


Article

Application of Machine Learning in Predicting Formation Condition of Multi-Gas Hydrate

Zimeng Yu and Hailong Tian * 

Key Laboratory of Groundwater Resources and Environment, Ministry of Education, Jilin University, Changchun 130021, China; yuzm21@mails.jlu.edu.cn

* Correspondence: thl@jlu.edu.cn

Abstract: Thermodynamic models are usually employed to predict formation condition of hydrates. However, these thermodynamic models usually require a large amount of calculations to approach phase equilibrium. Additionally, parameters included in the thermodynamic model need to be calibrated based on the experimental data, which leads to high uncertainties in the predicted results. With the rapid development of artificial intelligence (AI), machine learning as one of sub-discipline has been developed and been widely applied in various research area. In this work, machine learning was innovatively employed to predict the formation condition of natural gas hydrates to overcome the high computation cost and low accuracy. Three data-driven models, Random Forest (RF), Naive Bayes (NB), Support Vector Regression (SVR) were tentatively used to determine the formation condition of hydrate formed by pure and mixed gases. Experimental data reported in previous work were taken to train and test the machine learning models. As a representative thermodynamic model the Chen–Guo (C-G) model was used to analyze the computational efficiency and accuracy of machine learning models. The comparison of results predicted by C-G model and machine learning models with the experimental data indicated that the RF model performed better than the NB and SVR models on both computation speed and accuracy. According to the experimental data, the average AADP calculated by the C-G model is 7.62 times that calculated by the RF model. Meanwhile, the average time costed by the C-G model is 75.65 times that by the RF model. Compared with the other two machine learning models, the RF model is expected to be used in predicting the formation condition of natural gas hydrate under field conditions.

Keywords: gas hydrate; machine learning model; thermodynamic model; equation of state; Random Forest



Citation: Yu, Z.; Tian, H. Application of Machine Learning in Predicting Formation Condition of Multi-Gas Hydrate. *Energies* **2022**, *15*, 4719. <https://doi.org/10.3390/en15134719>

Academic Editor: Andrea De Pascale

Received: 10 April 2022

Accepted: 9 May 2022

Published: 28 June 2022

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



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

“Gas hydrates” are crystalline solid inclusion compounds, composed of multiple cavities caging individual gas molecules. The structure of the gas hydrate is determined by two aspects: the size of the “guest” gas molecule and the composition of gas mixture. Three types of hydrate structures have been determined, i.e., sI, sII, and sH [1]. Commonly encaged gases are methane, ethane, propane, carbon dioxide, and hydrogen sulfide, etc. Natural gas hydrate is a non-stoichiometric clathrate compound formed by water molecules and methane molecules under certain temperature and pressure conditions. Back in 2007, hydrate was considered to be a potential unconventional energy resource [2]. Shortly afterwards, the natural gas hydrate that dominates the hydrate provides a practical solution to the demand for alternative energy. Exploration and development of unconventional natural gas resources is considered an important step in the development and utilization of hydrates.

Exploration and development of unconventional natural gas resources, such as shale gas and natural gas hydrate, are important to the challenges of energy security faced by the world. As an energy storage medium, one m³ natural gas hydrate contains as much as

164~180 m³ of natural gas (standard temperature and pressure (STP)). A total of 97% of hydrate resources are explored in the submarine and the rest is buried in permafrost [3,4]. The natural gas hydrates are metastable, and its stability will be affected by pressure and temperature [5]. Moreover, the formation of hydrate is affected not only by temperature and pressure, but also the guest gases and their combinations [6]. Therefore, to predict the formation of hydrate formed by pure gases and their compositions is a challenging work.

Hydrate formation experiments are usually carried out to investigate the formation conditions of hydrates formed by pure gases and gas mixtures [7]. Additionally, based on the thermodynamic equilibrium theories, some thermodynamic statistical models were proposed to predict the hydrate stability conditions [8,9]. With the development of computing power, the thermodynamic models are increasingly employed to characterize the formation of hydrate due to their flexibility and liability. Among the thermodynamic models the vdW-P model [10] is the earliest and most widely used one to predict the hydrate formation conditions. Most of the existing thermodynamic models are derived from the vdW-P model, such as Parrish and Prausnitz [11], Ng and Robinson [12], Tohidi et al. [13], Chen and Guo [14], Klauda and Sandler [15], Ballard and Sloan [16], and Lee and Holder [17]. These thermodynamic models are based on three phase equilibrium in the bulk systems, and usually require large amount of computing time to approach the equilibrium state with different accuracies. Additionally, parameters embedded in the thermodynamic models need to be adjusted to match the experimental data well, which introduce human errors into the results. Therefore, it is necessary to propose a new model with high efficiency and high accuracy.

Machine learning as a branch of artificial intelligence is one of the fastest growing fields in computer science that are designed to emulate human intelligence by learning from the surrounding environment [18,19]. It is a data analysis method that automates the construction of analysis models. Machine learning can learn from data, recognize patterns, and make decisions with minimal human intervention [20]. There are many types of machine learning [18], including (a) Supervised Learning/Semi-Supervised Learning; (b) Unsupervised Learning; (c) Reinforced Learning [21]. Supervised Learning and Semi-Supervised Learning infer the underlying relationship between the observed data, when the amount of observed data labeled is greater than zero. They can solve the problems of “classification” and “regression”. Unsupervised learning applies to all cases where the data are unlabeled and aims to discover the hidden structure between the given data, which mainly solves the problem of “Clustering” and “outlier detection”. Reinforcement learning, another area of machine learning, involves exploration of an adaptive sequence of actions or behaviors by an intelligent agent (RL-agent) in a given environment. Its motivation is maximizing the cumulative reward given data, and is mostly used in decision-making problems (e.g., a computer playing chess) [22–24].

If the label in Supervised Learning is discrete, it can be handled by classification algorithms. In another case, the data with labels being continuous should be handled by regression algorithms. Classification is mostly used in prediction, pattern recognition, and outlier detection, whereas regression is used for prediction and ranking. The progress of selecting machine learning methods is shown in Figure 1. In pattern regression, the amount of data is a critical factor affecting the performance [25]. The degree of convergence of the algorithm depends largely on the amount of data and computational resources. In the practical case, limited data and other resources will cause the model to overfit the training data by losing its generalization [26]. Data augmentation can strengthen the learning of ontology features and prevent overfitting in machine learning, improve generalization [27].

However, there are a variety of machine learning models with high computational efficiency and accuracy, few works on the application of machine learning method to the prediction of hydrate formation were previously reported. The aspects of the machine learning model to be applied in this article belong to the category of “regression”. The objective of this work is to introduce the machine learning methods into calculating formation conditions of hydrate to improve the efficiency and accuracy of prediction. The Python

language was used to complete all the work. Three machine learning models, including Random Forest (RF), Naive Bayes (NB), and Support Vector Regression (SVR), were firstly selected for calculating the hydrate formation conditions. Subsequently, the machine learning models were trained and tested by experimental data reported in the previous works. The computational efficiency and accuracy of the machine learning models were compared with those of the conventional thermodynamic Chen–Guo model. By summarizing the discussion on the comparison conclusions were finally drawn. This work provides an efficient and accurate method for predicting the formation condition of hydrates formed by different guest gases and their mixtures, and guidance for the exploration of natural gas hydrate.

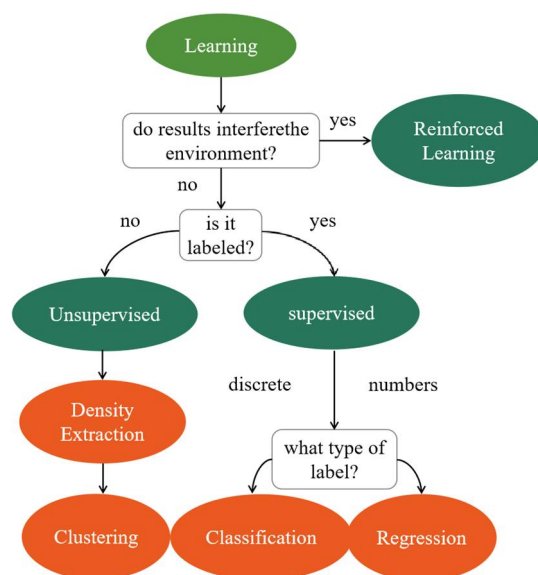


Figure 1. Machine learning method categorization.

2. Methodologies and Experimental Data

2.1. Machine Learning Model Selection

Machine learning is the study of computer algorithms that can be improved automatically through the given data or the past experiences by optimizing the performance of computer. The supervised and semi-supervised machine learning were chosen for the regression prediction analysis. A typical example of supervised learning and unsupervised learning is GBDT (Gradient Boosting Decision Tree), RF (Random Forest), SVM (Support Vector Machine), Naive Bayes (NB), LR (Logistic Regression), etc. [28]. Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM) were selected as statistical learning models to predict the formation conditions of hydrates.

2.1.1. Random Forest (RF)

By adding a layer of randomness to bagging (bootstrap aggregating [29]), Random Forest was proposed by Breiman [30] as a bagging algorithm in ensemble learning. RF is built on the basis of decision tree, which is different from the standard trees. Randomly select a part of the data from the original dataset, and construct a sub-dataset of the same length and size as the original dataset. The data in the sub-data set can be repeated. After that, use the sub-dataset to build a sub-decision tree. Select a part of the result features obtained by each decision tree, and finally select the optimal result feature from the randomly selected part of the result features [31].

Because the work performed in this paper is to predict the hydrate formation conditions, it belongs to the regression model category. The RF for regression is formed by

growing trees depending on a random vector, its predictor is formed by taking the average over k of the growing trees [30]. The mathematical model of RF is as follows:

$$G(x) = \frac{1}{m} \sum_{i=1}^m g_i(x) \quad (1)$$

where $g_i(x)$ represents the value of each base learner.

The model is not determined by specific eigenvalues or combinations of features, and the final prediction results are averaged, giving the overall model results generalization performance [32] and decreasing the average error of model results. In this paper, a total of 15 decision trees are used to aggregate into a RF model. Moreover, a strategy of eight-fold cross-validation was adopted during training models.

2.1.2. Naive Bayes (NB)

Naive Bayes (NB) is one of the most efficient and effective inductive learning algorithms for machine learning and data mining [33]. It relies on an assumption that the input variables are independent each other, but it performs well even under conditions where the algorithm is not ideal. Compared with NB's application in classification, it has more limitations in regression. Although its algorithm is relatively simple, NB is surprisingly effective and immensely appealing and performs well in most classification tasks due to its more accurate than other complex methods [34]. This is because its algorithm is no need to applicate complex iterative parameter estimation schemes to large datasets, which facilitates construction and use [24]. The algorithm of NB is briefly described by the following.

Y is defined as a numeric target value, and an example E is regard as one set consisting of m attributes X_1, X_2, \dots, X_m . When NB makes regression, each attribute is numeric, treated as a real; number or nominal. in which case it is a set of unordered values. $P(Y|E)$ represents the probability density function of the target value. NB can estimate it by applying Bayes' theorem and designating independence of the attributes X_1, X_2, \dots, X_m given the target value Y . Bayes' theorem states that:

$$p(Y|E) = \frac{p(E, Y)}{\int p(E, Y)dY} = \frac{p(E|Y)p(Y)}{\int p(E|Y)p(Y)dY} \quad (2)$$

where $P(E|Y)$ is the probability density function (pdf) of the example E at a given target value Y , and the prior $p(T)$ is the pdf of the target value before any examples have been seen.

2.1.3. Support Vector Regression (SVR)

Support Vector Machine (SVM) are supervised learning methods that analyze data and recognize patterns [24]. In its learning phase, it needs to solve the convex constrained quadratic programming (CCQP) problem to find a set of parameters [35]. We provide the equation of convex constrained quadratic programming below.

$$\min \frac{1}{2} \|\omega\|^2 + C \sum_i \xi_i \quad (3)$$

$$s.t. y_i(\omega \cdot x_i + b) \geq 1 - \xi_i \quad \xi_i \geq 0 \quad (4)$$

where $y_i = (\omega \cdot x_i + b)$ represents hyperplane for SVM; for SVR, it represents a straight line. C is the penalty factor that represents the contribution weight of points with abnormal distribution to the objective function, $C \sum_i \xi_i$ is the penalty function term, ξ_i is called slack variable. ω and b are the parameters to be solved.

Support Vector Regression (SVR) is a VC-theory based regression derived from Support Vector Machine (SVM) method which can overcome the difficulty of determining the structure of traditional Neural Network (NN) and the number of hidden neurons [36]. Among them, VC-theory was developed by Vapnik and Chervonenkis [37] and Vap-

nik [38,39] over the last three decades. VC-theory characterizes properties of learning machines, which enables learning machines to generalize well to unseen data [40]. SVR differs from SVM in kernels used, sparse solutions, VC control margins, and the number of support vectors.

SVR meta-parameters, the loss function ε and the error penalty factor C determine the quality of the SVR models. In addition, the kernel function also directly affects the end of the model. The commonly used polynomial kernel function (Equation (5)) was employed in this study.

$$K(x, x') = (\gamma x^T x' + r)^d, \gamma > 0 \quad (5)$$

The comparison of the three machine learning models is listed in Table 1.

Table 1. Characteristics of the three statistical learning methods.

Methods	RF	NB	SVR
Applicable issues	Multi-class classification, regression	Multi-class classification, regression	Two-class classification, regression
Model characteristics	Build multiple decision trees	Joint probability distribution of features and categories, conditional independence assumption	Separating hyperplane
Model type	Additive model	Generative model	Discriminant model
Learning strategy	Integrate multiple decision trees	Maximum likelihood estimation	from the misclassification point
Learned loss function	Square loss and exponential loss	Log likelihood loss	Hinge loss
Learning algorithm	Integrated algorithm composed of decision tree	Probability calculation formula, EM algorithm	Sequence Minimum Optimization algorithm (SMO)
Advantage	1. Training can be highly parallelized; 2. The trained model has small variance and strong generalization ability; 3. Insensitive to some missing data.	1. Fast and easy to train; 2. Not very sensitive to missing data.	1. Can make predictions on small sample data; 2. It can handle high-dimensional features; 3. The classification plane does not depend on all data, but is only related to a few support vectors.
Disadvantage	1. Easy to fall into overfitting; 2. Under special circumstances, the attribute weights produced by random forests are not credible.	If the input variables are related, problems will occur.	1. Difficult to train; 2. Need to find a suitable kernel function; 3. Sensitive to missing data.

2.2. Chen–Guo (C-G) Model

To compare the computational efficiency and accuracy of machine learning models with conventional thermodynamic models, the Chen–Guo model was selected as the representative of thermodynamic model. The Chen–Guo (C-G) model [41] is primarily a fugacity based approach for hydrate phase stability prediction. Suggesting that the adsorption process of gas molecules by water molecules is not similar to that of Langmuir isothermal adsorption, Chen and Guo proposed a new two-step hydrate formation mechanism. Equating the fugacity of guest gas in gas phase with that in hydrate phase can yield.

The left-side of Equation (6) f_i denotes the fugacity of gas species i in gas phase, and the right-side of Equation (6) means the fugacity of the corresponding gas i in hydrate phase; μ_w stands for the chemical potential of water; λ_2 represents a constant related only to the hydrate structure equaling 3/23 for structure I, and 1/17 for structure II, respectively;

the Langmuir constant C_i represents the interaction between the guest i in the small cage (5^{12}) and the water molecules of the surrounding cage; θ_i represents the fraction of the small cavities occupied by the gas species i , and is estimated by Equation (7); α represents hydrate structure parameters in the model, $\alpha = \frac{\lambda_1}{\lambda_2}$; x_i in Equation (6) denotes mole fraction of hydrate form by gas component i in large cavity, and is subject to Equation (8).

$$f_i = x_i \exp\left(\frac{\Delta\mu_w}{RT\lambda_2}\right) \times \frac{1}{C_{1i}} \times \left(1 - \sum_{i=1}^k \theta_i\right)^\alpha \quad (6)$$

$$\theta_i = \frac{C_i f_i}{1 + \sum_{i=1}^k C_i f_i} \quad (7)$$

$$\sum_{i=1}^k x_i = 1.0 \quad (8)$$

In Equation (6) the fugacity of gas species i f_i in gas phase can be calculated using the widely used cubic equation of state Soave–Redlich–Kwong equation (SRK [42]).

$$\ln\left(\frac{f_i}{\phi}\right) = \frac{b_i}{b_m}(Z_m - 1) - \ln(Z_m - B_m) + \frac{A_m}{B_m} \left(\frac{b_i}{b_m} - \frac{2}{a_m} \sum_j x_j a_{ij}\right) \ln\left(1 + \frac{B_m}{Z_m}\right) \quad (9)$$

where f_i also denotes the fugacity of species i in gas phase; Z_m means the compression factor for the m -component; x_j indicates the molar fraction of j -gas component; ϕ means the partial pressure, $\phi = (x_j \times P)$.

The coefficients A_m and B_m for a mixture of components in the vapor phase are determined using certain rules for mixing. They can calculate by the following expression:

$$A_m = \frac{a_m P}{(RT)^2} \quad (10)$$

$$B_m = \frac{b_m P}{RT} \quad (11)$$

$$a_m = \sum_i \sum_j y_i y_j a_{ij} \quad (12)$$

$$b_m = \sum_i y_i b_i \quad (13)$$

$$Z_m^3 - Z_m^2 + (A_m - B_m - B_m^2)Z_m - AB = 0 \quad (14)$$

where $a_{ij} = (a_i a_j)^{0.5} (1 - k_{ij})$. Z_m is the compressibility factor, its value is equal to the largest real root of Equation (14). k_{ij} is the symmetrical matrix containing binary interaction coefficients for the components of gas mixture. y_i and y_j are mole or volume fractions of component i, j in gas mixture, respectively. The coefficients a_i and b_i can be obtained by the following expression:

$$a_i = 0.42747 \alpha_i(T) \frac{P_{ri}}{T_{ri}^2} \quad (15)$$

$$b_i = 0.08664 \frac{P_{ri}}{T_{ri}} \quad (16)$$

where $T_{ri} = T/T_{ci}$ is the reduced temperature, $P_{ri} = P/P_{ci}$ is the reduced pressure. T, P are the temperature and pressure of the system, T_{ci}, P_{ci} are the critical temperature and pressure for the substance i . $\alpha_i(T)$ is the dimensionless coefficient which becomes unity at $T = T_{ci}$, and can be obtained by the following expression:

$$a_i(T) = \left[1 + m_i \left(1 - \sqrt{T_{ri}}\right)\right]^2 \quad (17)$$

where the parameter m_i is calculated by

$$m_i = 0.48 + 1.574\omega_i - 0.176\omega_i^2 \quad (18)$$

with as the acentric factor of a substance I , and is an indicator of the nonsphericity of the field of intermolecular forces.

The binary interaction parameters and related physical properties of N_2 , CO_2 , CH_4 , C_2H_6 , and C_3H_8 were listed in Tables 2 and 3.

Table 2. Binary interaction parameters in SRK-EoS.

k_{ij}	Component				
	N_2	CO_2	CH_4	C_2H_6	C_3H_8
N_2	0	−0.0171	0.031199	0.031899	0.0886
CO_2	−0.0171	0	0.0956	0.1401	0.1368
CH_4	0.031199	0.0956	0	0.002241	0.006829
C_2H_6	0.31899	0.1401	0.002241	0	0.001258
C_3H_8	0.0886	0.1368	0.006829	0.001258	0

Table 3. Related physical properties of components in this work.

Gas Name	Critical Temperature, T_c (K)	Critical Pressure, P_c (Mpa)	Acentric Factor, ω	Molar Weight, MW(g/mol)
N_2	−146.96	3394.37	0.04	−146.96
CO_2	30.95	7370	0.23894	30.95
CH_4	−82.45	4640.68	0.011498	−82.45
C_2H_6	32.38	4883.85	0.0986	32.38
C_3H_8	96.75	4256.66	0.1524	96.75

2.3. Data Preparation

In this work, the experimental data used to train and test the machine learning models were taken from a previous work [43]. The experiments were conducted in a high-pressure autoclave with a volume of about 500 mL. the container made from corrosion and acid-resistant stainless steel, was designed for pressures up to 25 MPa. The autoclave was mounted inside a pressureless heating sleeve filled with a circulating water/glycol mixture to keep temperature constant. The measuring principle selected was the isochoric method allowing hydrate formation conditions to be determined only from the measured pressure and temperature. The schematic of apparatus used in the work of Nixdorf and Oellrich [43] is shown in Figure 2.

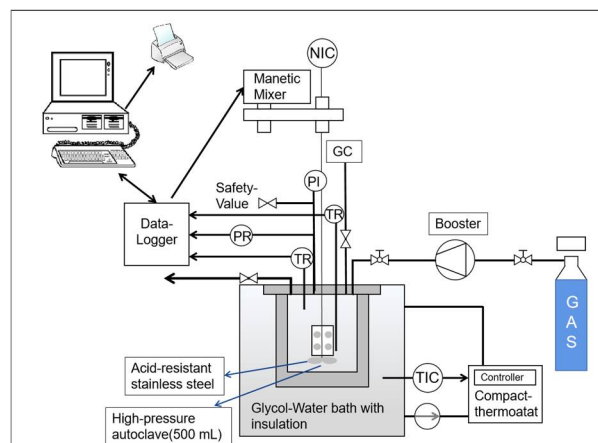


Figure 2. Schematic diagram of the experimental apparatus.

Pure CH₄, N₂, CO₂, C₂H₆, C₃H₈, and their combinations were used as guest gases to form hydrates. A total of nine cases were selected in this study, including two pure gases, two binary gas combinations, three ternary gas combinations, and one quaternary gas combination (details are listed in Table 4). To prevent the machine learning model from overfitting, we use data augmentation to increase the order of magnitude of training data for machine learning from 10² to 10³.

Table 4. Design of experiments. (The data are arranged in order, from left to right, from top to bottom, the data gradually becomes larger, where Peq. is the equilibrium pressure at which hydrate is formed at a given temperature).

System	T/K	Peq./MPa	T/K	Peq./MPa	T/K	Peq./MPa
P1: 100% CH ₄	273.49	2.716	281.12	5.822	285.99	9.874
	274.36	2.961	281.38	6	286.95	10.922
	275.11	3.18	282.07	6.428	287.85	12.314
	276.29	3.564	283.04	7.139	288.62	13.475
	277.46	3.998	283.39	7.43	289.44	15
	278	4.244	284.01	7.925	290.84	17.861
	278.25	4.348	284.17	7.972	291.57	19.165
	279.1	4.733	284.18	7.97	291.6	19.195
	280.16	5.304	285.08	8.928		
P2: 100% N ₂	273.67	16.9350	275.77	20.7480		
	274.07	17.6680	277.27	24.0920		
	275.11	19.5210				
B1: 10.74% N ₂ + 89.26% CH ₄	278.7	4.938	288.68	14.976		
	282.03	6.943	290.97	20.023		
	285.64	10.399	292.44	24.428		
B2: 9.53% C ₂ H ₆ + 90.47% CH ₄	278.21	2.254	288.12	7.208	294.63	19.891
	279.6	2.628	290.44	9.992	295.52	23.198
	283.69	4.191	292.97	14.849		
B3: 0.03% N ₂ + 5% CO ₂ + 94.97% CH ₄	276.85	3.454	287.41	10.935		
	279.95	4.868	290.76	16.827		
	283.49	7.035	293.41	23.979		
T1: 4.18% N ₂ + 4.89% C ₂ H ₆ + 90.93% CH ₄	276.85	3.454	287.41	10.935		
	279.95	4.868	290.76	16.827		
	283.49	7.035	293.41	23.979		
T2: 2.93% C ₃ H ₈ + 12.55% C ₂ H ₆ + 84.52% CH ₄	277.36	2.575	287.45	8.36	294.23	23.833
	280.91	3.803	290.93	13.806		
	284.9	6.096	292.82	18.82		
T3: 1.00% C ₃ H ₈ + 3.98% C ₂ H ₆ + 95.02% CH ₄	277.1	1.198	291.52	6.924	298.14	24.474
	281.56	1.982	294.32	11.207		
	287.42	3.999	296.14	17.034		
T4: 0.02% N ₂ + 5.13% C ₂ H ₆ + 5.25% CO ₂ + 89.60% CH ₄	279.01	2.964	290.04	11.876		
	283.54	5.003	292.28	17.515		
	287.08	7.735	294.21	24.326		

2.4. Error Analysis

In terms of the performance of machine learning models, the mean absolute percentage error (MAPE), mean absolute error (MAE), root mean square error (RMSE), mean squared error (MSE) were used to assess computational accuracy. Additionally, the coefficient of determination (R^2) was used to quantify the strength of relationship between the independent variables and the dependent variables. The definitions of MAPE, MAE, RMSE, MSE,

R^2 are given by Equations (19)–(23). Higher R^2 or lower $MAPE$, MAE , $RMSE$, MSE indicate higher precision and accuracy.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (19)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (20)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (21)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (22)$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (23)$$

where y_i is the predictive value, \hat{y}_i is the real value, \bar{y} is the average of real value, the full name of SSE is the Sum of Squares Error, and the full name of SST is the Sum of Squares Total.

3. Results and Discussion

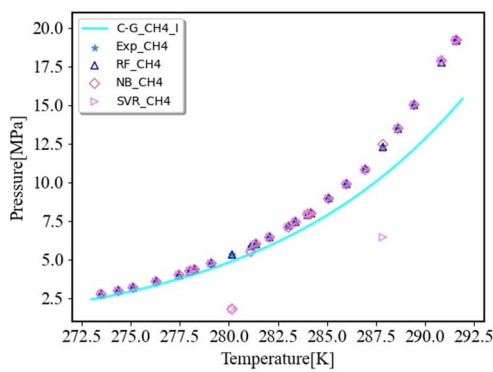
To investigate the feasibility of machine learning methods in predicting the hydrate formation conditions, the computational accuracy of machine learning models was tested based on hydrate formation experiments. In order to evaluate the computational efficiency, the computational time costed by the three machine learning models and the thermodynamic Chen–Guo model were compared.

Gas compositions and temperatures used in the hydrate formation experiments were taken as input data during training and testing the machine learning models. The pressure at which hydrate were formed at given temperature and gas composition was employed to validate the computational accuracy and efficiency of machine learning models. A total of 70% of the experimental data randomly selected from the total data were used to train the three machine learning models, and the rest were used to test the trained models.

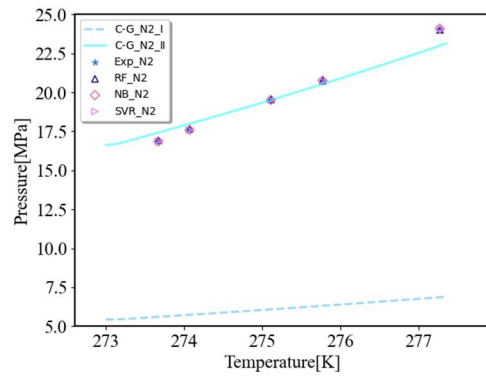
3.1. Computational Accuracy

3.1.1. Comparison of Machine Learning Models with Experimental Data

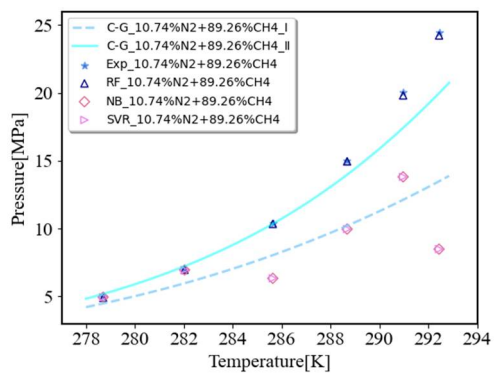
The predicted pressures of hydrate formation by the machine learning models are shown in Figure 3. From Figure 3a it can be found that for the pure methane hydrate the predicted formation pressure by machine learning models match the experimental data well. SVR and NB have almost the exact same behavior. At the same time, there is a large deviation from the experimental value. However, there are also slight deviations in the results of the two. In Figure 3a,d–f, we can see that they both show different results at individual prediction points. NB performs slightly better than SVR. SVR underperformed in two or three combinations of N_2 , CO_2 , CH_4 , C_2H_6 , and C_3H_8 , and NB underperformed in two or three combinations of N_2 , CH_4 , C_2H_6 , and C_3H_8 . The two models deviate most strongly for the combination of nitrogen and methane, see Figure 3c. The guess we provide is that the importance of nitrogen is higher than that of methane. When nitrogen and methane exist at the same time, the SVR and NB models focus more on following the law of nitrogen, which makes it inapplicable to the mixture of the two. The convincing explanation remains to be found.



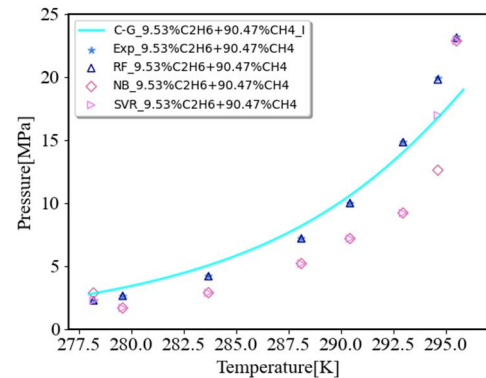
(a)



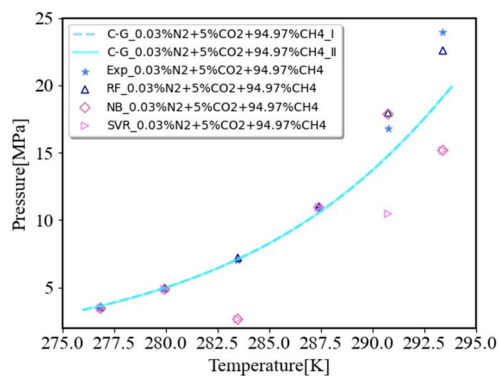
(b)



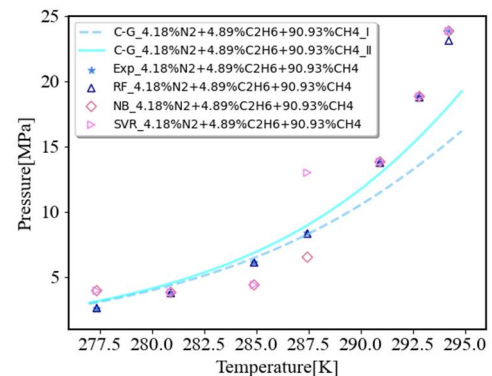
(c)



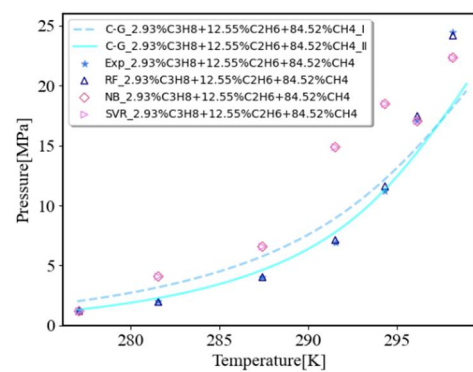
(d)



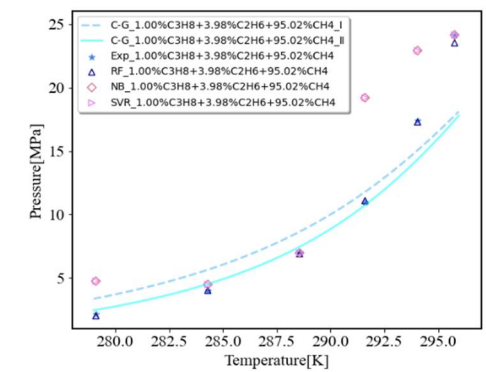
(e)



(f)

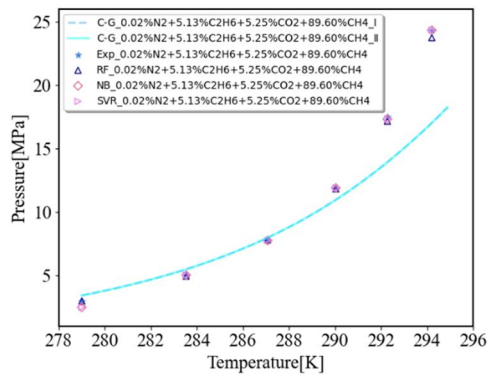


(g)



(h)

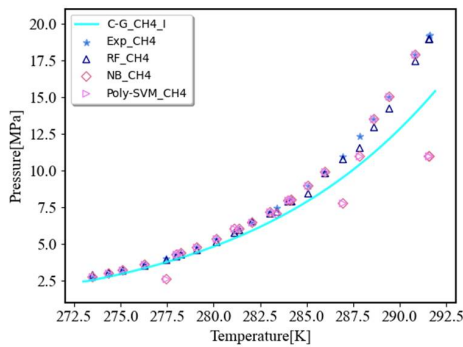
Figure 3. Cont.



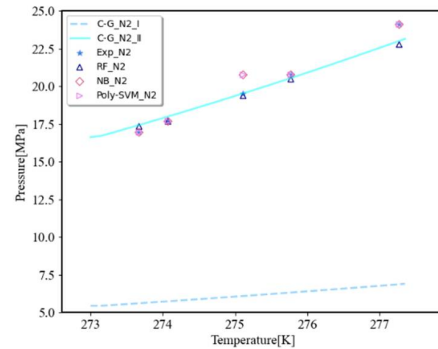
(i)

Figure 3. Comparison chart of prediction results using two prediction methods after data augmentation: (a) description of the P1 system; (b) description of the P2 system; (c) description of the B1 system; (d) description of the B2 system; (e) description of the B3 system; (f) description of the T1 system; (g) description of the T2 system; (h) description of the T3 system; and (i) description of the T4 system.

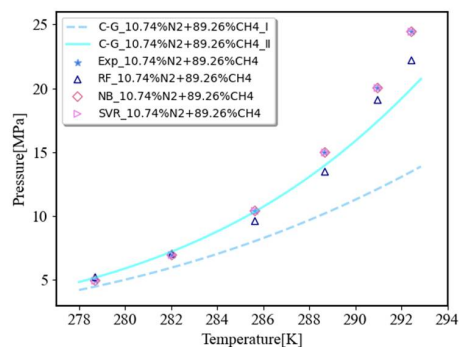
The reason for the large deviation of NB is that the data after data augmentation does not fully meet the independence assumption of NB. Regression performance of NB is mainly controlled by independence assumption [34]. After data augmentation, the independence between data is weakened, which leads to poor regression performance of NB. For a more intuitive explanation of the impact of independence assumption on NB prediction results, we draw a schematic diagram of machine learning models without data enhancement, as shown in Figure 4. Focus on NB prediction results in Figures 3c,d and 4c,d.



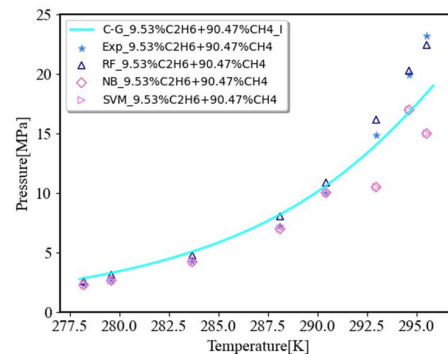
(a)



(b)



(c)



(d)

Figure 4. Cont.

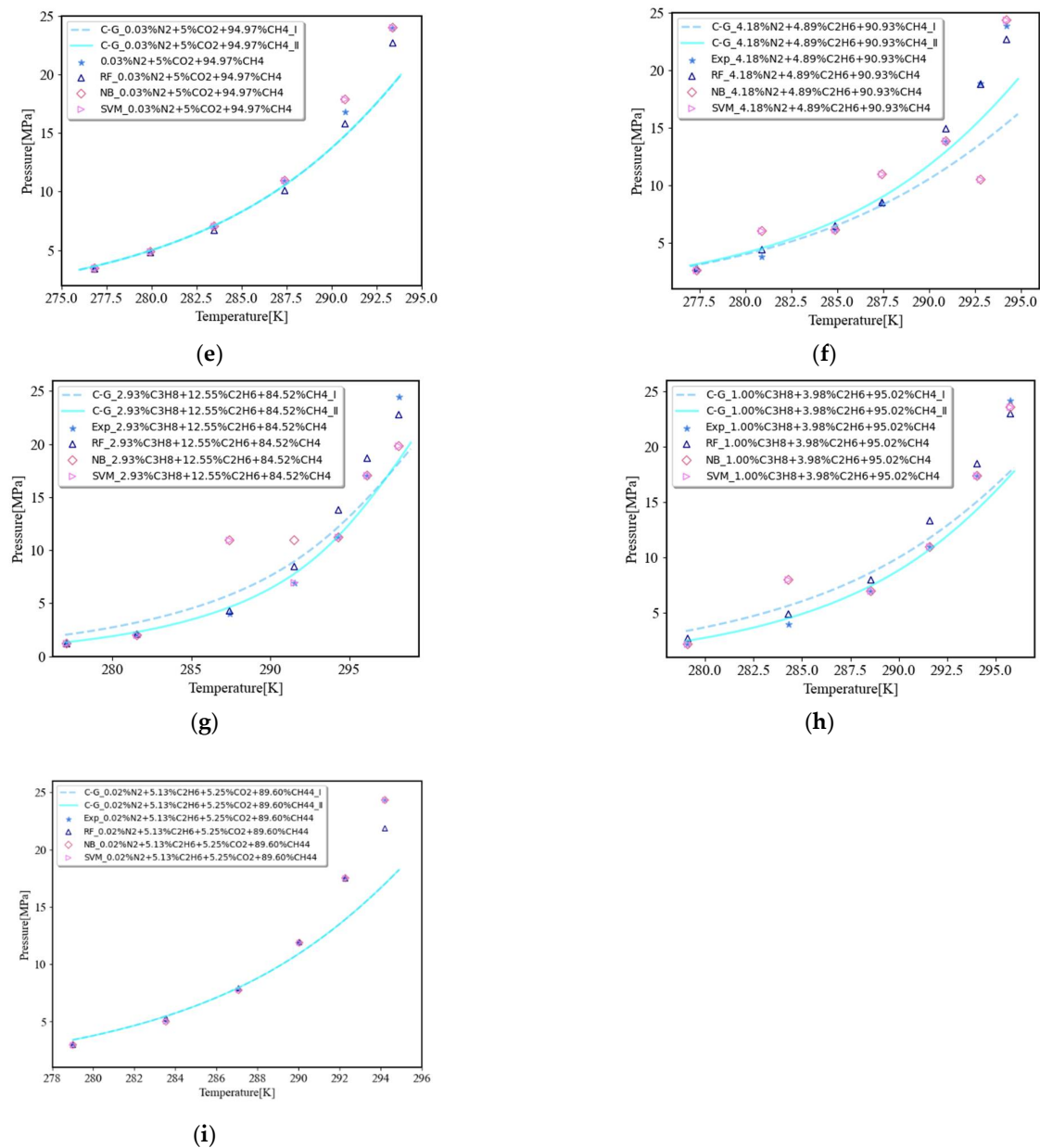


Figure 4. Comparison chart of prediction results using two prediction methods: (a) description of the P1 system; (b) description of the P2 system; (c) description of the B1 system; (d) description of the B2 system; (e) description of the B3 system; (f) description of the T1 system; (g) description of the T2 system; (h) description of the T3 system; and (i) description of the T4 system.

SVR also has a large deviation because SVR does not apply to situations where the sample size is particularly large [22]. We also obtain the performance of SVR in the original data (without data enhancement) training, see Figure 4. Compared with Figures 3c,d and 4c,d, we can find that SVR performs better when the sample size decreases and the kernel function remains unchanged.

RF performance is good in almost all plots, in Figure 3g–i, there is a slight deviation. Comparing Figures 3 and 4, it is not difficult to find that data augmentation is helpful to improve the accuracy of RF prediction results. This is most prominent in Figures 3c,i and 4c,i.

From all figures listed in Figure 3, we can find that the predicted results using the machine learning method are instable, some predicted points do not fit the trend of the curve, and the calculated accuracy depends on the number of sample training and the number of data.

RF uses an eight-fold cross-validation training method. Among the eight-fold results, the *MAPE* of the optimal result reaches 11.53%. Calculating the *MAPE* with all the data, the *MAPE* is obtained as 1.24%. The detailed eight-fold cross-validation training effect is shown in Figure 5. At the same time, RF can evaluate the importance of features. In the process of training three machine learning algorithms, six parameters are selected as the main features, which are the contents of N_2 , CO_2 , CH_4 , C_2H_6 , C_3H_8 in mixture and temperature at which hydrates are expected to be formed. The predicted value is equilibrium pressure of hydrate formation under given gas mixture and temperature. Compared with other four guest gases, the content of nitrogen is identified as the more important parameter (Figure 6).

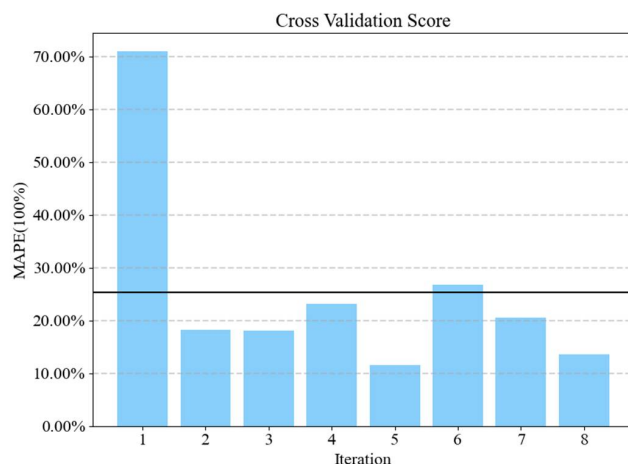


Figure 5. RF 8-fold training effect.

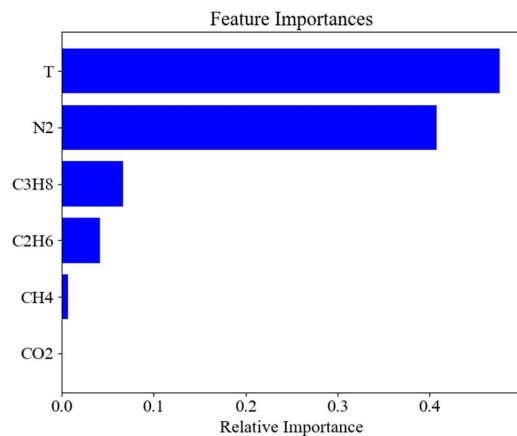


Figure 6. An example of the importance of a RF 8-fold sample.

This work employed three machine learning models, namely RF regression model, NB regression model and SVR model. *MAPE*, *MAE*, *RMSE*, *MSE*, R^2 were used to evaluate the machine learning models. The errors are listed in Table 5. It can be seen from the table that the results of RF are better than those of the other two, which implies that the RF model is a better solution.

Table 5. The error comparison of training samples.

Error Index	Methods		
	RF	NB	SVR
<i>MAPE</i>	1.24%	23.66%	24.21%
<i>MAE</i>	0.16	0.25	0.31
<i>RMSE</i>	0.23	0.32	0.45
<i>MSE</i>	0.10	0.18	0.62
R^2	1.0000	0.9993	0.9993

3.1.2. Comparison of Machine Learning Models with Chen-Guo Model

The formation conditions of hydrates formed by pure gas and mixed gases water were evaluated employing thermodynamic C-G model. The evaluated results are shown in Figure 3. Figure 3a,b show the predicted formation condition of methane and nitrogen hydrates using the Chen–Guo model. From Figure 3a, it can be seen that for pure methane hydrate, when the hydrate structure was Type I the evaluated formation pressures fitted the experimental data well. When the temperature is lower ($-6.66\text{ }^{\circ}\text{C}\sim 11.85\text{ }^{\circ}\text{C}$), the predicted data were in good agreement with the measured data, while when the temperature is higher ($>11.85\text{ }^{\circ}\text{C}$), the simulated value is lower than the measured value. The overall fit is better, and the AADP is smaller.

For pure nitrogen hydrate, we considered Type I and Type II hydrate structures. Figure 3b shows that the structure of nitrogen hydrate is II type. Because of the different Antoine parameters (type I and type II) in the Chen–Guo model, the overall rising speed of two type curve is different. The comparison of the calculated values with the experimental values of binary gas hydrate was illustrated by Figure 3c,d. The B1 case contains 10.74% N_2 and 89.26% CH_4 (Figure 3c). From Figure 3c,d we can find that when the formation structure is type II, both the curves and AADP of the calculated results indicate a better agreement with the measured data.

Compared with pure nitrogen hydrate, the formation pressure of hydrate formed by mixture composed of nitrogen and other gases is lower. The reason is that when the content of nitrogen in the mixture is higher than a certain value, the small molecule gas (methane) is encaged in the small pores, reducing the pressure of macromolecular gas (nitrogen). In the B2 case the gas mixture was composed of ethane and methane which both form type I hydrates. Figure 3d indicates that the mixture consisting of ethane and methane forms type I hydrate. As the temperature increases, the deviation of the simulated value from the experimental value increases. At low temperatures ($0.06\text{--}11.85\text{ }^{\circ}\text{C}$), the simulated results matched the measured data well.

The comparison of the calculated formation conditions of hydrate formed by ternary gas mixture with those of the experimental values are shown in Figure 3e–h. For the B3 case where the gas mixture is composed of N_2 , CO_2 , and CH_4 (Figure 3e). We can find that the fitting curves of the type I and type II structures basically overlap. The reason is that the nitrogen accounts for 0.03% of the total components, therefore, the effect of nitrogen on the structure of the type I hydrate is not obvious in this component. Since the type II hydrate AADP is smaller than that of the type I hydrate, we believe that this component forms a type II structure. It can be clearly seen from the image that the T1 component gas generates a type II structure (Figure 3f), and the curve fits well with experimental points. The experimental points are distributed on both sides of the curve, and the overall curve trend follows the trend of the experimental points. Both the T2 component gas (Figure 3g) and the T3 component gas (Figure 3h) showed that the fitting curve of type II structure fits better with the experimental points, but the AADP is larger than that of type I, due to the accurate hydrate formation conditions is more important, so the importance of the degree of image matching is higher than the size comparison of AADP. In summary, for the T2 and T3 components, we think that it generates a type II structure. Comparing the T2 component and the B2 component, it is not difficult to find that small gas molecules, such as ethane and methane, in the T2 component- $\text{C}_3\text{H}_8\text{-C}_2\text{H}_6\text{-CH}_4$ system dissolve in the small pores, reducing the generation of macromolecular gases (propane) pressure, so when there is a small amount of propane in the gas component, it is easier to form type II hydrate. The results of comparing the calculated values of the quaternary gas hydrate formation conditions with the experimental values are shown in Figure 3i. The T4 components also generate type II structures, and the type I and II structure curves of the T4 gas components overlap. Most of the fitting curves show the simulated formation pressure under high temperature conditions. The specific reasons need to be explored.

Compared with the machine learning models, the accuracy of the prediction result of the Chen–Guo model is lower. However, the overall trends of pressure curves match the experimental value well.

According to the results of machine learning and Chen–Guo model simulation, the comparison of computational accuracy with machine learning models was drawn in Figure 7. The AADP (average absolute deviation percent) is used to describe computational accuracy (see Equation (24)). For pure and quaternion cases, machine learning performed better than Chen–Guo model on computational accuracy. The RF, NB, and SVR models are 3–36 times more accurate than the Chen–Guo model (see Figure 7a–d). For binary cases, the RF performance is particularly outstanding, up to 40 times higher than the Chen–Guo model. Although NB and SVR perform poorly in binary cases, this situation can also be seen in the ternary cases (see Figure 7b,c). Although for the T1 in the ternary cases, the prediction accuracy of SVR and NB is higher than that of Chen–Guo model, but the performance is not outstanding. In general, the RF machine learning model is the most stable and accurate in predicting the formation pressures of hydrates formed by pure gases and their combinations.

$$\text{AADP} (\%) = \frac{1}{N_p} \sum_{i=1}^{N_p} \left| \frac{P_{cal} - P_{exp}}{P_{exp}} \right| \times 100 \quad (24)$$

where P_{cal} is the predictive value, P_{exp} is the experimental value, N_p is the total number of data points.

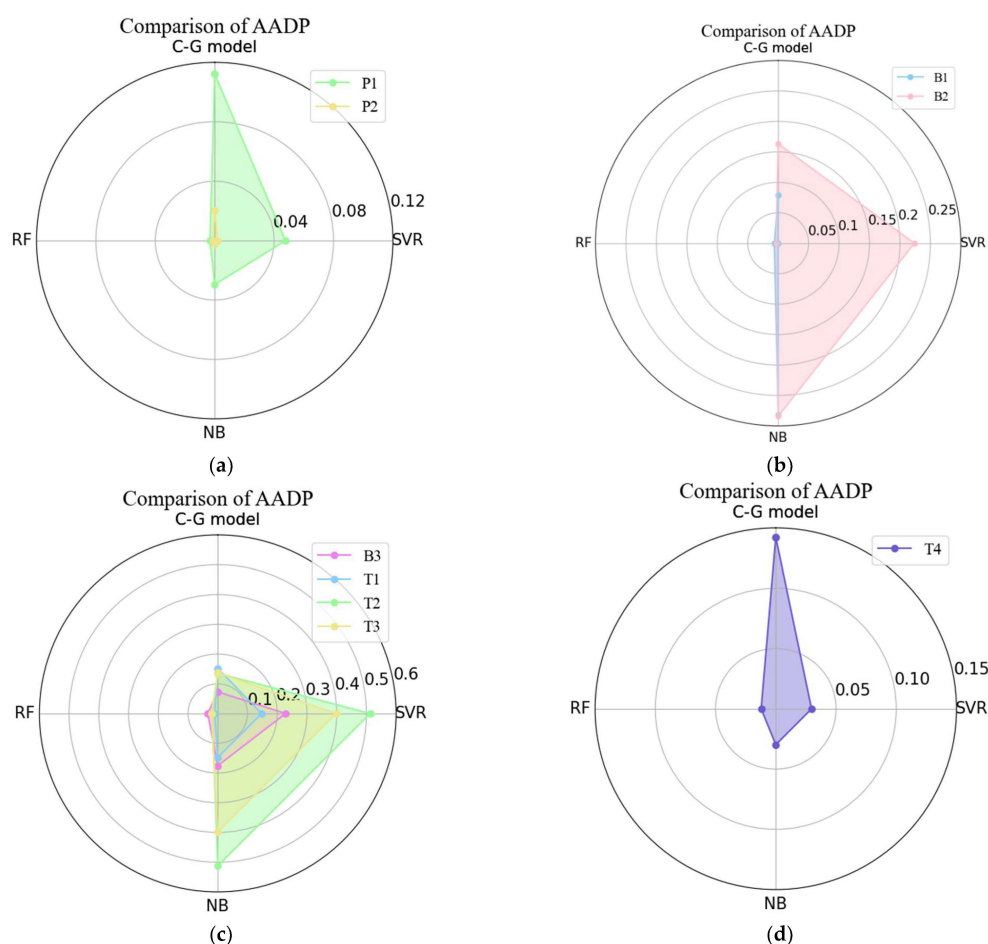


Figure 7. Accuracy of the models in calculating the formation pressure of hydrates formed by pure gases and their combinations: (a) description of pure gas system; (b) description of binary gas system; (c) description of ternary gas system; and (d) description of quaternion gas system.

3.2. Comparison of Computing Time

The Chen–Guo model is based on iterative calculations for approaching the thermodynamic equilibrium state. Each component needs to calculate the mole fraction x_i in hydrate phase, which requires a lot of computing time. To match the actual formation conditions measured by the experiment the Antoine parameters need to be adjusted. In terms of machine learning models, there is no need to adjust parameters and waste the time in iterative calculations. Without relying on the theory of thermodynamic equilibrium, the machine learning model finds laws from practical examples, consuming less time, and is highly optimizable.

Figure 8 shows the comparison of computing time between three machine learning models and Chen–Guo model. It can be seen that the computing time of the three machine learning models is similar, and for different cases, it varies from 0.3 s to 3.8 s. The first time-consuming model is RF, followed by SVR, and finally NB model. For RF model, B2 and T4 components consume the most computing time. They were found to have CH_4 and C_2H_6 , and the components containing both gases were selected from all components: B2, T1, T2, T3, and T4. It was found that T1, T2, and T3 did not take much time. There is no rule obtained from the computing time of RF model. The data provided does not satisfy the independence assumption of NB, and there is not so much time complexity in the prediction. We believe that this is also the reason for little time change. As for SVR model, the prediction performance is slightly worse than NB model, and the time consumption is slightly more. Due to iterative calculations, Chen–Guo model needs 80–240 s to predict formation pressure of different cases. Regular conclusion of calculation time costed by C-G model and machine learning models to predict hydrate formation in different cases with various gas mixtures was not obtained. The reason is inferred that the computing time is independent on the gas composition for C-G model and machine learning models.

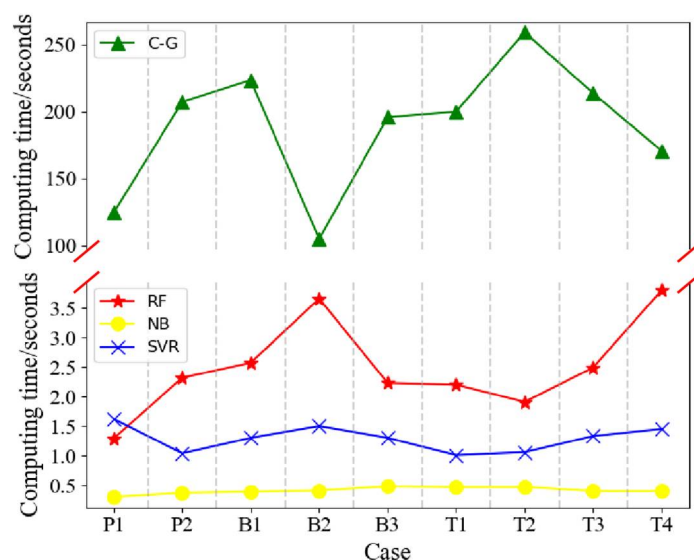


Figure 8. Comparison of the computing time needed in the four cases.

4. Conclusions

In this work, the machine learning method was applied in calculating the formation pressure of hydrate formed by pure gases and mixtures composed of different gases. Experimental data previously reported were used to train and test the learning models. A thermodynamic model, the Chen–Guo model was used to compare the computational efficiency and accuracy of machine learning models. By discussing the results the following conditions were drawn:

- (1) After data augmentation, support vector machine regression cannot achieve the desired prediction results directly. It also affects the independence of each feature and

indirectly affects the prediction accuracy of Naive Bayes. They perform well in small sample training. In this prediction, the independence between features decreases with the increase in the number of samples, and the independence between features is not enough to support Naive Bayes for accurate prediction.

- (2) Comparisons of predicted results indicate that Random Forest performs better in the stability and accuracy than the other two machine learning models. After data augmentation, the prediction accuracy of Random Forest is greatly improved. Data enhancements can be used for data preprocessing when forecasting later using Random Forest.
- (3) In terms of computation time, Naive Bayes and Support Vector Regression take the least computation time, followed by the Random Forest. The average time of Random Forests is 2.497 s being substantially less than that of the Chen–Guo model (189 s). In this work, the machine learning models show better performance in predicting the formation conditions of hydrates formed by pure gases and mixtures. However, the performance of machine learning model strongly depends on the training and testing steps, which requires large amounts of data. Therefore, the experimental data of hydrate formations are important to the performance of machine learning models.

Author Contributions: Conceptualization—Formal analysis and Software, H.T. and Z.Y.; Data curation—Investigation and Methodology, Z.Y.; Funding acquisition—Project administration and Resources, H.T.; Validation—Visualization and Writing—Original draft, Z.Y.; Supervision, H.T.; Writing—Review and editing, H.T. and Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was performed in support of the National Natural Science Foundation of China (Grant No. 41772247 and No. 41877185).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the Key Laboratory of Groundwater Resources and Environment, Ministry of Education, Jilin University for financial support and facilities, the National Natural Science Foundation of China for financial support.

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

References

1. Changyu, S.; Wenzhi, L.; Xin, Y.; Fengguang, L.; Qing, Y.; Liang, M.; Jun, C.; Bei, L.; Guangjin, C. Progress in research of gas hydrate. *Chin. J. Chem. Eng.* **2011**, *19*, 151–162.
2. Sloan, E.D., Jr.; Koh, C.A. *Clathrate Hydrates of Natural Gases*; CRC Press: Boca Raton, FL, USA, 2007.
3. Makogon, Y.F. Perspectives for the development of gas hydrate deposits, Gas hydrates and permafrost. In Proceedings of the 4th Canadian Permafrost Conference, Calgary, Alberta, 2–6 March 1982; pp. 299–304.
4. Makogon, Y.F. Natural gas hydrates—A promising source of energy. *J. Nat. Gas Sci. Eng.* **2010**, *2*, 49–59. [[CrossRef](#)]
5. Kvenvolden, K. A primer on the geological occurrence of gas hydrate. *Geol. Soc. Lond. Spec. Publ.* **1998**, *137*, 9–30. [[CrossRef](#)]
6. Moridis, G.J.; Freeman, C.M. The RealGas and RealGasH2O options of the TOUGH+ code for the simulation of coupled fluid and heat flow in tight/shale gas systems. *Comput. Geosci.* **2014**, *65*, 56–71. [[CrossRef](#)]
7. Eslamimanesh, A.; Mohammadi, A.H.; Richon, D.; Naidoo, P.; Ramjugernath, D. Application of gas hydrate formation in separation processes: A review of experimental studies. *J. Chem. Thermodyn.* **2012**, *46*, 62–71. [[CrossRef](#)]
8. Collett, T.S.; Kuuskraa, V.A. Hydrates contain vast store of world gas resources. *Oil Gas J.* **1998**, *96*, 90–95.
9. Khan, M.N.; Warriar, P.; Peters, C.J.; Koh, C.A. Review of vapor-liquid equilibria of gas hydrate formers and phase equilibria of hydrates. *J. Nat. Gas Sci. Eng.* **2016**, *35*, 1388–1404. [[CrossRef](#)]
10. Platteeuw, J.; Van der Waals, J. Thermodynamic properties of gas hydrates. *Mol. Phys.* **1958**, *1*, 91–96. [[CrossRef](#)]
11. Parrish, W.R.; Prausnitz, J.M. Dissociation pressures of gas hydrates formed by gas mixtures. *Ind. Eng. Chem. Process Des. Dev.* **1972**, *11*, 26–35. [[CrossRef](#)]
12. Ng, H.-J.; Robinson, D.B. The measurement and prediction of hydrate formation in liquid hydrocarbon-water systems. *Ind. Eng. Chem. Fundam.* **1976**, *15*, 293–298. [[CrossRef](#)]

13. Tohidi, B.; Danesh, A.; Todd, A. Modeling single and mixed electrolyte-solutions and its applications to gas hydrates. *Chem. Eng. Res. Des.* **1995**, *73*, 464–472.
14. Chen, G.-J.; Guo, T.-M. Thermodynamic modeling of hydrate formation based on new concepts. *Fluid Phase Equilibria* **1996**, *122*, 43–65. [[CrossRef](#)]
15. Klauda, J.B.; Sandler, S.I. A Fugacity Model for Gas Hydrate Phase Equilibria. *Ind. Eng. Chem. Res.* **2000**, *39*, 3377–3386. [[CrossRef](#)]
16. Ballard, A.; Sloan, E., Jr. The next generation of hydrate prediction: I. Hydrate standard states and incorporation of spectroscopy. *Fluid Phase Equilibria* **2002**, *194*, 371–383. [[CrossRef](#)]
17. Lee, S.Y.; Holder, G.D. Model for gas hydrate equilibria using a variable reference chemical potential: Part 1. *AIChE J.* **2002**, *48*, 161–167. [[CrossRef](#)]
18. Alpaydin, E. *Introduction to Machine Learning*; MIT Press: Cambridge, MA, USA, 2020.
19. El Naqa, I.; Murphy, M.J. What is machine learning? In *Machine Learning in Radiation Oncology*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 3–11.
20. Philbrick, K.A.; Weston, A.D.; Akkus, Z.; Kline, T.L.; Korfiatis, P.; Sakinis, T.; Kostandy, P.; Boonrod, A.; Zeinoddini, A.; Takahashi, N. RIL-contour: A medical imaging dataset annotation tool for and with deep learning. *J. Digit. Imaging* **2019**, *32*, 571–581. [[CrossRef](#)] [[PubMed](#)]
21. Ribeiro, M.; Grolinger, K.; Capretz, M.A. Mlaas: Machine learning as a service. In Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 9–11 December 2015; pp. 896–902.
22. Baştanlar, Y.; Özuysal, M. Introduction to machine learning. In *miRNomics: MicroRNA Biology and Computational Analysis*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 105–128.
23. Zhu, X.; Goldberg, A.B. Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2009**, *3*, 1–130. [[CrossRef](#)]
24. Awad, M.; Khanna, R. Support vector regression. In *Efficient Learning Machines*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 67–80.
25. Simard, P.Y.; LeCun, Y.A.; Denker, J.S.; Victorri, B. Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 239–274.
26. Kobayashi, S. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv* **2018**, arXiv:1805.06201.
27. Tanner, M.A.; Wong, W.H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **1987**, *82*, 528–540. [[CrossRef](#)]
28. Li, G.; Wen, C.; Huang, G.-B.; Chen, Y. Error tolerance based support vector machine for regression. *Neurocomputing* **2011**, *74*, 771–782. [[CrossRef](#)]
29. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [[CrossRef](#)]
30. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
31. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1999.
32. Yan, J.; Li, D.; Guo, R.; Yan, H.; Wang, Y. Classification of tooth marks and tongue based on deep learning and random forest. *Chin. J. Tradit. Chin. Med.* **2022**, *40*, 19–22, 259–261. (In Chinese)
33. Zhang, H. The optimality of naive Bayes. *Aa* **2004**, *1*, 3.
34. Frank, E.; Trigg, L.; Holmes, G. Technical Note: Naive Bayes for Regression. *Mach. Learn.* **2000**, *41*, 5–25. [[CrossRef](#)]
35. Lin, C.-J. Asymptotic convergence of an SMO algorithm without any assumptions. *IEEE Trans. Neural Netw.* **2002**, *13*, 248–250.
36. Courant, R.; Hilbert, D. *Methods of Mathematical Physics: Partial Differential Equations*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
37. Vapnik, V.N.; Chervonenkis, A.J. *Theory of Pattern Recognition*; Nauka: Moscow, Russia, 1974.
38. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: Berlin/Heidelberg, Germany, 1995.
39. Vapnik, V. *Estimation of Dependences Based on Empirical Data*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
40. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [[CrossRef](#)]
41. Chen, G.; Sun, C.; Ma, Q. *Gas Hydrate Science and Technology*; Chemical Industry Press: Beijing, China, 2008. (In Chinese)
42. Soave, G. Equilibrium constants from a modified Redlich-Kwong equation of state. *Chem. Eng. Sci.* **1972**, *27*, 1197–1203. [[CrossRef](#)]
43. Nixdorf, J.; Oellrich, L.R. Experimental determination of hydrate equilibrium conditions for pure gases, binary and ternary mixtures and natural gases. *Fluid Phase Equilibria* **1997**, *139*, 325–333. [[CrossRef](#)]