IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 7, 2020, accepted July 30, 2020, date of publication August 11, 2020, date of current version August 24, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3015875*

# Multisource Latent Feature Selective Ensemble Modeling Approach for Small-Sample High-Dimensional Process Data in Applications

**JIAN TANG[1,2], (Member, IEEE), JIAN ZHANG[3], GANG YU[4], WENPING ZHANG[5], AND WEN YU[6], (Senior Member, IEEE)**

[1]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[2]Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China
[3]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
[4]State (Beijing) Key Laboratory of Process Automation in Mining and Metallurgy, Beijing 102600, China
[5]Metallurgical Laboratory Branch, Shandong Gold Mining Technology Company Ltd., Jinan 250014, China
[6]Departamento de Control Automatico, CINVESTAV-IPN (National Polytechnic Institute), Mexico City 07360, Mexico

Corresponding author: Gang Yu (yugang@bgrimm.com)

This work was supported in part by the National Key Research and Development Program of the Ministry of Science and Technology under Grant 2018YFC1900800-5; in part by the National Science Foundation of China under Grant 61703089, Grant 61873009, Grant 61803191, and Grant 61973226; in part by the Beijing Natural Science Foundation under Grant 4192009; and in part by the National and Beijing Key Laboratory of Process Automation in Mining and Metallurgy under Grant BGRIMM-KZSKL-2020-02.

**ABSTRACT** Several difficult-to-measure production qualities or environment pollution indices of industrial process must be measured using offline laboratory instruments. Soft measurement method is often used to perform online prediction of such parameters. Only small-sample modeling data with high-dimensional input features can be obtained due to the limitations and complex characteristics of the measurement device and process, respectively. Therefore, a new multisource latent feature selective ensemble (SEN) modeling approach is proposed in this study. First, input features are divided into different subgroups according to the characteristics of the modeling data. Second, the extracted multisource latent features evolve from the multi-layered selection algorithms, which are specified by feature reduction ratio, feature contribution ratio and mutual information value orderly for each subgroup. Finally, in order to construct candidate sub-models, an adaptive hyper-parameter selection algorithm based on the multi-step grid search is employed in terms of the reduced features. Sequentially, the optimized ensemble submodels with their weighting strategies are adaptively determined to build the final SEN model. The proposed method is verified by using benchmark near-infrared data, high dimensional mechanical frequency spectrum data and industrial dioxin emission concentration data.

**INDEX TERMS** Multisource feature extraction, multi-layered feature selection, selective ensemble modeling, hyperparameter selection, high dimensional process data.

## I. INTRODUCTION

Reducing energy consumption and pollution emission of complex industrial processes by using control strategy to optimize operation is an open issue that needs to be solved [1], [2]. To achieve the above target, most key process parameters relative to production qualities or environment pollution indices of such industrial process should be measured online first [3]. Limited by the complexity and strong coupling characteristic of the industrial process, these process parameters must be measured using offline laboratory instruments. These parameters, such as the dioxin (DXN) emission concentration of the municipal solid waste incineration (MSWI) process [4] and the mill load parameter of the mineral grinding process [5], are difficult to measure. Offline methods based on domain expert estimation and laboratory analysis experience are difficult to assist the realization of operational optimization and control. Establishing soft measurement models for these difficult-to-measure parameters by using offline historical data effectively solves this problem [6]. Only a small sample

The associate editor coordinating the review of this manuscript and approving it for publication was Shih-Wei Lin.

of modeling data with high-dimensional input features can be obtained due to the limitations and complex characteristics of the measurement device and process, respectively. Thus, the number of input features is always greater than the number needed to build an efficient and concise model with characteristics that can be physically interpreted. Moreover, an effective strategy based on prior knowledge to determine the input features is difficult to obtain for industrial processes with complex multidisciplinary mechanisms, such as the MSWI process for measuring DXN emission concentration.

The increase in input features makes obtaining complete training samples difficult [7]. Thus, the dimensions must be reduced. Normally, feature selection and extraction methods can be used to achieve this target [8]. To improve the stability of feature selection, the number of modeling examples can be increased [9], [10] or the dimensionality of the input features can be reduced [11], [12]. This paper focuses on the issue of dimensionality reduction in a new perspective. The ratio of the training samples to the reduced features indicates that the value should meet the requirements of constructing a robust learning model. Normally, it must be satisfied. At present, the commonly used method is feature selection based on mechanism or experience. Thus, most features have to be dropped, and certain information may be lost. Moreover, input features in different regions may have various physical meanings [5]. In process industries, input features also correspond to different stages of the whole process. Hence, feature subgroups of the whole features have different meanings. Extracting the latent features' subsets that can represent different feature subgroups may be a good choice. Thus, feature selection may not be the best method for a small sample of high-dimensional dataset. Fortunately, the latent feature extraction method can extract implication changes of high-dimensional data, among which principal component analysis (PCA) is commonly used in modeling difficult-to-measure parameters of industrial process [13]. However, using principal component (PC) with low-contribution-rate modeling reduces the stability of model prediction. Moreover, the correlation between the above latent features and these key process parameters may be weak. Therefore, not only the latent features whose contribution rate meets the stable modeling requirement must be reselected, but also the relativity of these latent features to the predicted key process parameters should be concerned.

In theory, the support vector machine (SVM) algorithm based on the structural risk minimization criterion can effectively model small sample data [14], [15]. However, SVM needs to solve the quadratic programming (QP) problem, whose hyperparameters are difficult to be selected adaptively. Least squares–support vector machine (LS–SVM) overcomes the QP problem by solving linear equations [16]. However, the selection of hyperparameters is still an open issue at present. Although these parameters can be obtained by optimization algorithms such as genetic algorithm and difference evolution [17]–[19], they are time consuming and obtain only suboptimal solutions [20]. Therefore, the above researches lack an adaptive selection mechanism for hyperparameters.

Aiming at high-dimensional datasets, the latent features extracted from different subgroups can be represented as multisource local information. Similarly, latent features extracted from all input features can be used as global information. Thus, they can be used to construct different submodels with various prediction accuracies. In the soft measurement model based on selective ensemble (SEN), certain valuable submodels have better stability and robustness than the traditional single one. In theory, the performance of the SEN model relates to the diversities of the ensemble submodels. Ref. [21] reviews the diversity construction strategies, in which the predictive model based on feature space ensemble construction strategy has the best generalization performance. Moreover, the combination method must be selected carefully to obtain the most accurate and stable predictions [22]. However, the majority of present research focuses on classification problems, such as hierarchical ensemble methodology that promotes diversity among the elements of an ensemble [23], ensemble learning method based on dropout technique [24] that maximizes diversity by transformed ensemble learning [25], [26], and ensemble different fine-tuned convolutional neural networks with SVM [27]. For the regression modeling problem based on a small sample of multisource high-dimensional spectral data, Tang *et al.* proposed the SEN model based on selective fusion of multisource features and multicondition samples with adaptive weighting algorithm [28], [29]. Furthermore, a subspace-based general framework for ensemble learning is proposed [30]. Recently, ensemble learning model based on evolutionary algorithm and LS–SVM has been proposed [31], whose long offline training consumption is unavoidable. Different input features also relate to various hyperparameters of the SEN model and its ensemble submodels. Multiobjective evolutionary can be employed to address this problem [32]–[34]. However, an effective joint optimization strategy can also solve the above problems.

On the basis of the multiple region/stage characteristics of small-sample high-dimensional process data, a new multisource latent feature SEN modeling approach is proposed. By extracting the latent features of different subgroups from original input features and selecting them with three-layer feature selection method, dimensionality reduction is performed, so that the soft-sensor model with a hyperparameter adaptive selection and SEN modeling mechanism is then constructed. Simulation results verify that the proposed approach can achieves efficacy and effectiveness for near-infrared benchmark data, high dimensional mechanical vibration frequency spectrum data and actual dioxin emission concentration data.

This study has the following contributions. First, a new modeling approach for small high-dimensional process data is proposed, which can utilize all the input features without losing useful information. Second, the latent features, selected in three-layer feature selection methods which is

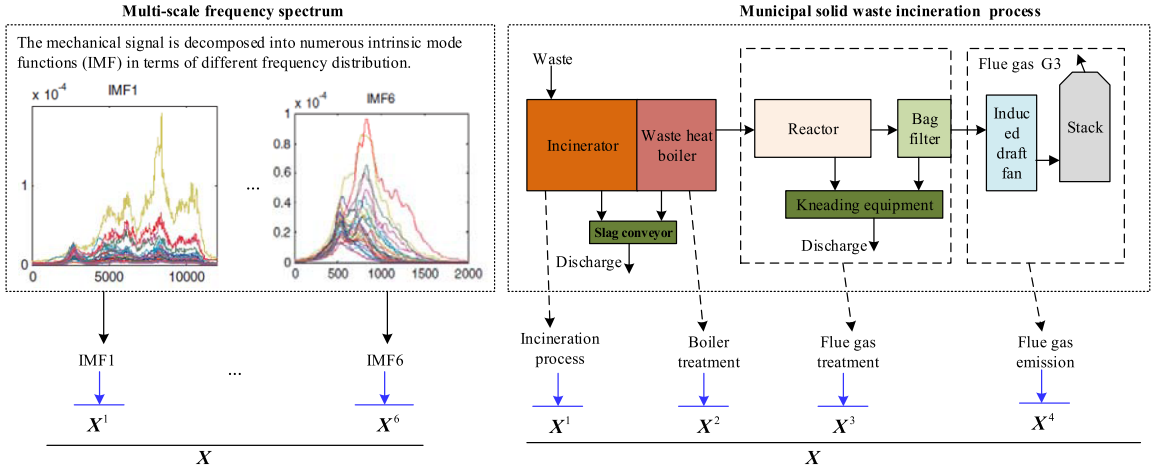


**FIGURE 1.** Example of different feature subgroups corresponding to different physical meanings.

derived based on feature selection ratio, contribution ratio and mutual information (MI) value. It can contribute to the requirement on the ratio of the number training samples to the reduced features as well as the modeling stability and the key process parameter relativity. Third, adaptively determining the hyperparameters as a joint optimization strategy is handled for the modeling data's characteristic, so does selecting the ensemble submodel's combination approach and ensemble size in terms of prediction performance. Thus, optimal ensemble submodels based on latent feature with complementary characteristics can be fused selectively.

The rest of this paper is organized as follows. Section II analyzes the small-sample high-dimensional process data modeling problem. Section III describes the proposed method in detail. Section IV gives the experimental results and discussions. Section V concludes the study and discusses recommendations for future work.

## II. SMALL-SAMPLE HIGH-DIMENSIONAL PROCESS DATA MODELING PROBLEM ANALYSIS

A sufficient number of modeling samples with complete coverage running conditions is important for building an effective soft measurement model. However, the definition of such modeling sample data have great relativity and subjectivity [7]. Several indices are proposed to determine the minimum number of training samples needed to obtain the necessary predictive performance [35], [36]. For the classification problem, the relationship among classification errors, number of training samples, input feature dimension, and classification algorithm complexity is studied [37]. In the field of pattern recognition, the expected ratio of the number of training samples to the input features $\alpha_{\text{ratio}}^{\text{ori}}$ can be calculated and set as

$$\alpha_{\text{ratio}}^{\text{ori}} = N_{\text{sample}} / P_{\text{feature}}, \quad \alpha_{\text{ratio}}^{\text{ori}} = 2, 5, 10 \tag{1}$$

where $N_{\text{sample}}$ and $P_{\text{feature}}$ represent the number of training samples and the input feature, respectively.

Given the high dimensional spectral data and the complex industrial process, the input features may number hundreds and thousands. Moreover, input features in different regions may have various physical meanings [5]. In process industries such as MSWI, these input features correspond to different process stages [38], [39]. They are shown in Fig. 1.

Fig. 1 shows that different subgroups have various meanings. Interesting information may be dropped only to reduce the dimension for the global input features. By contrast, the physical meanings of different subgroups are lost by using feature extraction method to the whole input features. The contributions of different parts are also unclear. Therefore, alternately, the latent features that represent different local and global information should be extracted. The $i$th subgroup can be denoted as $\boldsymbol{X}^i$, which results in the following feature extraction process:

$$\boldsymbol{X}^i \xrightarrow{\text{FeatureExtraction}} \boldsymbol{Z}^i \tag{2}$$

where $\boldsymbol{Z}^i$ indicates the extracted latent feature, and it's dimension is $P_{\text{feature-redu}}^i$. Thus, the new expected ratio of the number of training samples to the reduce features is,

$$\alpha_{\text{ratio}}^{\text{redu}} = N_{\text{sample}} / P_{\text{feature-redu}}^i \quad \alpha_{\text{ratio}}^{\text{redu}} = 2, 5, 10 \tag{3}$$

Note that (3) is more suitable for realization than (1).

Assume that $(I-1)$ subgroups are obtained. The global information is represented by taking the whole input features as the $I$th special subgroup. Thus, the total extracted feature subset can be denoted as $\{\boldsymbol{Z}^i\}_{i=1}^I$. The new latent feature set $\{\boldsymbol{Z}_{\text{sub}}^i\}_{i=1}^I$ is obtained by further selecting the latent feature in terms of improvement of predictive relevance. Then, we can build submodel $f^i(\cdot)$ based on the $i$th feature subset, whose predictive output is denoted as

$$\hat{\boldsymbol{y}}^i = f^i(\boldsymbol{Z}_{\text{sub}}^i, M_{\text{para}}^i) \tag{4}$$

where $M^i_{\text{para}}$ represents the hyperparameter. To effectively combine these submodels, the $I$ should comply with in (5),

$$\frac{N_{\text{sample}}}{I} \geq \alpha^{\text{ori}}_{\text{ratio}} \quad \alpha^{\text{ori}}_{\text{ratio}} = 2, 5, 10 \tag{5}$$

Thus, the output of the ensemble model can be denoted as

$$\hat{\boldsymbol{y}}_{\text{EN}} = f_{\text{EN}}(\{\hat{\boldsymbol{y}}^i\}_{i=1}^{I}) = f_{\text{EN}}(\{f^i(\boldsymbol{Z}^i_{\text{sub}}, M^i_{\text{para}})\}_{i=1}^{I}) \tag{6}$$

However, the above method cannot selectively fuse the subgroups with complementary characteristics. Thus, SEN method would be a good choice, whose output is denoted as

$$\hat{\boldsymbol{y}}_{\text{SEN}} = f_{\text{SEN}}(\{\hat{\boldsymbol{y}}^{i_{\text{sel}}}\}_{i_{\text{sel}}=1}^{I_{\text{sel}}}) = f_{\text{SEN}}(\{f^{i_{\text{sel}}}(\boldsymbol{Z}^{i_{\text{sel}}}_{\text{sub}}, M^{i_{\text{sel}}}_{\text{para}})\}_{i_{\text{sel}}=1}^{I_{\text{sel}}}) \tag{7}$$

Therefore, to model a small sample high-dimensional data, the following problems should be addressed: (1) how to effectively partition subgroups; (2) how to instruct the latent feature; (3) how to select the candidate submodels' hyperparameters; and (4) how to select and combine ensemble submodels.

## III. PROPOSED METHODOLOGY

### A. FEATURE GROUPING MODULE

The high-dimensional input data $\boldsymbol{X}$ include $N$ samples (rows) and $M$ input features (columns). On the basis of domain knowledge and the flowchart of the industrial process, or clustering algorithms, the original input features are divided into (I-1) subgroups, which represent different local information. Let $\boldsymbol{X}^i$ be the modeled data from the $i$th subgroup, then all features of (I-1) subgroups can be defined by

$$\{\boldsymbol{X}^i\}_{i=1}^{I-1} = [\boldsymbol{X}^1, \cdots, \boldsymbol{X}^i, \cdots, \boldsymbol{X}^{I-1}] = f_{\text{group}}(\boldsymbol{X}, Know) \tag{8}$$

where the argument $Know$ indicates a prior knowledge.

To represent global information, all input features are considered the $I$th subgroup in the broad sense. Thus, the following relationship exists:

$$M = M^1 + \cdots + M^i + \cdots + M^{I-1} = \sum\nolimits_{i=1}^{I-1} M^i \tag{9}$$

where $M^i$ indicates the number of input features contained in the $i$th subgroup. The output data $\boldsymbol{y} = \{y_n\}_{n=1}^{N}$ consists of $N$ samples (rows), which are usually obtained from offline assay data. Therefore, $N \ll M$.

### B. LATENT FEATURE EXTRACTION AND MULTI-LAYERED FEATURE SELECTION MODULE

For the $i$th subgroup, the feature extraction algorithm is used to extract latent features. After being normalized, the input data $\boldsymbol{X}^i$ is decomposed as follows

$$\mathbf{X}^i = \sum_{m^i_{\text{FeAll}}}^{M^i_{\text{FeAll}}} \mathbf{t}^i_{m^i_{\text{FeAll}}} (\mathbf{p}^i_{m^i_{\text{FeAll}}})^{\text{T}} \tag{10}$$

where $\mathbf{t}^i_{m^i_{\text{FeAll}}}$ and $\mathbf{p}^i_{m^i_{\text{FeAll}}}$ denote the score and loading vectors, respectively, and $M^i_{\text{FeAll}} = rank(\mathbf{X}^i)$ is the number of extracted latent features.

Therefore, all latent features extracted can be dentoed as

$$\mathbf{T}^i = [\mathbf{t}^i_{1_{\text{FeAll}}}, \cdots, \mathbf{t}^i_{m^i_{\text{FeAll}}}, \cdots, \mathbf{t}^i_{M^i_{\text{FeAll}}}] \tag{11}$$

where $\mathbf{T}^i \in R^{N \times M^i_{\text{FeAll}}}$ represents the score matrix, which is the orthogonal projection of $\boldsymbol{X}^i$ in the direction of the load matrix $\mathbf{P}^i \in R^{M \times M^i_{\text{FeAll}}}$. The latter can be represented as

$$\mathbf{P}^i = [\mathbf{p}^i_{1_{\text{FeAll}}}, \cdots, \mathbf{p}^i_{m^i_{\text{FeAll}}}, \cdots, \mathbf{p}^i_{M^i_{\text{FeAll}}}] \tag{12}$$

Thus, the latent features extracted can be expressed as

$$\begin{aligned}
&\mathbf{Z}^i_{\text{FeAll}} \\
&= \mathbf{T}^i = \mathbf{X}^i\mathbf{P}^i \\
&= [\mathbf{z}^i_{1_{\text{FeAll}}}, \cdots, \mathbf{z}^i_{m^i_{\text{FeAll}}}, \cdots, \mathbf{z}^i_{M^i_{\text{FeAll}}} \\
&= [\{(z^i_{1_{\text{FeAll}}})_n\}_{n=1}^{N}, \cdots, \{(z^i_{m^i_{\text{FeAll}}})_n\}_{n=1}^{N}, \cdots, \{(z^i_{M^i_{\text{FeAll}}})_n\}_{n=1}^{N} \\
&= \{(\mathbf{z}^i_{\text{FeAll}})_n\}_{n=1}^{N}
\end{aligned} \tag{13}$$

In order to satisfy (3), the first layer feature selection based on feature reduction ratio is made. The expected latent feature number is calcuated by,

$$M^i_{\text{FeAllSel\_1st}} = N/\alpha^{\text{redu}}_{\text{ratio}} \tag{14}$$

Further, selected latent features in the first layer can be denoted by

$$\begin{aligned}
\mathbf{Z}^i_{\text{FeSelst}} &= \mathbf{Z}^i_{\text{FeAll}}(:, 1 : M^i_{\text{FeAllSel_1st}}) = \{(\mathbf{z}^i_{\text{FeSelst}})_n\}_{n=1}^{N} \\
&= [\{(z^i_{1_{\text{FeSelst}}})_n\}_{n=1}^{N}, \cdots, \{(z^i_{m^i_{\text{FeSelst}}})_n\}_{n=1}^{N}, \\
&\quad \cdots, \{(z^i_{M^i_{\text{FeSelst}}})_n\}_{n=1}^{N}
\end{aligned} \tag{15}$$

However, the latent feature with low contribution ratio can inevitably cause the instability problem in terms of the final predicition performance. In the seond layer, the feature selection strategy based on contribution ratio is determined as follows. Here, the eigenvalue corresponding to the $m^i_{\text{FeAll}}$th loading vector is labeled as $\lambda^i_{m^i_{\text{FeAll}}}$, and then the contribution rate $\theta^i_{m^i_{\text{FeAll}}}$ of the $m^i_{\text{FeAll}}$th latent feature can be calculated by

$$\theta^i_{m^i_{\text{FeAll}}} = \frac{\lambda^i_{m^i_{\text{FeAll}}}}{\sum_{m^i_{\text{FeAll}}=1}^{M^i_{\text{FeAll}}} \lambda^i_{m^i_{\text{FeAll}}}} \times 100 \tag{16}$$

All the contribution rates can be denoted as $\{\theta^i_{m^i_{\text{FeAll}}}\}_{m^i_{\text{FeAll}}=1}^{M^i_{\text{FeAll}}}$. Thus, the selected latent features in the fisrt layer is the former $M^i_{\text{FeAllSel\_1st}}$ elememt, which is depicted as $\{\theta^i_{m^i_{\text{FeAll}}}\}_{m^i_{\text{FeAll\_1st}}=1}^{M^i_{\text{FeAll\_1st}}}$. Through defining the threshold $\theta_{\text{Contri}}$ selected by expert experience, the criterion to select latent features from $\mathbf{Z}^i_{\text{FeSelst}}$ in the second layer is given by

$$\xi^i_{m^i_{\text{FeAll\_1st}}} = \begin{cases} 1, & \text{if } \theta^i_{m^i_{\text{FeAll\_1st}}} \geq \theta_{\text{Contri}} \\ 0, & \text{else } \theta^i_{m^i_{\text{FeAll\_1st}}} < \theta_{\text{Contri}} \end{cases} \tag{17}$$

As a matter of fact, $\xi^i_{m^i_{\text{FeAll\_1st}}}$ is an indicator function, which means whether the $m^i_{\text{FeAll_1st}}$th latent feature is selected or

not. Here, value 1 indicates that this latent feature is selected, otherwise, it is not selected.

Therefore, selected latent features in the second layer for the $i$th subgroup can be demonstrated as

$$\begin{aligned}\mathbf{Z}^i_{\text{FeSe2nd}} &= [\mathbf{z}^i_{1_{\text{FeSe2nd}}}, \cdots, \mathbf{z}^i_{m^i_{\text{FeSe2nd}}}, \cdots, \mathbf{z}^i_{M^i_{\text{FeSe2nd}}}] \\ &= [\{(z^i_{1_{\text{FeSe2nd}}})_n\}_{n=1}^N, \cdots, \{(z^i_{m^i_{\text{FeSe2nd}}})_n\}_{n=1}^N, \\ &\quad \cdots, \{(z^i_{M^i_{\text{FeSe2nd}}})_n\}_{n=1}^N] \\ &= \{(\mathbf{z}^i_{\text{FeSe2nd}})_n\}_{n=1}^N \end{aligned} \tag{18}$$

These selected latent features in the second layer are independent in terms of the high contribution rate without considering the correlation between these features and the process parameter. Thus, their MI values are calculated by,

$$\xi^{m^i_{\text{FeSe2nd}}}_{\text{MI}} = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{z}^i_{m^i_{\text{FeSe2nd}}}) \tag{19}$$

where $H(\mathbf{y}|\mathbf{z}^i_{m^i_{\text{FeSe2nd}}})$ is the conditional entropy, and $H(\mathbf{y})$ indicates the information entropy.

Specially, the threshold value for latent feature selection in the third layer is ranged from $\theta^i_{\text{MIUP}}$ and $\theta^i_{\text{MIDN}}$ for each subgroup, which is adaptively calculated by

$$\theta^i_{\text{MI}} = \frac{n_{\text{MI}} \cdot (\theta^i_{\text{MIUP}} - \theta^i_{\text{MIDN}})}{N^{\text{Step}}_{\text{MI}}} + \theta^i_{\text{MIDN}} \tag{20}$$

$$\theta^i_{\text{MIUP}} = \max(\xi^{m^i_{\text{FeSe2nd}}}_{\text{MI}}) \tag{21}$$

$$\theta^i_{\text{MIDN}} = \min(\xi^{m^i_{\text{FeSe2nd}}}_{\text{MI}}) \tag{22}$$

where $N^{\text{Step}}_{\text{MI}}$ represents the number of candidate MI threshold steps, and $n_{\text{MI}}$ is the one of selected MI threshold times based on priro knowledge for all subgroups.

The following criterion reveals the latent feature selection strategy in the third layer

$$\beta^i_{m^i_{\text{FeSe2nd}}} = \begin{cases} 1, & \text{if } \xi^{m^i_{\text{FeSe2nd}}}_{\text{MI}} \geq \theta^i_{\text{MI}} \\ 0, & \text{else } \xi^{m^i_{\text{FeSe2nd}}}_{\text{MI}} < \theta^i_{\text{MI}} \end{cases} \tag{23}$$

Note that $\beta^i_{m^i_{\text{FeSe2nd}}}$ is also an indicator function, which means whether the $m^i_{\text{FeSe2nd}}$th latent feature in the second layer is selected or not. Explicitly, value 1 indicates that this latent feature is selected again. Thus, the latent feature of the $i$th subgroup in the third layer can be rewritten as

$$\begin{aligned}\mathbf{Z}^i_{\text{FeSe3rd}} &= [\mathbf{z}^i_{1_{\text{FeSe3rd}}}, \cdots, \mathbf{z}^i_{m^i_{\text{FeSe3rd}}}, \cdots, \mathbf{z}^i_{M^i_{\text{FeSe3rd}}}] \\ &= [\{(z^i_{1_{\text{FeSe3rd}}})_n\}_{n=1}^N, \cdots, \{(z^i_{m^i_{\text{FeSe3rd}}})_n\}_{n=1}^N, \\ &\quad \cdots, \{(z^i_{M^i_{\text{FeSe3rd}}})_n\}_{n=1}^N] \\ &= \{(\mathbf{z}^i_{\text{FeSe3rd}})_n\}_{n=1}^N \end{aligned} \tag{24}$$

Therefore, all the selected features can be accordingly determined by

$$\begin{aligned}\mathbf{Z}_{\text{FeSe3rd}} &= [\mathbf{Z}^1_{\text{FeSe3rd}}, \cdots, \mathbf{Z}^i_{\text{FeSe3rd}}, \cdots, \mathbf{Z}^I_{\text{FeSe3rd}}] \\ &= \{\mathbf{Z}^i_{\text{FeSe3rd}}\}_{i=1}^I \end{aligned} \tag{25}$$

### C. SEN MODELING MODULE BASED ON HYPERPARAMETER ADAPTIV SELECTION

On the basis of the minimized root-mean-square error (RMSE) criterion, the SEN modeling process referring to the adaptive selection of sub-model hyperparameter, ensemble sub-models, and their combination strategy can be formulated as the following minimality problem:

$$\begin{aligned} Min\ E^{\text{SEN}}_{\text{RMSE}} &= \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n - \hat{y}_n\right)^2} \\ &= \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n - f_{\text{SEN}}\left(\{\hat{y}_n^{i_{\text{sel}}}\}_{i_{\text{sel}}=1}^{I_{\text{sel}}}\right)\right)^2} \\ &= \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n - f_{\text{SEN}}\left(\left\{f^{i_{\text{sel}}}(\mathbf{Z}^{i_{\text{sel}}}_{\text{FeSe3rd}}, \{K^{i_{\text{sel}}}_{\text{er}}, R^{i_{\text{sel}}}_{\text{eg}}\}\right\}_{i_{\text{sel}}=1}^{I_{\text{sel}}}\right)\right)^2} \\ s.t.\ &\begin{cases} 2 \leq I_{\text{sel}} \leq I-1 \\ \{K^{i_{\text{sel}}}_{\text{er}}, R^{i_{\text{sel}}}_{\text{eg}}\} \in M_{\text{para}} \\ \{f^{i_{\text{sel}}}(\cdot)\}_{i_{\text{sel}}=1}^{I_{\text{sel}}} \in \{f^i(\cdot)\}_{i=1}^I \\ f_{\text{SEN}}(\cdot) \in \{\text{AWF, PLS, Entropy}, \cdots\} \end{cases} \end{aligned} \tag{26}$$

where $f^i(\cdot)$ is the candidate submodel constructed by the $i$th group latent feature $\mathbf{Z}^i_{\text{FeSe3rd}}$; $f^{i_{\text{sel}}}(\cdot)$ denotes the ensemble submodel; $I_{\text{sel}}$ indicates the ensemble size; $\{K^{i_{\text{sel}}}_{\text{er}}, R^{i_{\text{sel}}}_{\text{eg}}\}$ is the hyperparameter for the $i_{\text{sel}}$th ensemble sub-model; $M_{\text{para}}$ signifies the candidate hyperparameter matrix; $f_{\text{SEN}}(\cdot)$ is the combination method or model, which involves adaptive weighting fusion (AWF), prediction error entropy weighting (Entropy), partial least squares (PLS), or other linear/nonlinear mapping algorithms.

Taking the $i$th subgroup as an example, we illustrate the process of constructing candidate submodel according to adaptive selection strategies for hyperparameter pair $\{K^i_{\text{er}}, R^i_{\text{eg}}\}$. First, with the use of $\varphi(\cdot)$, the $\{(\mathbf{z}^i_{\text{FeSe3rd}})_n\}_{n=1}^N$ is transformed into a high-dimensional feature space to solve the following optimization problem:

$$\begin{cases} \min\limits_{\mathbf{w}^i, b^i} O_{\text{LS-SVM}} = \frac{1}{2}(\mathbf{w}^i)^{\text{T}}\mathbf{w}^i + \frac{1}{2}R^i_{\text{eg}}\sum\limits_{n=1}^{N}(\zeta^i_n)^2 \\ s.t:\ \hat{y}^i_n = (\mathbf{w}^i)^{\text{T}}\varphi((\mathbf{z}^i_{\text{FeSe3rd}})_n) + b^i + \zeta^i_n \end{cases} \tag{27}$$

where $\mathbf{w}^i$ is the weight coefficient, $b^i$ is the bias, and $\zeta^i_n$ is the prediction error of the $n$th sample. The following equation can be obtained via the Lagrangian method:

$$\begin{aligned} &\text{L}^i(\mathbf{w}^i, b^i, \boldsymbol{\zeta}^i, \boldsymbol{\beta}^i) \\ &= \frac{1}{2}(\mathbf{w}^i)^T\mathbf{w}^i + \frac{1}{2}\sum_{n=1}^{N}(\zeta^i_n)^2 \\ &\quad - \sum_{n=1}^{N}\beta^i_n[(\mathbf{w}^i)^{\text{T}}\varphi((\mathbf{z}^i_{\text{FeSe3rd}})_n) + b^i + \zeta^i_n - \hat{y}^i_n] \end{aligned} \tag{28}$$

where $\boldsymbol{\beta}^i = [\beta^i_1, \cdots, \beta^i_n, \cdots, \beta^i_N]$ represents the Lagrangian operator vector, and $\boldsymbol{\zeta}^i = [\zeta^i_1, \cdots, \zeta^i_n, \cdots, \zeta^i_N]$ represents

the prediction error vector. The above equation is solved with

$$\frac{\partial L^i}{\partial \mathbf{w}^i}=0,\quad \frac{\partial L^i}{\partial b^i}=0,\quad \frac{\partial L^i}{\partial \boldsymbol{\xi}^i}=0,\quad \frac{\partial L^i}{\partial \boldsymbol{\beta}^i}=0 \tag{29}$$

The kernel function used is expressed as follows:

$$\Omega^i_{\mathrm{ker}}(\mathbf{z}^i_{\mathrm{FeSe3rd}},(\mathbf{z}^i_{\mathrm{FeSe3rd}})_n)=<\varphi(\mathbf{z}^i_{\mathrm{FeSe3rd}})\cdot\varphi((\mathbf{z}^i_{\mathrm{FeSe3rd}})_n)> \tag{30}$$

The LS–SVM problem is converted to solve the following linear equation system:

$$\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & \Omega^i_{\mathrm{ker}}(\cdot)_{11}+\dfrac{1}{R^i_{\mathrm{eg}}} & \cdots & \Omega^i_{\mathrm{ker}}(\cdot)_{1N} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \Omega^i_{\mathrm{ker}}(\cdot)_{N1} & \cdots & \Omega^i_{\mathrm{ker}}(\cdot)_{NN}+\dfrac{1}{R^i_{\mathrm{eg}}} \end{bmatrix} \cdot \begin{bmatrix} b^i \\ \beta^i_1 \\ \vdots \\ \beta^i_N \end{bmatrix} = \begin{bmatrix} 1 \\ y^i_1 \\ \vdots \\ y^i_N \end{bmatrix} \tag{31}$$

By solving the above formula, we obtain $\boldsymbol{\beta}^i$ and $b^i$.

A candidate submodel constructed based on LS–SVM for the $i$th subgroup can be expressed as

$$\hat{\mathbf{y}}^i=\sum_{n=1}^{N}\beta^i_n\cdot\Omega^i_{\mathrm{ker}}\left(z^i_{\mathrm{FeSe3rd}},(z^i_{\mathrm{FeSe3rd}})_n\right)+b^i \tag{32}$$

The adaptive hyperparameter selection mechanism of the above candidate submodel is implemented by the following $N_{\mathrm{grid}}$-times grid search method. In the first step, the grid search strategy adaptively selects the initial hyperparameter pair $\{(K^{\mathrm{initial}}_{\mathrm{er}})^i,(R^{\mathrm{initial}}_{\mathrm{eg}})^i\}$ in the candidate hyperparameter matrix, which can be represented as

$$M_{\mathrm{para}}=\begin{bmatrix} [K^1_{\mathrm{er}},R^1_{\mathrm{eg}}] & \cdots & [K^1_{\mathrm{er}},R^r_{\mathrm{eg}}] & \cdots & [K^1_{\mathrm{er}},R^R_{\mathrm{eg}}] \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ [K^k_{\mathrm{er}},R^1_{\mathrm{eg}}] & \cdots & [K^k_{\mathrm{er}},R^r_{\mathrm{eg}}] & \cdots & [K^k_{\mathrm{er}},R^R_{\mathrm{eg}}] \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ [K^K_{\mathrm{er}},R^1_{\mathrm{eg}}] & \cdots & [K^K_{\mathrm{er}},R^r_{\mathrm{eg}}] & \cdots & [K^K_{\mathrm{er}},R^R_{\mathrm{eg}}] \end{bmatrix}_{K\times R} \tag{33}$$

where $K$ represents the number of candidate kernel parameter; $R$ represents the number of candidate regularization parameters; $[K^k_{\mathrm{er}},R^r_{\mathrm{eg}}]$ represents a hyperparameter pair consisting of the $k$th kernel parameter and the $r$th regularization parameter, which is the $j$th element, i.e., $M^j_{\mathrm{para}}=[K^k_{\mathrm{er}},R^r_{\mathrm{eg}}]$; and, $J=K\times R$ represents the number of candidate hyperparameter pairs. Therefore, the hyperparameter pair selected for the $i$th candidate submodel is one of the elements of $M_{\mathrm{para}}$.

For the last result $\{(K^{\mathrm{G}_{n-1}}_{\mathrm{er}})^i,(R^{\mathrm{G}_{n-1}}_{\mathrm{eg}})^i\}$ in the remaining $(N_{\mathrm{grid}}-1)$-time grid search, the following equation obtains a new candidate kernel parameter vector $(\boldsymbol{K}^{\mathrm{V}_n}_{\mathrm{er}})^i$ and a new regularization parameter vector $(\boldsymbol{R}^{\mathrm{V}_n}_{\mathrm{eg}})^i$:

$$(\boldsymbol{K}^{\mathrm{V}_n}_{\mathrm{er}})^i=\frac{(K^{\mathrm{G}_{n-1}}_{\mathrm{er}})^i}{k^{\mathrm{down}}_{\mathrm{supara}}}:\left(k^{\mathrm{up}}_{\mathrm{supara}}\cdot(K^{\mathrm{G}_{n-1}}_{\mathrm{er}})^i-\frac{(K^{\mathrm{G}_{n-1}}_{\mathrm{er}})^i}{k^{\mathrm{down}}_{\mathrm{supara}}}\Big/N_{\mathrm{ker}}\right):\left((K^{\mathrm{G}_{n-1}}_{\mathrm{er}})^i\cdot k^{\mathrm{up}}_{\mathrm{supara}}\right) \tag{34}$$

$$(\boldsymbol{R}^{\mathrm{V}_n}_{\mathrm{er}})^i=\frac{(R^{\mathrm{G}_{n-1}}_{\mathrm{eg}})^i}{k^{\mathrm{down}}_{\mathrm{supara}}}:\left(k^{\mathrm{up}}_{\mathrm{supara}}\times(R^{\mathrm{G}_{n-1}}_{\mathrm{eg}})^i-\frac{(R^{\mathrm{G}_{n-1}}_{\mathrm{eg}})^i}{k^{\mathrm{down}}_{\mathrm{supara}}}\right)\Big/N_{\mathrm{ker}}:\left(k^{\mathrm{up}}_{\mathrm{supara}}\times(R^{\mathrm{G}_{n-1}}_{\mathrm{eg}})^i\right) \tag{35}$$

where $N_{\mathrm{ker}}$ and $N_{\mathrm{reg}}$ represent the number of new candidate kernel parameters and the number of regularization parameters, respectively, and $k^{\mathrm{down}}_{\mathrm{supara}}$ and $k^{\mathrm{up}}_{\mathrm{supara}}$ are the hyperparameters' shrink and expansion factors, respectively. Then, the grid search strategy is employed again to calculate the hyperparameter pairs $\{K^i_{\mathrm{er}},R^i_{\mathrm{eg}}\}$ of the $i$th candidate submodel.

Due to the above process performed on the selected latent features of different subgroups, the prediction output set of the candidate submodel can be defined by

$$\begin{aligned}\hat{Y}&=[\hat{\mathbf{y}}^1,\cdots,\hat{\mathbf{y}}^i,\cdots,\hat{\mathbf{y}}^I]\\&=\{\hat{\mathbf{y}}^i\}^I_{i=1}=\{f^i(K^i_{\mathrm{er}},R^i_{\mathrm{eg}},\mathbf{Z}^i_{\mathrm{FeSe3rd}})\}^I_{i=1}\end{aligned} \tag{36}$$

Upon giving the candidate submodel and the combination algorithm, the selection of the ensemble submodel is similar to the optimal feature subset selection problem. In this study, the number of the candidate submodels is limited. Thus, the commonly adopted strategy is to use the multiple coupling operation branch-and-bound optimization algorithm and submodel combination algorithm to construct multiple SEN models with an ensemble size of 2 to $(I-1)$. The preferred SEN model is obtained based on prediction performance ranking. The pseudocode of this branch-and-bound-based algorithm is shown in [40].

Assume that the ensemble size of the final soft measurement model is $I_{\mathrm{sel}}$. The final predicted output $\hat{y}$ is calculated by using

$$\begin{aligned}\hat{y}&=f_{\mathrm{SEN}}\left(\{\hat{y}^{i_{\mathrm{sel}}}\}^{I_{\mathrm{sel}}}_{i_{\mathrm{sel}}=1}\right)\\&=f_{\mathrm{SEN}}\left(\left\{f^{i_{\mathrm{sel}}}(\mathbf{Z}^{i_{\mathrm{sel}}}_{\mathrm{FeSe3rd}},\{K^{i_{\mathrm{sel}}}_{\mathrm{er}},R^{i_{\mathrm{sel}}}_{\mathrm{eg}}\}\right\}^{I_{\mathrm{sel}}}_{i_{\mathrm{sel}}=1}\right)\end{aligned} \tag{37}$$

where $f_{\mathrm{SEN}}(\cdot)$ is a combination method or a mapping model.

### D. FLOW CHART OF THE PROPOSED METHOD

The flow chart is shown in Fig. 2.

Fig. 2 shows that the proposed algorithm should meet the (3) and (1) in the canddiate submodel construction phase and ensemble submodel combination phase, respectively. The elaborate analysis and proof are shown in Marks I and II, respectively.

#### 1) MARK I

The proposed method can come to the requirement of (3) in terms of feature reduction based on feature extraction

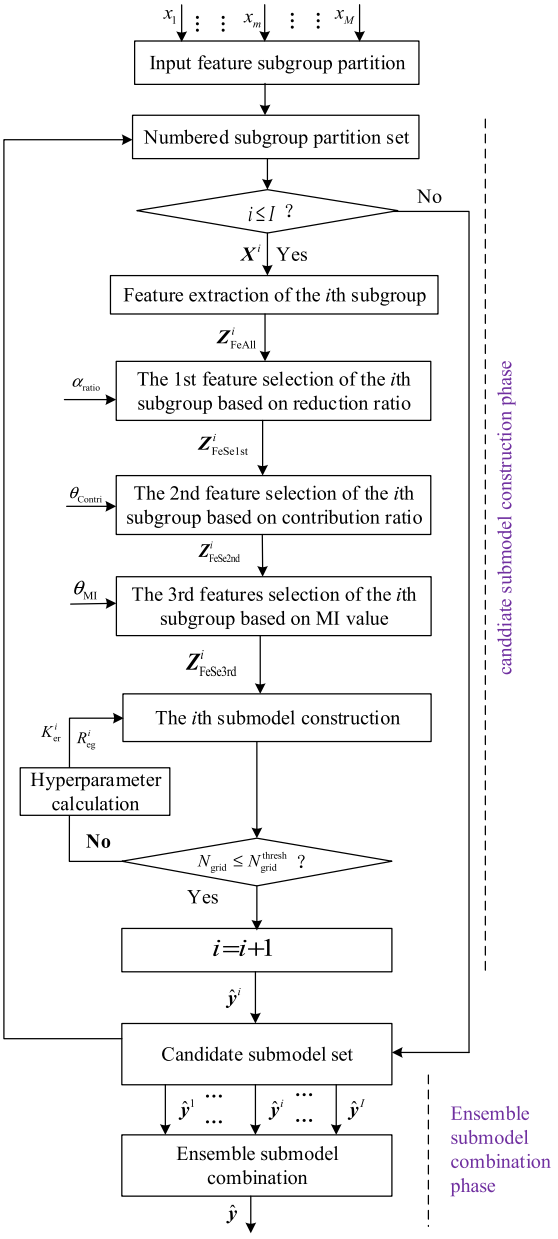


**FIGURE 2.** Flow chart of the proposed method.

and multi-layered feature selection. The detailed analysis is shown as follows.

The input feature is partitioned into $(I-1)$ subgroups. The total input feature is denoted as the $I$th subgoup. Normally, the symbol $I$ satifies (5), which is much samller than the number of input features, i.e., $I \ll M$.

For example, the number of the extracted feature $P^I_{\text{feature-extra}}$ can be denoted as,

$$P^I_{\text{Feature-Extr}} = f_{\text{extr}}(\boldsymbol{X}^I), \quad P^I_{\text{Feature-Extr}} \le M \tag{38}$$

With the pre-set expected feature reduction ratio $\alpha^{\text{redu}}_{\text{ratio}}$, the number of feature selection strategies in the first layer is described as,

$$P^I_{\text{Feature-1stSel}} = P^I_{\text{Feature-Extr}}/\alpha^{\text{redu}}_{\text{ratio}}, \quad \alpha^{\text{redu}}_{\text{ratio}} \ge 2 \tag{39}$$

Similarly, with the pre-set contribution ratio $\theta_{\text{Contri}}$, the number of ones in the second layer is illustrated as,

$$P^I_{\text{Feature-2ndSel}} = f_{\text{2ndSel}}(P^I_{\text{Feature-1stSel}}, \theta_{\text{Contri}}) \tag{40}$$

where $P^I_{\text{Feature-2ndSel}} \le P^I_{\text{Feature-1stSel}}$.

Accordingly, with the pre-set MI threshold-step time $n_{\text{MI}}$, the number of ones in the third layer is denoted as,

$$P^I_{\text{Feature-3rdSel}} = f_{\text{3rdSel}}(P^I_{\text{Feature-2ndSel}}, n_{\text{MI}}), \quad 1 \le n_{\text{MI}} \le N^{\text{Step}}_{\text{MI}} \tag{41}$$

where $P^I_{\text{Feature-3rdSel}} \le P^I_{\text{Feature-2ndSel}}$.

Therefore, the actual feature reduction ratio contributes to the following result,

$$\begin{aligned}
\alpha^{\text{I*}}_{\text{ratio}} &= \frac{N_{\text{sample}}}{P^I_{\text{Feature-3rdSel}}} = \frac{N_{\text{sample}}}{f_{\text{3rdSel}}(P^I_{\text{Feature-2ndSel}}, n_{\text{MI}})} \\
&\ge \frac{N_{\text{sample}}}{P^I_{\text{Feature-2nd}}} = \frac{N_{\text{sample}}}{f_{\text{2ndSel}}(P^I_{\text{Feature-1stSel}}, \theta_{\text{Contri}})} \\
&\ge \frac{N_{\text{sample}}}{P^I_{\text{Feature-1stSel}}} = \frac{N_{\text{sample}}}{P^I_{\text{Feature-Extr}}/\alpha^{\text{redu}}_{\text{ratio}}} \\
&= \frac{N_{\text{sample}}}{P^I_{\text{Feature-Extr}}}\alpha^{\text{redu}}_{\text{ratio}} \ge \frac{N_{\text{sample}}}{M}\alpha^{\text{redu}}_{\text{ratio}}
\end{aligned} \tag{42}$$

For small sample modeling data, there exits $M \gg N_{\text{sample}}$. So $\alpha^{\text{I*}}_{\text{ratio}} > \alpha^{\text{redu}}_{\text{ratio}}$ is reasonable.

As other subgroups have less input feature number than the $I$th subgoup's, the above relationship is also satified.

Thereby, the requirement of the ratio of the number of training samples to the input features in (3) can be satisfied.

### 2) MARK II

The proposed method can ensure the requirement of (1) through combining ensembel submodels with the combination method or mapping model and obtain the final SEN model. The analysis is shown in detail as follows.

The ratio of the number of training samples to the input features (i.e., ensembel submodel's predciton output) can be denoted as

$$\alpha^{\text{Combin}}_{\text{ratio}} = \frac{N}{I_{\text{sel}}} \quad 2 \le I_{\text{sel}} \le I \tag{43}$$

According to (5), $\frac{N_{\text{sample}}}{I} \ge \alpha^{\text{ori}}_{\text{ratio}}$ can be changed into $I \le \frac{N_{\text{sample}}}{\alpha^{\text{ori}}_{\text{ratio}}}$, which further leads to the following result:

$$\begin{aligned}
\alpha^{\text{Combine}}_{\text{ratio}} &= \frac{N_{\text{sample}}}{I_{\text{sel}}} \\
&\ge \frac{N_{\text{sample}}}{I} \ge \frac{N_{\text{sample}}}{\frac{N_{\text{sample}}}{\alpha^{\text{ori}}_{\text{ratio}}}} = \alpha^{\text{ori}}_{\text{ratio}}
\end{aligned} \tag{44}$$

Therefore, the above Marks show that the proposed algorithnm satisfis the (3) and (1) in different phases of the

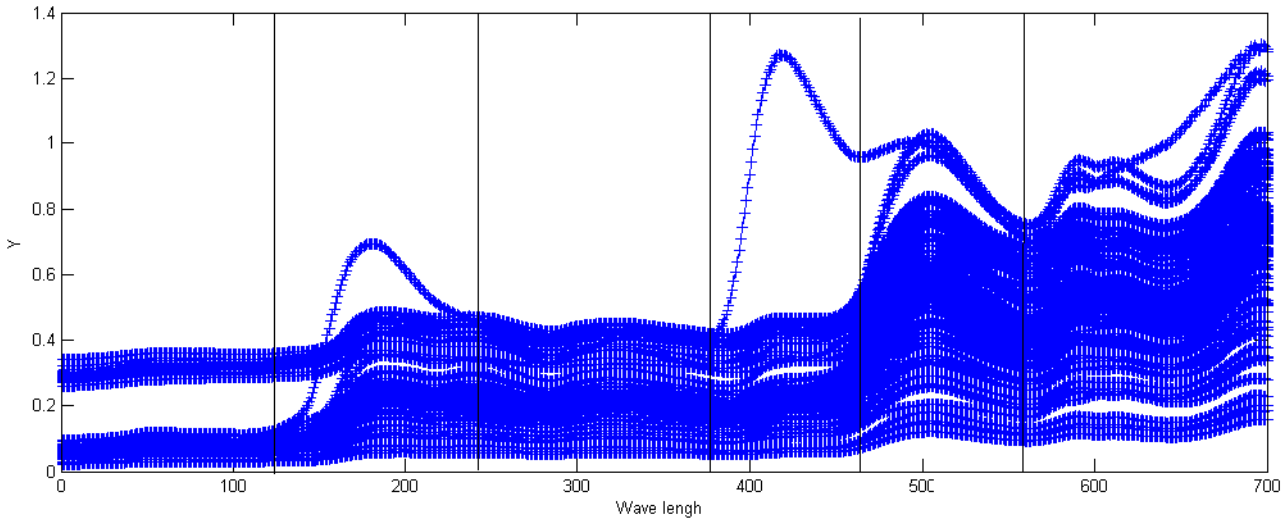


**FIGURE 3.** The NIR training data set and input feature subgroups.

proposed modeling method. And the proposed SEN modeling strategy can well solve modeling problems of the small-sample and high-dimensional process data in the view of subgroups' input feature reduction and ensemble submodels' combination.

## IV. EXPERIMENT RESEARCH

In this section, the LS–SVM algorithm with radius basis kernel function is used to build candidate submodel. The set of candidate regularization parameters and kernel parameters are preset as {0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 2000, 4000, 6000, 8000, 10000, 20000, 40000, 60000, 80000, 160000} and {0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 1600, 3200, 6400, 12800, 25600, 51200, 102400}, respectively. Three candidate ensemble submodel combination approaches, i.e., adaptive weighting fusion (AWF), entropy weighting based on prediction error (entropy), and partial least squares (PLS), are carried out.

### A. MODELING DATA DESCRIPTION

#### 1) HIGH-DIMENSIONAL NEAR INFRARED (NIR) SPECTRA DATA

They are used to estimate the saccharose level of orange juice. The NIR data come from http://www.ucl.ac.be, in which the data sizes for training set and testing set are 150 and 68, respectively. Thus, the number of input features, i.e., 700, is four times as many as that of the training samples. Additionally, different regions of the input features have different curve shapes.

#### 2) HIGH-DIMENSIONAL MECHANICAL VIBRATION FREQUENCY SPECTRA (MVF) DATA

They can estimate the mill load parameters (i.e., material to ball volume ration (MBVR), pulp density (PD), charge volume ratio (CVR)) inside of ball mill. Supposed that only CVR is modeled and the experiments are performed on a laboratory-scale ball mill. The mechanical vibration signal is measured by an accelerometer located in the surface of the mill shell. Based on four mill rotating periods data, MVF data are obtained according to empirical mode decomposition and FFT technologies [40], [29]. In this process, only IMF2-IMF8 are selected to construct CVR soft measurement model for simplification, and the number of training samples is 26. The number of input features ranges from 250 to 12000.

#### 3) HIGH-DIMENSIONAL DIOXIDE (DXN) EMISSION CONCENTRATION DATA

The DXN data come from an MSWI plant in China, which covers 39 samples of DXN emission concentration from 2012 to 2018 [41]. Deleting several variables with incomplete data, the total number of input features is 286. It shows that the dimension of the input features far exceeds the number of training samples. Thus, dimension reduction is necessary.

### B. MODELING RESULTS

#### 1) FEATURE GROUPING RESULTS

The NIR data are divided into six subgroups according to the profile of the training data, whose ranges are 1–120, 121–240, 241–380, 381–470, 471–560, and 561–700. The MVF data consist of seven IMFs which are different sub-signals of the original mechanical shell vibration signal with different frequency distributions. The DXN data are divided into six subgroups according to the flow chart of MSWI process, i.e., the incinerator, boiler, flue gas, steam, stack, and common. These subgroups contain differentiated local information, which can be considered different multisource information. The simplified curves of MVF data and the flow chart of MSWI process are shown in Fig. 1, while the curve of the NIR data is shown in Fig. 3.

**TABLE 1.** Number of the input feature for different subgroups.

| Subgroup | NIR | MVF | DXN |
|---|---|---|---|
| I | 120 | 12000 | 79 |
| II | 120 | 6000 | 14 |
| III | 140 | 3600 | 20 |
| IV | 90 | 2000 | 53 |
| V | 90 | 1000 | 6 |
| VI | 140 | 500 | 115 |
| VII | 700 | 250 | 287 |

**TABLE 2.** Cumulative contribution rate of the former six PCs.

| Subgroups | NIR | MVF | DXN |
|---|---|---|---|
| I | 99.9998% | 99.7300 | 84.0764 |
| II | 99.9987% | 99.6209 | 99.2771 |
| III | 99.9994% | 99.2664 | 92.8777 |
| IV | 99.9993% | 99.6408 | 96.3519 |
| V | 99.9991% | 99.8038 | 100 |
| VI | 99.9947% | 99.6577 | 86.4119 |
| VII | 99.9838% | 99.9796 | 82.9931 |

Therefore, according to all the original input features that characterize the global information, seven generalized subgroups are numbered from I to VII for three datasets. And the input feature number of different subgroups for three datasets are shown in Table 1.

#### 2) FEATURE EXTRACTION AND SELECTION RESULTS

The latent features of the seven subgroups for different datasets are extracted by PCA. Then the cumulative contribution rate of the first sixth PCs are demonstrated in Table 2.

Table 2 shows the results as follows. First, the cumulative contribution rates of the former six PCs in all subgroups of NIR and MVF datasets come to nearly 100%, which implies strong collinearity between input features. Second, the VII subgroup of DXN dataset represents that the global information is only extracted by 82.9931% in terms of contribution ratio, while other subgroups illustrate better performance than the VII one. Thus, latent features should be extracted from different input feature subgroups.

Let the feature reduction ratio, contribution ratio and mutual information threshold-step times for NIR, MVF and DXN datasets be (2, 0.001, 4), (2, 0.5, 7) and (6, 1, 10), respectively. With the proposed method, the selected feature numbers of three different layers are calculated and listed in Table 3.

Table 3 shows that the feature number is further reduced after the 1st-layer feature selection. Thus, the contribution ratio or mutual information values are different with each other for different latent features. However, the 2nd-layer feature selection cannot further reduce the number of input feature for DXN data. Thus, the industrial process data's characteristic is different from that of high dimensional spectra data.

**TABLE 3.** Selected feature number results of three layers.

| Subgroups | Feature selection | NIR data | MVF data | DXN data |
|---|---|---|---|---|
| I | 1st | 75 | 7 | 3 |
| I | 2nd | 4 | 3 | 3 |
| **I** | **3rd** | **1** | **2** | **1** |
| II | 1st | 75 | 7 | 3 |
| II | 2nd | 5 | 3 | 3 |
| **II** | **3rd** | **4** | **1** | **1** |
| III | 1st | 75 | 7 | 3 |
| III | 2nd | 4 | 4 | 3 |
| **III** | **3rd** | **3** | **2** | **2** |
| IV | 1st | 75 | 7 | 3 |
| IV | 2nd | 5 | 4 | 3 |
| **IV** | **3rd** | **3** | **1** | **1** |
| V | 1st | 75 | 7 | 3 |
| V | 2nd | 4 | 4 | 3 |
| **V** | **3rd** | **3** | **1** | **1** |
| VI | 1st | 75 | 7 | 3 |
| VI | 2nd | 4 | 3 | 3 |
| **VI** | **3rd** | **1** | **2** | **1** |
| VII | 1st | 75 | 7 | 3 |
| VII | 2nd | 9 | 4 | 3 |
| **VII** | **3rd** | **8** | **2** | **1** |

#### 3) SEN MODELING RESULTS

The grid search times are set 5, 3 and 2 for NIR, MVF and DXN datasets, respectively. Then, the results of different hyperparameters are shown in Table 4.

Table 4 shows the following results. First, different data need different grid search times, which is necessary to set for different datasets. Second, the hyperparameters' values of different subgroups don't have the same character, which makes them different from each other. Third, due to the huge scale of most hyperparameters' values, these small datasets correspondingly have the large distribution range.

Based on the above candidate sub-models, different ensemble sub-models and combination methods are optimally selected. For NIR dataset, all the sub-models are fused by PLS algorithm in terms of 7 latent features, whose RMSE is 4.4806. For MVF dataset, three sub-modes based on IMF2, IMF4 and IMF7 are fused by Entropy algorithm with RMSE 0.1091. And for DXN dataset, PLS algorithm same as NIR dataset is used to fuse all candidate sub-models with RMSE 0.01172. The extracted contribute ratio for input and output data blocks with 2 latent features are 71.06% and 87.88%, respectively.

### C. COMPARISON RESULTS

The proposed method is compared with the baseline methods (PLS and RWNN) and differential modeling approaches for NIR, MVF and DNB data. RWNN is a single hidden layer feed-forward network. Its output weights are computed analytically via Moore-Penrose generalized inverse method. Thus, it has unstable prediction performance. The decorrelated ensemble algorithm based RWNN is used to overcome this disadvantage, wherein learning parameters are selected by GA [42]. The results are shown in Table 5.

**TABLE 4. Grid search results of hyperparameter pairs.**

| Sub-group | Feature selection | NIR data | MVF data | DXN data |
|---|---|---|---|---|
| I | 1st | {1e3,100} | {10,1} | {100,0.1} |
| I | 2nd | {1e4,208} | {100,2.08} | {10,0.0595} |
| I | 3rd | {9.0100e4,432.6} | {158.5,2.267} | -- |
| I | 4th | {8.1180e5,899.9} | -- | -- |
| I | 5th | {7.3143e6,1.8718e3} | -- | -- |
| II | 1st | {1,10} | {10,10} | {6e4,1600} |
| II | 2nd | {1.5850,10.9} | {15.85,10.9} | {6e5,4912} |
| II | 3rd | {0.9531,6.4855} | {17.276,11.881} | -- |
| II | 4th | {1.0280,7.0692} | -- | -- |
| II | 5th | {1.1205,7.7054} | -- | -- |
| III | 1st | {1,10} | {1.6e5,0.1} | {8e3,10} |
| III | 2nd | {0.5950,5.95} | {1.6e6,0.208} | {4.43e4,15.85} |
| III | 3rd | {0.3540,3.5403} | {1.6e7,0.2267} | -- |
| III | 4th | {0.3859,3.8589} | -- | -- |
| III | 5th | {0.4206,4.2062} | -- | -- |
| IV | 1st | {1,10} | {1,0.01} | {1,0.1} |
| IV | 2nd | {1.0900,15.85} | {1.09,0.0109} | {0.595,0.0595} |
| IV | 3rd | {1.1881,17.2765} | {1.1881,0.0119} | -- |
| IV | 4th | {1.2950,18.8314} | -- | -- |
| IV | 5th | {0.7705,11.2047} | -- | -- |
| V | 1st | {100,1} | {1.6e5,2.56e4} | {1.6e5,2.56e4} |
| V | 2nd | {10,0.5950} | {1.6e6,9.12e4} | {1.6e6,1.16e5} |
| V | 3rd | {10.9,0.6486} | {1.6e7,4.15e5} | -- |
| V | 4th | {11.881,0.7069} | -- | -- |
| V | 5th | {18.8314,0.7705} | -- | -- |
| VI | 1st | {100,100} | {10,1} | {1000,1} |
| VI | 2nd | {109,109} | {100,0.5950} | {9509,1.5850} |
| VI | 3rd | {118.81,118.81} | {1e4,0.6486} | -- |
| VI | 4th | {70.6919,70.6919} | -- | -- |
| VI | 5th | {42.0617,77.0542} | -- | -- |
| VII | 1st | {1,10} | {10,0.1} | {8e4,1} |
| VII | 2nd | {2.5750,5.95} | {100,0.0595} | {2.06e5,1.09} |
| VII | 3rd | {2.8067,6.4855} | {1e4,0.0649} | -- |
| VII | 4th | {3.0594,7.0692} | -- | -- |
| VII | 5th | {1.8203,7.7054} | -- | -- |

**TABLE 5. Statistical results of different modeling methods for NIR, MVF and DXN datasets.**

| Method | NIR | MVF | DXN | Note |
|---|---|---|---|---|
| PLS | 7.154 | 0.2022 | 0.01790 | Single model |
| RWNN | 12.41 ± 3.328 | 0.1658± 0.008351 | 0.02697 ± 0.001782 | Single model |
| PCA-MI-LSSVM | 5.1115 | 0.1231 | 0.01563 | Single model |
| Ref [42] | 6.015 ± 0.6696 | -- | -- | EN model,GA |
| Ref [40] | -- | 0.1386 | -- | SEN model, AWF weighting |
| Ref [41] | -- | -- | 0.01332 | SEN model, Entropy weighting |
| This paper | 4.4806 | 0.1091 | 0.01172 | PLS, Entropy and PLS weighting |

Table 5 shows the following:

(1) For the single model, the linear PLS and nonlinear RWNN models have the worst prediction performance. The RWNN approach does not have optimized input features and hidden layer node selection, thereby having poor prediction stability. Bu using PCA, the extracted latent features are independent of each other. Thus, the PCA–MI–LSSVM method constructed with all input features has the best prediction performance among all single modeling methods.

(2) For the ensemble (EN) model, the input features and learning parameters of RWNN submodels are jointly selected by GA-based optimization method. Compared to PCA–MI–LSSVM method, the smaller RMSE (4.801) is obtained. However, this method also has disadvantages (i.e., long optimization time and poor prediction stability) which are fatal for small sample dataset. Therefore, it is not useful to model MVF and DXN datasets with GA-based optimization method.

(3) The proposed SEN method has the best performance in terms of RMSE for NIR, MVF and DXN datasets. For MVF dataset in Ref [29], more IMFs' frequency spectrum data are utilized to construed SEN model through combination method based on adaptive weighting fusion (AWF) algorithm. However, AWF algorithm can't extract the independent latent features with high contribution ratio and relativity of mill load parameter. Moreover, the proposed method inclusively makes an optimally selection of weighting methods. For DXN dataset, the method in Ref [41] has three shortcomings, which are: the first-layer latent feature selection algorithm does not consider the feature reduction ratio; the mutual information based on feature selection in the second layer doesn't consider the differences among subgroups; and only a fixed threshold value is set for all subgroups. Moreover, the first-layer latent feature is just selected according to the prediction performance so that overfitting problem cannot be avoided. Another problem is that only entropy-based weighting method is used to combine the ensemble sub-models. To overcome above deficiencies, the proposed method is suitable for modeling a small sample of high-dimensional process data.

In summary, the proposed method has the best predictive performance, which can effectively and selectively fuse the latent features of the feature subgroups that represent local information and global information.

### D. DISCUSSIONs ON LEARING PARAMETERS

The above-mentioned simulation results show the effectiveness of the proposed method. However, several learning parameters, such as feature reduction ratio, contribution ratio, mutual information threshold-step time and grid search time, have to be properly set. Thus, it is of necessity for them to analyze and discuss separately.

Here, the candidate sets for feature reduction ratio, contribution ratio, mutual information threshold-step time and grid search time are configured as 2-7, {0.00001, 0.00005, 0.00010, 0.00050, 0.00100, 0.00500, 0.01000}, 1-10 and 2-7, respectively. In the simulation process, just one learning parameter is changed at each simulation test. And the default values of the four learning parameters are set as 2, 0.001, 5 and 3, respectively.

The relationships between these learning parameters and RMSEs are shown in Figs. 4-6 for NIR, MVF and DXN

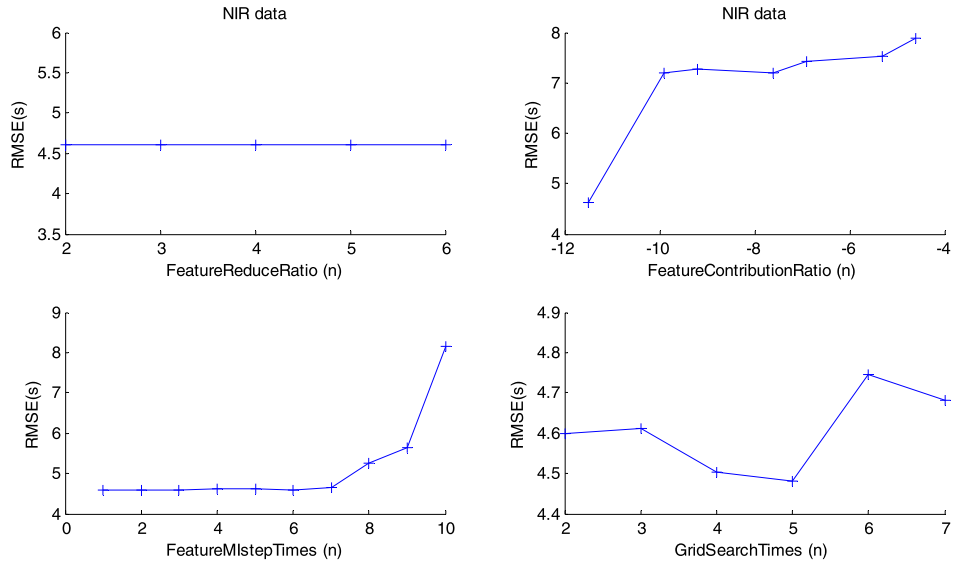


**FIGURE 4.** Relationships between four learning parameters and RMSEs for NIR data.

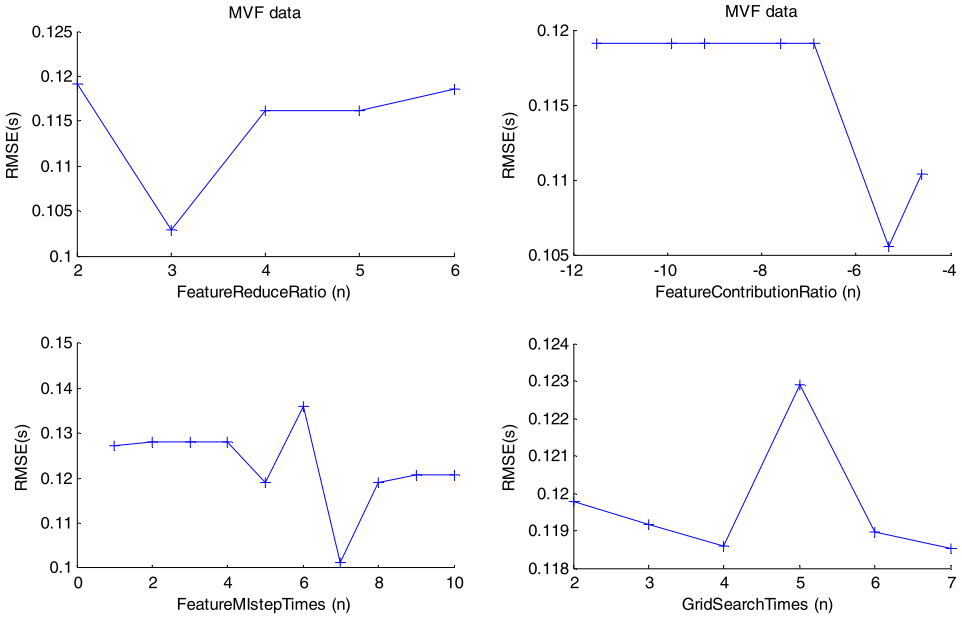


**FIGURE 5.** Relationships between four learning parameters and RMSEs for MVF data.

datasets, respectively. In order to show the feature contribution ratio clearly, base-10 logarithm is used to preprocess it.

Fig. 4 shows that there are minimum values for three learning parameters except for the feature reduction ratio. The reason is that not only the training sample number is not enough small but also the high collinearly exits among wavelength of NIR dataset. For example, if feature reduction ratio is 6, the expected latent feature number will be $150/6 = 25$. However, Table 1 shows that the first six latent feature have been captured nearly up to 100%. Thus, the RMSEs with different feature reduction ratios are the same, which also shows the effectiveness of the proposed method.

For MVF and DXN datasets, the number of training samples is rather small. Figs. 5 and 6 show that all four learning parameters have the minimum extreme values, while Figs. 4-6 show that there are a little insensitive fluctuation for these learning parameters. In a word, they have the property of data dependence. As the former learning parameters directly impact on the latter ones, a jointly optimization strategy should be chosen to solve the problem of interaction effectiveness.

**Mark3:** Compared with the single-model method, the proposed SEN model is a time-consuming process because it uses the grid search approach to optimize the learning

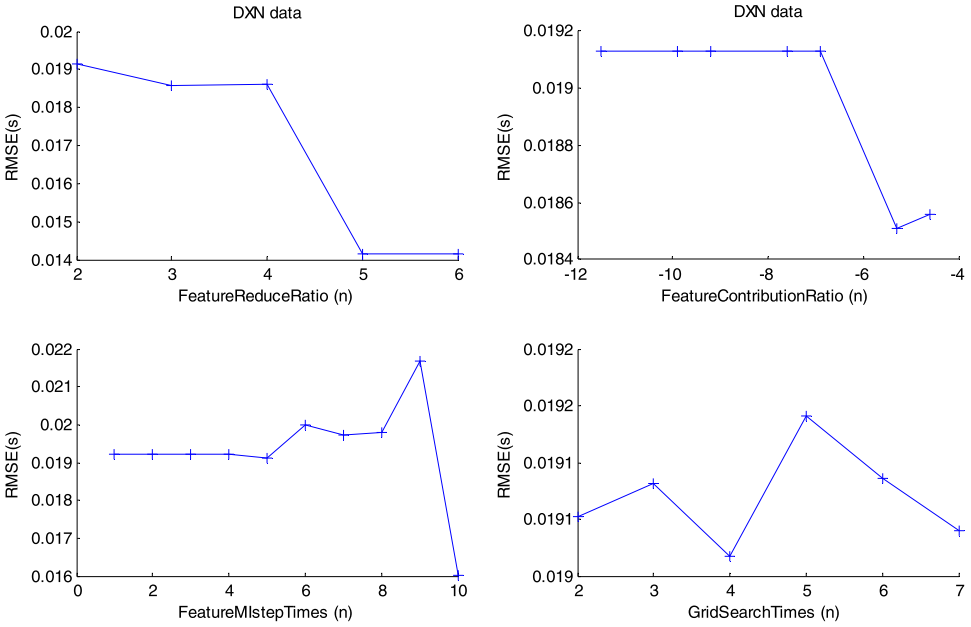


**FIGURE 6.** Relationships between four learning parameters and RMSEs for NIR data.

parameters. In the actual industry, the characteristics of the modeling process have to drift. Thus, the old SEN model must be updated or retrained. In this online updating condition, certain history knowledge from the old model can be transferred. For example, the number of the ensemble submodels and the range of the learning parameters have a limited range. Thus, the number of grid cells and the consumed time are reduced.

## V. CONCLUSION

To address the difficulty of modeling small-sample high- dimensional process data, this study proposes a soft measurement method based on multisource latent feature SEN modeling. The original input features are divided into multiple subgroups that represent local information. All input features that represent global information are taken as one special subgroup. After the latent features being extracted from different subgroups, a three-layer feature selection strategy is proposed. Concretely, the first-layer feature selection strategy based on feature reduction ratio is employed to guarantee the expected ratio of the number of the training samples to the input features. The second-layer feature selection strategy based on contribution ratio ensures the prediction stability. The third-layer feature selection strategy based on mutual information contributes to the relativity of the predicted process parameter. Moreover, a multiage grid search method for hyperparameter adaptive optimization is proposed. Based on the branch-and-bound-based SEN modeling mechanism, the adaptive selection strategy in terms of ensemble size, ensemble submodels, and their combination method is realized. The proposed method can not only avoid the loss of valuable information during the feature selection process but adaptively and selectively fuse the information from subgroups with complementary characteristics. The effectiveness of the proposed method is verified by the benchmark NIR data, high-dimensional MVF data and the industrial DXN emission concentration data, which can also be further extended to a general modeling framework.

**TABLE 6.** Abbreviations and their meanings.

| | Abbreviations | Meanings |
|---|---|---|
| 1 | SEN | Selective ensemble |
| 2 | MSWI | Municipal solid waste incineration |
| 3 | DXN | Dioxin |
| 4 | PCA | Principal component analysis |
| 5 | PC | Principal component |
| 6 | LS–SVM | Least squares–support vector machine |
| 7 | NIR | Near infrared |
| 8 | MI | Mutual information |
| 9 | RMSE | Root-mean-square error |
| 10 | AWF | Adaptive weighting fusion |
| 11 | PLS | Partial least squares |
| 12 | RWNN | Random weight neural network |
| 13 | LV | Latent variable |
| 14 | EN | Ensemble |

## APPENDIX

See Table 6.

## REFERENCES

[1] T. Y. Chai, "Operational optimization and feedback control for complex industrial processes," *Acta Automatica Sinica*, vol. 39, no. 11, pp. 1744–1757, Jan. 2013.
[2] T. Ko and H. Kim, "Fault classification in high-dimensional complex processes using semi-supervised deep convolutional generative models," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2868–2877, Apr. 2020.

[3] W. Shao, Z. Ge, Z. Song, and J. Wang, "Semisupervised robust modeling of multimode industrial processes for quality variable prediction based on Student's t mixture model," *IEEE Trans. Ind. Informat.*, vol. 16, no. 5, pp. 2965–2976, May 2020.

[4] J.-F. Qiao, Z.-G. Guo, and J. Tang, "Dioxin emission concentration measurement approaches for municipal solid wastes incineration process: A survey," *Acta Automatica Sinica*, vol. 46, no. 6, pp. 1063–1089, 2020, doi: 10.16383/.j.aas.c190005.

[5] J. Tang, J. Qiao, Z. Liu, X. Zhou, G. Yu, and J. Zhao, "Mechanism characteristic analysis and soft measuring method review for ball mill load based on mechanical vibration and acoustic signals in the grinding process," *Minerals Eng.*, vol. 128, pp. 294–311, Nov. 2018.

[6] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven soft sensors in the process industry," *Comput. Chem. Eng.*, vol. 33, no. 4, pp. 795–814, Apr. 2009.

[7] D.-C. Li and C.-W. Liu, "Extending attribute information for small data set classification," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 452–464, Mar. 2012.

[8] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ.-Comput. Inf. Sci.*, Jun. 2019, doi: 10.1016/j.jksuci.2019.06.012.

[9] Y.-S. Lin, "Small sample regression: Modeling with insufficient data," in *Proc. 40th Int. Conf. Comput. Ind. Eng.*, Awaji, Japan, Jul. 2010, pp. 1–7.

[10] Y. Lin, "Modeling with insufficient data to increase prediction stability," in *Proc. 5th IIAI Int. Congr. Adv. Appl. Informat. (IIAI-AAI)*, Kumamoto, Japan, 2016, pp. 719–724.

[11] V. Junttila and M. Laine, "Bayesian principal component regression model with spatial effects for forest inventory variables under small field sample size," *Remote Sens. Environ.*, vol. 192, pp. 45–57, Apr. 2017.

[12] D. Dernoncourt, B. Hanczar, and J.-D. Zucker, "Analysis of feature selection stability on high dimension and small sample data," *Comput. Statist. Data Anal.*, vol. 71, pp. 681–693, Mar. 2014.

[13] J. Tang, Z. Liu, J. Zhang, Z. Wu, T. Chai, and W. Yu, "Kernel latent features adaptive extraction and selection method for multi-component non-stationary signal of industrial mechanical device," *Neurocomputing*, vol. 216, pp. 296–309, Dec. 2016.

[14] T. A. F. Gomes, R. B. C. Prudêncio, C. Soares, A. L. D. Rossi, and A. Carvalho, "Combining meta-learning and search techniques to select parameters for support vector machines," *Neurocomputing*, vol. 75, no. 1, pp. 3–13, Jan. 2012.

[15] X. D. Xiao, J. W. Lu, and J. Hai, "Prediction of dioxin emissions in flue gas from waste incineration based on support vector regression," *Renew. Energy Resour.*, vol. 35, no. 8, pp. 1107–1114, Aug. 2017.

[16] X. Lu, W. Liu, C. Zhou, and M. Huang, "Robust least-squares support vector machine with minimization of mean and variance of modeling error," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2909–2920, Jul. 2018.

[17] P. B. C. Miranda, R. B. C. Prudêncio, A. P. L. F. de Carvalho, and C. Soares, "A hybrid meta-learning architecture for multi-objective optimization of SVM parameters," *Neurocomputing*, vol. 143, pp. 27–43, Nov. 2014.

[18] G. Yu, T. Chai, and X. Luo, "Multiobjective production planning optimization using hybrid evolutionary algorithms for mineral processing," *IEEE Trans. Evol. Comput.*, vol. 15, no. 4, pp. 487–514, Aug. 2011.

[19] C. Liu, L. Tang, and J. Liu, "Least squares support vector machine with self-organizing multiple kernel learning and sparsity," *Neurocomputing*, vol. 331, pp. 493–504, Feb. 2019.

[20] S. Yin and J. Yin, "Tuning kernel parameters for SVM based on expected square distance ratio," *Inf. Sci.*, vols. 370–371, pp. 92–102, Nov. 2016.

[21] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorisation," *Inf. Fusion*, vol. 6, no. 1, pp. 5–20, Mar. 2005.

[22] D. Sovilj, K.-M. Björk, and A. Lendasse, "Comparison of combining methods using extreme learning machines under small sample scenario," *Neurocomputing*, vol. 174, no. 22, pp. 4–17, Jan. 2016.

[23] C. Perales-González, M. Carbonero-Ruz, D. Becerra-Alonso, J. Pérez-Rodríguez, and F. Fernández-Navarro, "Regularized ensemble neural networks models in the extreme learning machine framework," *Neurocomputing*, vol. 361, pp. 196–211, Oct. 2019.

[24] J. Zhai, L. Zang, and Z. Zhou, "Ensemble dropout extreme learning machine via fuzzy integral for data classification," *Neurocomputing*, vol. 275, pp. 1043–1052, Jan. 2018.

[25] S. S. Mao, J. W. Chen, L. C. Jiao, S. P. Gou, and R. F. Wang, "Maximizing diversity by transformed ensemble learning," *Appl. Soft Comput.*, vol. 82, pp. 105–110, Sep. 2019.

[26] S. Mao, W. Lin, L. Jiao, S. Gou, and J.-W. Chen, "End-to-end ensemble learning by exploiting the correlation between individuals and weights," *IEEE Trans. Cybern.*, early access, Aug. 14, 2020, doi: 10.1109/TCYB.2019.2931071.

[27] K. Singh, S. Rajora, G. Tripathi, D. K. Vishwakarma, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned convNets," *Neurocomputing*, vol. 371, pp. 188–198, Jan. 2020.

[28] J. Tang, T. Chai, W. Yu, Z. Liu, and X. Zhou, "A comparative study that measures ball mill load parameters through different single-scale and multiscale frequency spectra-based approaches," *IEEE Trans. Ind. Informat.*, vol. 12, no. 6, pp. 2008–2019, Dec. 2016.

[29] J. Tang, J. Qiao, Z. Wu, T. Chai, J. Zhang, and W. Yu, "Vibration and acoustic frequency spectra for industrial process modeling using selective fusion multi-condition samples and multi-source features," *Mech. Syst. Signal Process.*, vol. 99, pp. 142–168, Jan. 2018.

[30] G. Ma, Y. Wang, and L. Wu, "Subspace ensemble learning via totally-corrective boosting for gait recognition," *Neurocomputing*, vol. 224, pp. 119–127, Feb. 2017.

[31] X. P. Wang, Y. Zhang, Z. Wang, and L. X. Tang, "Naphtha pyrolysis process modeling based on ensemble learning with LSSVM," *Comput. Aided Chem. Eng.*, vol. 44, pp. 2035–2040, 2018.

[32] Q. Chen, J. Ding, S. Yang, and T. Chai, "A novel evolutionary algorithm for dynamic constrained multiobjective optimization problems," *IEEE Trans. Evol. Comput.*, vol. 24, no. 4, pp. 792–806, Aug. 2020.

[33] C. Yang, J. Ding, Y. Jin, C. Wang, and T. Chai, "Multitasking multiobjective evolutionary operational indices optimization of beneficiation processes," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 3, pp. 1046–1057, Jul. 2019.

[34] Q. Chen, J. Ding, S. Yang, and T. Chai, "Constrained operational optimization of a distillation unit in refineries with varying feedstock properties," *IEEE Trans. Control Syst. Technol.*, early access, Nov. 21, 2019, doi: 10.1109/TCST.2019.2944342.

[35] J. Shawe-Taylor, M. Anthony, and N. L. Biggs, "Bounding sample size with the Vapnik-Chervonenkis dimension," *Discrete Appl. Math.*, vol. 42, no. 1, pp. 65–73, Feb. 1993.

[36] Y. Muto and Y. Hamamoto, "Improvement of the parzen classifier in small training sample size situations," *Intell. Data Anal.*, vol. 5, no. 6, pp. 477–490, Dec. 2001.

[37] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, Mar. 1991.

[38] Z. H. Guo, J. Tang, and J. F. Qiao, "Mathematic simulation model of dioxin emission concentration of municipal solid waste incineration based on aspen-plus software," in *Proc. 37th Chin. Control Conf.*, Wuhan, China, Jul. 2018, pp. 3502–3507.

[39] J. Tang and J. F. Qiao, "Dioxin emission concentration soft measuring approach of municipal solid waste incineration based on selective ensemble kernel learning algorithm," *J. Chem. Ind. Eng. (China)*, vol. 70, no. 2, pp. 696–706, Jan. 2019.

[40] J. Tang, W. Yu, T. Chai, Z. Liu, and X. Zhou, "Selective ensemble modeling load parameters of ball mill based on multi-scale frequency spectral features and sphere criterion," *Mech. Syst. Signal Process.*, vols. 66–67, pp. 485–504, Jan. 2016.

[41] J. Tang, J. F. Qiao, and Z. H. Guo, "Dioxin emission concentration soft measurement based on multi-source latent feature selective ensemble modeling for municipal solid waste incineration process," *Acta Automatica Sinica*, Jun. 2020, doi: 10.16383/j.aas.c190254.

[42] J. Tang, J. Qiao, J. Zhang, Z. Wu, T. Chai, and W. Yu, "Combinatorial optimization of input features and learning parameters for decorrelated neural network ensemble-based soft measuring model," *Neurocomputing*, vol. 275, pp. 1426–1440, Jan. 2018.

**JIAN TANG** (Member, IEEE) received the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2012. He is currently a Professor with the Beijing University of Technology. His research interests include small sample data intelligent modeling and intelligent modeling and control of municipal solid waste incineration process.

**JIAN ZHANG** received the M.S. degree in applied mathematics from Liaoning University, in 2008, and the Ph.D. degree in pattern recognition and intelligent systems from Northeastern University, China, in 2012. He is currently working with the Nanjing University of Information Science and Technology, China. His research interests include wireless sensor networks, edge computing, and machine learning.

**GANG YU** received the M.S. and Ph.D. degrees in control theory and engineering from Northeastern University, Shenyang, China, in 2006 and 2013, respectively. He is currently a Senior Engineer with the State (Beijing) Key Laboratory of Process Automation in Mining & Metallurgy. His current research interests include modeling and optimization for the complex industrial systems, planning and scheduling, and intelligent optimization methods.

**WENPING ZHANG** received the M.S. degree in mineral processing engineering from the Shandong University of Science and Technology, in 2009. He is currently a first-level Research and Development Engineer and a Project Team with Shandong Gold Mining Technology Company Ltd. His current research interests include the process mineralogy parameter detection and analysis, and so on.

**WEN YU** (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Northeastern University, Shenyang, China, in 1995. Since 1996, he has been with the Centrode Investigación yde Estudios Avanzados, National Polytechnic Institute (CINVESTAV-IPN), Mexico City, Mexico, where he is currently a Professor with the Departamento de Control Automatico. Since 2006, he has been a Visiting Professor with Northeastern University. He serves as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS, *Neurocomputing*, and the *Journal of Intelligent and Fuzzy Systems*.