A Novel Heterogeneous Ensemble Approach to Variable Selection For Gas-Liquid Two-Phase CO_2 Flow Metering

Caiying Sun[1,3], Lijuan Wang[2], Yong Yan[2,*], Wenbiao Zhang[1], Ding Shao[1]

1 School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, P.R. China
2 School of Engineering and Digital Arts, University of Kent, Canterbury, Kent CT2 7NT, U.K.
3 School of Information Engineering, Inner Mongolia University of Science & Technology, Baotou 014010, P.R. China
* Correspondent author: y.yan@kent.ac.uk

Abstract

Variable selection is an important preprocessing step in the development of effective data-driven models for CO_2 flow measurement in carbon capture and storage systems. In order to effectively quantify the importance of potential input variables to the desired output, ensemble learning is proposed and incorporated into variable selection methodology. This paper presents a tree-based heterogeneous ensemble approach to variable selection and its application to gas-liquid two-phase CO_2 flow measurement. The importance of each variable is determined through combining the importance scores from four tree-based algorithms, including decision tree regression, bootstrap aggregating of regression trees, gradient boosting decision tree and gradient boosting random forest. Then the backward elimination algorithm is applied to remove the relatively less important variables and hence a small set of input variables for data-driven models. The selection results demonstrate that the significant variables for CO_2 mass flow measurement include *apparent mass flow rate, time shift, differential pressure* and *pressure drop* while *observed density, density drop, observed flow velocity* and *outlet temperature* for prediction of gas volume fraction. To assess the validity of the selected variables, data-driven models based on gradient boosting random forest are developed. Results suggest that the relative error of the model output is mostly within 1% for CO_2 mass flowrate measurement and 5% for gas volume fraction prediction by taking the selected variables as model inputs.

Keywords: carbon capture and storage, gas-liquid two-phase CO_2, variable selection, heterogeneous ensemble approach, data-driven models

1. Introduction

With the rapid development of machine learning technology, variable selection becomes more and more important in data analysis and data-driven modelling (Wang et al., 2013; Xin et al., 2012; Nan et al., 2014; Zhang et al., 2018). A variety of variable selection methods have been developed over the past few years (Tuv et al., 2009; Zhang et al., 2015; Zhang et al., 2017). However, there is no common rule to determine which method is suitable for a particular application. It is normally determined by balancing the computational cost and the accuracy of the output from the data-driven model. Zhu et al. (2006) extended the application of ensemble learning methods (Mendes-Moreira et al., 2012; AL-Qutami et al., 2018) from the prediction ensemble to the variable selection ensemble. In general, the ensemble approach for variable selection can be classified into two categories: homogeneous and heterogeneous approaches. (Zhu et al., 2011; Zhou., 2012; Li et al., 2017). The homogeneous ensemable approach is to use the same selection method on different datasets while the heterogeneous ensemble approach is to train different selection algorithms on the same dataset.

CO_2 flow in carbon capture and storage (CCS) systems is of complex nature (Wang et al., 2018; Zhang et al., 2018; Shao et al., 2020) and it is thus challenging to measure its dynamic characteristics. To measure mass flowrate and gas volume fraction of multi-phase flow, data-driven modelling has been considered as an efficient and cost-effective way (Yan et al., 2018). Applications of Coriolis flowmeters to gas-liquid two-phase flow measurement have been attempted by using prototype transmitters and investigating into the use of the internal parameters (Green et al., 2008; Kunze et al., 2014; Li et al., 2018; Li et al., 2019). Coriolis flowmeters incorporating data-driven modelling algorithms have demonstrated a potential for multiphase flow measurement (Wang et al. 2017). One ket feature of this approach is to minimise the hardware modification and enable commercial Coriolis flowmeters to work under two-phase flow conditions by simply adding a software module. In order to develop optimal data-driven models for Coriolis flowmeters under two-phase flow conditions and quantify the parametric dependency among the input variables and their significance to the desired outputs, Wang et al. (2017) compared three input variable selection methods, partial mutual information, genetic algorithm-artificial neural network, and tree-based iterative input selection. It is found that a single tree-based selection method can generate varying

results for different datasets in the variable selection process.

To improve the performance of the tree-based selection method, a heterogeneous ensemble approach is introduced to the variable selection process in this paper. A total of four different tree-based selection methods, including decision tree (DT) regression, bootstrap aggregating (Bagging) of regression trees, gradient boosting decision tree (GBDT) and gradient boosting random forest (GBRF) are implemented with the same dataset and fused importance score is calculated for each variable. This paper aims to propose an approach to input variable selection and data-driven modelling for two-phase flow measurement and test the developed models on the same type of flowmeters and transmitters. Experimental tests were conducted with gas-liquid two-phase CO_2 flow. The validity of the selected variables is verified by assessing the performance of GBRF based data-driven models.

2. Methodology

The structure of the ensemble variable selection method is shown in Fig. 1. The dataset is acquired from multiple sensors including a Coriolis mass flowmeter, a DP transducer, two pressure transducers and two temperature sensors. The first step is to generate variable selectors with optimized parameters based on different tree-based algorithms. In this step, the importance scores of variables are obtained from each selector. The second step is to combine the variable importance derived from different variable selectors and then remove the less important variables through backward elimination. Therefore, a set of variables which has significant effect on the mass flow rate measurement and gas volume fraction (GVF) prediction of two-phase CO_2 flow is obtained, respectively. The third step is to develop data-driven models based on the selected variables to produce mass flow rate and GVF.

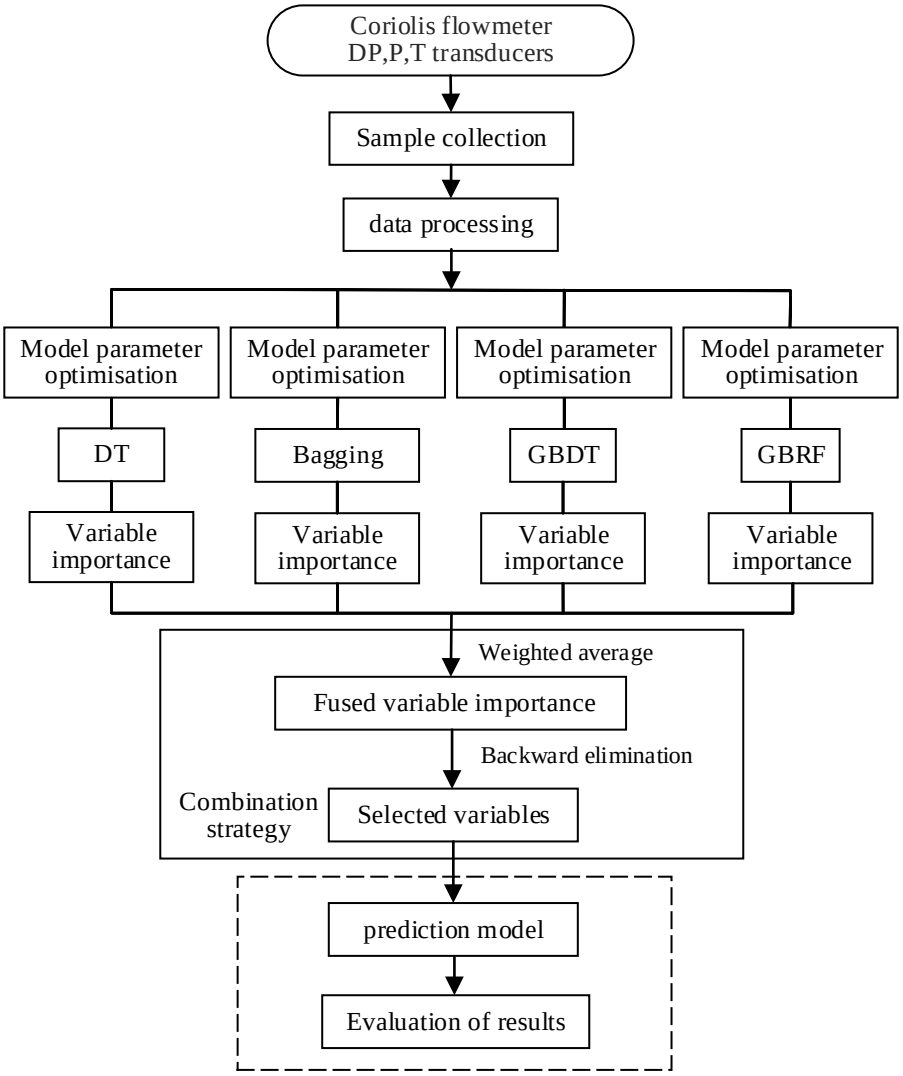


Fig. 1 Structure of the ensemble variable selection method

2.1 Variable selectors

Four tree-based models, including DT, Bagging, GBDT and GBRF, are taken as four individual selectors. These are commonly used tree based algorithms and are effective in the variable selection. The importance score of each variable can be quantified by all tree-based algorithms, respectively. The weighted average method is then used to combine the importance scores from different tree based models. The application of ensemble learning is to improve the reliability of variable selection. As shown in Fig. 2, the DT algorithm (Zhou., 2012) traverses all the input variables in each iteration and take the input variable that produces the minimum MSE (Mean Squared Error) value of the prediction result as the split point of the node. The process continues recursively until the row arrives at a terminal (leaf) node where a prediction value is assigned to the row. The value assigned to the terminal

node is the mean of the outcomes of all training observations that wound up in the leaf node. These observation results are the predicted values corresponding to the predicted target parameters. In this paper, the observation results are mass flow rate and GVF of gas-liquid two-phase CO_2 flow.

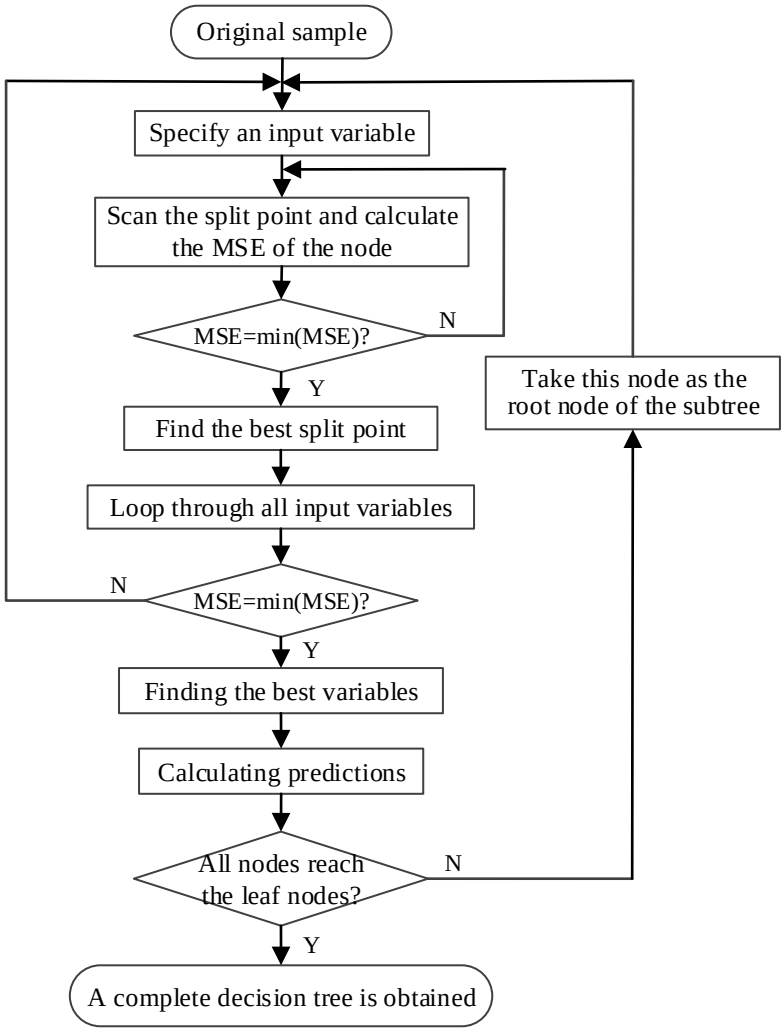


Fig. 2 Decision tree flowchart

For a single decision tree, the more important the variable is, the earlier it is used. The variable importance is quantified as (Tuv et al., 2009):

$$VI(X_i,T)=\sum_{t\in T}\Delta I(X_i,t) \qquad (1)$$

where $\Delta I_i\ (X_i,t)$ is the reduction in impurity due to an actual (or potential) split on the variable X_i at the node t of the optimally pruned tree T. Node impurity $I(t)$ in this paper is MSE with a node t. After normalizing the VI of each variable, the final importance score will be obtained.

Bagging algorithm (Zhang et al., 2015; Zhu et al., 2011) uses bootstrap sampling to obtain the data

subsets for training the base learners (i.e. decision trees). In Fig. 3, φ_i is the prediction result of the i^{th} decision tree for the test samples. The output *f(x)* in regression is the average of the results from all trees. For tree-based ensembles, the importance score of a variable is the average value derived from all the trees.

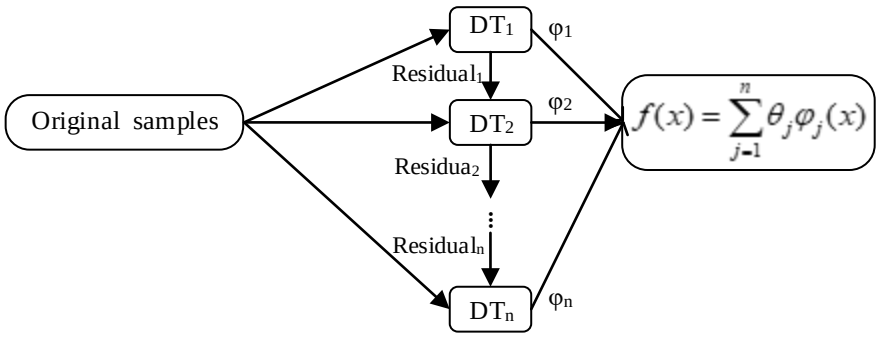


Fig. 3 Block diagram of Bagging algorithm

The GBDT (Zhang et al., 2015; Tuv et al., 2009) algorithm is a decision tree ensemble learning algorithm based on the gradient boosting framework. It uses the original training samples for learning. The i^{th} residual in Fig. 4 is the difference between the predicted result of the i^{th} tree and the target value. It is taken as a new target for the next tree to achieve the concatenation of GBDT algorithm. The output *f(x)* of the algorithm is obtained by summing the prediction results of individual trees. $f_0(x)$ is the initial value of the learner. *lr* is the learning rate, which is used to control the step size and ensure the convergence of the algorithm during the iteration.

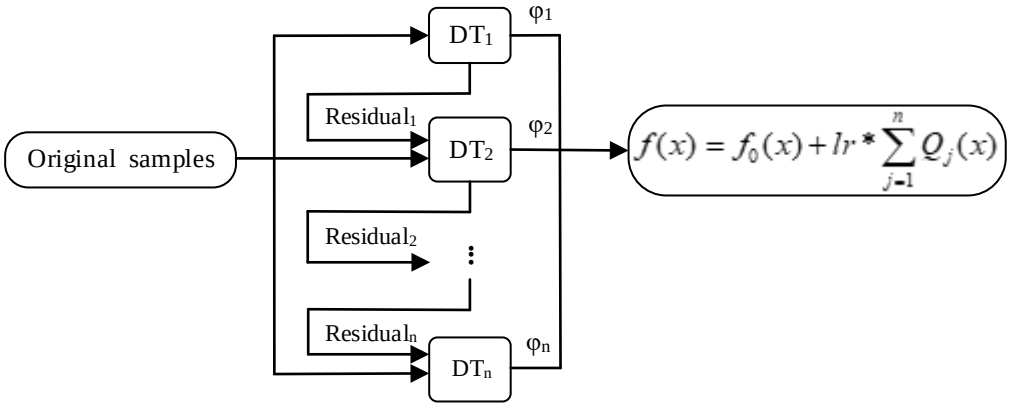


Fig. 4 Block diagram of GBDT algorithm

GBRF algorithm (Tuv et al., 2009) is a mixture of gradient boosting and random forest algorithms. When dividing the samples at each node of the tree in GBRF algorithm, it only takes max-features attributes at random rather than all attributes. The GBRF algorithm introduces random selection of input variables in the splitting process and thus the correlation between the single models is further reduced. The structure of the GBRF algorithm is the same as that of GBDT. By changing the value of different coefficients in the model, different ranking results of variables are obtained.

2.2 Combination strategy

Combination strategy which is used to fuse importance score from individual selectors usually includes averaging method, voting method and learning method. Stacking (Zhou., 2012; Breiman., 1996), as a typical learning method, is to train the first-level selectors using the original training dataset. The fused importance *FI* for a particular variable is the weighted average of importance scores from individual selectors and defined as:

$$FI = \sum_{i=1}^{4} W_i \times VI_i \tag{2}$$

where VI_i is the importance score of the variable from selector i (i=1,2,3,4). W_i is the weighting factor based on prediction accuracy for each selection algorithm and determined by:

$$W_i = \frac{MAPE_{\max} - MAPE_i}{MAPE_{\max} - MAPE_{\min}} \tag{3}$$

where $MAPE_i$ is the prediction error in terms of mean absolute percentage error based on the i^{th} selector. $MAPE_{\max}$ and $MAPE_{\min}$ are the maximum and minimum prediction errors from the four selectors, respectively. The definition of MAPE is shown in equation 4.

Backward elimination algorithm is applied to remove irrelevant and less important variables. The resulted variables are regarded as the input variable for training the next-level selectors. Fig. 5 shows the flow chart of the stacking framework in ensemble variable selection. The selection process repeats until a stop condition is met, either the prediction accuracy approaches the goal or the maximum number of epochs is reached. In this case, the optimal input variables for data-driven modelling of gas-liquid CO2 flow measurement are obtained.

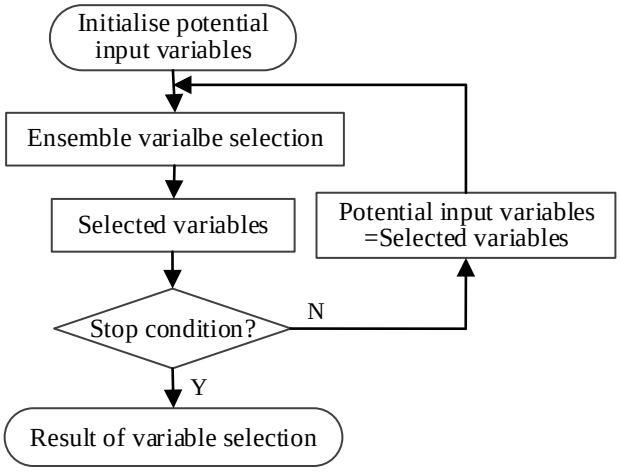


Fig.5 Block diagram of combination strategy based on stacking

3. Experimental results and discussion

3.1 Experimental conditions

The test rig used in this study for gas-liquid two-phase CO_2 flow is shown in Fig. 6. The single gas phase and single liquid phase CO_2 flows are mixed through the mixer and form two-phase flow at the horizontal test section. During the mixing process, the temperature and pressure were kept around the gas/liquid transition line according to the phase diagram of CO_2 in order to achieve gas/liquid two-phase flow conditions. Meanwhile, the temperature and pressure were maintained at constant values via the control system to reduce the likelihood of state change during each test run. Meanwhile, the flowmeter under test was installed at only 1.6 m away from the gas-liquid mixer and the difference in temperature between the mixer and the test section was less than 1℃, so the state change between gas CO_2 and liquid CO_2 was unlikely. Two separate Coriolis flowmeters were installed on the single liquid phase pipeline and single gas phase pipeline, respectively, to provide reference mass flowrate of CO_2 liquid and gas phases. The uncertainty of the flowmeters for the mass flowrate metering of liquid phase and gas phase are 0.16% and 0.35%, respectively. The accuracy of these meters is high enough to be used to obtain reliable reference values. At the test section, a Coriolis flowmeter (KROHNE OPTIMASS 6400 S15) is used as the target instrument to evaluate the proposed method. Pressure, temperature and differential-pressure transducers are added to capture additional information about the flow.

Experimental tests were conducted under the condition of total mass flowrate from 150～3500 kg/h

and the GVF from 1.11%～88.44%. A total number of 541 experimental data were obtained. The time duration under each experimental condition is 100s. The temperature observed at the meter under test over all experiments ranged between 18℃ and 25℃and the pressure was from 5.4 MPa ～ 6.5 MPa.

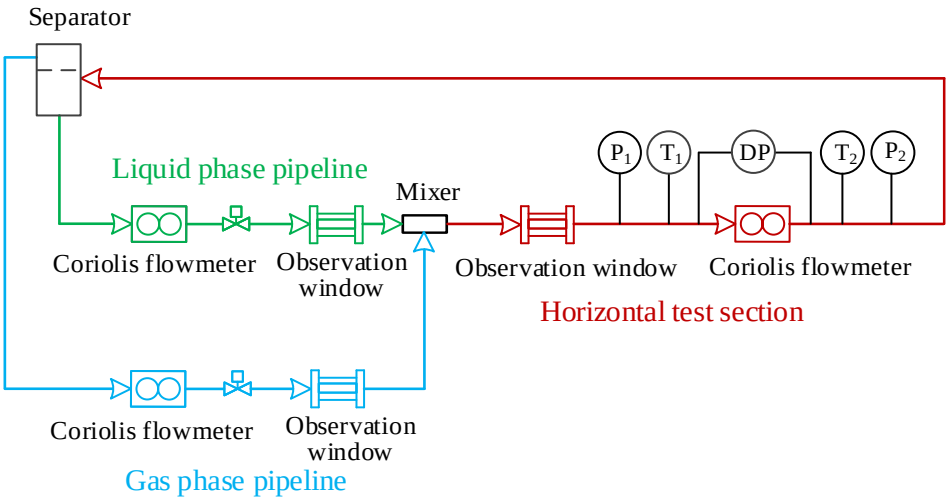


Fig. 6 Test rig for gas-liquid two-phase CO_2 flow

All the variables derived directly or indirectly from sensor signals are listed in Table 1. *x1- x16* are the original attributes collected by the sensors. As the data from the Coriolis flowmeter under test were updated around every 40 ms, *x1-x10* were acquired via the General Device Concept (GDC) protocol with sampling rate of 48 Hz as per the sampling theorem. No filtering or limiting is applied to the data from the meter. Variables *x11-x16* were acquired via an NI (National Instrument) data acquisition card from transducers with a sampling rate of 30 Hz. *x17- x30* are the extended attributes including temperature difference, relative pressure difference and some statistical values of some original attributes. The extended attributes are regarded as potential variables, which contain useful information about the two-phase flow.

Table 1 Input variables and their corresponding physical definitions

ID	Variable name	Physical definition
x1	Apparent mass flowrate ($\dot{q}$)	The mass flowrate reading from the Coriolis flowmeter at the test section
x2	Process temperature (*T*)	The temperature reading from the Coriolis flowmeter at the test section
x3	Observed density (ρ_1)	The density reading from the Coriolis flowmeter at the test

		section
x4	Tube frequency (*f*)	The oscillation frequency reading from the Coriolis measuring tube inside the Coriolis flowmeter
x5	Two phase indicator	An indicator for the detection of a two-phase
x6	Time shift (t_d)	The time delay between the signals reading from the two motion sensors
x7	Observed flow velocity (*v*)	The flow velocity reading from the Coriolis flowmeter at the test section
x8	Sensor A level (V_A/V_{MAX})	The relative voltage amplitude of signals from the motion sensor A
x9	Sensor B level (V_B/V_{MAX})	The relative voltage amplitude of signals from the motion sensor B
x10	Drive level (I_D/I_{MAX})	The relative current amplitude of the driver output
x11	Inlet pressure (P_1)	The pressure of the fluid at the inlet of the Coriolis flowmeter
x12	Inlet temperature (T_1)	The temperature of the fluid at the inlet of the Coriolis flowmeter
x13	Outlet pressure (P_2)	The pressure of the fluid at the outlet of the Coriolis flowmeter
x14	Outlet temperature (T_2)	The temperature of the fluid at the outlet of the Coriolis flowmeter
x15	Temperature different(ΔT)	The temperature difference across the Coriolis flowmeter
x16	Differential pressure(*DP*)	The differential pressure across the Coriolis flowmeter
x17	Relative variance of *DP*	Variance/*differential pressure*
x18	Sensor level different (ΔV)	The relative amplitude difference
x19	Pressure drop (DP/P_1)	Relative ratio of the pressure differential
x20	Damping (*x10/x8*)	Damping factor of the Coriolis measuring tubes
x21	Variance of flow velocity	The variance of the flow velocity
x22	Relative variance of flow velocity	Variance/ observed flow velocity
x23	Skewness of flow velocity	Skewness of flow velocity
x24	Variance of mass flowrate	Variance of mass flowrate
x25	Relative variance of mass flowrate	Variance/ apparent mass flowrate
x26	Skewness of mass flowrate	Skewness of mass flowrate
x27	Variance of density	Variance of density
x28	Relative variance of density	Variance/*observed density*
x29	Skewness of density	Skewness of density
x30	Density drop $(\rho_0-\rho_1)/\rho_0$	Relative ratio of measuring section density to liquid density (ρ_0 is theoretical density of CO_2 liquid phase at certain temperature and pressure)

3.2 Parameter optimization of individual selectors

To improve the performance of selectors there are several parameters need to be optimized. As for the

DT algorithm, the parameter tree depth is to be optimised. Bagging algorithm requires to determine the tree depth and the total number of training trees (i.e. model size). Apart from the tree depth and model size, GBDT and GBRF algorithms need to optimise the parameters of learning rate. As GBRF algorithm has the characteristics of random forest algorithm, the number of maximum input features is another parameter to be optimised. All the parameters are optimised through a trial-and-error approach.

During the process of parameter optimization, MAPE is used to evaluate the performance of each selector and defined as

$$MAPE = \sum_{i=1}^{n} \left| \frac{\hat{Y}_i - Y}{Y} \right| \times \frac{100\%}{n} \quad (4)$$

where Y is the desired value of the target variable; $\hat{Y}_i$ is the predicted value obtained by the i^{th} basic learner, and n is the number of all samples. All the MAPE values in the paper are the average prediction accuracy from 10-fold cross validation.

When the target variable is the mass flowrate of gas-liquid two-phase CO_2, the desired value Y is determined by q_m, the sum of liquid CO_2 mass flow (q_{ml}) and gas CO_2 mass flow (q_{mg}):

$$q_m = q_{ml} + q_{mg} \quad (5)$$

When the target variable is the GVF of gas-liquid two-phase CO_2, the desired value Y is equal to α:

$$\alpha = \frac{q_{vg}}{q_{vl} + q_{vg}} \times 100\% \quad (6)$$

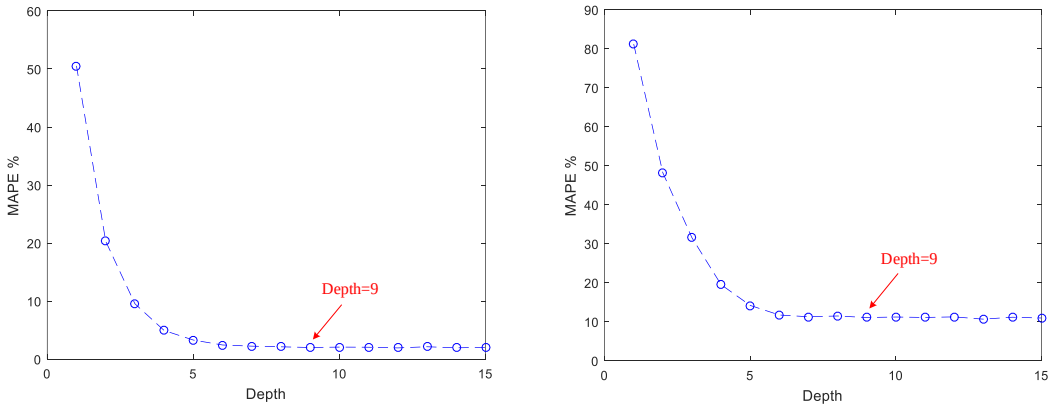
where q_{vl} and q_{vg} are calculated volume flowrates of the liquid and gas phases, repsectively.

The reference value of GVF calculated from equation (6) is based on the assumption that there is no state change between the reference meters and the meter under test.

1) Parameter optimisation for DT and Bagging models

DT models are developed respectively with the tree depth from 1 to 15. As shown in Fig. 7(a) and 7(b),

the MAPE decrease for both mass flowrate measurement and GVF prediction as the tree depth increases. When the depth is greater than 9, the prediction error does not change significantly. In order to avoid over-fitting, the model training depth is set to 9 in single decision tree models.



(a) Performance of mass flowrate models (b) Performance of GVF models

Fig.7 Optimal depth in single tree models

As shown in Fig. 8, for the tree-based ensemble algorithms (Bagging, GBDT and GBRF), the prediction error of the target variable reaches a stable value (about 1.64% for mass flowrate) when the number of training trees is 500. Moreover, prediction error is no longer reduced as the number of training trees increases. Therefore, the number of training trees is set to 500 in the ensemble algorithms.

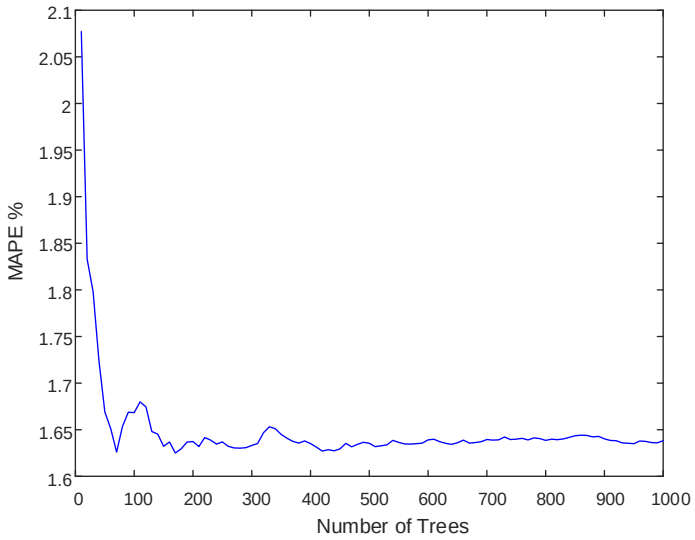
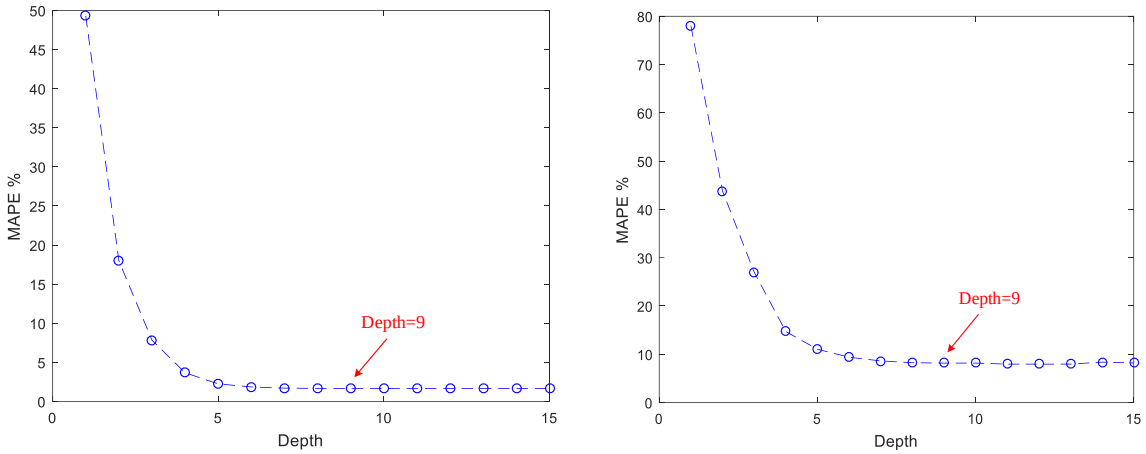


Fig.8 Optimal number of trees in the tree-based ensemble algorithms

Fig. 9 shows the process of determination of optimal depth in bagging algorithms which are used to

predict mass flowrate and GVF, respectively. In these bagging based models, the number of trees is set to 500. It is obvious that the MAPEs value of the prediction is approaching to the minimum of 1.64% for mass flowrate and 8.17% for GVF at a depth of 9. It also verifies the depth selection result in a single decision tree models. Compared with the prediction error of the single decision tree in Fig. 8, Bagging model with the same depth performs better as the result of ensemble learning.

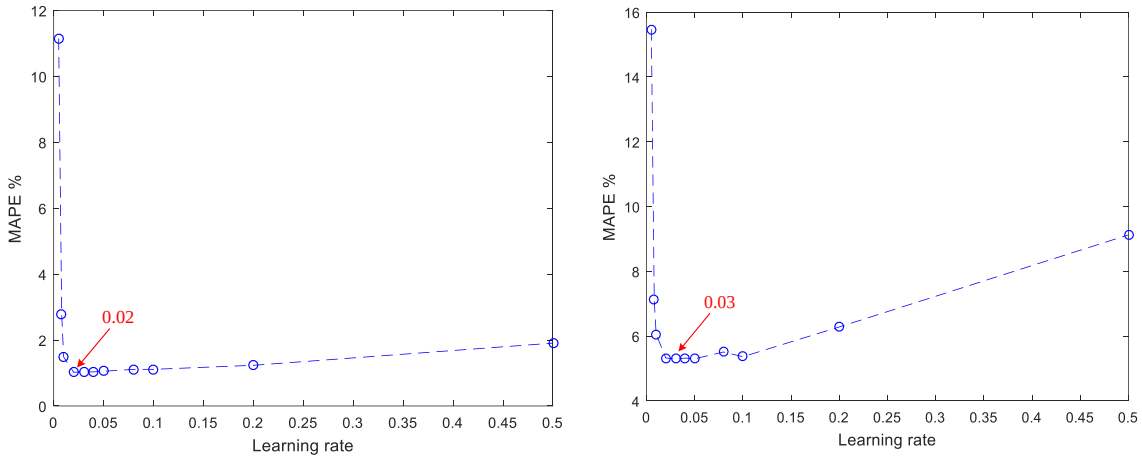


(a) Performance of mass flowrate models (b) Performance of GVF models

Fig.9 Optimal depth in Bagging algorithm

2) Parameter optimisation for GBDT and GBRF models

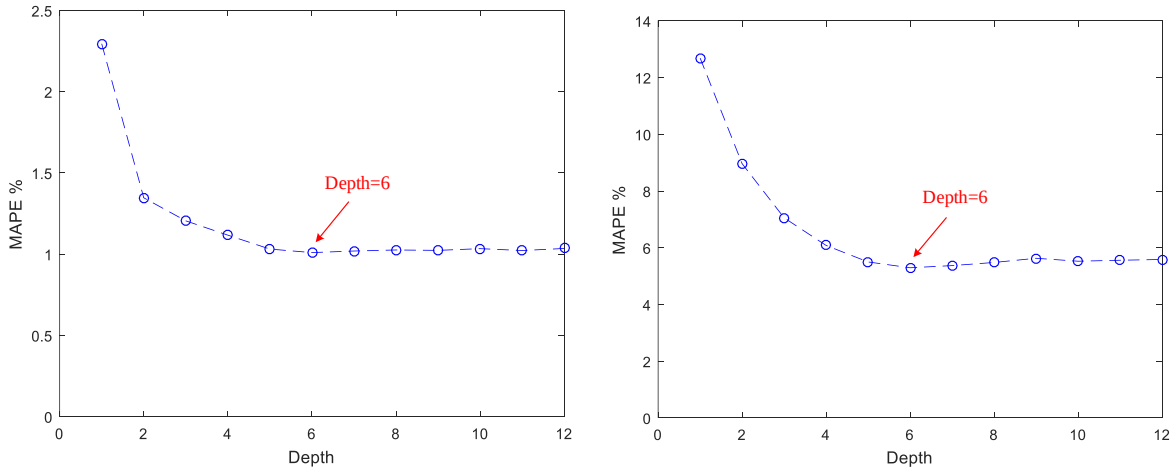
Different from bagging and RF, Gradient boosting (GB) is a serial ensemble and able to reduce bias and variance. GB often has low error values with stumps (a decision tree with a depth of 1) in deeper trees. Before determining their depth, it is necessary to determine the learning rate. At this time, the model is still trained with the depth of 9. The performance of GB models with respect to different learning rates is depicted in Fig.10. When the learning rate varies from 0.001～0.5, 500 trees are trained using GB algorithms at each learning rate. The prediction results of mass flowrate and GVF using GB method with different learning rates are shown in Fig. 11(a) and (b), respectively. When the learning rate is equal to 0.02, MAPE of mass flowrate models reaches the minimum value of 1.02%. For GVF models, the optimized learning rate is 0.03 to achieve a minimal MAPE of 5.3%.



(a) Performance of mass flowrate models (b) Performance of GVF models

Fig.10 Optimal learning rate in GB algorithm

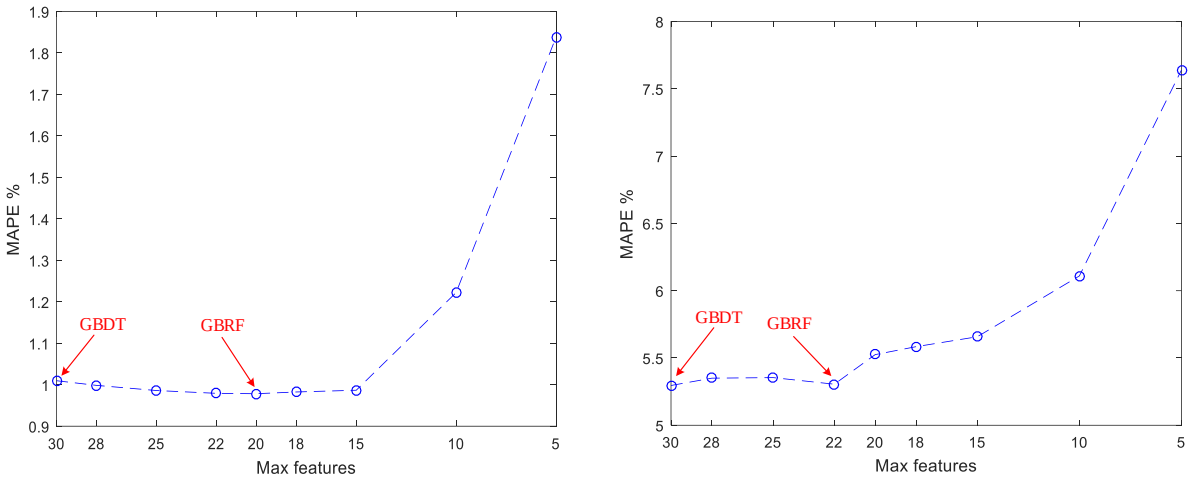
Fig. 11 shows the performance of mass flowrate models and GVF models when the depth of the tree changes from 1 to 12 with a constant learning rate of 0.02 and 0.03. The MAPE value of models reaches the minimum of 1.03% for mass flowrate and 5.3% for GVF, when the depth of the tree is 6. For the serial gradient boosting based models, the model performance can be greatly improved by slightly increasing the depth of the decision tree due to the interaction among the potential variables. As there are some important variables with strong correlation in the potential input variables, it is necessary to apply the gradient lifting algorithm since both GBDT and GBRF algorithms play in a serial ensemble to the basic learners. They have the same depth parameter of 6.



(a) Performance of mass flowrate models (b) Performance of GVF models

Fig.11 Optimal depth in GB algorithm

Since GBRF is a combination of GB and random forest (RF) algorithms, it also need to select the maximum number of features (maxFeature) for model training. The maxFeature means the number of input variables. The learning rate and depth of GBRF are set to the determined values, respectively. As shown in Fig. 12, for the measurement of mass flowrate, the prediction error of the GBRF model reaches the minimum 0.98% when the maximum number of input features is 20. For the prediction of GVF, the GBRF model produce minimum MAPE of 5.3% when the maximum number of input features is 22. However, the optimal maxFeature for GBDT algorithm is 30 which is the total number of all possible input variables.



(a) Performance of mass flowrate models (b) Performance of GVF models

Fig.12 Optimal max features in GBDT and GBRF algorithms

3) Summary of optimal parameters

According to the above analysis, the optimal parameters of the tree-based algorithms are obtained and summarized in Table 2.

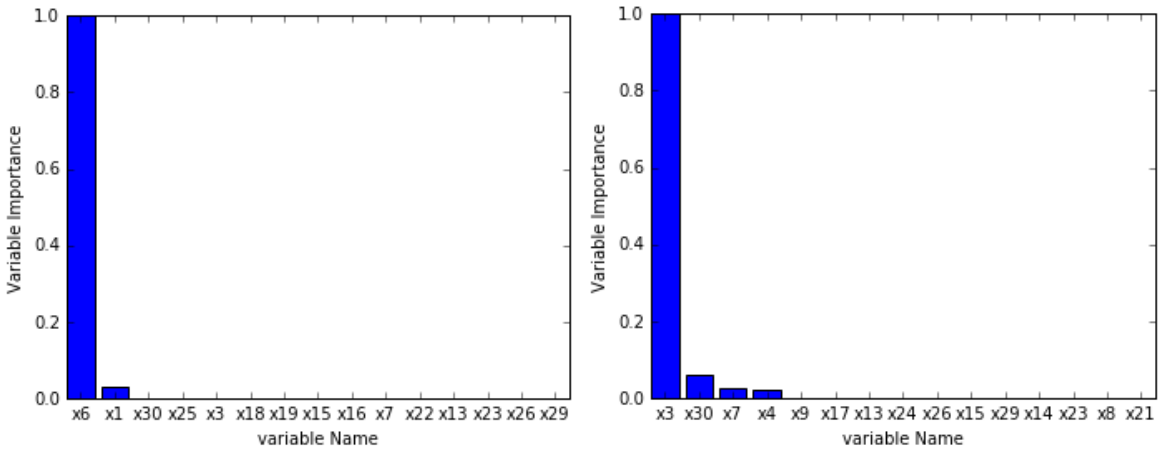
Table 2 Optimal parameter of the basic learners

Parameter		DT	Bagging	GBDT	GBRF
Depth		9	9	6	6
The number of trees		1	500	500	500
Max number of input variables	Mass flowrate	30	30	30	20
	GVF	30	30	30	22
Learning rate	Mass flowrate	/	/	0.02	0.02
	GVF	/	/	0.03	0.03

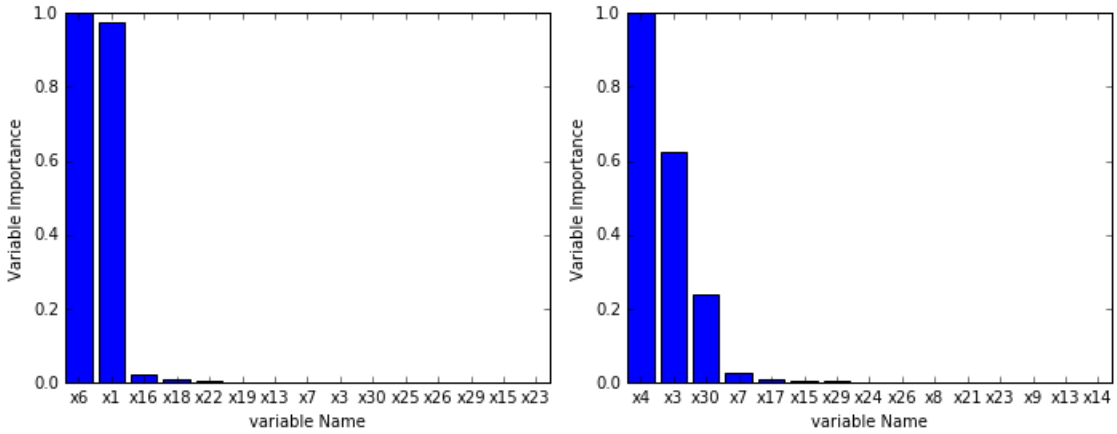
3.3 Implementation of ensemble variable selection

Each of the algorithms is implemented based on the optimal parameters outlined in Table 2. The corresponding results of variable sorting and variable importance scores are obtained. The relative importance of the tree model is represented by the reduction in impurities due to the split on a specific variable set. For ensembles, the metric is averaged over the collection of base learners (Zhang et al., 2017), weighted average is used to derive the combined variable importance.

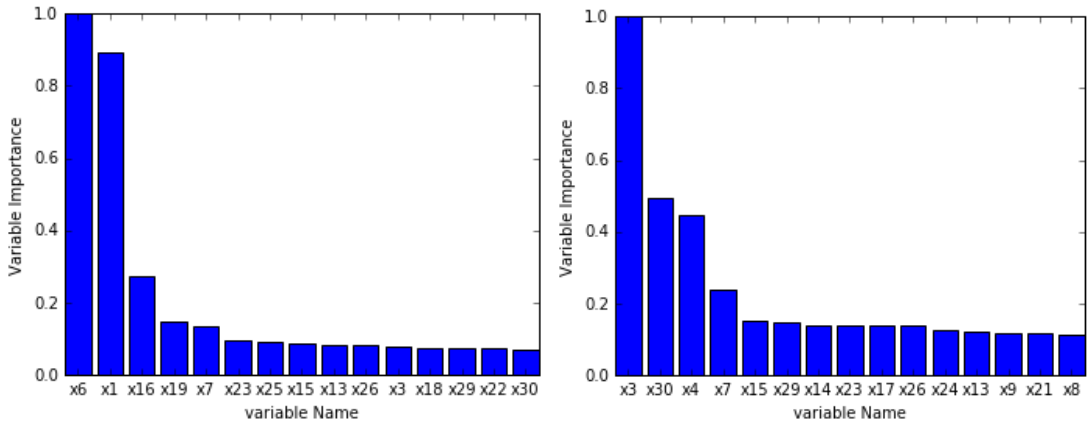
Fig.13 shows the order of first 15 variables obtained by each algorithm according to the normalized variable importance score. It can be seen from each figure, when the number of preselected input variables is 15, the variable importance of each algorithm varies. It is essential to effectively fuse and trim the sorting results of different algorithms to obtain the most accurate and concise set of input variables.



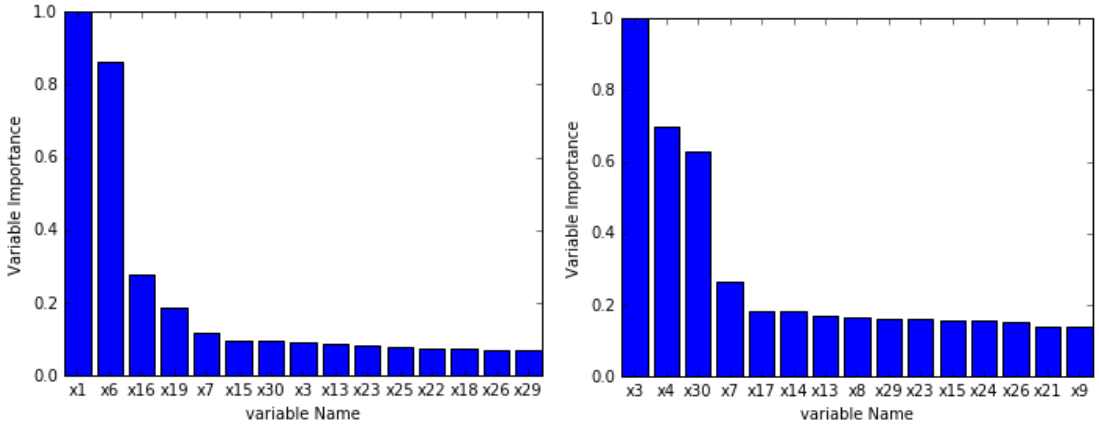
(a) Variable importance to mass flowrate using DT (b) Variable importance to GVF using DT



(c) Variable importance to mass flowrate using Bagging (d) Variable importance to GVF using Bagging



(e) Variable importance to mass flowrate using GBDT (f) Variable importance to GVF using GBDT



(g) Variable importance to mass flowrate using GBRF (h) Variable importance to GVF using GBRF

Fig.13 Variable importance to mass flowrate and GVF based on basics selectors

As shown in Fig. 13(a) and (b), the single algorithm DT can only select the first few most important variables, but cannot distinguish the importance of other variables. In Fig. 13 (c) and (d), the Bagging algorithm is an ensemble of 500 trees and the model performance is the average result of 500 trees. The algorithm can improve the polarization of importance and select more important variables. Therefore, prediction accuracy based on Bagging models is improved compared to single tree algorithm. In Fig. 13 (e) and (f), the use of GBDT algorithm is effective to get rid of the polarization phenomenon. The importance of all variables is smoothly changed. Therefore, the algorithm is superior to DT and Bagging algorithms, and has higher prediction accuracy. In Fig.13 (g) and (h), GBRF algorithm is a combination of gradient lifting and random forest algorithm, which further narrows the gap of the importance score of variables and further improves the prediction accuracy of target variables.

Tables 3 and 4 summarize the results of the combined importance scores of the potential input variables. The algorithm performs 13 iterations of variable ranking and 12 iterations of variable selection. Backward elimination algorithm is applied to remove less important variables. For different prediction target, the results of variable selection are quite different. Finally, the validity of the selected variables is verified through assessing the performance of data-driven models.

Table 3 Results of variable selection at all levels for mass flowrate

The order of importance of tree-based feature selection													
Number / Index	30	25	20	15	10	8	7	6	5	4	3	2	1
x1	*x1*	*x6*	*x6*	*x1*	*x6*	*x6*	*x6*	*x6*	*x6*	*x1*	*x1*	*x6*	*x6*
x2	*x6*	*x1*	*x1*	*x6*	*x1*	*x1*	*x1*	*x1*	*x1*	*x6*	*x6*	*x1*	
x3	*x16*	*x16*	*x16*	*x16*	*x16*	*x16*	*x16*	*x16*	*x16*	*x16*	*x16*		
x4	*x19*	*x19*	*x19*	*x19*	*x19*	*x19*	*x19*	*x19*	*x19*	*x19*			
x5	*x7*	*x25*	*x7*	*x7*	*x7*	*x7*	*x7*	*x7*	*x7*				
x6	*x25*	*x9*	*x25*	*x13*	*x25*	*x13*	*x25*	*x25*					
x7	*x22*	*x7*	*x15*	*x25*	*x13*	*x25*	*x13*						
x8	*x23*	*x22*	*x30*	*x15*	*x15*	*x15*							
x9	*x26*	*x23*	*x3*	*x23*	*x22*								
x10	*x29*	*x29*	*x26*	*x22*	*x23*								
x11	*x15*	*x30*	*x22*	*x30*									
x12	*x21*	*x24*	*x23*	*x26*									
x13	*x18*	*x26*	*x13*	*x18*									
x14	*x24*	*x15*	*x18*	*x29*									
x15	*x8*	*x18*	*x29*	*x3*									
x16	*x28*	*x8*	*x21*										
x17	*x27*	*x3*	*x2*										
x18	*x17*	*x13*	*x24*										

x19	*x9*	*x21*	*x9*										
x20	*x30*	*x2*	*x8*										
x21	*x2*	*x28*											
x22	*x3*	*x17*											
x23	*x13*	*x20*											
x24	*x12*	*x12*											
x25	*x20*	*x27*											
x26	*x4*												
x27	*x14*												
x28	*x11*												
x29	*x10*												
x30	*x5*												

342

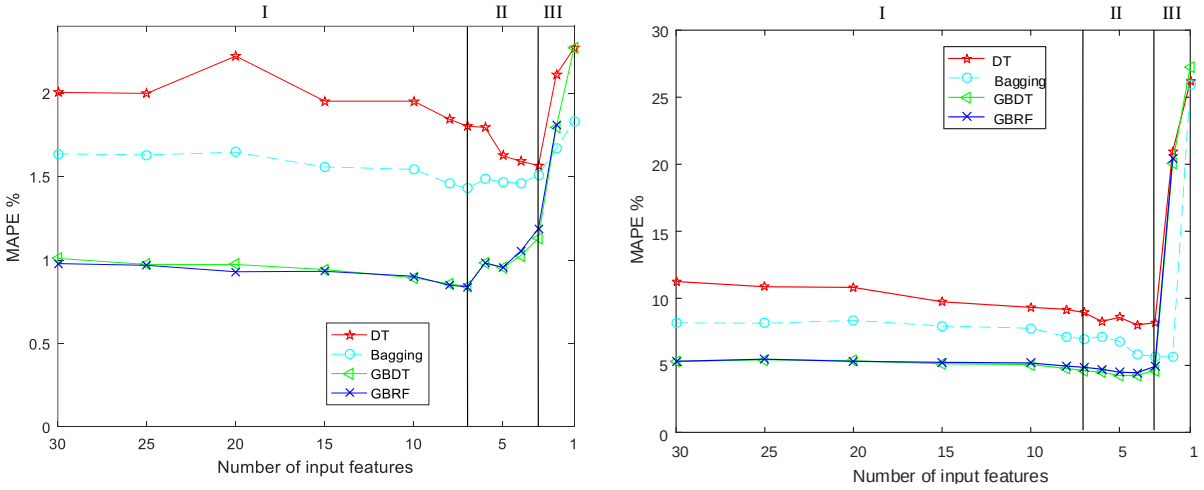
343 Table 4 Results of variable selection at all levels for GVF

The order of importance of tree-based feature selection													
Number / Index	30	25	20	15	10	8	7	6	5	4	3	2	1
x1	*x3*	*x3*	*x3*	*x3*	*x3*	*x3*	*x3*	*x3*	*x3*	*x3*	*x3*	*x3*	*x3*
x2	*x30*	*x30*	*x30*	*x30*	*x30*	*x30*	*x30*	*x30*	*x30*	*x30*	*x30*	*x30*	
x3	*x4*	*x4*	*x4*	*x4*	*x4*	*x4*	*x7*	*x7*	*x7*	*x7*	*x7*		
x4	*x29*	*x23*	*x29*	*x7*	*x7*	*x7*	*x4*	*x4*	*x14*	*x14*			
x5	*x23*	*x29*	*x23*	*x29*	*x14*	*x24*	*x14*	*x14*	*x4*				
x6	*x26*	*x26*	*x26*	*x24*	*x24*	*x14*	*x24*	*x24*					
x7	*x24*	*x15*	*x15*	*x26*	*x23*	*x23*	*x23*						
x8	*x15*	*x24*	*x24*	*x14*	*x29*	*x29*							
x9	*x17*	*x8*	*x17*	*x23*	*x26*								
x10	*x21*	*x17*	*x13*	*x15*	*x15*								
x11	*x8*	*x9*	*x7*	*x17*									
x12	*x9*	*x18*	*x8*	*x8*									

x13	*x27*	*x27*	*x14*	*x13*									
x14	*x18*	*x21*	*x21*	*x9*									
x15	*x7*	*x7*	*x9*	*x21*									
x16	*x28*	*x25*	*x27*										
x17	*x14*	*x14*	*x18*										
x18	*x25*	*x28*	*x25*										
x19	*x11*	*x13*	*x28*										
x20	*x19*	*x16*	*x16*										
x21	*x12*	*x22*											
x22	*x13*	*x12*											
x23	*x22*	*x11*											
x24	*x2*	*x19*											
x25	*x16*	*x2*											
x26	*x20*												
x27	*x5*												
x28	*x10*												
x29	*x6*												
x30	*x1*												

In the process of variable selection, the sooner the feature is removed, the less important it is. As shown in Fig.10, this paper divides the elimination process of feature selection into three stages. At the first stage, the range of input features retained after feature selection 7～30. At this stage, unimportant variables are gradually eliminated until 7 variables are retained. Because the removed variables are insignificant to the target variable, the prediction accuracy of the model is gradually improved. At the second stage, the range of input variables retains after feature selection is 3～7. At this stage, it can be seen that the lack of sub-important variables leads to a small increase in the prediction error, but it does not have much impact on the overall performance. As shown in Fig.14(a), the prediction performance of mass flowrate models is to be improved when the number of selected features is 5. This is because the extended variables of important variables play an important role in predicting mass

flowrate. Fig. 14(b) shows the performance of GVF models. The MAPE value tends to increase when the number of selected features is more than 3. At the third stage, the range of input features retained after feature selection is 1～3. At this stage, one of the important variables required by the model is removed by fusion method, which leads to a rapid decline in the prediction accuracy of the model. Therefore, the optimal number of input variables for the prediction model should be determined at the second stage.



(a) CO_2 mass flowrate models (b) CO_2 GVF models

Fig.14 Performance comparison of variable selection at different levels

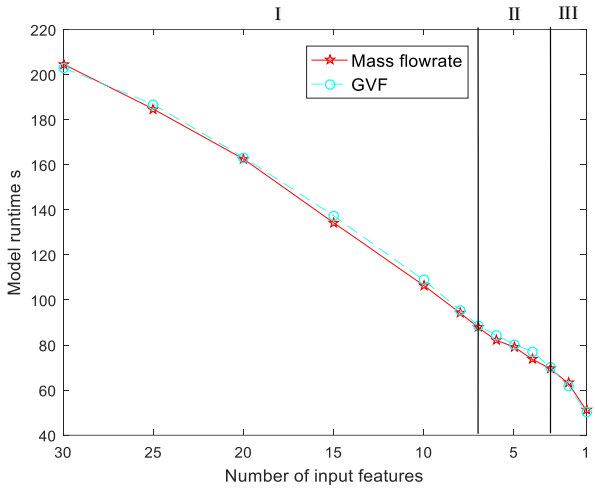


Fig.15 Model efficiency comparison of mass flowrate and GVF at different levels

The model runtime in Fig. 15 is obtained, which represents the average run time of 54 test data (The processor of the computer used is Intel Xeon E5-2640 V4 CPU @ 2.4 GHz). The total running time of

each iteration is the sum of the running time of the four selection algorithms and their combination time. As shown in Fig. 16, as the number of input features is gradually reduced, the complexity of the model is reduced, and the time efficiency of the model is greatly improved. When the number of input features of the prediction model is 4, the efficiency of the GBRF algorithm is increased by 3 times compared with the feature number of 30. When the feature selection process reaches the second stage, the removal of a large number of variables unrelated to the predicted target value improves the model efficiency rapidly. This point indicates that the fusion variable selection method in this paper is very effective for large data measurement systems with a large number of input variables. Under the same number of input features, the running time of GVF prediction model is slightly longer than that of mass prediction model. This is because the relationship between the target variable GVF and the variable candidates is more complex and difficult to find.

The ensemble tree algorithm can not only select variables, but also derive the predicted results of target variables. Three indicators are integrated to determine the final variable selection results: prediction accuracy (MAPE), model efficiency (running time) and model complexity (number of input variables) in Figs 14 and 15. In Fig.14 (a), the MAPE of the three ensemble algorithms reaches the minimum value when seven variables are used as model inputs. In Fig. 14 (b), the MAPE of the three ensemble algorithms reaches the minimum value when four variables are used as model inputs. Based on the analysis results of the above three evaluation indicators, four input variables are determined as the final selection results for both mass flowrate and GVF prediction models. The results of variable selection are shown in table 5. The MAPE value increases slightly for mass flowrate when the number of model input variables is 4. However, the model efficiency is greatly improved and the model complexity is reduced.

The input variable for the mass flowrate prediction model is determined as{*x1 x6 x16 x19*} while the input variable for the GVF prediction model is {*x3 x30 x7 x14*}.The input variables selection results using the proposed the tree-based heterogeneous ensemble approach is summarized in Table 5.

Table 5 Result of selection of four input variables

Index	Mass flowrate	GVF

1	*x1*-Apparent mass flowrate ($\dot{m}$)	*x3*-Observed density (ρ_1)
2	*x6*-Time shift (t_d)	*x30*-Density drop ((ρ_0-ρ_1)/ρ_0)
3	*x16*-Differential pressure (*DP)*	*x7*-Observed flow velocity (*v*)
4	*x19*- Pressure drop (*DP /P*$_1$)	*x14*-Outlet temperature (T_2)

According to the physical meaning of the input variables and previous theoretical and experimental study (Henry et al., 2006; Li et al., 2018), the selection results are further analyzed. Apparent mass flowrate ($\dot{m}$) *x1* is measured in the horizontal test section, even though the CMF produces large errors in measuring the mass flowrate of two-phase flow (the original measurement error). It is still related to the desired CO_2 mass flowrate. Time shift (t_d) *x6* and apparent mass flowrate has a functional relationship, so time shift is also correlated with the desired CO_2 mass flowrate. As the liquid CO_2 flowing through the meter with various gas CO_2 entrainment, the pressure difference *x*16 across the Coriolis flowmeter and relative pressure difference *x*19 can characterise the mixed CO_2 flow to some extent.

For variable selection of GVF, *x*3 is the observed density of gas-liquid mixture phase, and density drop *x30* is derived from the observed density and the liquid phase density, which can somehow reflect GVF. Although *x7* observed flow velocity is not accurate two-phase flow velocity, it still can reflect the variation of mixed flow in the pipe. As the physical properties of CO_2 are very sensitive to the variations in fluid temperature and pressure, so when temperature increases, phase change from liquid to gas may occur and hence increasing GVF. Therefore, x14 temperature is also an important variable for GVF prediction.

In Table 5, *x1* (apparent mass flowrate), *x6* (time shift), *x16* (differential pressure) and *x19* (pressure drop) are selected for mass flowrate measurement. *x3* (observed density), *x30* (density drop), *x7* (observed flow velocity) and *x14* (outlet temperature) are selected for GVF prediction. Variables *x1* & *x6*, *x3* & *x30*, and *x16* & *x19* look like highly redundant pairs. However, the mathematical relationship between each pair is complex particularly in the case of gas-liquid two-phase flow. For instance, *x1* is derived from *x6*, but their exact relationship depends on the fluid temperature and material properties of the sensing tube (Wang et al 2017). Since *x1* includes temperature compenstation and material property effect, both *x1* and *x6* are selected in this case. *x19* is the ratio of differential pressure (*x16*) to

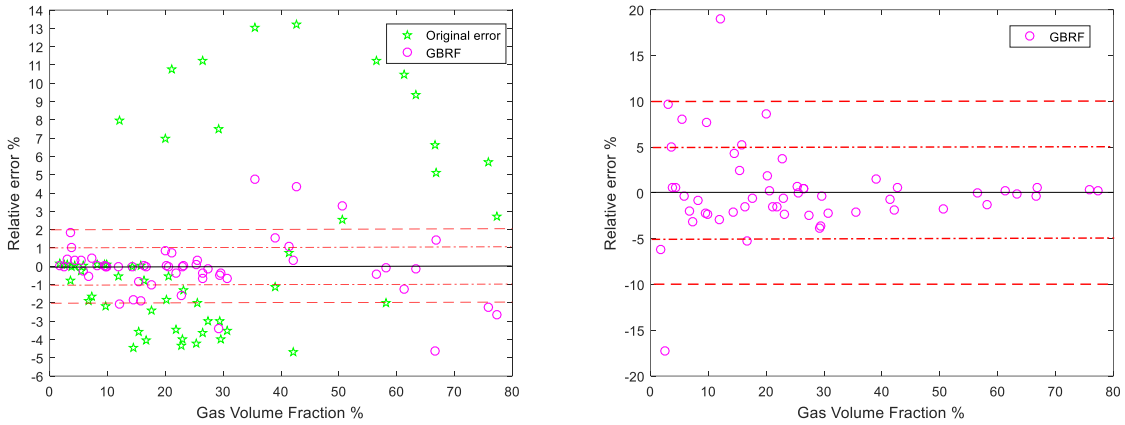
the inlet pressure (*x11*). *x30* is the relative ratio of the observed density (*x3*) to liquid density ρ_0 (ρ_0 is the theoretical density of CO_2 liquid phase at certain temperature and pressure). Additional fluid information is included in *x6*, *x19* and *x30* than those in *x1*, *x16* and *x3*.

The selection processes produce different importance scores for these variables. In consideration of the prediction accuracy and model complexity of the data-driven models, the combination of the variables outlined in table 5 are taken as the 'optimal' inputs to the tree-based models. If both variables from each pair are used as input features, the MAPE values of the tree-based models will be reduced by at least 0.5% for mass flowrate and 7% for GVF. In this case, these pairs provide complementary information for the data-driven models.

3.4 Evaluation of ensemble variable selection

GBRF models were developed based on the selected variables. The model performance was assessed with 54 test samples (mass flowrate range: 212 kg/h～3449 kg/h and GVF range: 1.82%～77.29%). As can be seen from Fig.16(a), the method proposed in this paper has improved the measurement accuracy over the direct mass flow measurements. The prediction accuracy of GVF from the GBRF model with selected input variables is shown in Fig.16(b). It turns out that the relative error in mass flowrate measurement is mostly within ±1.0% and GVF prediction mostly within ±5.0%.

When GVF is low, the gas entrainment has less effect on the vibration of the flowtubes in the Coriolis mass flowmeter, which makes the apparent mass flowrate still very close to the true mass flowrate. In this case, it is easier for the data-driven models to correct the errors. As GVF increases, the large bubbles in the mixed flow lead to larger difference and nonlinear relationship between the apparent mass flowrate and true mass flowrate. In this case, it is more challenging for the data-driven models to derive the relationship and hence relative large errors than low GVF conditions.



(a) Mass flowrate (b) GVF

Fig.16 Model performance with the selected variables

To further verify the selected results, a comparative experiment was conducted. A model is established with two input variables *x1* (apparent mass flowrate) and *x30* (density drop), respectively, for mass flow measurement and GVF prediction. A total of 108 samples for mass flowrate measurement and GVF prediction are tested. The performance of the models with two inputs and the selected inputs are summerised in Table 6. The results demonstrate that more test results from the model with selected inputs lie in the expected error range. 73% of test data produce relative error within ±1.2% for mass flow measurement and 80% of test data produce relative error within ±4% for GVF prediction.

Table 6 Performance comparison of data-driven models with different inputs

Comparison of mass flow prediction		
	{*x1*,*x30*}	{*x1*,*x6*,*x16*,*x19*}
Relative error within±1.2%	70%	73%
Comparison of GVF prediction		
	{*x1*,*x30*}	{*x3*,*x7*,*x14*,*x30*}
Relative error within±4%	69%	80%

4. Conclusions

A tree based heterogeneous embedded ensemble approach has been proposed for variable selection and applied to gas-liquid CO_2 two-phase flow measurement in this paper. Based on the combination

strategy of stacking and weighted averaging, the proposed method fuses the variable selection results from four single selectors. At the same time, mass flowrate measurement and GVF prediction of gas-liquid two-phase CO_2 flow have been carried out using the selected variables as inputs to the GBRF models. The relative error of mass flowrate from the GBRF model is mostly within 1% with the selected input variables (*apparent mass flow rate, time shift, differential pressure* and *pressure drop*). The prediction error of GVF is mostly less than 5% using the selected input variables (*observed density, density drop, observed flow velocity, outlet temperature*). The outcome from such modelling research will help to enhance the understanding of two-phase flow measurement. Meanwhile, the results presented in the paper demonstrate that the proposed heterogeneous ensemble approach is capable of providing a small number of input variables and developing effective data-driven models for multiphase flow measurement. In the further, more effort will be made to improve the transferability of the developed data-driven model.

Engineering judgement here is still important as we have some knowledge of the two-phase flow and Coriolis sensing process. Meanwhile, research is ongoing through analytical modelling of the gas-liquid two-phase flow, which is a related area of research we are working on. The results from such modelling research will help enhance engineering judgement. However, the variable selection as reported in this paper will assist the optimisation of the machine learning models significantly. The results presented in the paper demonstrate that the proposed heterogeneous ensemble approach is capable of providing a small number of input variables and developing effective data-driven models for multiphase flow measurement.

Acknowledgements

The authors would like to acknowledge the financial support of the National Natural Science Foundation of China (No. 61973113 and No. 62073135). This work is also supported by Fundamental Research Funds for the Central Universities (2020MS015) and by the UK CCS Research Centre (www.ukccsrc.ac.uk). The UK CCSRC is funded by the EPSRC as part of the RCUK Energy Programme.

500 **References**

501 [1] AL-Qutami, T. A., Ibrahim, R., Ismail, I., Ishak, M. A., 2018. Virtual multiphase flowmetering
502 using diverse neural network ensemble and adaptive simulated annealing. Expert Syst. Appl. 93,
503 72-85.

504 [2] Breiman, L., 1996. Stacked regressions. Mach. Learn. 24(1),49-64.

505 Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32(2),
506 407-499.

507 [3] Green, T., Reese, M., Henry, M., 2008. Two-phase CO_2 measurement and control in the Yates oil
508 field. Measurement and Control , 41 (7), 205-207.

509 [4] Henry, M., Tombs, M., Duta, M., Zhou, F.B., Mercado, R., Kenyery, F., Langansan, R., 2006.
510 Two-phase flow metering of heavy oil using a Coriolis mass flow meter: A case study. Flow Meas.
511 Instrum. 17 (6), 399-413.

512 [5] Kunze, J.W., Storm, R., Wang, T., 2014. Coriolis mass flow measurement with entrained gas.
513 Sensors and Measuring Systems. 17. ITG/GMA Symposium. VDE, 1-6.

514 [6] Li, J.L., Zhang, C.X., 2017. Ensembling variable selectors by stability selection for the Cox model.
515 In 2017 International Conference on Machine Learning and Cybernetics (ICMLC). 1, 35-41. IEEE.

516 [7] Li, M., Henry, M., Zhou, F., Tombs, M., 2019. Two-phase flow experiments with Coriolis Mass
517 Flow Metering using complex signal processing. Flow Meas. Instrum. 69 , 101613.

518 [8] Li, M., Henry, M., 2018. Complex signal processing for Coriolis mass flow metering in two-phase
519 flow，Flow Meas. Instrum. 64, 104-115.

520 [9] Mendes-Moreira, J., Soares, C., Jorge, A. M., Sousa, J. F. D., 2012. Ensemble approaches for
521 regression: A survey. ACM Comput. Surv. 45(1), 10.

522 [10] Nan,. Y.,Yang, Y., 2014.Variable selection diagnostics measures for high-dimensional
523 regression. J. Comput. Graph. Stat. 23(3),636-656.

524 [11] Shao, D., Yan, Y., Zhang, W.B, Sun, S.J, Sun, C.Y, Xu, L.J，2020. Dynamic measurement of gas
525 volume fraction in a CO_2 pipeline through capacitive sensing and data driven modelling. Int. J. of
526 Greenh. Gas Control. 94, 102950.

527 [12] Tuv, E., Borisov, A., Runger, G., Torkkola, K. , 2009. Feature selection with ensembles,artificial
528 variables, and redundancy elimination. J. Mach. Learn. Res.10(Jul), 1341-1366.

529 [13] Wang, L.J, Yan, Y., Wang, X., Wang, T., 2017. Input variable selection for data-drivenmodels of

Coriolis flowmeters for two-phase flow measurement. Meas. Sci. Technol. 28(3), 035305.

[14] Wang, L. J, Yan, Y., Wang, X., Wang, T., Duan, Q., Zhang, W., 2018. Mass flow measurement of gas-liquid two-phase CO_2 in CCS transportation pipelines using Coriolis flowmeters. Int. J. of Greenh. Gas Control. 68, 269-275.

[15] Wang, J.Z., Wu, L.S., Kong, J., Li, Y.X., Zhang, B.X., 2013. Maximum weight and minimum redundancy: a novel framework for feature subset selection. Pattern Recognit. 46(6), 1616-1627.

[16] Xin, L., Zhu, M., 2012. Stochastic stepwise ensembles for variable selection. J. Comput. Graph. Stat. 21(2), 275-294.

[17] Yan, Y., Wang, L. J, Wang, T., Wang, X., Hu, Y., Duan, Q., 2018. Application of softcomputing techniques to multiphase flow measurement: A review. Flow Meas. Instrum. 60, 30-43.

[18] Zhang, C.X., Zhang, J.S., Yin, Q.Y., 2018. Early stopping aggregation in selectivevariable selection ensembles for high-dimensional linear regression models. Knowledge-Based Syst. 153,1-11.

[19] Zhang, C.X., Wang, G.W., Liu, J.M., 2015. RandGA: injecting randomness into parallelgenetic algorithm for variable selection. J. Appl. Stat. 42(3), 630-647.

[20] Zhang, C.X., Zhang, J.S., Yin, Q.Y., 2017. A ranking-based strategy to prune variable selection ensembles. Knowledge-Based Syst. 125, 13-25.

[21] Zhu, M., Chipman, H.A., 2006. Darwinian evolution in parallel universes: A parallelgenetic algorithm for variable selection. Technometrics 48(4), 491-502.

[22] Zhu, M., Fan, G., 2011. Variable selection by ensembles for the Cox model. J. Stat. Comput. Simul. 81(12), 1983-1992.

[23] Zhou, Z.H., 2012. Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.

[24] Zhang, W.B, Shao, D., Yan, Y., Liu, S., Wang, T., 2018. Experimental investigationsintothe transient behaviours of CO_2 in a horizontal pipeline during flexible CCS operations. Int. J. of Greenh. Gas Control. 79, 193-199.