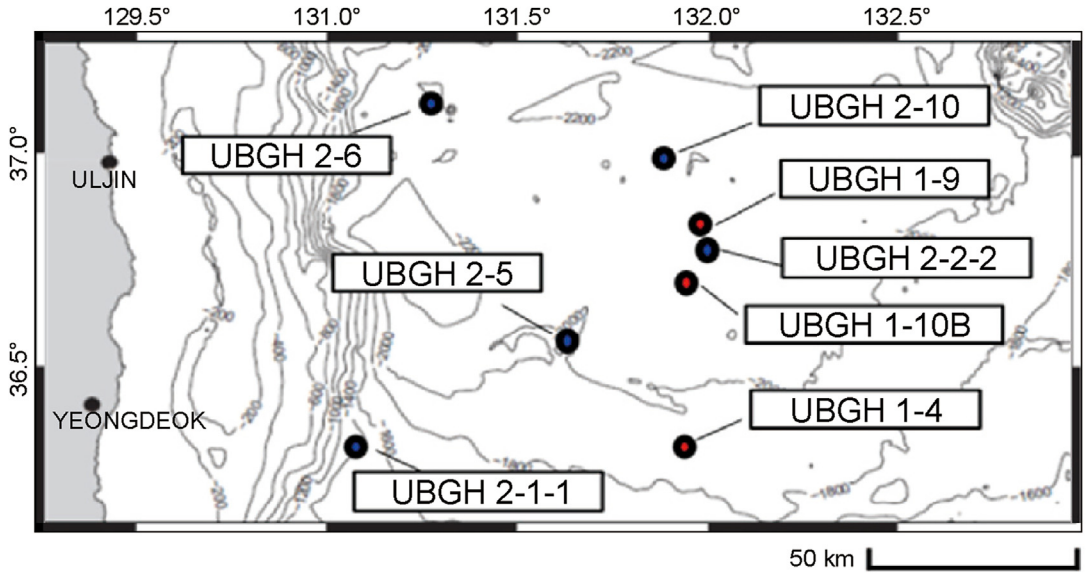

Fig. 1. Location of UBGH-1 (1-4, 1-9, and 1-10B) and UBGH-2 (2-1-1, 2-2-2, 2-5, 2-6, and 2-10) wellbores (modified from Park et al. (2022)).

Table 1
Number of samples for the 8 boreholes of UBGH-1 and UBGH-2 (modified from Park et al. (2022)).

	UBGH-1 (94)			UBGH-2 (394)				
Boreholes name	1-4	1-9	1-10B	2-1-1	2-2-2	2-5	2-6	2-10
Number of samples	17	40	37	91	55	87	87	74
Depth of the topmost sample, mbsf	0.0	0.1	0.5	2.6	0.5	1.7	0.3	1.0
Depth of bottommost sample, mbsf	186.5	174.5	204.8	219.4	178.9	217.4	227.0	204.9
Water depth, m	1841.4	2099.1	2077.0	1534.4	2097.9	1973.8	2156.9	2148.0

* mbsf: meter below sea floor.

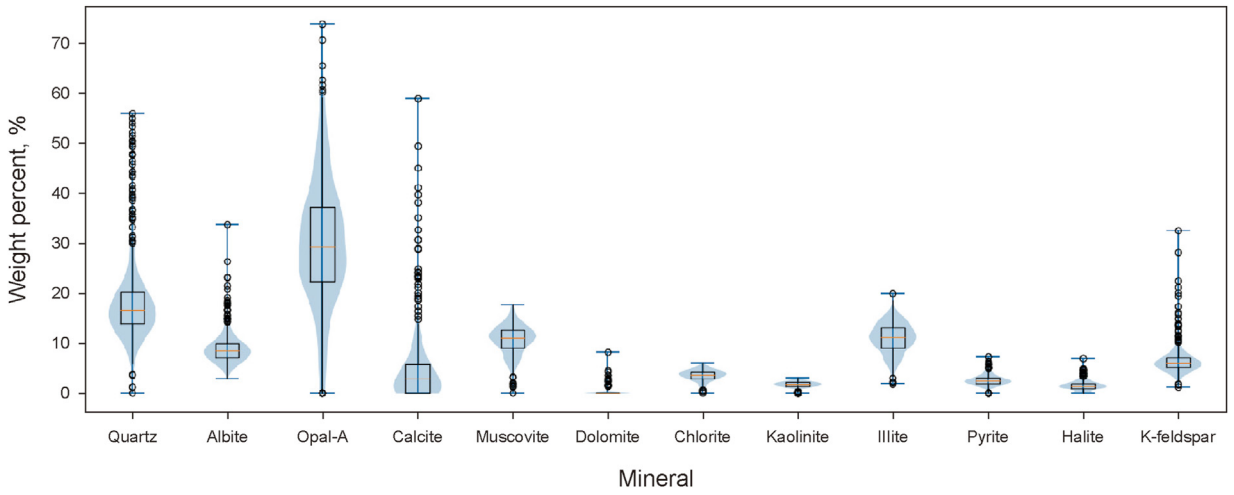

Fig. 2. Mineral composition of samples from XRD experiment analysis.

albite, K-feldspar, and calcite had average compositions of less than 10%, few data points showed extreme outlier values greater than 30%. Albite and K-feldspar had several outliers with distributions of 15–33.7% and 10–32.5%, respectively. In the case of calcite, although 165 data points had 0% composition, few data points were high (58.9 %). The dolomite content was 0% in most samples except for 19 data points. The halite was also observed in all samples, originating from seawater. This is because halite precipitated as the seawater in the core evaporated during the process of drying the core samples for XRD experiment.

2.2. Data preprocessing

When analyzing the XRD results, the absolute value and exact angles of the intensity peak are not essential. Instead, the ratio and shape of the intensities are important for estimating mineral composition. Table 3 indicates that samples #85 and #269 have similar mineral compositions and that the quartz composition is

Table 2
Maximum, minimum, average, and standard deviation values of mineral composition for total data and test dataset (Park et al., 2022).

(Unit: %)

Mineral	Total data				Test data			
	Max.	Min.	Ave.	Std.	Max.	Min.	Ave.	Std.
Quartz	55.90	0.00	18.77	8.96	54.10	8.10	19.06	9.65
Albite	33.70	2.90	8.96	3.19	33.70	3.30	9.94	5.39
Opal-A	73.70	0.00	30.11	12.57	57.50	0.00	29.21	12.15
Calcite	58.90	0.00	4.70	7.20	45.00	0.00	4.48	7.20
Muscovite	17.60	0.00	10.64	2.96	17.60	4.40	10.81	2.49
Dolomite	8.10	0.00	0.11	0.62	1.50	0.00	0.03	0.21
Chlorite	6.00	0.10	3.57	1.06	6.00	1.10	3.57	1.07
Kaolinite	3.10	0.00	1.69	0.63	3.10	0.00	1.67	0.63
Illite	19.90	1.90	10.96	3.01	15.80	2.20	10.56	2.80
Pyrite	7.20	0.00	2.49	1.11	5.70	0.00	2.68	1.26
Halite	7.00	0.00	1.42	0.95	3.00	0.00	1.39	0.87
K-feldspar	32.50	1.20	6.56	2.92	11.40	3.10	6.60	1.89

Table 3
Mineral composition of samples #85 and #269.

(Unit: %)

Sample	Quartz	Albite	Opal-A	Calcite	Muscovite	Dolomite
#85	16.6	8.2	32.6	0	13.3	0
#269	16.6	7.8	31.8	0	13.5	0
Sample	**Chlorite**	**Kaolinite**	**Illite**	**Pyrite**	**Halite**	**K-feldspar**
#85	3.6	2.2	12.6	2.8	1.2	6.9
#269	4.4	2.1	14.5	2	1.4	5.8

Table 4
Maximum intensity peak values and their angles of samples #85 and #269.

Sample	Intensity peak angle	
	26.57°	26.59°
#85	799.0	771.0
#269	208.6	234.1

the same (16.6%). However, Fig. 3 indicates that the intensity profiles of the two samples differ. Specifically, the intensity peak for sample #85 is 799.0 cps at 26.57° and 234.1 cps for sample #269 at 26.59° (Table 4). This is because the intensity profiles can be affected by various factors like particle arrangement and size, atomic scattering factor, and multiplicity factor (Park et al., 2022). Applying these intensity profiles to machine learning model without normalization negatively affects the model's performance. Therefore, process of data preprocessing using normalization is necessary before training the model.

However, the data preprocessing of Park et al. (2022) ignored the characteristics of the XRD data. Fig. 4 illustrates an example of preprocessing applied to the intensity profile. The original intensity profile in Fig. 4(a) was preprocessed for each feature (Fig. 4(b)) in the previous method and for each sample (Fig. 4(c)) in the proposed method. In general, normalization is applied to each feature (i.e., each 2θ angle) to overcome the scaling issue in machine learning. Park et al. (2022) used this approach. However, owing to variations in the peak position and scale for each XRD intensity profile (Fig. 3), the shape of the normalized intensity failed to preserve the trend of the original data, as shown in Fig. 4b, which is the most important factor in interpreting mineral composition.

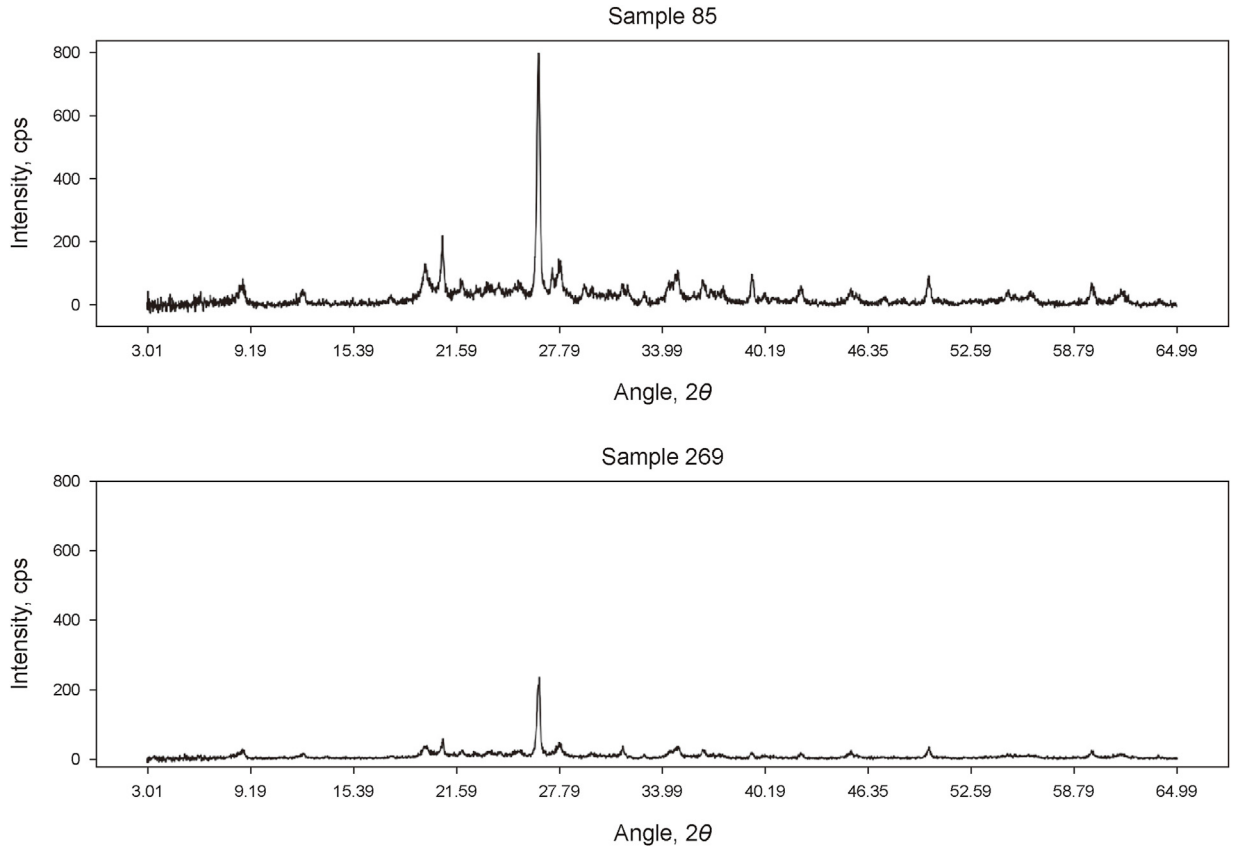

Fig. 3. Intensity profile of samples #85 and #269.

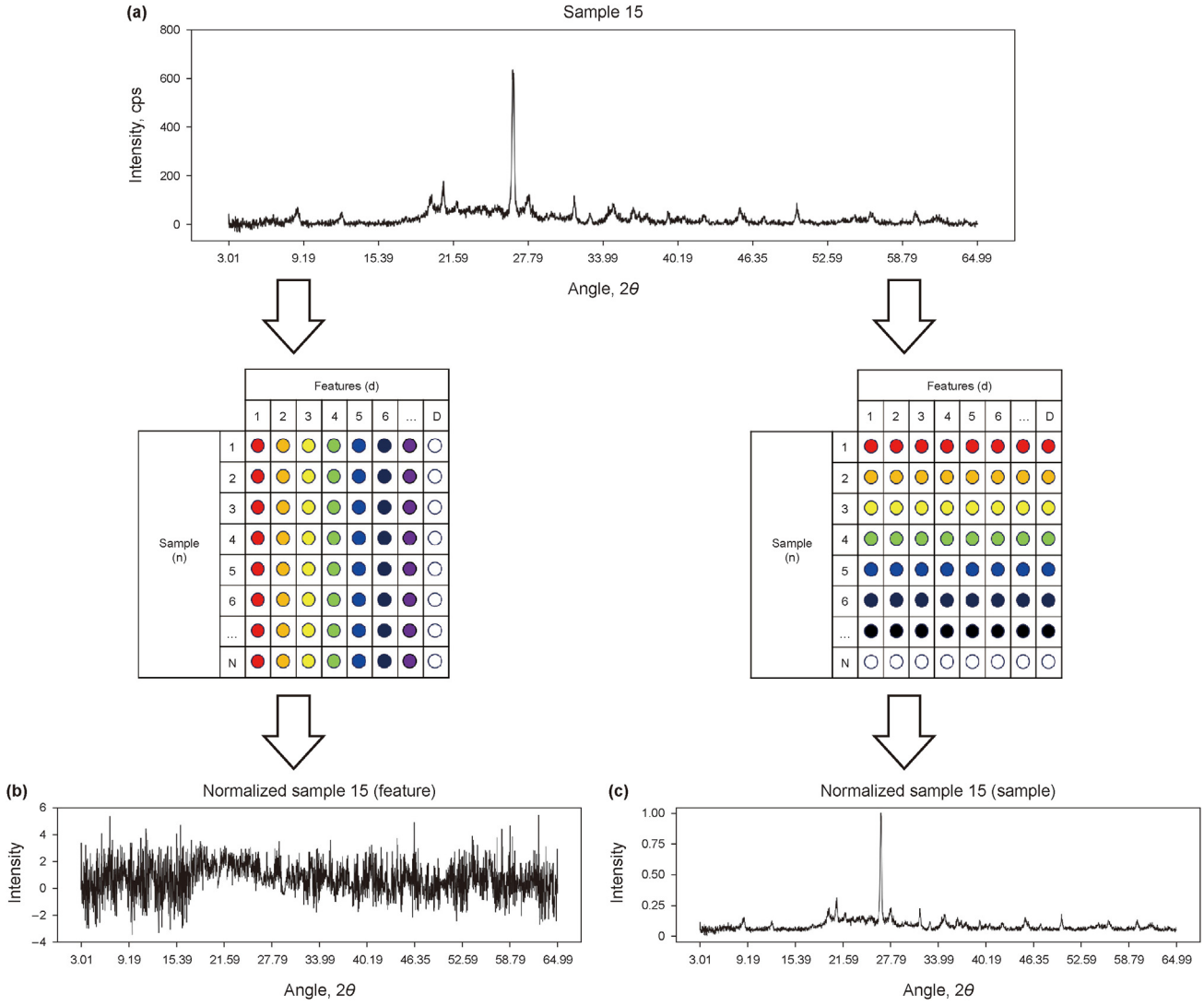

Fig. 4. Intensity profile of sample #15: (**a**) raw profile and normalized profiles (**b**) by each feature and (**c**) by each sample.

To preserve the relative scale and shape of the intensity peaks, a min-max scaler was applied to each sample (Eq. (1)). This preprocessing approach helps alleviate the effects of scale and preserve the characteristics of the XRD experimental data. Fig. 4(c) shows an example of the application of the min-max scaler to the samples. The shape of the raw data is retained, and the scale is corrected to a range of 0–1. This result indicates that all 488 samples have the same range.

$$Intensity\_scaled^{j}{}_{i} = \frac{Intensity^{j}_{i} - Intensity^{j}{}_{\text{minimum}}}{Intensity^{j}{}_{\text{maximum}} - Intensity^{j}{}_{\text{minimum}}} \quad (1)$$

$$i = 1, \ldots, 3100,\ j = 1, \ldots, 488$$

where subscripts i, maximum, and minimum represent the i-th, maximum, and minimum intensities of the j-th data, respectively. Superscript j represents the j-th data point among the 488 data points.

When preprocessing for machine learning, test data are not used to calculate the normalization factor because they are unseen data. Thus, the mean and standard deviation parameters from the training data are applied to the test data. However, sharing these parameters is unnecessary because normalization is performed within each sample.

Data splitting is a crucial step in machine learning that involves dividing the data into training data for model learning, validation data to prevent overfitting, and test data for verification. Fig. 2 shows that the mineral composition in the 488 data points was not evenly distributed but concentrated within a narrow range. Park et al. (2022) randomly split the total of data (488) into a 9:1 ratio for a train dataset (439) and test dataset (49). Then, the train dataset was split into an 8:2 ratio for train (307) and validation (132). If data split is not performed properly, accurately assessing the performance of the developed machine learning model can be difficult. In the literature, the statistical values of the training and test mineral data were compared to ensure unbiased data split (Table 2). In this study, the same data were used to determine the effects of the preprocessing method.

2.3. Machine learning process

The workflow of machine learning is shown in Fig. 5. Data preprocessing is shown in the blue box, and model training is shown in the green box. The procedure is the same as that in Park

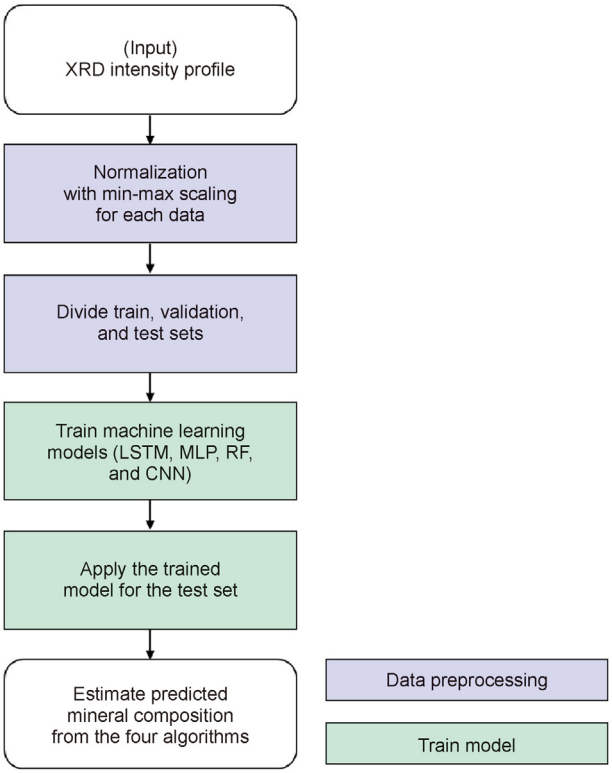

Fig. 5. Workflow for predicting mineral composition using machine learning algorithms.

et al. (2022), except for the normalization. Because feature-based preprocessing is not suitable for XRD data owing to the unique characteristics discussed in Subsection 2.2, sample-based preprocessing was used (Fig. 4).

After preprocessing, the data were divided into training, validation, and test sets. The training dataset was then applied to long short-term memory (LSTM), multilayer perceptron (MLP), random forest (RF), and CNN. The learning mechanism and structure of the models were the same as those in Park et al. (2022). Details are presented in Tables A1, A2, and A3 in Appendix A.

3. Results

3.1. Prediction results for the 49 test data using machine learning

The preprocessed data were utilized to train various machine learning algorithms, namely, LSTM, MLP, RF, and CNN. Table 5 displays the average and standard deviation of the mean absolute error (MAE) between the predicted and labeled mineral compositions for the 49 test data points, as given by Eq. (2).

Table 5
Average and standard deviation of error between original and predicted mineral compositions of the 49 test data by the two preprocessing methods and four algorithms.

Preprocess method	Algorithm	Ave. MAE	(Unit: %) Std. MAE
For each sample	**LSTM**	2.11	1.14
	MLP	0.87	0.80
	RF	1.31	1.43
	CNN	0.68	0.45
For each feature	**LSTM**	2.53	1.13
	MLP	1.03	0.72
	RF	1.35	1.72
	CNN	1.16	0.88

$$\text{Ave. } MAE\ [\%] = \frac{1}{N_{te}} \frac{1}{N_{mineral}} \sum_{i=1}^{N_{te}} \sum_{j=1}^{N_{mineral}} \left| y_i^j - \hat{y}_i^j \right| \tag{2}$$

$$i = 1, \ldots, 49,\ j = 1, \ldots, 12$$

where N_{te} and $N_{mineral}$ are the number of test data points (49) and minerals (12), respectively. y_i^j and $\hat{y}_i^j$ represent the true and predicted compositions (%) of the j-th mineral of i-th test data, respectively.

Table 5 illustrates that when sample-based preprocessing is applied to the CNN, both the MAE and its standard deviation are the lowest among all cases. A CNN with the sample-based preprocessing resulted in a 41% improvement in MAE compared with that of a CNN with the feature-based preprocessing. A detailed analysis was conducted to examine the impact of the preprocessing method on CNN performance. The results for LSTM, MLP, and RF are detailed in Appendices B and C.

The scatter plot and coefficient of determination (R^2) of the predicted mineral compositions for the 49 test datasets using the CNN are presented in Fig. 6. The scatter plot and the calculation of R^2 were conducted after normalizing each mineral because the mineral composition distribution varied with each mineral. Mineral normalization was calculated using Eq. (3), and R^2 was calculated using Eq. (4).

$$Mineral\_scaled^j{}_i = \frac{Mineral\ composition_i^j - \mu_j}{\sigma_j} \tag{3}$$

$$\mu_j = \frac{1}{N_{tr} + N_{val}} \sum_{k=1}^{N_{tr}+N_{val}} mineral\ compostion_k^j, i = 1, \ldots, 49,$$

$$j = 1, \ldots, 12, k = 1, \ldots, 439$$

$$\sigma_j = \sqrt{var\left(mineral\ composition^j\right)}$$

where the subscripts i, k, tr, and val represent the i-th test data, k-th training and validation data, and training and validation data, respectively. Superscript j denotes mineral. The parameters from the training and validation datasets were used to scale the test data.

$$R^2 = 1 - \frac{\sum_{i=1}^{N_{te}} \sum_{j=1}^{N_{mineral}} \left(y_i^j - \hat{y}_i^j \right)^2}{\sum_{i=1}^{N_{te}} \sum_{j=1}^{N_{mineral}} \left(y_i^j - \bar{y}^j \right)^2} \tag{4}$$

$$\bar{y}^j = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} mineral\ compostion_i^j$$

$$i = 1, \ldots, 49, j = 1, \ldots, 12$$

where N_{te} and $N_{mineral}$ are the numbers of test data points (49) and minerals (12), respectively. y_i^j and $\hat{y}_i^j$ represent the normalized true and predicted compositions of the j-th mineral of i-th test data, respectively. $\bar{y}^j$ is the average normalized j-th mineral composition.

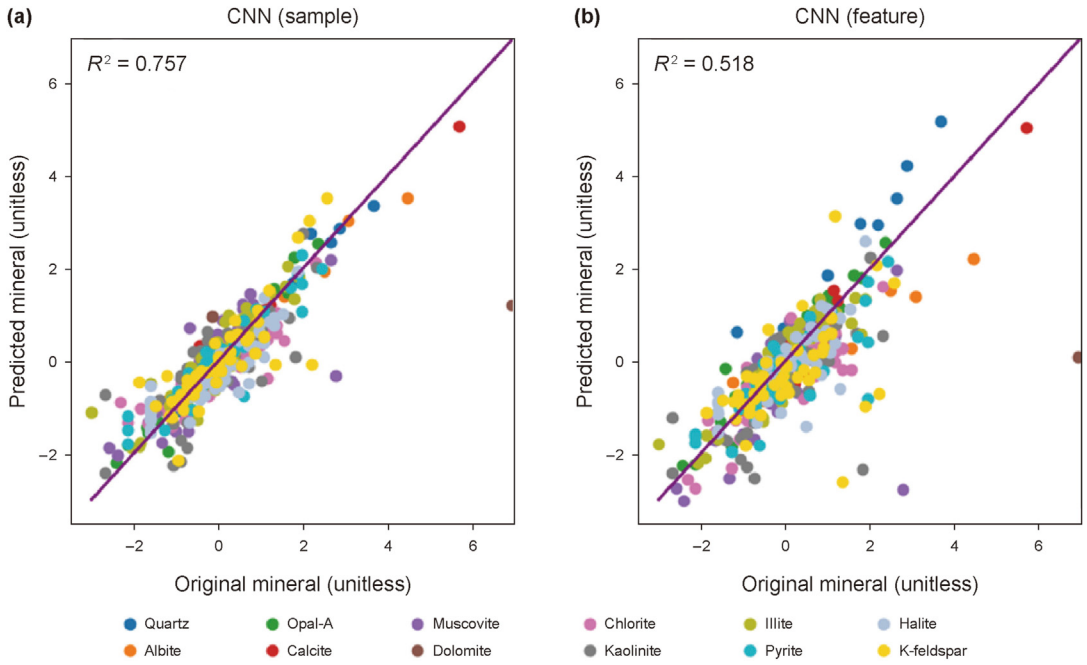

Fig. 6. Scatter plot and coefficient of determination of 49 the test data using CNN preprocessed by each (**a**) sample and (**b**) feature.

The R^2 of the CNN (sample) and CNN (feature) were 0.757 and 0.518, respectively (Fig. 6(a) and (b)). The R^2 improved by approximately 46% through the preprocessing of each sample. Fig. 6(a) shows that the predicted values agreed well with the original values. However, the prediction result for dolomite was underestimated because most of the dolomite had zero composition (Fig. 2 and Table 2). If additional data is obtained from additional experiments or augmentation approaches with a high dolomite composition, the machine learning model may be retrained to predict the outlier values for dolomite.

3.2. Further analysis of each test data

The results of the mineral prediction, such as the average and variance of MAE and the R^2 for the 49 test data, indicate that the CNN algorithm with sample-based preprocessing was the best model. For a detailed analysis, one sample with a general composition (#76) and three samples with outlier compositional distributions (#1, #408, and #483) were selected from the 49 test data. Sample #1 has an unusually high albite composition of 37.7%, which is the highest value among all 488 samples (Table 2). Sample #408 has a high quartz composition of 45% and opal-A does not even exist. Also, Sample #483 has an outlier mineral composition for calcite about 45%.

Fig. 7 shows the results of the mineral compositions predicted using the two CNN models depending on the preprocessing methods for the four examples (#1, #76, #408, and #483). The red bar indicates the mineral composition estimated by an expert (label data); the blue and green bars represent the mineral compositions predicted by the CNN with the sample and feature, respectively.

Sample #1 is an outlier compositional distribution with the highest albite composition (33.7%) among all 488 samples, as shown in Fig. 2 and Table 2. Notably, opal-A is approximately 10% composition. The CNN (sample) model predicted an albite content close to 30% and other minerals, such as quartz and opal-A, within acceptable difference with the label data. However, the CNN (feature) model showed poor prediction accuracy, with higher predictions for quartz and opal-A than for albite (Fig. 7(a)).

Sample #76 in Fig. 7(b) shows the general mineral composition of the 488 samples: approximately 35% opal-A; 17% quartz; and 10% albite, muscovite, and illite. Both CNN models successfully assessed mineral composition trends. The CNN (sample) model provides a more precise prediction than the CNN (feature) model (Fig. 7(b)).

Sample #408 in Fig. 7(c) is an outlier in which the composition of quartz is high at approximately 45%, and opal-A is not present. Both CNN models reliably predicted the overall trend of high quartz and extremely low opal-A compositions. However, the CNN (feature) model had a problem: overestimating quartz and underestimating albite. By contrast, the CNN (sample) model exhibited satisfactory performance for all minerals (Fig. 7(c)).

Sample #483 in Fig. 7(d) has an outlier composition: a high calcite composition, 45%; quartz and opal-A, approximately 10%; and muscovite, approximately 15%. The CNN (feature) model predicted the calcite composition to be approximately 40%, similar to the label data, but overpredicted quartz, albite, opal-A, and illite and underpredicted muscovite and K-feldspar. The CNN (sample) model predicted a composition similar to that of the CNN (features) for calcite. The predictions for quartz, albite, and muscovite using the CNN (sample) were improved compared with those of the CNN (feature) model (Fig. 7(d)). However, the prediction accuracy for other minerals remains poor, which might be solved using additional outlier samples for calcite.

Table 6 lists the MAE values for the four samples shown in Fig. 7. For all samples, the prediction performance of the CNN (sample) model was better than that of the CNN (feature) model. The MAE of sample #76, which was a general composition, was the lowest at 0.72 and 1.19, respectively. The MAE of sample #408, which has high quartz composition, showed the second-lowest error with 0.77 and 2.51, respectively. It showed a lower error than for samples #1 and #483 because the intensity of quartz was more distinct than that of the other minerals at approximately 26.5° (Fig. 3). Additionally, the training data had a higher quartz composition than the samples with other outlier compositions did (Fig. 2).

The results for sample #1 showed a high error of 1.61 and 4.20.

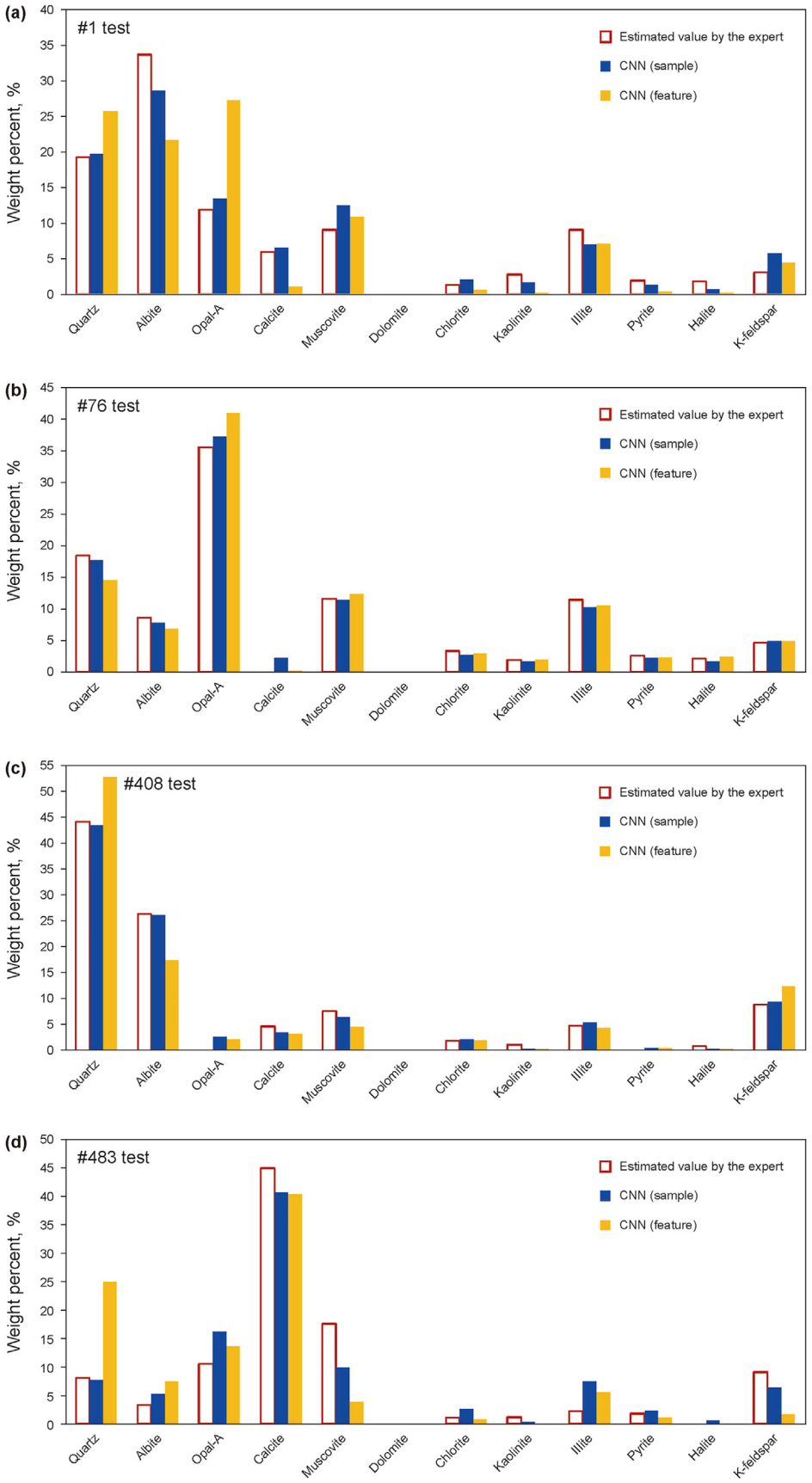

Fig. 7. Results for the CNN models for the four samples among the 49 test data: (**a**) #1, (**b**) #76, (**c**) #408, and (**d**) #483.

Table 6
Errors the four test data between original and predicted mineral composition.

(Unit: %)

MAE	# 1	# 76	# 408	# 483
CNN (sample)	1.61	0.72	0.77	2.61
CNN (feature)	4.20	1.19	2.51	4.61

Because machine learning is a data-driven technique, solving the extrapolation problem can be difficult. However, the proposed model accurately replicated the compositions of albite, quartz, and opal-A, although the MAE was double that of samples #76 and #408.

The errors for sample #483 were 2.61 and 4.61, respectively, which were the largest MAE. As shown in Fig. 2 and Table 2, calcite typically has a composition of less than 10% in most of the 488 samples. However, the maximum calcite content was 58.9%, and sample #483 contained approximately 45% calcite. Therefore, interpreting the intensity profile of this sample is difficult owing to insufficient data on calcite compositions between 10% and 58.9%.

4. Conclusion

We developed a highly reliable deep learning model to predict the composition of 12 minerals using XRD experimental data from the UB. A novel preprocessing approach was implemented using a CNN algorithm, considering the unique characteristics of the XRD experimental data.

(1) The XRD intensity data showed clear differences in the peak angles and absolute scales, although the samples had similar mineral compositions. However, for the analysis of the XRD intensity profiles, these differences were not critical for interpreting the mineral composition. The relative scale to the peaks and the intensity trend were more important than absolute intensity value and exact peak angle. Therefore, the data were preprocessed by applying the min-max scaler to each sample and not to each feature. This method not only adjusts the range of each sample from 0 to 1 but also preserves its intensity trend, including the ratio and the shape of peaks.

(2) In this study, the LSTM, MLP, RF, and CNN models were compared to determine a suitable algorithm for the newly proposed preprocessing method. The CNN with the proposed preprocessing showed an improvement of 41% (1.16–0.68) and 46% (0.518–0.757) in MAE and R^2 for the 49 test data compared with those of the CNN with feature-based preprocessing (Park et al., 2022). Additionally, the CNN with the proposed preprocessing significantly improved the prediction performance for samples with outlier compositions, such as high albite, quartz, and calcite, which could not be assessed by Park et al. (2022).

(3) In this study, we confirmed that the CNN model with sample-based preprocessing is an appropriate approach for predicting mineral composition using XRD intensity data. Unlike other algorithms, such as LSTM, MLP, and RF, CNN can easily learn the spatial trends of the intensity profile using the concept of a receptive field. The developed CNN model can be applied to new XRD data near the UB, and the workflow of this study can be applied to XRD data from other regions.

(4) CNN (sample) exhibits larger prediction errors for few minerals compared to CNN (feature). For example, in case of K-feldspar, CNN (sample) showed smaller error in samples # 480 and #483 than CNN (feature) but the opposite result was shown in sample #1. This issue may be caused by the insufficiency of training data, especially for the imbalanced distribution of mineral composition. In future research, the proposed method can be improved by securing additional training data with diverse mineral compositions.

CRediT authorship contribution statement

Hongkeun Jin: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Ju Young Park:** Writing – review & editing, Visualization, Investigation, Formal analysis, Conceptualization. **Sun Young Park:** Software, Resources, Investigation, Formal analysis, Data curation, Conceptualization. **Byeong-Kook Son:** Software, Resources, Data curation, Conceptualization. **Baehyun Min:** Methodology, Formal analysis, Conceptualization. **Kyungbook Lee:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This study was supported by the Gas Hydrate R&D Organization and the Korea Institute of Geoscience and Mineral Resources (KIGAM) (GP2021–010). This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2021R1C1C1004460) and Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korean government (MOTIE) (20214000000500, Training Program of CCUS for Green Growth).

Abbreviations

3D, three-dimensional; CNN, convolutional neural network; GH, gas hydrate; IQR, interquartile range; LSTM, long short-term memory; MAE, mean absolute error; MLP, multi-layer perceptron; RF, random forest; RNN, recurrent neural network; UB, Ulleung Basin; UBGH, Ulleung Basin gas hydrate drilling expedition; XRD, X-ray diffraction.

Appendix A. Machine learning model structure

Tables A1-A3 present the structures of the LSTM, MLP, RF, and CNN models used in the study. The hyperparameters of the neural network algorithms and RF were determined through a sensitivity analysis. For the neural network algorithms, the number of layers, the activation function, and the optimizer were set as hyperparameter. For RF, the depth and number of trees were used as hyperparameter (Park et al., 2022).

Table A1
Design of mineral composition prediction model using LSTM and MLP.

		LSTM	MLP
Input		3100 intensity profile	
Output		12 minerals composition	
Number of total input datasets	Training	(3,100,307)	
	Validation	(3,100, 132)	
	Test	(3,100, 49)	
Number of hidden layers		1	4
Number of nodes		20	1000, 300, 100, 30
Activation functions		tanh	sigmoid
Loss function		Categorical crossentropy	
Optimizer		Adam	
Learning rate		0.001	
Maximum number of epochs		200	
Early stopping		Yes	
Performance indicators		MAE, R^2	

Table A2
Design of mineral composition prediction model using CNN.

		CNN
Input		3100 intensity profile
Output		12 minerals composition
Number of total input datasets	Training	(3,100,307)
	Validation	(3,100, 132)
	Test	(3,100, 49)
Convolution layers		3 (50,32−50,64-50,128)
Maxpooling layers		3 (4-4-4)
Fully connected layer		100
Activation functions		ReLU
Loss function		Categorical crossentropy
Optimizer		Adam
Learning rate		0.001
Maximum number of epochs		200
Early stopping		Yes
Performance indicators		MAE, R^2

Table A3
Design of mineral composition prediction model using RF.

		RF
Input		3100 intensity profile
Output		12 minerals composition
Number of total input datasets	Training	(3,100, 439)
	Test	(3,100, 49)
Max_depth		10
Number of decision trees		200
Learning rate		0.001
Performance indicators		MAE, R^2

Appendix B. Prediction results using LSTM, MLP, and RF

Fig. B1 shows the scatter plot and R^2 of the prediction results obtained using the LSTM, MLP, and RF models. The R^2 of LSTM (sample) and LSTM (feature) were unreliably low at 0.133 and 0.225, respectively (Fig. B1a and B1b). LSTM also predicted each mineral composition as an average value regardless of the actual mineral values, as shown by the horizontal trend of the predicted mineral in Fig. B1. The reason for this phenomenon is that the 3100 intensity data points are too long for LSTM to learn the intensity peaks (Gers et al., 1999). In addition, LSTM is efficient for sequential data and assumes that the final data are the most important. However, the importance of intensity data does not increase as the incident angle increases because each mineral has a certain peak region based on Bragg's law.

The MAE of MLP and RF improved when the preprocessed data for each sample were used (Table 5); however, the R^2 was similar (Fig. B1c to B1f). Although CNN can capture the overall trend in intensity profile using the concept of a receptive field, MLP and RF consider 3100 intensities as an individual feature. Therefore, MLP and RF could not understand the continuous changes in the XRD pattern. A critical limitation is that the positions of the intensity

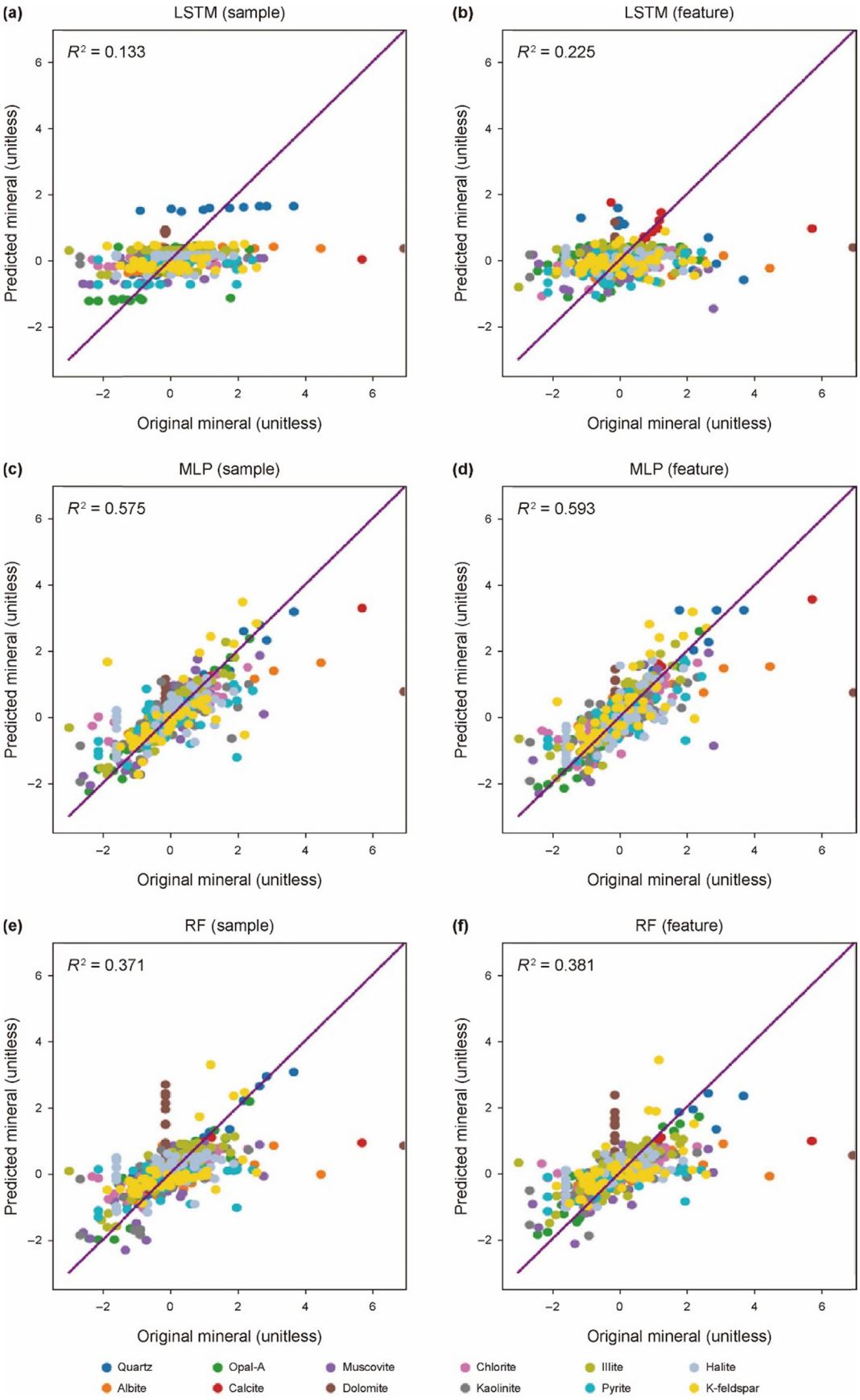

Fig. B1. Scatter plot and coefficient of determination of the 49 test data: LSTM with (a) sample, (b) feature, MLP with (c) sample, (d) feature, RF with (e) sample, and (f) feature.

peaks can differ slightly (Fig. 3 and Table 4).

Appendix C. Results for 5-fold cross-validation

The mineral composition data used in this study exhibits an uneven distribution as shown in Fig. 2. Therefore, we conducted 5-fold cross-validation to assess the impact of data split on each algorithm's performance. Table C1 presents MAE across the 5-folds and their average and standard deviation. The proposed method, Sample_CNN, showed the lowest MAE for each case. Also, standard deviation of CNN's MAEs presented the lowest value indicating a stable performance regarding the data split issue. Because the results by the 5-fold is similar with the results in Table 5, we can state that CNN is the superior algorithm in this study among the four algorithms.

Table C1
MAE of 5-fold cross-validation for each algorithm including LSTM, MLP, RF, and CNN.

(Unit: %)

Fold number	1	2	3	4	5	Ave.	Std.
Sample_LSTM	2.43	2.20	2.23	1.86	2.22	2.19	0.20
Sample_MLP	0.89	0.87	0.82	0.78	0.82	0.86	0.06
Sample_RF	1.29	1.08	1.15	1.24	1.23	1.20	0.08
Sample_CNN	0.66	0.73	0.64	0.73	0.66	0.68	0.04

References

Ainane, A., Taleb, M., El-Hajjaji, F., Hammouti, B., Chatouani, A., Ainane, T., 2021. Study of dependence between two types of most abundant natural clays in Bejaad province (Central Morocco) using a statistical approach. Moroc. J. Chem. 9, 210–220. https://doi.org/10.48317/IMIST.PRSM/morjchem-v9i2.22438.

Domínguez-Olmedo, J.L., Toscano, M., Mata, J., 2020. Application of classification trees for improving optical identification of common opaque minerals. Comput. Geosci. 140, 104480. https://doi.org/10.1016/j.cageo.2020.104480.

Dong, R., Zhao, Y., Song, Y., Fu, N., Omee, S.S., Dey, S., Li, Q., Wei, L., Hu, J., 2022. DeepXRD, a deep learning model for predicting XRD spectrum from material composition. ACS Appl. Mater. Interfaces 14, 40102–40115. https://doi.org/10.48550/arXiv.2203.14326.

Gers, F.A., Schmidhuber, J., Cummins, F., 1999. Learning to forget: continual prediction with LSTM. Ninth Int. Conf. Artific. Neural Networks ICANN 99 (2), 850–855. https://doi.org/10.1049/cp:19991218.

Ji, M.S., Kwon, S.Y., Kim, M., Kim, S.I., Min, B.H., 2022. Generation of synthetic compressional wave velocity based on deep learning: a case study of Ulleung Basin gas hydrate in the Republic of Korea. Appl. Sci. 12, 8775. https://doi.org/10.3390/app12178775.

Kim, D.K., Choi, J.H., Kim, D.W., Byun, J.M., 2020. Predicting mineralogy by integrating core and well log data using a deep neural network. J. Petrol. Sci. Eng. 195, 107838. https://doi.org/10.1016/j.petrol.2020.107838.

Kvenvolden, K.A., Ginsburg, G.D., Soloviev, V.A., 1993. Worldwide distribution of subaquatic gas hydrates. Geo Mar. Lett. 13, 32–40. https://doi.org/10.1007/BF01204390.

Lee, J.-W., Park, W.B., Lee, J.H., Singh, S.P., Sohn, K.-S., 2020. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. Nat. Commun. 11, 704. https://doi.org/10.1038/s41467-020-14512-9.

Li, Y., Liu, L., Jin, Y., Wu, N., 2021. Characterization and development of natural gas hydrate in marine clayey–silt reservoirs: a review and discussion. Adv. Geo-Energy Res. 5, 75–86. https://doi.org/10.46690/ager.2021.01.08.

McLennan, S.M., Hemming, S., McDaniel, D.K., Hanson, G.N., 1993. Geochemical approaches to sedimentation, provenance and tectonics. In: Johnsson, M.J., Basu, A. (Eds.), Processes Controlling the Composition of Clastic Sediments, vol. 284. Spec. Pap. Geol. Soc. Am., pp. 21–40. https://doi.org/10.1130/SPE284-p21

Okada, N., Maekawa, Y., Owada, N., Haga, K., Shibayama, A., Kawamura, Y., 2020. Automated identification of mineral types and grain size using hyperspectral imaging and deep learning for mineral processing. Minerals 10, 809. https://doi.org/10.3390/min10090809.

Park, S.Y., Son, B.-K., Choi, J.Y., Jin, H.K., Lee, K.B., 2022. Application of machine learning to quantification of mineral composition on gas hydrate-bearing sediments, Ulleung Basin, Korea. J. Petrol. Sci. Eng. 209, 109840. https://doi.org/10.1016/j.petrol.2021.109840.

Rietveld, H.M., 1969. A profile refinement method for nuclear and magnetic structures. J. Appl. Crystallogr. 2, 65–71. https://doi.org/10.1107/S0021889869006558.

Ryu, B.-J., Collett, T.S., Riedel, M., Kim, G.Y., Chun, J.-H., Bahk, J.-J., Lee, J.Y., Kim, J.-H., Yoo, D.-G., 2013. Scientific results of the second gas hydrate drilling expedition in the Ulleung Basin (UBGH2). Mar. Petrol. Geol. 47, 1–20. https://doi.org/10.1016/j.marpetgeo.2013.07.007.

Schuetzke, J., Benedix, A., Mikut, R., Reischl, M., 2021. Enhancing deep-learning training for phase identification in powder X-ray diffractograms. IUCrJ 8, 408–420. https://doi.org/10.1107/S2052252521002402.

Shahnazar, S., Hasan, N., 2014. Gas hydrate formation condition: review on experimental and modeling approaches. Fluid Phase Equil. 379, 72–85. https://doi.org/10.1016/j.fluid.2014.07.012.

Sietronics, 1996. SIROQUANT: A Quantitative XRD Software. Sietronics Pty Limited, Belconnen ACT, Australia.

Taylor, J.C., 1991. Computer programs for standardless quantitative analysis of minerals using the full powder diffraction profile. Powder Diffr. 6, 2–9. https://doi.org/10.1017/S0885715600016778.

Taylor, J.C., Matulis, C.E., 1994. A new method of rietveld clay analysis. Part I. Use of a universal measured standard profile for rietveld quantification of montmorillonite. Power Diffract. 9, 119–123. https://doi.org/10.1017/S0885715600014093.

Wang, R., Shi, W., Xie, X., Zhang, W., Qin, S., Liu, K., Busbey, A.B., 2020. Clay mineral content, type, and their effects on pore throat structure and reservoir properties: Insight from the Permian tight sandstones in the Hangjinqi area, north Ordos Basin, China. Mar. Petrol. Geol. 115, 104281. https://doi.org/10.1016/j.marpetgeo.2020.104281.

Zeng, X., Xiao, Y., Ji, X., Wang, G., 2021. Mineral identification based on deep learning that combines image and mohs hardness. Minerals 11, 506. https://doi.org/10.3390/min11050506.

Zhang, F., An, M., Zhang, L., Fang, Y., Elsworth, D., 2019. The role of mineral composition on the frictional and stability properties of powdered reservoir rocks. J. Geophys. Res. Solid Earth 124, 1480–1497. https://doi.org/10.1029/2018JB016174.