energies

MDPI

Article

Data-Driven Three-Phase Saturation Identification from X-ray CT Images with Critical Gas Hydrate Saturation

Sungil Kim [1], Kyungbook Lee [1,2,*], Minhui Lee [3] and Taewoong Ahn [1]

[1] Petroleum and Marine Research Division, Korea Institute of Geoscience and Mineral Resources, Daejeon 34132, Korea; skim@kigam.re.kr (S.K.); twahn@kigam.re.kr (T.A.)

[2] Department of Geoenvironmental Sciences, Kongju National University, Gongju, Chungnam 32588, Korea

[3] GEOLAB Co., Ltd., Sejong 30121, Korea; geolab@geolab.co.kr

* Correspondence: kblee@kongju.ac.kr; Tel.: +82-41-850-8511

Received: 19 September 2020; Accepted: 6 November 2020; Published: 9 November 2020

Abstract: This study proposes three-phase saturation identification using X-ray computerized tomography (CT) images of gas hydrate (GH) experiments considering critical GH saturation ($S_{GH,C}$) based on the machine-learning method of random forest. Eight GH samples were categorized into three low and five high GH saturation (S_{GH}) groups. Mean square error of test results in the low and the high groups showed decreases of 37% and 33%, respectively, compared to that of the total eight. Additionally, a universal test set was configured from the total eight and tested with two trained machines for the low and high GH groups. Results revealed a boundary at ~50% of S_{GH} signifying different saturation identification performance and the ~50% was estimated as $S_{GH,C}$ in this study. The trained machines for the low and high S_{GH} groups had less performance on the larger and smaller values, respectively, of $S_{GH,C}$. These findings conclude that we can take advantage of suitable separation of obtained training data, such as GH CT images, under the criteria of $S_{GH,C}$. Moreover, the proposed data-driven method not only serves as a saturation identification method for GH samples in real time, but also provides a guideline to make decisions for data acquirement priorities.

Keywords: X-ray CT image; critical gas hydrate saturation; saturation identification; random forest; data management; machine-learning

1. Introduction

Naturally reserved gas hydrate (GH) has high uncertainty regarding its kinetic behavior, geomechanical stability, and economic feasibility. For these reasons, multiple countries such as South Korea, Japan, the United States, and India [1–6] encounter difficulties in pursuing research and development (R&D) or holding their test production in fields, despite having conducted related R&D. In particular, one location containing reserved GH—the East Sea in South Korea—is a challenging field for producing GH owing to its sparse GH distribution, well stability problems, uncertain GH dissociation, and the current energy ecosystem being incompatible for GH [7].

Concerning these issues, many investigations into the behaviors of GH and reducing uncertainty of general characteristics of GH have been carried out [8–12]. One method, X-ray computerized tomography (CT), involves scanning out of a target GH sample during experiments to infer how inner fluids behave in porous media to address the difficulty of understanding what happens in a GH sample directly [13–15]. In the GH experimental environment, production rates are difficult to measure accurately due to either the dead volume between measurement equipment and a GH sample, or flow delay in a GH sample, or a flow line. By addressing these issues, X-ray CT scanning could be an appropriate method to investigate fluid workings in a GH sample and infer its approximate trend.

The Korea Institute of Geoscience and Mineral Resources (KIGAM) utilized X-ray CT images to quantitatively analyze depressurization velocity and critical GH saturation ($S_{GH,C}$) during application of depressurization in a GH sample. In the KIGAM experiments, normalization of CT values was effective for quantitative analysis of approximate GH behavior, but the CT values included three phase behaviors (i.e., water, gas, and GH) such that no significant difference between the different depressing velocities could be determined.

Therefore, identification of each phase is needed for more accurate analysis in GH experiments. In particular, it is key to distinguish water from GH owing to their similar densities, which causes difficulty in their identification. Dependable identification of GH saturation will, in turn, lead to identifying an optimal depressurization parameter with higher reliability than using only normalized CT values. Our previous study has shown reliable applicability of machine-learning for GH saturation identification based on X-ray CT images [16]. In that study, the machine-learning methods utilized CT images for input and saturation values for output; in particular, random forest (RF) brought over 95% correlation between the original and predicted data for both training and test data. However, that study used only 960 items of training data without thorough filtering or selection of CT images for each of the phase saturations. Furthermore, such an amount of data seems insufficient to cover overall types of GH trends.

In most cases, the number of training data is several hundred or at most a few thousand for machine-learning applications in petroleum engineering [17–28], which was also noted in a previous paper [16]. Contrastingly, the number of training data is over ten thousand or even up to one million in computer-science-engineering-centered applications [29]. In spite of that, the number of training data does not necessarily guarantee reliability of training performance when inappropriate data is combined with the entire data pool. Either unqualified data should be removed, or differing types of data should be properly separated based on domain knowledge for better machine-learning [28].

$S_{GH,C}$ can be one suitable standard to separate or categorize given GH data because $S_{GH,C}$ highly distinguishes GH behavior. In Gil et al. [30] and KIGAM [31], numerical analysis of GH dissociation behavior was conducted to find $S_{GH,C}$. In Gil et al. [30], the $S_{GH,C}$ was suggested as ~50–60% according to the simulation results, mimicking the environment of Ulleung Basin, East Sea, South Korea. Although the $S_{GH,C}$ has been narrowed down to a certain range of GH saturation, more accuracy is still required to better describe GH saturation for Ulleung Basin.

We need to conduct reliable saturation identification of GH, thereby leading to more accurate $S_{GH,C}$ measurements and descriptions of GH behaviors. Eventually, it will be necessary to apply proper machine-learning, owing to its ability to draw meaningful conclusions or lessons from given data. Considering the previous machine-learning applications for petroleum engineering and necessity of proper data construction, this paper will suggest how to separate given X-ray CT images for machine-learning applications with $S_{GH,C}$. In addition, it will analyze how data quantity and quality (or construction) function in terms of machine-learning performance.

Section 2 explains how GH experiments were conducted and how X-ray CT was scanned for the experiments. In addition, the procedures of data acquisition and data preprocessing will be described with the applied machine-learning method, RF. Section 3 presents the data of the given CT images and compares with previous data [16]. Section 4 reports the saturation identification results pictured for four cases to analyze the effect of data construction in terms of quantity and quality. Section 5 presents conclusions and identifies valuable guidelines for the next steps of GH saturation identification.

2. Methodology

2.1. Experiments of Gas Hydrate (GH) with X-Ray Computerized Tomography (CT) Scanning

KIGAM introduced a visualization system to experiment with different GH production options and a variety of production-related parameters. The visualization system consists of four main parts: GH sample installation, fluid injection, fluid production, and X-ray CT equipment, as presented in Figure 1.

X-ray CT scanning is continuously performed during all experimental procedures. The high-pressure cell is composed of glass fiber and aluminum complex to minimize X-ray diffraction, with a diameter and length of 1 and 2 inches, respectively (Figure 1a). The X-ray CT scanning equipment (General Electricity OPTIMA 660) can scan at a rate of up to 96 mm/sec and can also conduct quantitative analysis. Other details about the equipment and its specifications are explained in Kim et al. [16].

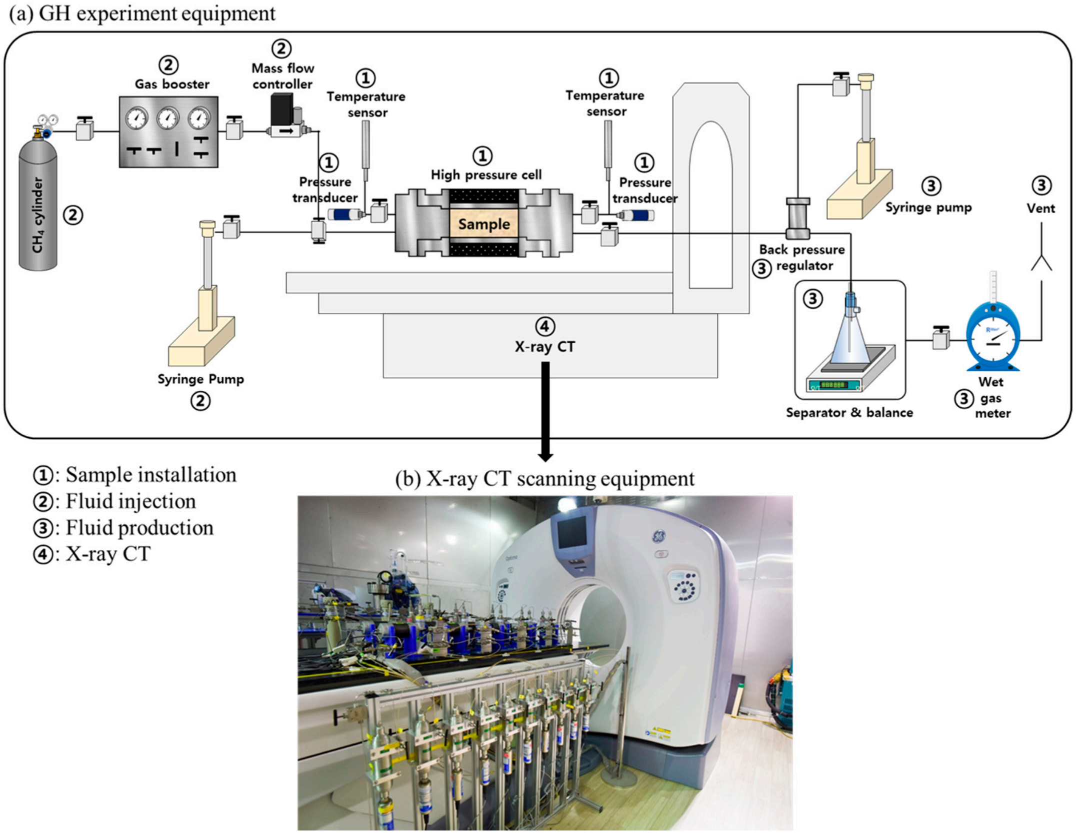

Figure 1. (**a**) Schematic diagram of the gas hydrate (GH) visualization system at Korea Institute of Geoscience and Mineral Resources (KIGAM) [31]; (**b**) the installed X-ray computerized tomography (CT) equipment for the system [31].

In this study, GH generation is followed by GH sample charge, initial water saturation setting, and pressurization by CH_4. In particular, the pressurization is performed by excess gas method, which forms GH showing grain-coating or cementing-type habit [32,33]. Excess gas method is to increase pressure and decrease temperature by methane gas injection in partial water saturation condition. As presented in Figure 2, the experiment consists of five stages before the GH depressurization stage: "DRY", "SAT", "IWS", "GH", and "GTW", described as follows. X-ray CT was scanned for every stage and the scanned CT images corresponded to the five stages. Only one scan is sufficient to determine the internal state of a target GH sample in each of the five stages because there is no critical change within the same stage. In the DRY stage, CT images are scanned immediately after the charge with sand to check whether the distribution of sand particles is regular and that those CT values are smallest for the entire experiment because a sand sample is saturated only with the air. Before SAT stage, the entire experiment system is inspected to ensure that the system is stably operated and each compartment is tightly connected through valves. In the SAT stage, a GH sample is 100% saturated with water, which has the largest CT values in this experiment. Therefore, the smallest and largest CT values are utilized to both normalize and quantitatively analyze CT images, as follows:

$$CT_{norm} = \frac{CT_{STAGE} - CT_{DRY}}{CT_{SAT} - CT_{DRY}}, \quad (1)$$

where CT_{norm} is the normalized CT value; CT_{STAGE} is the CT value from the IWS, GH, or GTW stages. Note that more experimental explanation can be reviewed in detail in [16].

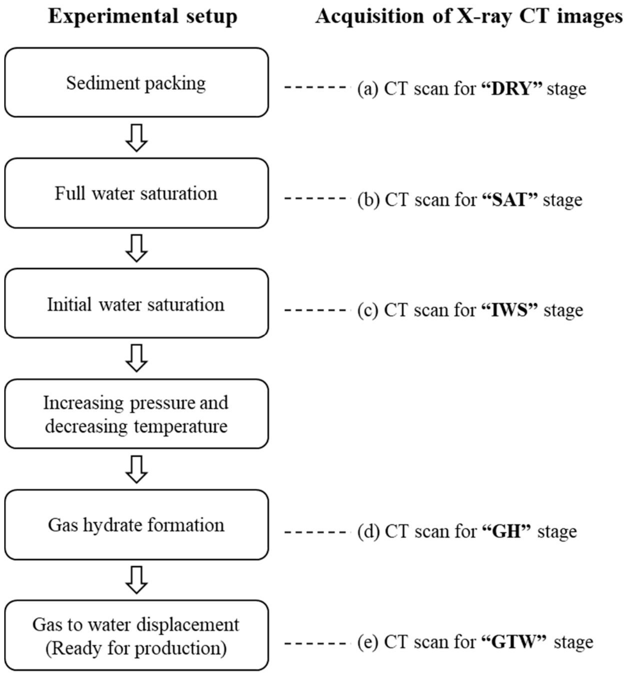
Experimental setup
Acquisition of X-ray CT images
Sediment packing
(a) CT scan for "DRY" stage
Full water saturation
(b) CT scan for "SAT" stage
Initial water saturation
(c) CT scan for "IWS" stage
Increasing pressure and decreasing temperature
Gas hydrate formation
(d) CT scan for "GH" stage
Gas to water displacement (Ready for production)
(e) CT scan for "GTW" stage

Figure 2. Flow chart of the gas hydrate (GH) generation experiment with computerized tomography (CT) scanning for obtaining the machine-learning data. (**a**) A sand sample made in DRY stage; (**b**) water saturated sample in SAT stage; (**c**) gas and water saturated status in IWS stage; (**d**) GH generation by pressurization in GH stage; (**e**) water replaces gas in GTW stage.

In the IWS stage, the sample has initial water saturation with CH_4, wherein the GH sample is pressurized to make GH (GH stage of Figure 2d). Methane gas is injected with air-cooling operation until it reaches the suitable high pressure and low temperature condition for GH formation. Then, remaining gas in the GH sample is replaced by 3% salinity brine to build the most similar environment of naturally reserved GH (GTW stage of Figure 2e). According to Equation (1), the normalized CT values will be 0 and 1 for the images from the DRY and SAT stages, respectively. In terms of the IWS, GH, or GTW stages, the normalized CT values will be between 0 and 1.

2.2. Data Acquisition and Preprocessing

The CT images resolution is 512 × 512 × 96 and the size of each pixel is ~660 µm and 100 µm in the vertical and horizontal directions, respectively. High CT values due to the end-piece, including the temperature sensor, were removed in order to acquire proper CT values for training the machine-learning models. Thus, a total of 96 slices were obtained; however, only the front 64 slices were utilized as training data. According to Equation (1), given CT values were normalized and the related procedures are shown in Figure 3. Row CT images were obtained by CT scanning (Figure 3a) and the useless parts were discarded (Figure 3b). Target images for normalization were put to the CT slot as described in Figure 3c and this procedure makes image data standardized for versatility between other GH sample experiments. Finally, the normalized CT images and three phase saturations comprised the data pairs for input and output of training data, respectively.

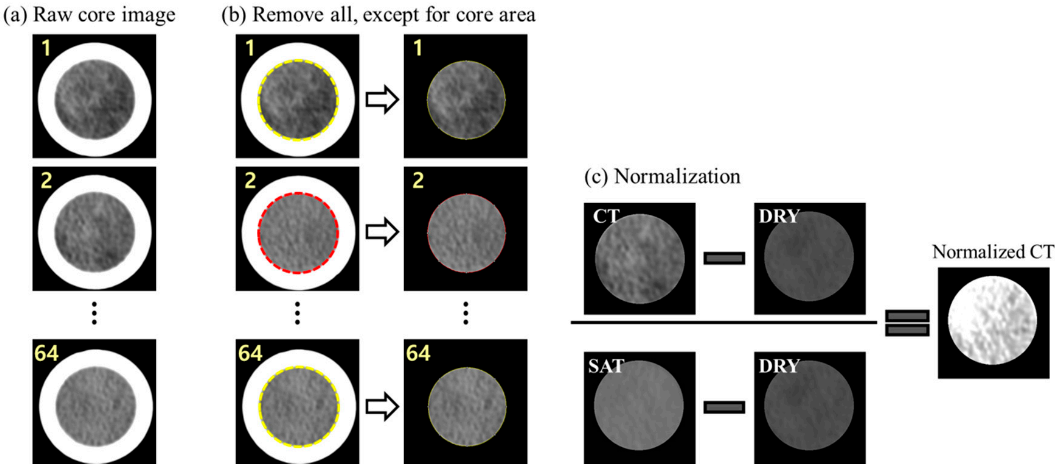

Figure 3. Modification and normalization of the CT images using ImageJ software [16]. (**a**) Acquisition of raw core images from the first to the sixty-fourth slice; (**b**) editing given images to acquire the target circle area; (**c**) normalization with CT images of DRY and SAT for standardization.

Saturation values of water, GH, and gas were generated, corresponding to the normalized GH CT images, a process known as “labeling” for building training data for supervised learning [34]. CT images of the DRY and SAT stages technically have fixed values such as $S_G = 1$ or $S_W = 1$, respectively. The IWS and GH stages would clearly show $S_{GH} = 0$ or $S_W = 0$, respectively. In addition, S_{GH} of the GTW stage would be the same as that of the GH stage. We can calculate three phase saturations according to those conditions, given an experimental environment according to Table 1 and Equations (2) and (3):

$$S_{\mathrm{W,STAGE}} \times d_{\mathrm{W,STAGE}} + S_{\mathrm{GH,STAGE}} \times d_{\mathrm{GH,STAGE}} + S_{\mathrm{G,STAGE}} \times d_{\mathrm{G,STAGE}} = CT^{\mathrm{avg}}_{\mathrm{norm}}, \tag{2}$$

$$CT^{\mathrm{avg}}_{\mathrm{norm},\ j} = \frac{1}{n_{\mathrm{c}}} \sum\nolimits_{i=1}^{n_{\mathrm{c}}} CT_{\mathrm{norm},j,\ i} \text{ for } j = 1, \ldots, n_{\mathrm{s}}\ , \tag{3}$$

where S and d indicate saturation and density, respectively. The subscripts W, GH, and G of S are water, gas hydrate, and gas, respectively. The second subscript STAGE represents the given experimental stage (IWS, GH, or GTW). $CT^{\mathrm{avg}}_{\mathrm{norm}}$ is the average, normalized CT value from all data pixels (n_c) of the actual GH sample in one slice image (i.e., the circle-shaped area in the right column of Figure 3b). Here, n_{c} is 47,992 at the experimental environment in this study. n_{s} is the number of slices of a GH sample; 64 in this experiment setting. Consequently, one CT slice image is paired with three saturation values, becoming one sample for training data.

Table 1. Saturation and density conditions for labeling in the three experimental stages.

Experimental Stage	S_W (Density, g/cc)	S_{GH} (Density, g/cc)	S_G (Density, g/cc)
“IWS” (14.7 psi, 18 °C)	$S_{\mathrm{W,IWS}}$ ($d_{\mathrm{W,IWS}} = 1$)	$S_{\mathrm{GH,IWS}} = 0$	$1 - S_{\mathrm{W,IWS}}$ ($d_{\mathrm{G,IWS}} = 0.000678$)
“GH” (2500 psi, 16 °C)	$S_{\mathrm{W,GH}} = 0$	$S_{\mathrm{GH,GH}}$ ($d_{\mathrm{GH,GH}} = 0.91$)	$1 - S_{\mathrm{GH,GH}}$ ($d_{\mathrm{G,GH}} = 0.143$)
“GTW” (2900 psi, 16 °C)	$S_{\mathrm{W,GTW}}$ ($d_{\mathrm{W,GTW}} = 1.008$)	$S_{\mathrm{GH,GTW}} = S_{\mathrm{GH,GH}}$	$1 - S_{\mathrm{W,GTW}} - S_{\mathrm{GH,GTW}}$ ($d_{\mathrm{G,GTW}} = 0.167$)

2.3. Machine-Learning Methodology: Random Forest (RF)

As an ensemble machine-learning method, RF consists of multiple decision trees. Figure 4 describes how a decision tree is constructed when it has n sample data of the property of d. First, the data portioning procedure divides the given data into m_1 and m_2, locating the single decision boundary which minimizes the average of two mean square variances, MSV_1 and MSV_2 (Figure 4a). The decided boundary then becomes a standard for dividing into two branches (Figure 4b). This procedure is repeated until reaching a specific criterion, which is usually set by a user. After that, pruning is conducted to prevent the final decision tree from overfitting (Figure 4c).

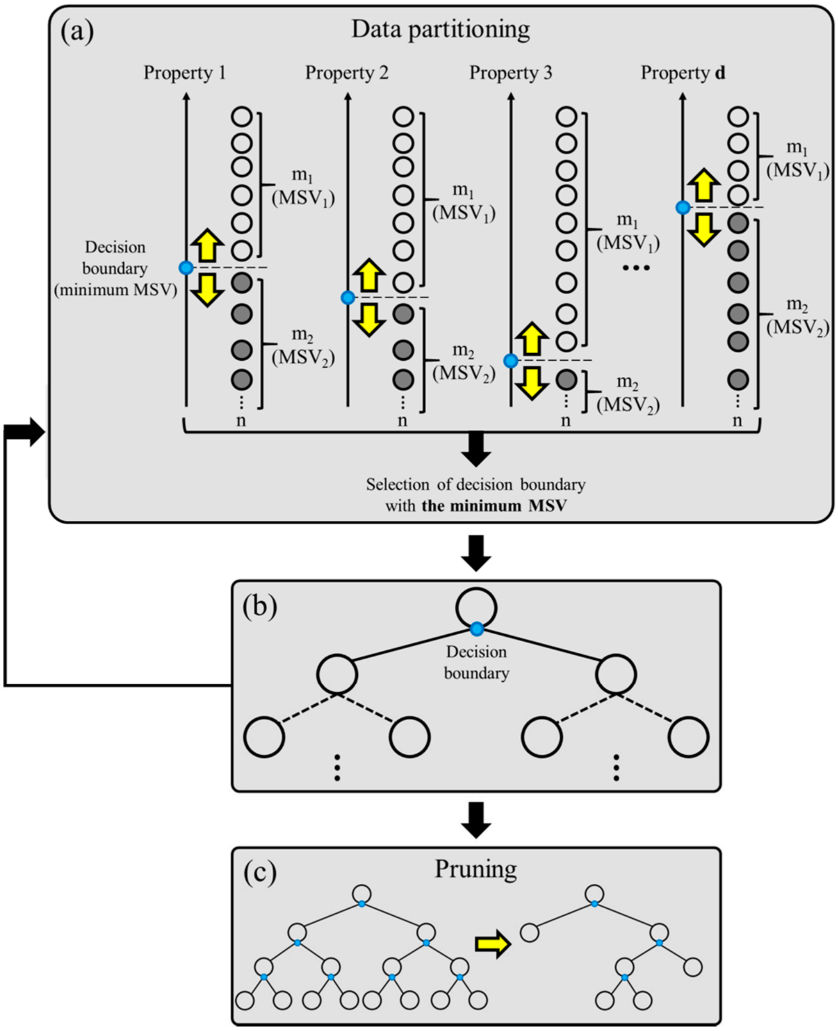

Figure 4. Construction of a decision tree with recursive partitioning and pruning based on n sample data of the d property [16]. (**a**) Search for decision boundary to minimize data impurity; (**b**) make branches out with decision boundaries; (**c**) remove excessive branches according to user-set standard.

Figure 5 shows RF composed of multiple decision trees. First, subsequent multiple data groups (k) are constructed by random selection from the entire data pool, known as bootstrapping (Figure 5b). Each decision tree (Figure 4) is then trained according to the multiple data groups (Figure 5c), yielding different looks due to the different data compositions from bootstrapping. In bootstrap aggregating, an output (saturation) is estimated based on where the given value (normalized CT image) is assigned to (Figure 5d,e), indicated according to the multiple decision trees. According to the principle of RF, the larger the number of decision trees, the more stable the performance shown by RF. Typically, the number of decision trees is set on the level of a couple of hundreds, at most thousands. Under the circumstances of given computing power (Intel Xeon Gold 6136 central processing unit with 3 and 2.99 GHz processors and 128 GB random-access memory), a couple of hundred trees is sufficiently affordable. Besides, the computational cost issue is out of scope from this study. Therefore, that issue will not be handled in detail.

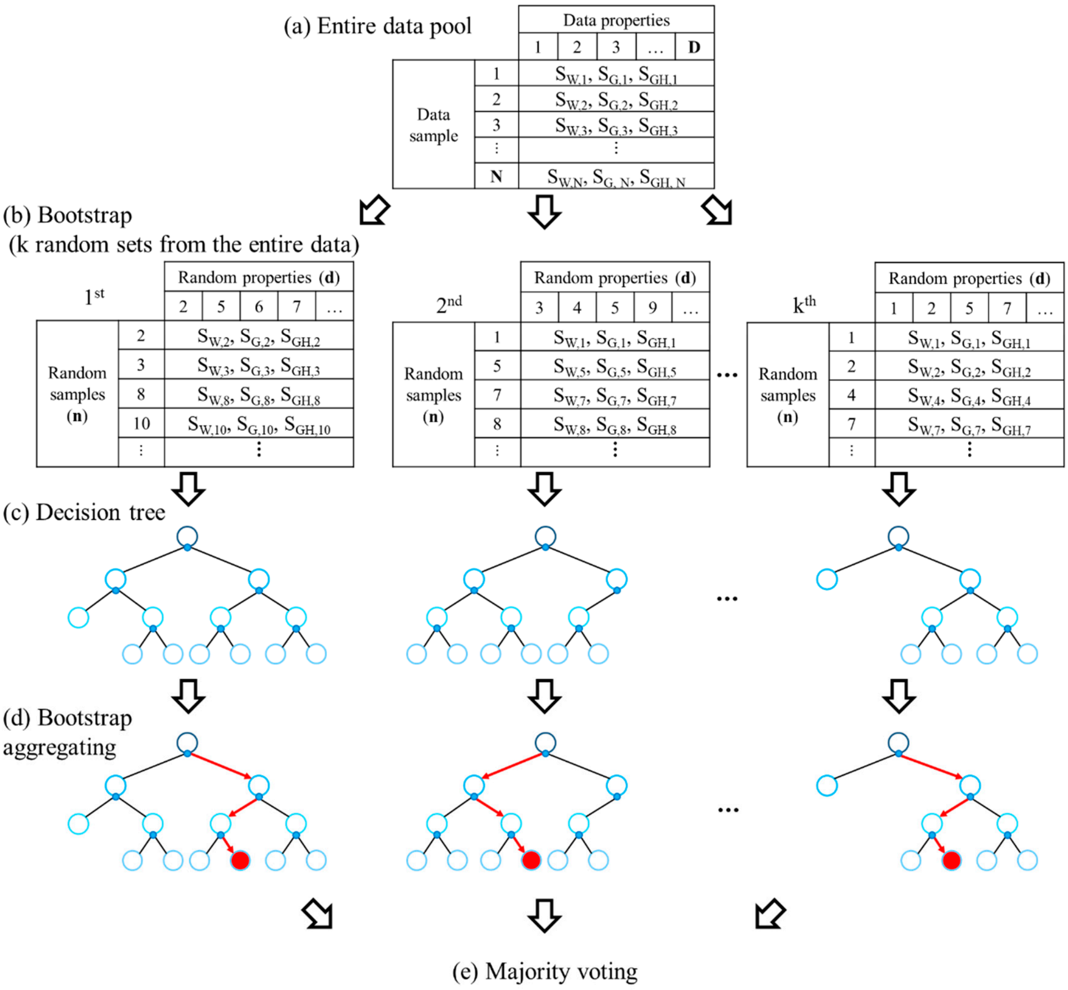
(a) Entire data pool
Data properties
1 | 2 | 3 | … | D
Data sample
1 $S_{W,1}, S_{G,1}, S_{GH,1}$
2 $S_{W,2}, S_{G,2}, S_{GH,2}$
3 $S_{W,3}, S_{G,3}, S_{GH,3}$
⋮ ⋮
N $S_{W,N}, S_{G,N}, S_{GH,N}$
(b) Bootstrap
(k random sets from the entire data)
1st
Random properties (d)
2 | 5 | 6 | 7 | …
Random samples (n)
2 $S_{W,2}, S_{G,2}, S_{GH,2}$
3 $S_{W,3}, S_{G,3}, S_{GH,3}$
8 $S_{W,8}, S_{G,8}, S_{GH,8}$
10 $S_{W,10}, S_{G,10}, S_{GH,10}$
⋮ ⋮
2nd
Random properties (d)
3 | 4 | 5 | 9 | …
Random samples (n)
1 $S_{W,1}, S_{G,1}, S_{GH,1}$
5 $S_{W,5}, S_{G,5}, S_{GH,5}$
7 $S_{W,7}, S_{G,7}, S_{GH,7}$
8 $S_{W,8}, S_{G,8}, S_{GH,8}$
⋮ ⋮
…
kth
Random properties (d)
1 | 2 | 5 | 7 | …
Random samples (n)
1 $S_{W,1}, S_{G,1}, S_{GH,1}$
2 $S_{W,2}, S_{G,2}, S_{GH,2}$
4 $S_{W,4}, S_{G,4}, S_{GH,4}$
7 $S_{W,7}, S_{G,7}, S_{GH,7}$
⋮ ⋮
(c) Decision tree
…
(d) Bootstrap aggregating
…
(e) Majority voting

Figure 5. Random forest with multiple decision trees for the prediction of the target parameters [16]. (**a**) The whole data with properties of D and data sample of N; (**b**) random data choice to construct data groups of k; (**c**) generation of k multiple trees from each of data groups; (**d**) summation of decisions according to multiple trees; (**e**) final decision making for given input.

3. Construction of Training Data

Figure 6 shows an example of normalized CT values and their distribution in the five experimental stages. In the DRY and SAT stages, only 0 or 1 values are present for all normalized CT values according to Equation (1) and Figure 3 [16]. On the contrary, individual distributions in the IWS, GH, and GTW stages are present, shown in histogram form as Figure 6b. The GH sample is marked with the yellow dotted circles in three of the stages and each histogram covers only that circle area. From IWS to GTW, the averaged, normalized CT values become larger than the previous stage because the overall density increases corresponding to the experiment design due to the five stages (Figure 2). The normalized CT values become closer to 0.9–1, similar to the density of GH and water (Table 1).

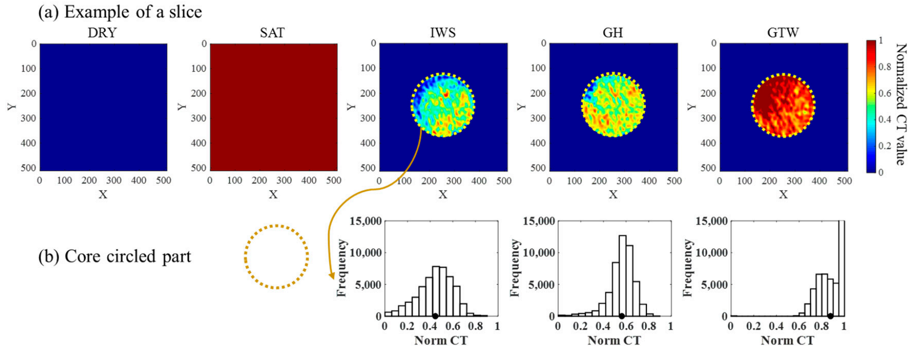

Figure 6. Normalized CT values in the first slice of the fourth GH sample presented with a color scale. (**a**) Normalized CT values for each experimental stage; (**b**) histogram of core circled part for the three experimental stages—IWS, GH, and GTW.

Figure 7 describes the eight total GH samples for the experiment and their distributions of normalized CT values of the GH stage. The first to third samples are categorized into the low GH saturation group, while the fourth to eighth samples are assigned to the high GH saturation group. There was no difficulty with the separation into these two groups because of the obvious difference of S_{GH} values—whereby values were around 40% and 50%, respectively. Figure 7b presents the distributions of normalized CT values for the first slice of eight GH samples, with black dots indicating the means of each distribution. Generally, the low GH saturation group has low averages of the normalized CT values and the high GH saturation group gives higher average values. We can distinguish S_{GH} values into the forty- and fifty-percent groups which are taken as the conventional values and critical values, respectively, because $S_{GH,C}$ was expected to be ~50% for the target GH sample in this study. For values close to that of $S_{GH,C}$, production efficiency of GH can be drastically reduced due to the slow pressure propagation in porous media.

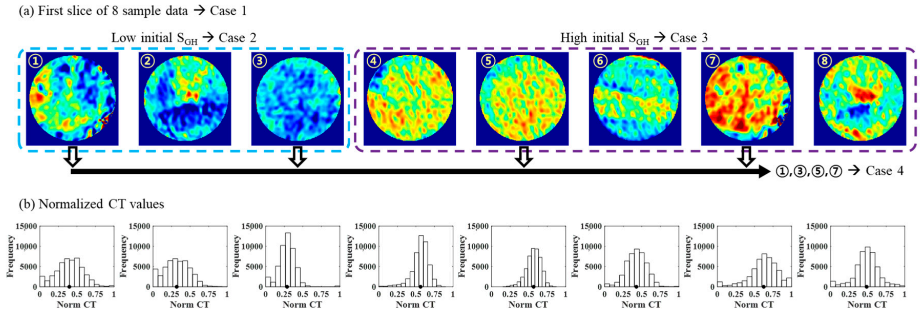

Figure 7. Normalized CT values of the eight GH samples in the "GH" stage for construction of the four cases. (**a**) Combination of the samples for each case; (**b**) normalized CT values of the first slice image in histogram.

The eight GH samples are randomly indexed after they are categorized into the two groups. In this study, the total eight GH samples are divided into four cases (Cases 1–4) for analysis of data construction and machine-learning performance. Case 1 represents the total eight GH samples and Cases 2 and 3 are the low and high S_{GH}, respectively. Case 4 consists of the odd-numbered GH samples—1, 3, 5, and 7—for randomly constructed data collection with low and high S_{GH} evenly. Case 4 is set to have the four selected GH samples in order to produce the fairest comparison with Cases 2 and 3 by fitting to the amount of training data.

Table 2 shows averaged GH saturation and CT values of each GH sample and Cases 1–4. Orange-colored cells indicate the utilized GH sample for each of the four cases. The means that Cases 1 and 4 are similar to each other because they are composed of combined GH samples from the low and high S_{GH} groups. The difference between averaged S_{GH} of Cases 2 and 3 is ~11%, which could give significant contrast of GH behaviors.

Table 2. Gas hydrate (GH) saturation and computerized tomography (CT) values in the "GH" stage for eight GH samples.

GH Sample No.	S_{GH}, %				
		Average			
	Each	**Case 1**	**Case 2**	**Case 3**	**Case 4**
		48.6	**41.5**	**52.9**	**47.4**
1	40.7	🟨	🟨	☒	🟨
2	43.8	🟨	🟨	☒	☒
3	40.0	🟨	🟨	☒	🟨
4	53.3	🟨	☒	🟨	☒
5	54.3	🟨	☒	🟨	🟨
6	50.5	🟨	☒	🟨	☒
7	54.4	🟨	☒	🟨	🟨
8	52.0	🟨	☒	🟨	☒

Figure 8 and Table 3 explain how the training and test data are divided and constructed for Cases 1–4. Figure 8 illustrates that four test sets are randomly selected from each data pool as a conventional machine-learning procedure. Case 1 uses all eight GH samples, and its test set consists of a random 10% of these. This test set of Case 1 is set as the universal test set, which is utilized as the common standard to compare all four cases with each other. The test sets for Cases 2, 3, and 4 are random 10% sets from each entire pool, represented by green, orange, and blue colored lines, respectively, in the figure. Training of RF is conducted for all four cases and the number of used training data is presented in Table 3. The eight GH samples have 320 data points due to the multiplication of 5 stages and 64 slices (Figures 2 and 3). Case 1 has 2560 due to the multiplication of 8 GH samples and 320 slices. The number of training data should then be 2304, which is 2560 subtracted by 256, and the same procedure is carried out for the rest of the three cases. The detailed training conditions of RF are shown in Table 3. Regarding number of properties, 219 is determined by the root of the total number of CT values, 47,996.

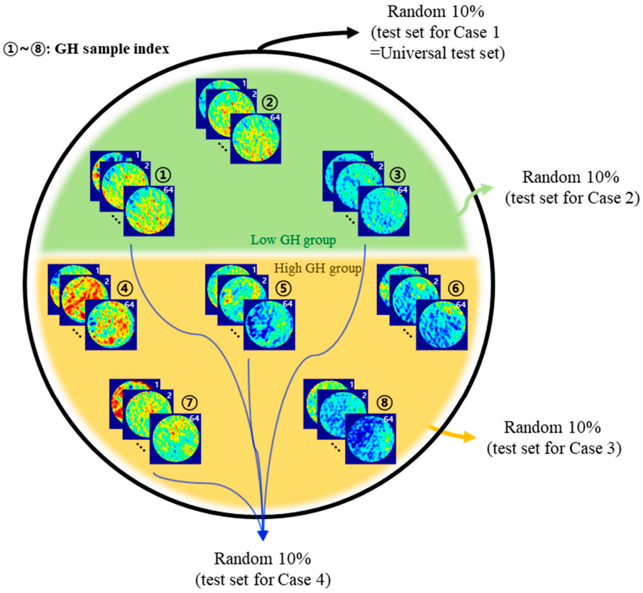

Figure 8. Two test sets: random 10% test from each of the four cases and universal test set from the eight GH samples.

Table 3. Conditions and settings for Cases 1–4 with random forest.

Method	List	Case 1	Case 2	Case 3	Case 4
Number of data	Entire data	2560	960	1600	1280
	Training data	2304	864	1440	1152
	Test data (10% of the entire data)	256	96	160	128
Training condition of RF	Maximum depth	10			
	Number of trees	200			
	Number of properties	219			
	Data size of one sample	47,996			

4. Results and Discussion

Figures 9–12 are the results of RF for the training and test sets, respectively. Figures 9a, 10a, 11a and 12a are the training and Figures 9b, 10b, 11b and 12b are the test set, while each column lists S_W, S_{GH}, and S_G in order. In both (a) and (b) panels, the first row is the scatterplot and the second row displays the same results in a histogram. In the first row, the X-axis indicates the original value of saturations and the Y-axis indicates the predicted (modeled) values. Blue dots indicate individual data samples and increasing darkness of the color indicates increasingly scattered data relative to a certain position. Therefore, although some of data seem to be deviated from the diagonal line, it can give a high coefficient of determination (S_W of Figure 9b). Correlation coefficients are calculated and presented at the top of all charts. In the histograms of the second row, the blue dotted box means the predicted saturation values and the red solid-line box presents the original saturations.

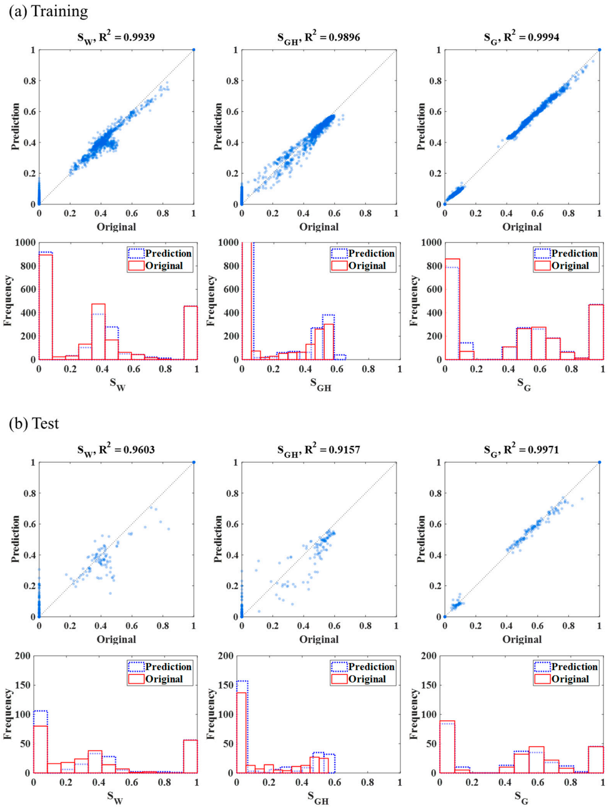

Figure 9. (**a**) Training and (**b**) test results of Case 1, each column means each saturation, water, gas hydrate, and gas with coefficient of determination. In the scatter plots, the X and Y axes indicate the given original data and the predicted value by random forest, respectively. In the histogram pictures, the X and Y axes mean each saturation and frequency, respectively.

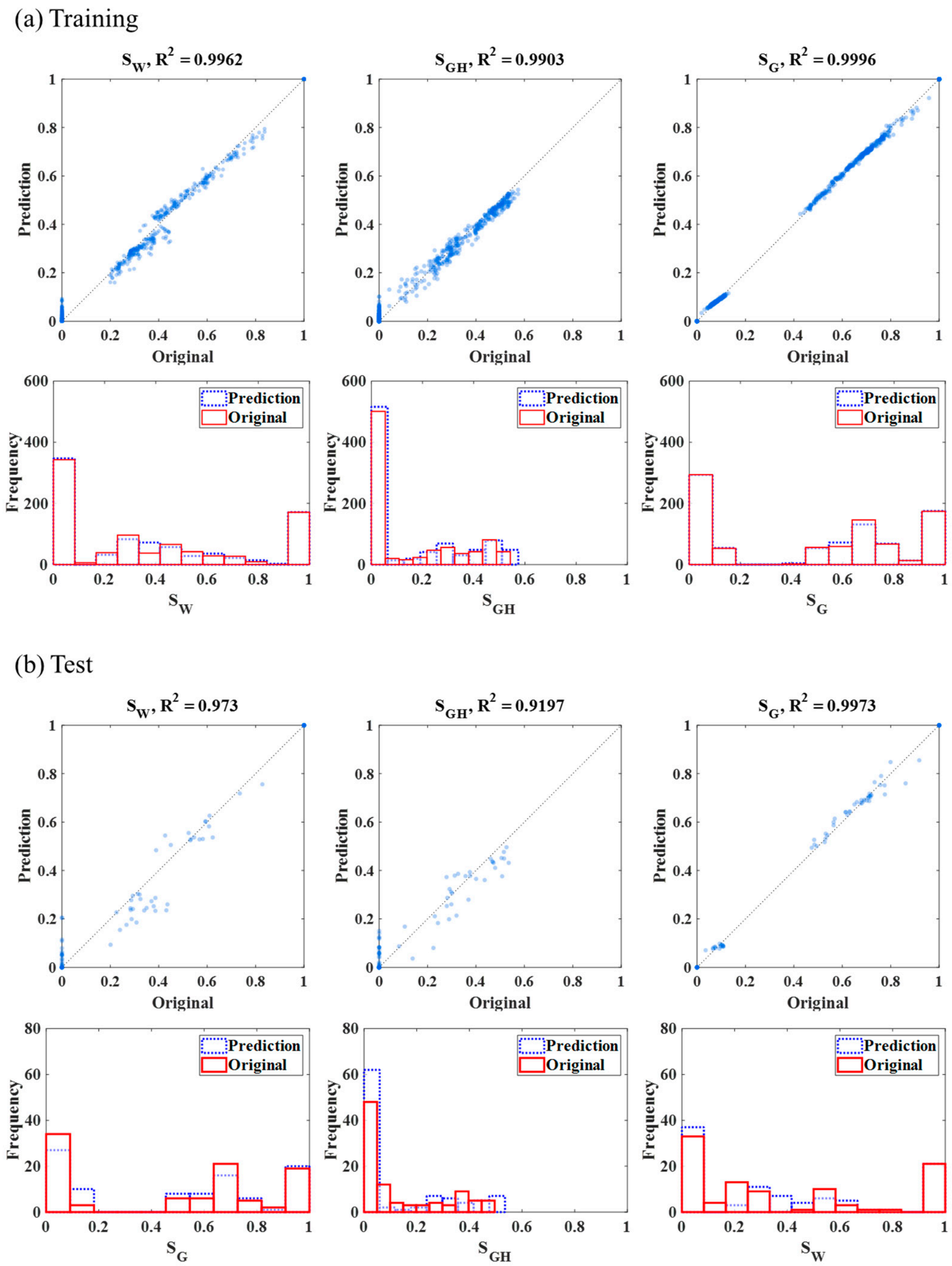

Figure 10. (**a**) Training and (**b**) test results of Case 2, each column means each saturation, water, gas hydrate, and gas with coefficient of determination. In the scatter plots, the X and Y axes indicate the given original data and the predicted value by random forest, respectively. In the histogram pictures, the X and Y axes mean each saturation and frequency, respectively.

Figure 9 illustrates the largest number and widest range of data because all eight GH samples are included. The coefficient of determination (R^2) of the training data is ~0.99 for all of the variables—S_W, S_{GH}, and S_G. In particular, S_G shows considerable fitting between the original and the predicted values in comparison with S_W and S_{GH}. This is because the density of water and GH are relatively similar, which leads to little difference in X-ray CT images and normalized CT values. However, the density of gas is much lower than that of water or GH, thereby causing certain changes in CT values. The certain difference of densities between gas and the others causes a clear discrepancy between normalized CT values of gas and the others. It makes the prediction of gas saturation easier than the prediction of water and GH saturations. This phenomenon was identified in our previous study [16].

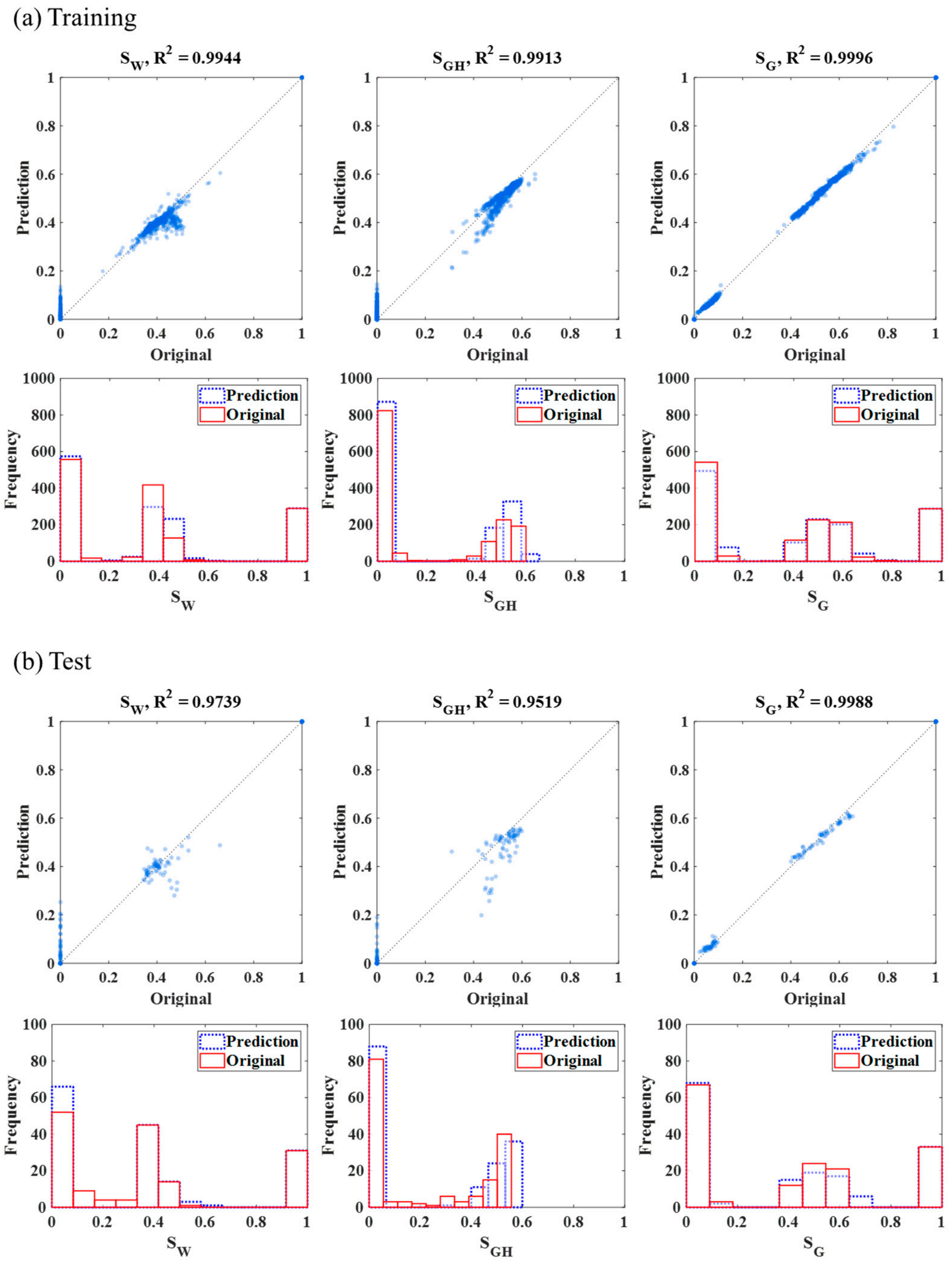

Figure 11. (**a**) Training and (**b**) test results of Case 3, each column means each saturation, water, gas hydrate, and gas with coefficient of determination. In the scatter plots, the X and Y axes indicate the given original data and the predicted value by random forest, respectively. In the histogram pictures, the X and Y axes mean each saturation and frequency, respectively.

Figures 10 and 11 contrast each other in terms of S_{GH}. It was expected for these charts to show the different distribution of $\mathrm{S_{GH}}$ because they are separated with S_{GH} criteria, showing the S_{GH} results. In both Cases 2 and 3, the overall machine-learning performances are suitable considering that all R^2 values are greater than 0.99 and the scattered dots are positioned on the diagonal line in an orderly fashion. However, in terms of S_{GH}, Case 2 has a relatively wide range of 0–0.6, whereas Case 3 mostly shows either 0 or ~0.5. In Figure 11, the scattered dots are deviated from the diagonal line especially near 0.4, which is a comparably low S_{GH}. It would be expected that Case 3 was trained for large S_{GH} values compared to Case 2, which is the reason why Case 3 functions this way according to the given data composition.

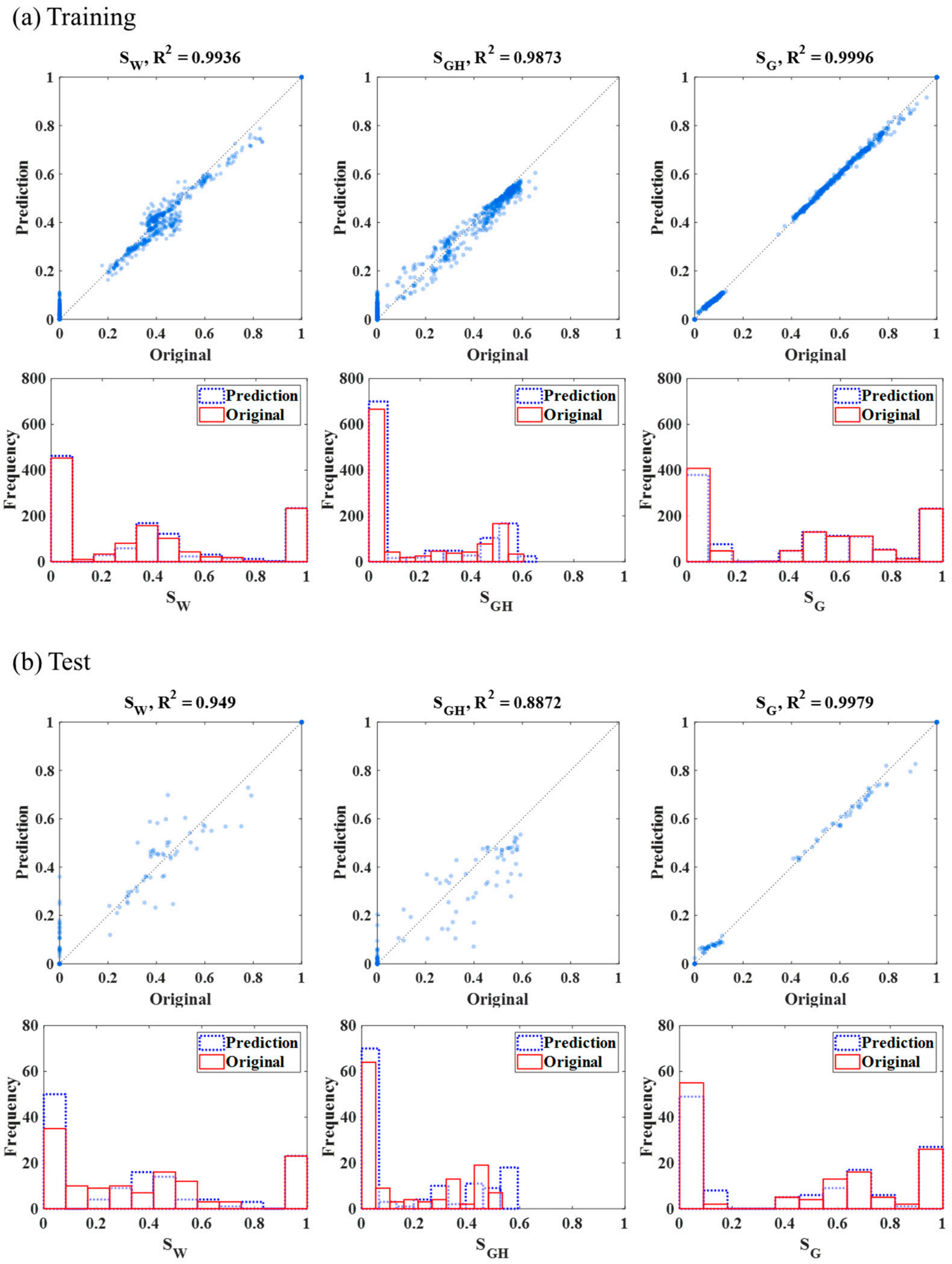

Figure 12. (**a**) Training and (**b**) test results of Case 4, each column means each saturation, water, gas hydrate, and gas with coefficient of determination. In the scatter plots, the X and Y axes indicate the given original data and the predicted value by random forest, respectively. In the histogram pictures, the X and Y axes mean each saturation and frequency, respectively.

Figure 12 shows overall decent performance except for the test set of S_{GH} whose R^2 is 0.89, the lowest value. Case 1 has the smallest R^2 value of 0.92 in the test set for S_{GH}. In Cases 2 and 3, S_{GH} shows the smallest R^2, seemingly indicating that the most challenging component of the process is the identification of S_{GH} for the test sets. The density of water is maintained at approximately 1 g/cc regardless of experimental stage. However, the density of GH highly depends on the given pressure and temperature of the three experimental stages—IWS, GH, and GTW (i.e., the last column of Table 1). Therefore, S_{GH} differently affects the normalized CT values according to these experimental stages. This phenomenon of S_{GH} could further lead to more complex relationships between $\mathrm{S_{GH}}$ and the normalized CT values, and consequently, results in higher difficulty of the machine-learning training. For this reason, the methodology introduced in this study should continue to be conducted.

In Figures 13 and 14, the four trained RF models correspond to the four cases, and those four RF models are tested using the universal test. The test results shown in Figures 13a and 14a are identical with those of Figure 9b. In most machine-learning-related studies, its trained performance is mainly evaluated with errors and correlation coefficients between original and predicted data in the test data set. In this study, the four, learned, RF models are tested together with the universal test set for a consistent analysis. According to one previous study, it was estimated that $S_{GH,C}$ might be somewhere between approximately 50–60% [31].

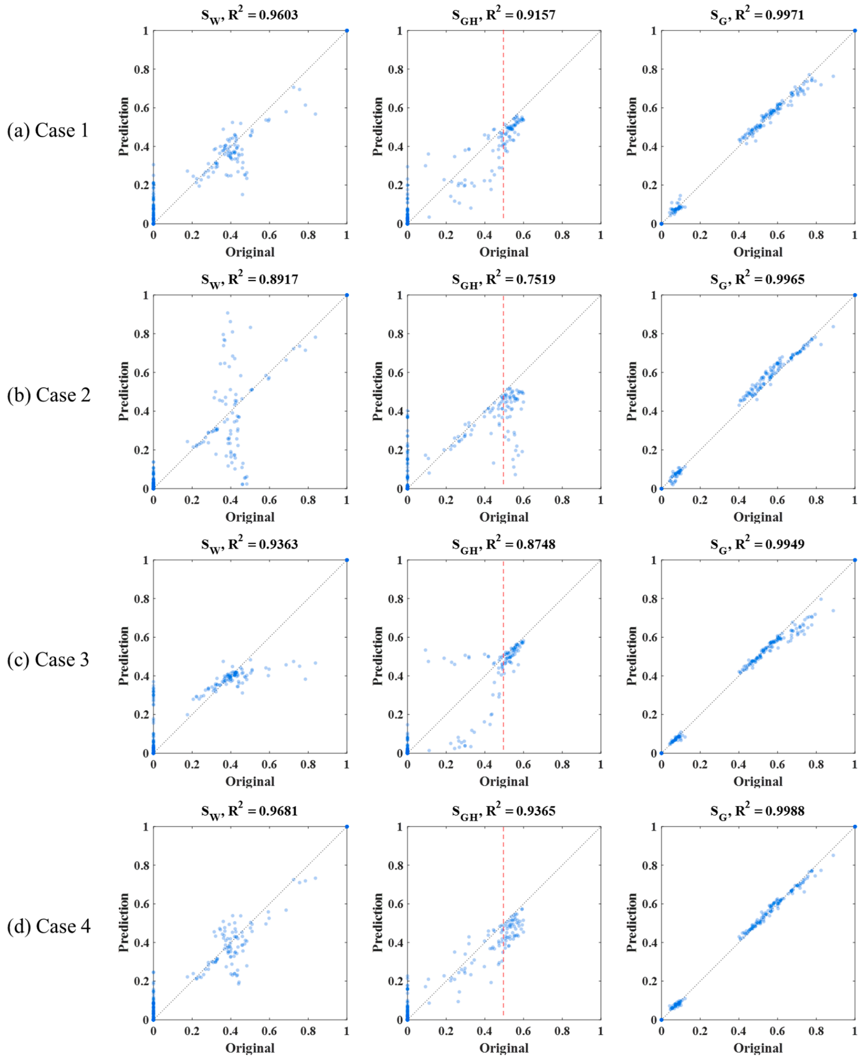

Figure 13. Universal test set results of Cases 1–4 corresponding to (**a**)–(**d**) in order. (**a**) Case 1 of the total eight GH samples; (**b**) Case 2 of the three low S_{GH} samples; (**c**) Case 3 of the five high S_{GH} samples; (**d**) Case 4 composed of the two low S_{GH} and the two high S_{GH} samples; each column represents each saturation, water, gas hydrate, and gas with coefficient of determination; the X and Y axes indicate the given original data and the predicted value by random forest; the red dotted line is drawn at 50% of S_{GH} to clarify given data trend.

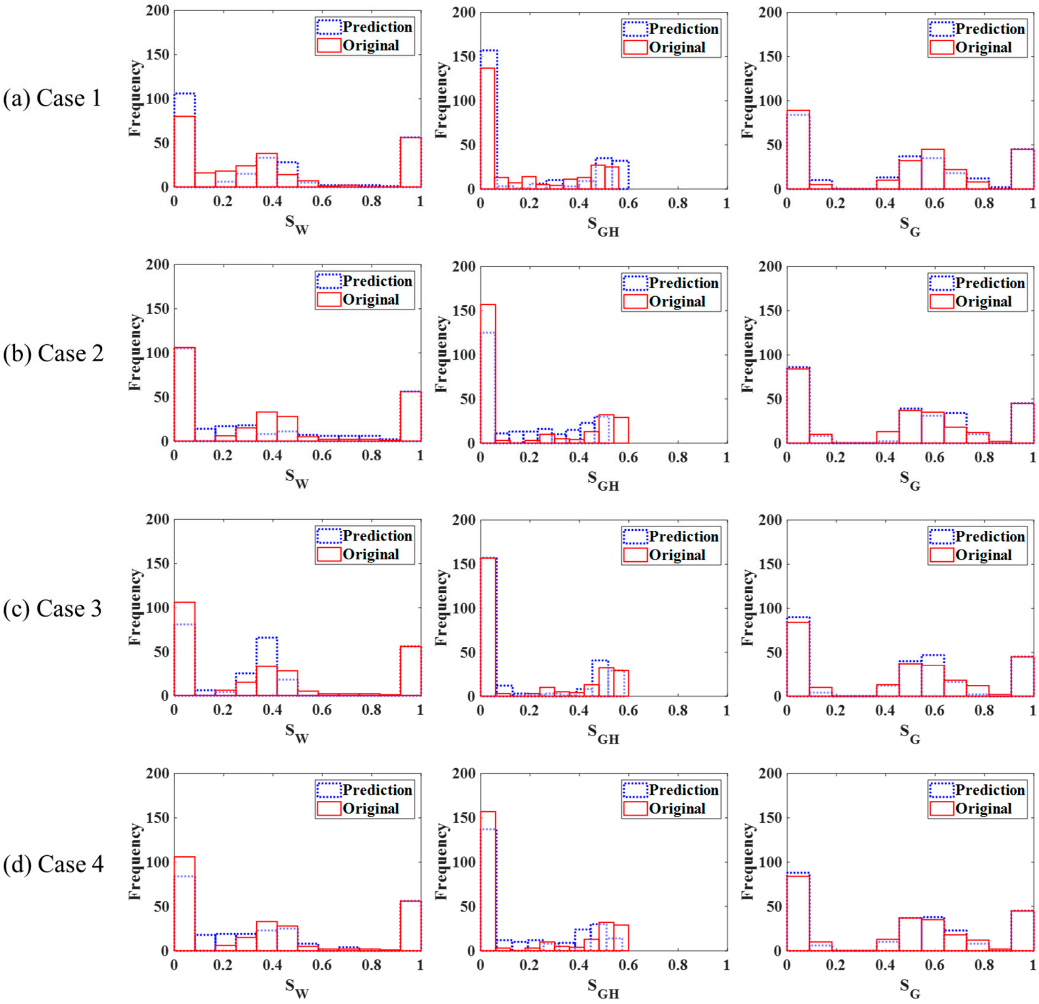

Figure 14. Universal test set results of Cases 1–4 in histogram corresponding to (**a**)–(**d**) in order. (**a**) Case 1 of the total eight GH samples; (**b**) Case 2 of the three low S_{GH} samples; (**c**) Case 3 of the five high S_{GH} samples; (**d**) Case 4 composed of the two low S_{GH} and the two high S_{GH} samples; each column means each saturation, water, gas hydrate, and gas with coefficient of determination; the X axis is saturation value of each phase and the Y axis indicates the frequency corresponding to saturation values. The red solid line box is the original data, and the blue dotted box is the predicted result.

Interestingly, according to the $\mathrm{S_{GH}}$ results shown in Figure 13b,c, an obvious boundary is shown at ~50% S_{GH}, where the red dotted lines indicate the validation as to whether there is any trend related to $S_{GH,C}$. In Figure 13b, the S_{GH} values over 50% are poorly matched in performance compared to Figure 13a,c,d. On the other hand, to the left of the red line in the S_{GH} results shown in Figure 13b, the S_{GH} values less than 50% have comparatively good fitting results. This is further emphasized when viewing the left of the red line for $\mathrm{S_{GH}}$ results of Figure 13c.

Furthermore, it should be noted that Case 2 shows deviated results for over 50% S_{GH} (Figures 13b and 14b), even though some training data nearby showed 50% S_{GH} (Figure 10a). This is an indication of an additional effect other than the range of training data values, "critical gas saturation". Thus, we can expect certain distinguishing behaviors of GH samples according to the different $S_{GH,C}$ values setting. Although there could be some S_{GH} values near 40% or 50% in one specific GH sample, the trend of GH behaviors would highly depend on the decided $S_{GH,C}$ as an experimental condition. Based on that possibility, we can infer that there must be a certain radical change of GH behavior from $S_{GH,C}$, which is evaluated as ~50% in this study.

Cases 2, 3, and 4 are relatively comparable to each other in terms of the absolute number of training data—864, 1440, and 1152, respectively (Table 3)—all of which are close to the value of 1000. Considering this, Case 4 has relatively less-biased results of S_{GH} compared to Cases 2 and 3 (Figure 13d). On both the left and right sides of the red line, data are generally positioned following the diagonal line.

Table 4 organizes the mean square error (MSE) results of Cases 1–4 for both the training and test sets and the universal test set. The MSEs are computed as follows:

$$MSE = \frac{1}{n}\sum\nolimits_{i=1}^{n}(Orig_i - Pred_i)^2, \tag{4}$$

where n is the number of training or test data, $Orig_i$ is the original ith data, and $Pred_i$ represents the predicted data for the ith original data.

Table 4. Mean square error (MSE) of training and test sets for Cases 1–4.

<table>
<tr><th rowspan="2">(×10^{−4})</th><th colspan="3">Training</th><th colspan="3">Random 10% Test</th><th colspan="3">Universal Test</th></tr>
<tr><th>S_w</th><th>S_{GH}</th><th>S_G</th><th>S_w</th><th>S_{GH}</th><th>S_G</th><th>S_w</th><th>S_{GH}</th><th>S_G</th></tr>
<tr><td>Case 1: 8 total</td><td>9</td><td>7</td><td>0.8</td><td>61</td><td>52</td><td>4</td><td>61</td><td>52</td><td>4</td></tr>
<tr><td>Case 2: 3 low S_{GH}</td><td>5.6</td><td>4.8</td><td>5.5</td><td>41</td><td>29</td><td>4</td><td>170</td><td>143</td><td>5.5</td></tr>
<tr><td>Case 3: 5 high S_{GH}</td><td>8</td><td>7</td><td>0.5</td><td>38</td><td>38</td><td>1.8</td><td>100</td><td>71</td><td>8</td></tr>
<tr><td>Case 4: 2 + 2 S_{GH}</td><td>9.6</td><td>8.6</td><td>0.6</td><td>75</td><td>71</td><td>3.7</td><td>49</td><td>44</td><td>1.6</td></tr>
<tr><td></td><td colspan="9">Average of S_W, S_{GH}, and S_G</td></tr>
<tr><td>Case 1: 8 total</td><td colspan="3">5.6</td><td colspan="6">39</td></tr>
<tr><td>Case 2: 3 low S_{GH}</td><td colspan="3">5.3</td><td colspan="3">24.7</td><td colspan="3">106.2</td></tr>
<tr><td>Case 3: 5 high S_{GH}</td><td colspan="3">5.2</td><td colspan="3">25.9</td><td colspan="3">60.0</td></tr>
<tr><td>Case 4: 2 + 2 S_{GH}</td><td colspan="3">6.3</td><td colspan="3">49.9</td><td colspan="3">31.5</td></tr>
</table>

Table 4 presents the MSEs corresponding to each data set, fluid phase, and case, and also shows the averaged MSEs for an overall comparison of training data, random 10% data, and the universal test. In terms of each fluid phase, S_G has the lowest errors among the three phase saturations. As shown in Figures 9–12 and Figure 13, S_W and S_{GH} have the larger MSEs. As conventional machine-learning shows, the MSEs of the training data are lower than those of the two test sets.

Meaningful lessons can be understood from the comparison between all four cases. First, the absolute number of data substantially affects machine-learning performance (Cases 1 and 4). Typically, a higher number of training data would be expected to have better performance, as long as other conditions such as features and algorithms are sufficiently appropriate [35,36]. However, Cases 1 and 4 have two times the difference in the number of data; however, the MSEs are nearly in the same scale without critical difference (39 and 31.5 in the universal test). Therefore, the results of Cases 1 and 4 of this study indicate that distributions of these cases must be similar to each other, such that they also produce similar-scaled MSEs. This indicates that the process of how data is constructed is as important as the absolute number of data for machine-learning performance. Second, if it were possible to obtain a limited number of data, it would be ideal to focus on the specific S_{GH} range. Cases 2 and 3 show lower MSEs compared to Cases 1 and 4 in random 10% test with the similar number of training data, which means specialized targeted data construction can be strategically advantageous for machine-learning performance. Third, the ratio of MSEs from Cases 2 and 3 is about 5:3, which is similar to a ratio of the rest of the data for Cases 2 and 3, respectively—in that, Case 2 has only three GH samples among the total eight GH samples and the rest has five GH samples. For Case 3, the rest of the data has three GH samples. The less related the data, the larger the MSE.

5. Conclusions

This paper proposed the saturation identification of water, GH, and the gas phase in GH samples based on a machine-learning method in consideration of $S_{GH,C}$. Moreover, the effects of training data quantity and quality were analyzed for RF utilization in the four cases. Compared to our previous related study [16], this study utilized five additional GH samples, whose number of data was 1600. Owing to this extra data, we could categorize samples into low and high S_{GH} groups and determine how the number of data and $S_{GH,C}$ affect the overall machine-learning performance.

This study validated the significant influence of $S_{GH,C}$ in cases where training data consists of low and high S_{GH} groups. The average MSE differences of random 10% test (Table 4) between Cases 1 and 4 (10.9) was larger than that between Cases 2 and 3 (1.2), indicating that S_{GH} can be a highly important standard for saturation identification in GH formation and dissociation experiments. In particular, $S_{GH,C}$ can be an important criterion to divide training data when any machine-learning technique is applied to given CT images in a GH experiment (refer to Cases 2 and 3). Thus, the separation of CT images according to $S_{GH,C}$ can be an appropriate option for constructing training data, leading to obtaining reliably specialized machine-learning models.

In conclusion, it is important to acquire a sufficiently high number of data in order to carry out trustworthy application of machine-learning; however, proper data construction should also be considered. It was expected that one specific standard for data building would be identified from the essential factors of interested behaviors based on domain knowledge, and it was verified to be $S_{GH,C}$ from this study. Therefore, if obtainment of data was restricted to some specific type or quantity of data, the first order of business would be selection of GH experiment to be conducted first according to a value of S_{GH}. After that, GH experiments could be intensively performed to preferentially obtain data of CT images and saturations based on a target field condition of $S_{GH,C}$. Accordingly, $S_{GH,C}$ would be the optimal guideline for training data building.

In future studies, additionally acquired GH CT images would be assigned to one of low or high S_{GH} groups whose criterion is $S_{GH,C}$. According to that, two machines could then be trained with those two categorized data, respectively, so as to produce two customized machine-learning models. After construction of the reliable machine-learning models based on qualitatively and quantitatively sufficient data, those models could be utilized to identify saturation values during the dissociation stage of GH sample experiments with depressurization. Saturation identification of GH samples in real time is expected to be a powerful tool to help determine general GH behaviors and conduct a variety of experiments for optimization of the parameters of GH production by depressurization.

Author Contributions: Conceptualization, K.L. and T.A.; methodology, S.K., K.L., and M.L.; validation, S.K., K.L., and M.L.; formal Analysis, S.K., K.L., and M.L.; investigation, S.K., K.L., and T.A.; writing–original draft preparation, S.K.; writing–review & editing, S.K., K.L., and M.L.; project administration, S.K. and T.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the MINISTRY OF TRADE, INDUSTRY AND ENERGY (MOTIE), Korea, the GAS HYDRATE R&D ORGANIZATION (GHDO), and the KOREA INSTITUTE OF GEOSCIENCE AND MINERAL RESOURCES (KIGAM), grant number GP2020-037. K. Lee was supported by the Energy Efficiency & Resources of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the KOREA GOVERNMENT MINISTRY OF TRADE, INDUSTRY & ENERGY, grant number 20162010201980.

Acknowledgments: The authors are grateful to Hwasoo Suk and Jaehyung Lee for providing the experimental data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kumar, P.; Collett, T.S.; Shukla, K.M.; Yadav, U.S.; Lall, M.V.; Vishwanath, K.; Yamada, Y. India National Gas Hydrate Program Expedition-02: Operational and technical summary. *Mar. Petrol. Geol.* **2019**, *108*, 3–38. [CrossRef]
2. Li, J.; Ye, J.; Qin, X.; Qiu, H.J.; Wu, N.Y.; Lu, H.L.; Xie, W.W.; Lu, J.A.; Peng, F.; Xu, Z.Q.; et al. The first offshore natural gas hydrate production test in South China Sea. *China Geol.* **2018**, *1*, 5–16. [CrossRef]
3. Haines, S.S.; Hart, P.E.; Collett, T.S.; Shedd, W.; Frye, M.; Weimer, P.; Boswell, R. High-resolution seismic characterization of the gas and gas hydrate system at Green Canyon 955, Gulf of Mexico, USA. *Mar. Petrol. Geol.* **2017**, *82*, 220–237. [CrossRef]
4. Chong, Z.R.; Yang, S.H.B.; Babu, P.; Linga, P.; Li, X.-S. Review of natural gas hydrates as an energy resource: Prospects and challenges. *Appl. Energy* **2016**, *162*, 1633–1652. [CrossRef]

5. Ito, T.; Komatsu, Y.; Fujii, T.; Suzuki, K.; Egawa, K.; Nakatsuka, Y.; Konno, Y.; Yoneda, J.; Jin, Y.; Kida, M.; et al. Lithological features of hydrate-bearing sediments and their relationship with gas hydrate saturation in the eastern Nankai Trough, Japan. *Mar. Petrol. Geol.* **2015**, *66*, 368–378. [CrossRef]
6. Lee, J.Y.; Jung, J.W.; Lee, M.H.; Bahk, J.J.; Choi, J.; Ryu, B.J.; Schultheiss, P. Pressure core based study of gas hydrates in the Ulleung Basin and implication for geomechanical controls on gas hydrate occurrence. *Mar. Petrol. Geol.* **2013**, *47*, 85–98. [CrossRef]
7. Huh, D.; Lee, J.Y. Overview of gas hydrates R&D. *J. Korean Soc. Miner. Energy Resour. Eng.* **2017**, *54*, 201–214.
8. Boswell, R.; Schoderbek, D.; Collett, T.S.; Ohtsuki, S.; White, M.; Anderson, B.J. The Iġnik Sikumi field experiment, Alaska North Slope: Design, operations, and implications for CO_2-CH_4 exchange in gas hydrate reservoirs. *Energy Fuel.* **2017**, *31*, 140–153. [CrossRef]
9. Koh, D.Y.; Kang, H.; Lee, J.W.; Park, Y.; Kim, S.J.; Lee, J.; Lee, J.Y.; Lee, H. Energy-efficient natural gas hydrate production using gas exchange. *Appl. Energy* **2016**, *162*, 114–130. [CrossRef]
10. Zhao, J.; Zhu, Z.; Song, Y.; Liu, W.; Zhang, Y.; Wang, D. Analyzing the process of gas production for natural gas hydrate using depressurization. *Appl. Energy* **2015**, *142*, 125–134. [CrossRef]
11. Anderson, B.J.; Kurihara, M.; White, M.D.; Moridis, G.J.; Wilson, S.J.; Pooladi-Darvish, M.; Gaddipati, M.; Masuda, Y.; Collett, T.S.; Hunter, R.B.; et al. Regional long-term production modeling from a single well test, Mount Elbert Gas Hydrate Stratigraphic Test Well, Alaska North Slope. *Mar. Petrol. Geol.* **2011**, *28*, 493–501. [CrossRef]
12. Tang, L.G.; Li, X.S.; Feng, Z.P.; Li, G.; Fan, S.S. Control mechanisms for gas hydrate production by depressurization in different scale hydrate reservoirs. *Energy Fuel.* **2007**, *21*, 227–233. [CrossRef]
13. Lee, M.; Suk, H.; Lee, J.; Lee, J. Quantitative analysis for gas hydrate production by depressurization using X-ray CT. In Proceedings of the 2018 Joint International Conference of the Geological Science & Technology of Korea, KSEEG, Busan, Korea, 17–20 April 2018; p. 363.
14. Wang, J.; Zhao, J.; Yang, M.; Li, Y.; Liu, W.; Song, Y. Permeability of laboratory-formed porous media containing methane hydrate: Observations using X-ray computed tomography and simulations with pore network models. *Fuel* **2015**, *145*, 170–179. [CrossRef]
15. Mikami, J.; Masuda, Y.; Uchida, T.; Satoh, T.; Takeda, H. Dissociation of natural gas hydrate observed by X-ray CT scanner. *Ann. N. Y. Acad. Sci.* **2006**, *912*. [CrossRef]
16. Kim, S.; Lee, K.; Lee, M.; Ahn, T.; Lee, J.; Suk, H.; Ning, F. Saturation modeling of gas hydrate using machine learning with X-ray CT images. *Energies* **2020**, *13*, 5032. [CrossRef]
17. Alhashem, M. Supervised machine learning in predicting multiphase flow regimes in horizontal pipes. In Proceedings of the Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, UAE, 11–14 November 2019. [CrossRef]
18. Singh, A.; Ojha, M.; Sain, K. Predicting lithology using neural networks from downhole data of a gas hydrate reservoir in the Krishna-Godavari basin, eastern Indian offshore. *Geophys. J. Int.* **2020**, *220*, 1813–1837. [CrossRef]
19. Kim, S.; Kim, K.H.; Min, B.; Lim, J.; Lee, K. Generation of synthetic density log data using deep learning algorithm at the Golden field in Alberta, Canada. *Geofluids* **2020**, *2020*, 5387183. [CrossRef]
20. Kim, S.; Lee, K.; Lim, J.; Jeong, H.; Min, B. Development of ensemble smoother-neural network and its application to history matching of channelized reservoir. *J. Petrol. Sci. Eng.* **2020**, *191*, 107159. [CrossRef]
21. Kim, S.; Min, B.; Kwon, S.; Chu, M. History matching of a channelized reservoir using a serial denoising autoencoder integrated with ES-MDA. *Geofluids* **2019**, *2019*, 3280961. [CrossRef]
22. Kim, S.; Min, B.; Lee, K.; Jeong, H. Integration of an iterative update of sparse geologic dictionaries with ES-MDA for history matching of channelized reservoir. *Geofluids* **2018**, *2018*, 1532868. [CrossRef]
23. Chen, B.; Harp, D.R.; Lin, Y.; Keating, E.H.; Pawar, R.J. Geologic CO_2 sequestration monitoring design: A machine learning and uncertainty quantification based approach. *Appl. Energy* **2018**, *225*, 332–345. [CrossRef]
24. Esmaili, S.; Mohaghegh, S.D. Full field reservoir modeling of shale assets using advanced data-driven analytics. *Geosci. Front.* **2016**, *7*, 11–20. [CrossRef]
25. Lee, K.; Lim, J.; Yoon, D.; Jung, H. Prediction of shale gas production at Duvernay Formation using deep-learning algorithm. *SPE J.* **2019**, *24*, 2423–2437. [CrossRef]
26. Kim, J.; Kim, S.; Park, C.; Lee, K. Construction of prior models for ES-MDA by a deep neural network with a stacked autoencoder for predicting reservoir production. *J. Petrol. Sci. Eng.* **2020**, *187*, 106800. [CrossRef]

27. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
28. Kim, T.; Kim, S.; Lim, J. Modeling and prediction of slug characteristics utilizing data-driven machine-learning methodology. *J. Petrol. Sci. Eng.* **2020**, *195*, 107712. [CrossRef]
29. Such, F.P.; Peri, D.; Brockler, F.; Hutkowski, P.; Ptucha, R.; Alaris, K. Fully convolutional networks for handwriting recognition. In Proceedings of the 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, 5–8 August 2018; pp. 86–91. [CrossRef]
30. Gil, S.M.; Shin, H.J.; Lim, J.S.; Lee, J. Numerical analysis of dissociation behavior at critical gas hydrate saturation using depressurization method. *J. Geophys. Res. Sol. Ea.* **2019**, *124*, 1222–1235. [CrossRef]
31. KIGAM. *Gas Hydrate Exploration and Production Study*; GP2016-027-2017(2); KIGAM: Daejeon, Korea, 2017; pp. 164–199.
32. Ta, X.H.; Yun, T.S.; Muhunthan, B.; Kwon, T. Observations of pore-scale growth patterns of carbon dioxide hydrate using X-ray computed microtomography. *Geochem. Geophys. Geosyst.* **2015**, *16*, 912–924. [CrossRef]
33. Waite, W.F.; Santamarina, J.C.; Cortes, D.D.; Dugan, B.; Espinoza, D.N.; Germaine, J.; Jang, J.; Jung, J.W.; Kneafsey, T.J.; Shin, H.; et al. Physical properties of hydrate-bearing sediments. *Rev. Geophys.* **2009**, *47*, RG4003. [CrossRef]
34. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
35. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55*, 78–87. [CrossRef]
36. Kang, B.; Lee, K. Managing Uncertainty in Geological Scenarios Using Machine Learning-Based Classification Model on Production Data. *Geofluids* **2020**, *2020*, 8892556. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.