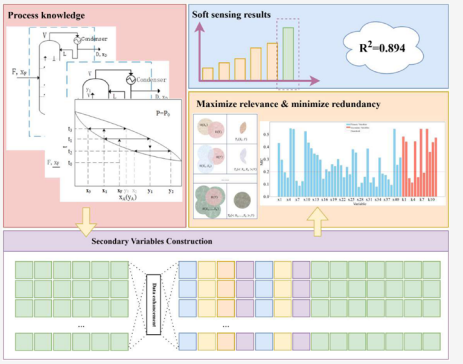
I&EC research
Industrial & Engineering Chemistry Research
pubs.acs.org/IECR
Article

Harnessing Process Knowledge via Soft Sensing for Accurate Online Monitoring: A Study on the Dow Data Challenge Problem

Wantong Fang,[§] Cheng Ji,[§] Fangyuan Ma, Zheyu Jiang, Jingde Wang,* and Wei Sun*

Cite This: *Ind. Eng. Chem. Res.* 2025, 64, 15363−15376

ABSTRACT: Process monitoring is a critical component in modern manufacturing facilities to guarantee process safety, ensure product quality, and improve process operability. In this work, we present a lightweight yet effective process monitoring framework that synergistically incorporates useful process knowledge into online monitoring. This new framework can greatly enhance the monitoring capabilities of existing sensing infrastructure without installing new physical sensors. The idea is to smartly construct essential secondary variables based on domain knowledge to uncover underlying process behaviors, followed by carefully selecting process variables with minimal redundancy for soft sensing. Both secondary variable construction and variable selection approaches are part of data preprocessing steps and thus can be easily done offline and integrated with a range of soft sensing algorithms. Compared to existing soft sensor modeling approaches that rely on complex deep learning architectures, our framework does not require extensive training and can be generalized to other monitoring applications. We also introduce and incorporate an adaptive variable selection mechanism based on the concept of mutual information to select original and secondary variables for monitoring, which significantly reduces measurement redundancy and further improves computational efficiency and monitoring accuracy. Together, these innovations lead to the highest accuracy (R^2 = 0.894) ever reported in the literature in online estimation of product impurity concentration in the Dow data challenge problem. Overall, our proposed monitoring framework offers a scalable, lightweight, and explainable solution for real-time process monitoring and quality control of manufacturing processes.



INTRODUCTION

Fast, accurate, and reliable process monitoring is essential to ensuring process safety, improving product quality, and reducing operating costs of modern manufacturing facilities as their scale and complexity continue to expand. Historically, process monitoring is enabled by measuring and analyzing online process data produced by physical sensors installed in the process units. However, solely relying on primary process data measured by physical sensors can limit process monitoring capabilities and lead to subpar performance due to several reasons. First, many process variables that directly reflect operating status and system health may not even be measurable by physical sensors.[1] And, even if they are, they are often not accessible in real time. For example, composition measurement in most chemical plants and refineries generally requires a sample to be taken by an operator and sent to an onsite laboratory for analysis while the process is running, thereby causing considerable time delay and measurement error in the analyzed results. Although inline composition analyzers are being developed, most of them are still premature and expensive to install and maintain, which is a significant bottleneck for large-scale deployment. As a result, industrial practitioners have been actively seeking ways to enhance process monitoring capabilities by leveraging existing sensing infrastructure without investing heavily in new, expensive sensors. In other words, new techniques need to be developed to complement physical sensor measurements to synergistically enhance process monitoring performance.[2]

Along this line, soft sensing is an emerging concept that shows great promise in harnessing process knowledge and uncovering the underlying process dynamics that cannot be clearly elucidated using only physical sensors. Soft sensing complements conventional physical sensing by estimating important yet hard-to-measure target variables using easy-to-measure primary process variables produced by physical sensors. Such estimation can be accomplished by two approaches. One is through the use of first-principle mathematical models that explicitly describe the functional relationship between hard-to-measure variables and easy-to-measure variables, and the other approach uses data-driven

Received: April 22, 2025
Revised: July 9, 2025
Accepted: July 14, 2025
Published: July 24, 2025

methods to implicitly learn the relationship between the two sets of process variables. Clearly, the choice between these two approaches and the effectiveness of soft sensing in harnessing process knowledge and dynamics highly depends on, first, which secondary variables are constructed and added to the list of monitoring variables and, second, whether explicit functional relationships exist between the primary and secondary variables. These criteria have directed soft sensing research and development efforts over the years.

Specifically, compared to data-driven methods, research on first-principle methods for soft sensing has been quite limited. This is mainly because industrial process dynamics are typically highly nonlinear, time-varying, and correlated, and thus it is nearly impossible to develop soft sensing methods straight from first-principle mathematical models. Instead, prevailing soft sensing research has been primarily focusing on developing data-driven techniques that do not rely on expert knowledge or understanding of the process being monitored.[3] Instead, the idea is to extract and learn the complex process dynamics from historical process data, from which secondary process variables can be inferred and predicted. These data-driven soft sensing methods range from classic statistical algorithms, such as partial least-squares[4] and support vector regression,[5] to more advanced machine learning and deep learning approaches, such as denoised autoencoders,[6] convolutional neural network (CNN),[7] CNN-based Gaussian process regression[8] and hybrid sequence-to-sequence recurrent neural network-deep neural network (Seq2Seq RNN-DNN).[9] However, these purely data-driven approaches face some common limitations. First, due to the lack of process knowledge, purely data-driven soft sensing faces interpretability and generalization challenges, which would become barriers to their widespread adoption among plant operators and process engineers. Furthermore, the monitoring performance of purely data-driven soft sensing methods depends on the availability of relevant historical process data. Thus, these methods typically do not perform well for new operating conditions and unexpected disturbances or faults that have not been encountered before.

To address these intrinsic limitations of purely data-driven approaches, recent soft sensing research efforts have started to incorporate process knowledge and insights, including mass, energy and momentum balances, thermodynamic and kinetic laws, as well as spatiotemporal process dynamics, into a data-driven modeling framework to enhance its performance. These efforts can be classified into two broad categories. The first category directly integrates partial/ordinary differential algebraic equations governing into the loss function of data-driven soft sensing methods in the form of physics-informed neural networks (PINNs). For instance, to monitor β-carotene biosynthesis in *Saccharomyces cerevisiae* fermentation process, Bangi et al.[10] encode mass conservation relationships and reaction kinetics using neural ordinary differential equation, which is then embedded in a data-driven soft sensing framework. The resulting hybrid soft sensing method leads to notable monitoring performance enhancement.

Despite showing promising potential, the implementation of PINN in soft sensing for process monitoring inevitably faces the trade-off between model accuracy and complexity. In particular, any PINN model that attempts to capture the nonlinear and coupling process dynamics to a reasonable degree of accuracy will become quite complex and computationally expensive to solve, thereby posing a need to carefully design tailored PINN architectures.[11] However, this can only be done in a trial-and-error procedure with a deep understanding of both PINN and detailed process dynamics. Furthermore, during neural network training, residual loss functions involving partial/ordinary differential equations are prone to gradient instability and local optimality issues and thus require meticulous hyperparameter tuning and substantial computational resources.[12,13] Last but not least, the accuracy of PINN also depends on the accuracy of model parameters, many of which are hard to obtain in practice. For example, characterizing the heat and mass transfer occurring in a distillation tray using rigorous partial differential equation models requires precise tray efficiency correlations as well as vapor−liquid mass transfer and equilibrium relationships, all of which involve several parameters that may not be available to industrial practitioners. Overall, these implementation issues pose several challenges in leveraging PINN for soft sensing.

The second category for introducing process knowledge in data-driven soft sensing features indirect consideration of spatial and/or temporal correlations embedded in process dynamics using primary sensor measurements collected at different locations. The idea is that, based on the expert knowledge of a given process, specific physical laws and process dynamics can potentially be revealed if sensors are arranged or distributed at strategic locations of the process units. In other words, process knowledge (mass, energy, momentum balances and thermodynamic and kinetic laws) can be inferred by analyzing the spatiotemporal correlations of multisensor measurements. In this regard, Ma et al.[14] introduced a novel spatial feature extraction method to model the temperature distribution across a prereforming reactor in a hydrogen production unit. This results in a novel input feature map, which is then used by a convolutional autoencoder method for fault detection. Similarly, Chen et al.[15] proposed a spatial self-attention mechanism, which can be encoded by graph convolution operations, to discover process knowledge from sensor data and construct soft sensors to predict butane content in the bottoms product of a debutanizer distillation column. Results show that the process knowledge obtained by this soft sensing approach is mostly consistent with expert knowledge. More recently, Ma et al.[16] proposed a multiblock monitoring structure that categorizes the process variables into multiple blocks by leveraging expert process knowledge about their associations with the overall process. This multiblock structure integrated with an orthogonal long short-term memory autoencoder to enhance process monitoring performance. These research efforts highlight the feasibility and potential of bringing process knowledge to elevate soft sensing performance. Nevertheless, it is important to note that, in these methods, process knowledge is only indirectly inferred, rather than directly embedded, from spatial and/or temporal correlations of multisensor data. Furthermore, these spatial and/or temporal correlations may be subtle and may not always be observable by sensor measurements, which are subject to various uncertainties. These issues could potentially limit the effectiveness of the resulting soft sensing techniques.

In summary, we conclude that, despite significant advancements in soft sensing in recent years, questions remain on how to seamlessly and explicitly integrate process knowledge with data-driven soft sensing methods to address the aforementioned issues of prevailing approaches and create an accurate, effective process monitoring framework. To this end, we introduce a lightweight yet powerful soft sensing framework

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

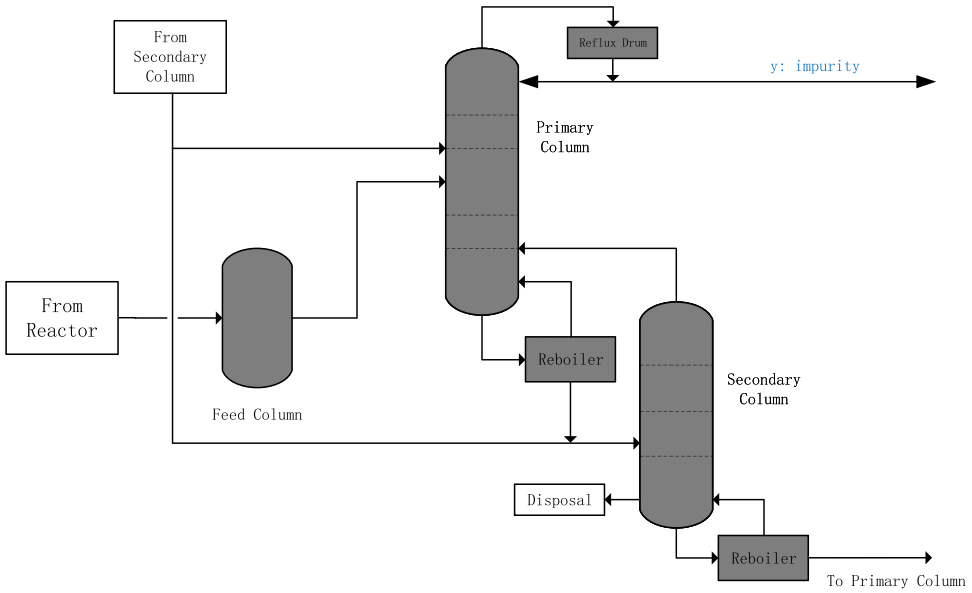


Figure 1. Block flow diagram of the Dow data challenge problem considered for the industrial data analytics study. This figure is referenced from the original work of Qin et al.[18] (Reproduced or adapted with permission from Qin et al.[18] Copyright 2021 Elsevier.).

that incorporates only the most essential process knowledge that is easy to capture, model, and incorporate. Compared to the existing state of the arts, our new framework has the following advantages that favor generalization and widespread adoption. First, rather than pursuing the existing routes of developing complex and computationally expensive PINNs or indirectly inferring process knowledge from sensor data, we propose a more natural, lightweight, and explainable approach to incorporate process knowledge in the form of secondary variables that are directly constructed from simple design equations and shortcut mathematical models. These standard equations and models are available in most undergraduate engineering textbooks, thereby allowing plant operators and process engineers to easily understand and adopt our proposed framework, as well as to make changes on their end. Second, to ensure computational efficiency and scalability, we develop and incorporate a highly effective variable selection method in our soft sensing framework based on mutual information to minimize redundancy of information gain from primary and secondary variables. Third, with strategic secondary variable construction and variable selection strategy in place to incorporate process knowledge, we can use standard machine/deep learning methods for soft sensing, thereby bypassing any complex deep learning architecture. In other words, these innovations come together nicely to form a holistic process monitoring framework that shows outstanding process monitoring performance. Furthermore, since both variable construction and variable selection procedures are part of data preprocessing steps, they can be done offline, making online monitoring very computationally efficient. Specifically, in the Dow data challenge problem, our proposed algorithm achieves the highest accuracy ($R^2 = 0.894$) ever reported in the literature in the online estimation of concentration of product impurity.

The rest of this article is organized as follows. In The Dow Data Challenge Problem section, we give an overview of the Dow data challenge problem as a new, realistic benchmark for process monitoring and industrial data analytics. We review recent breakthroughs in tackling this benchmark problem. Next, in Harnessing Process Knowledge via Secondary Variables Construction and Variable Selection section, we construct secondary variables and incorporate process knowledge for the Dow data challenge problem. We also introduce our adaptive variable selection algorithm to minimize variable redundancy. In Results and Discussion section, we apply these proposed methods to the Dow data challenge problem and systematically compare our monitoring results with existing approaches. Finally, we make some concluding remarks and insights and discuss future work directions in Conclusions section.

THE DOW DATA CHALLENGE PROBLEM

Overview of Dow Data Challenge Problem. The Dow data challenge problem was introduced by Braun et al.[17] at the Dow Chemical Company as one of the first industrial benchmark data sets for process monitoring and data analytics. The problem provides masked data sets obtained from one of Dow Chemical's facilities shown in Figure 1. This process exhibits slow dynamics as a result of impurity accumulation due to accelerated catalyst aging. All process variables presented in the data set (see Table 1) are taken from the plant's separation section, which consists of a feed column (FC), a primary column (PC), and a secondary column (SC). These distillation columns are interconnected through materials streams and recycles. In terms of the data set, the training set covers actual process data collected over a one-year period, whereas the validation data set covers a nine-month period after the training data set was collected. The data set is split into 8800 samples for training, 2200 samples for validation, and 1000 samples for testing. The main focus is to develop a reliable and robust model to predict the impurity concentration in the PC overhead product. For more detailed explanation and instructions of the data challenge, readers can refer to Braun et al.[17] and https://dschemaiandstats.github.io/dowdatasciencechallenge.

Prior Soft Sensing Research in Dow Data Challenge Problem. Since the Dow data challenge problem was first introduced in 2020, it has attracted considerable attention and

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

Table 1. Summary of Process Variables Included in the Dow Data Challenge Problem[a]

unit	variables	unit	variables
primary	x1: PC reflux flow	secondary	x22: SC base concentration
column (PC)	x2: PC tails flow	column (SC)	x23: flow from Input to SC
	x3: input to PC bed 3 flow		x24: SC tails flow
	x4: input to PC bed 2 flow		x25: SC tray DP
	x5: PC feed flow from FC		x26: SC head pressure
	x6: PC make flow		x27: SC base pressure
	x7: PC base level		x28: SC base temperature
	x8: PC reflux drum pressure		x29: SC tray 3 temperature
	x9: PC condenser reflux drum level		x30: SC bed 1 temperature
	x10: PC bed 1 DP		x31: SC bed 2 temperature
	x11: PC bed 2 DP		x32: SC tray 2 temperature
	x12: PC bed 3 DP		x33: SC tray 1 temperature
	x13: PC bed 4 DP		x34: SC tails temperature
	x14: PC base pressure		x35: SC tails concentration
	x15: PC head pressure	feed	x36: FC recycle flow
	x16: PC tails temperature	column (FC)	x37: FC tails flow to PC
	x17: PC tails temperature 1		x38: FC calculated DP
	x18: PC bed 4 temperature		x39: FC steam flow
	x19: PC bed 3 temperature		x40: FC tails flow
	x20: PC bed 2 temperature		
	x21: PC bed 1 temperature		
	x43: PC reflux/feed ratio		
	x44: PC make/reflux ratio		
	y: impurity		

[a]Reproduced or adapted with permission from Qin et al.[18] Copyright 2021 Elsevier.

interest among the process monitoring community as it offers one of the first industrial data sets that can be used as benchmark for researcher and students to develop and experiment various data analytics tools in a realistic setting. As previously discussed, most research studying the Dow data challenge problem can be classified into purely or knowledge-integrated data-driven approaches. For purely data-driven soft sensing, some of the notable works in statistical machine learning include Qin and Liu[19] and Liu and Qin,[20] who proposed a robust variable selection method as well as a two-step sparse learning approach for variable selection and model parameter estimation with optimally tuned hyperparameters in each step. The effectiveness of the proposed approach was demonstrated using the Dow data challenge data set. Subsequently, Qin et al.[21] developed a steepest descent PLS algorithm that leverages the iterative nature of the steepest descent method for more granular regularization path and compared its performance in the Dow data challenge problem with other regularized algorithms. Meanwhile, Barton and Lennox[22] successfully designed a stacked ensemble learning model to improve prediction stability and generalizability of each individual base models consisting of PLS, Lasso, random forest and XGBoost. In addition to statistical machine learning, deep learning approaches have also been developed to study the Dow data challenge problem. For example, Zhu et al.[23] proposed a spatiotemporal stacked autoencoder algorithm, which employs a CNN-LSTM-self-attention architecture to extract spatiotemporal features. More recently, Xu et al.[24] integrated slow feature analysis and LSTM (SLSTM) to capture gradual process variations in the Dow data challenge data set arising from the process's slow dynamics. Meanwhile, Meng et al.[25] introduced a robust dual-rate dynamic data modeling method based on hint convolutional neural network to make full use of dynamic process data sampled at different rates. A common feature of these deep learning methods is that they tend to use fairly complex, stacked neural network architectures to enable hierarchical processing of process data. However, this also comes with the price of increasing model complexity and training efforts. Furthermore, these complex neural networks could suffer from overfitting and poor generalizability, making them limited in handling new fault scenarios, changing operating condition and unforeseen system dynamics.

For knowledge-integrated data-driven soft sensing, some of the recent works include Qin et al.[18] who proposed a statistical learning procedure to integrate process knowledge in all steps from preprocessing to model interpretation for the Dow data challenge problem. An accurate inferential sensor model was built to predict the impurity concentration. Next, Liu et al.[26] introduced a two-step learning approach that runs a knowledge-informed Lasso algorithm[20] twice to identify and preserve key, knowledge-informed process variables followed by building an inferential model using these variables. This method employs cross-validation for secondary hyperparameter optimization, and the computational complexity scales quadratically with respect to the number of variables, making it challenging to adopt in large-scale manufacturing facilities equipped with numerous sensors.

In Table 2, we summarize the impurity concentration estimation accuracy in terms of coefficient of determination

Table 2. Summary of Impurity Concentration Estimation Accuracy for Existing Methods in the Literature

literature	algorithm	R^2	MSE	RMSE
Qin et al.[19]	stable lasso	0.683	–	–
Barton et al.[22]	stacked ensemble model	0.655	–	0.612
Liu et al.[20]	two-step sparse learning	<0.8	[0.375, 0.4]	–
Qin et al.[18]	knowledge-integrated statistical machine learning	0.88	–	–
Qin et al.[21]	steepest descent alternative to the PLS	0.48358	–	–
Liu et al.[26]	knowledge-informed lasso, KILasso	0.7564	0.2286	–
Zhu et al.[23]	spatiotemporal stacked autoencoder	0.885	–	–
Meng et al.[25]	dynamic data denoising generative adversarial imputation network-hint CNN	0.28838	–	–

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

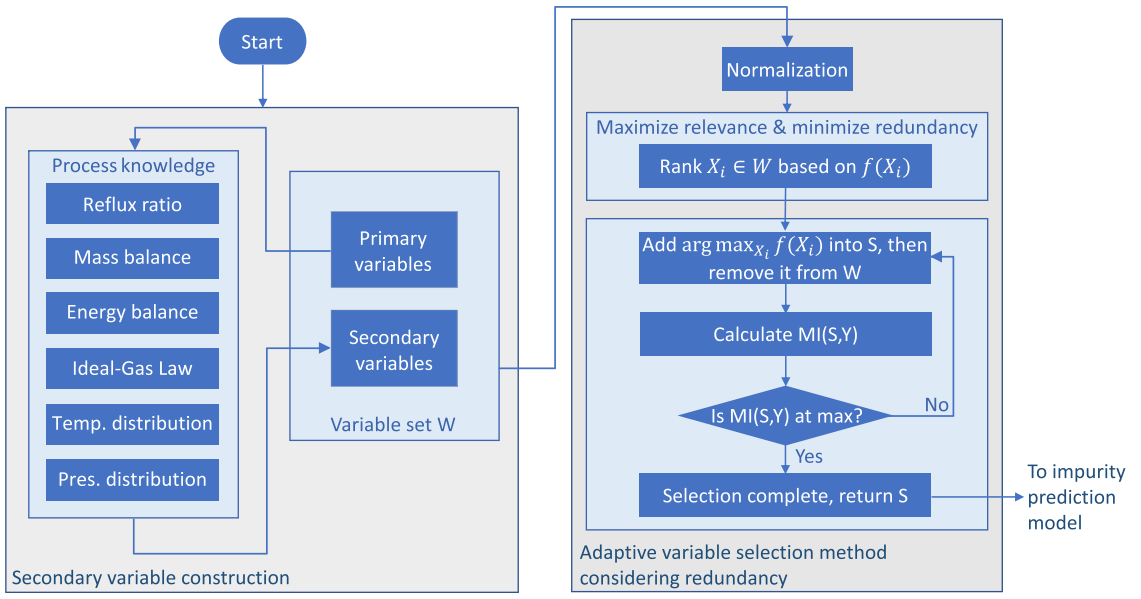


Figure 2. Our proposed framework for efficiently and effectively extracting process knowledge for the Dow data challenge problem. This framework first constructs a set of secondary variables from existing sensor measurements (see Table 1), followed by selecting a subset of primary and secondary variables based on a novel redundancy function (to be defined in later) for soft sensing to estimate impurity concentration.

(R^2) for all works aforementioned. Up to now, the highest R^2 reported in the literature is 0.885,[23] which is achieved by stacking multiple spatiotemporal autoencoders, each consisting of a sequence of CNN layers, LSTM layers, and self-attention layers, for feature/knowledge extraction. In this work, we propose an entirely different approach to harness process knowledge and integrate it with data-driven soft sensing that does not require sophisticated, stacked design of neural networks. Specifically, for the Dow data challenge problem which consists of three distillation columns, we find out that process knowledge can be successfully harnessed by incorporating a few standard distillation design equations, which have been surprisingly overlooked in previous works. This leads to a lightweight yet powerful knowledge-integrated soft sensing framework that outperforms the state-of-the-art methods in terms of R^2 result.

HARNESSING PROCESS KNOWLEDGE VIA SECONDARY VARIABLES CONSTRUCTION AND VARIABLE SELECTION

As shown in the flowchart in Figure 2, process knowledge is acquired and embedded into the data-driven soft sensing algorithm via two modules, namely the secondary variable construction module and adaptive variable selection module. Both modules are part of the data preprocessing step and can be done offline. This saves significant computational time and resources in online monitoring, making our soft sensing framework highly efficient and lightweight to implement. To construct secondary variables for the Dow data challenge problem, we recognize that all unit operations involved in the problem are distillation columns. Therefore, we identify a set of distillation design parameters and operational properties, namely reflux ratio, mass and energy balances, temperature and pressure distribution, and equation of state for the vapor traffic inside the columns.[27] Secondary process variables are then derived from these design parameters and operational properties, which are chosen based on the following criteria. First, they must be directly representable by primary process variables, so that secondary process variables can be synthesized and determined from existing sensor measurements. Second, to ensure computational efficiency and generalizability, each parameter must lead to a simple and relatively accurate shortcut mathematical model that relates a secondary process variable with primary process variables. Again, we emphasize here that, rather than relying on complex deep learning approaches, we adopt a much simpler, lightweight approach of directly implementing already existing and well established distillation modeling equations to uncover the underlying relationships and/or spatiotemporal dynamics between primary and secondary variables.

Once a set of secondary process variables are derived from these design parameters and operational properties, we perform an adaptive screening process select a subset of primary and secondary variables to be included in the soft sensing framework for monitoring. This procedure, which we call as adaptive variable selection, aims to minimize redundancy, reduce computational and data transmission burden during online monitoring, and improve process monitoring performance. By tracking which variables are selected, plant operators and engineers can also gain valuable insights regarding process dynamics and fault propagation. Next, we will discuss each module in detail.

Constructing Secondary Process Variables from Key Design Parameters. As discussed above, for the Dow data challenge problem, secondary process variables are constructed for each key design parameter using simple mathematical models which are readily available in chemical engineering textbooks. And they are directly or indirectly related to the target variable via fundamental physical laws and chemical engineering principles. This separates our proposed method from existing approaches in terms of integrating process knowledge for soft sensing. In Table 3, we summarize the key design parameters and operational properties as well as all 12 secondary process variables derived from these parameters and properties.

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

Table 3. List of Secondary Process Variables Derived from Key Design Parameters and Operational Properties[a]

process parameters	secondary variable	definition	MIC
temperature distribution	k1	x20 − x19	0.4824
	k2	x19 − x18	0.4417
	k3	x30 − x31	0.1476
	k4	x32 − x29	0.1124
pressure distribution	k5	x14 − x15	0.4438
	k6	x27 − x26	0.1557
reflux ratio	k7	$\frac{x5}{x36}$	0.5442
mass balance	k8	$\frac{x23}{x23-x24}$	0.1914
	k9	$\frac{x40}{x37}$	0.5444
ideal gas law	k10	$\frac{x27}{x28+273.15}$	0.3586
energy balance	k11	x1 × (x20 − x19)	0.4374
	k12	x1 × (x19 − x18)	0.4741

[a]The secondary variables are related to primary variables via elementary manipulations, thereby ensuring computational efficiency and robustness. The maximum information coefficient (MIC) values for secondary variables with respect to the target variable (impurity concentration in primary column distillate product) are also calculated.

Secondary Variable Constructed from Reflux Ratio. Reflux ratio is one of the key design parameter and is directly related to the operation and control of distillation column. By tuning the reflux ratio, operators can adjust product purity and energy consumption of a distillation column. From Figure 3a, by combining overall mass balance and component mass balance around the rectifying section of the column, we obtain the operating line equation represented in terms of reflux ratio R_D as eq 1

$$y_{i,n+1} = \frac{R_D}{R_D + 1} x_{i,n} + \frac{x_{i,D}}{R_D + 1} \tag{1}$$

where $x_{i,n}$ (respectively $y_{i,n+1}$) represent the liquid (respectively vapor) molar composition of component i leaving stage n (respectively stage $n + 1$), and $x_{i,D}$ is the composition of distillate product. From the operating line equation, one can see that, when fixing the distillate flow rate, as R_D increases, the distillate product composition for the more volatile component will increase. In other words, for primary column, the impurity concentration increases as $\frac{1}{R_D}$ increases. Therefore, $\frac{1}{R_D}$ for the primary column, namely k7 in Table 3, is chosen as one of the secondary process variables.

Secondary Variables Constructed from Mass Balance. From Figure 3b, by combining overall and component mass balances around the entire column, we obtain the following relationship connecting an intensive property D/F with compositions as eq 2

$$\frac{F}{D} = \frac{x_{i,D} - x_{i,B}}{z_{i,F} - x_{i,B}} \tag{2}$$

where D and F stand for distillate and feed flow rates, respectively, and $z_{i,F}$ and $x_{i,B}$ represent the molar composition of component i in the feed and bottoms product streams, respectively. Similar to the reflux ratio, the F/D ratio also reflects impurity concentration. Specifically, since secondary column and feed column are both connected to the primary

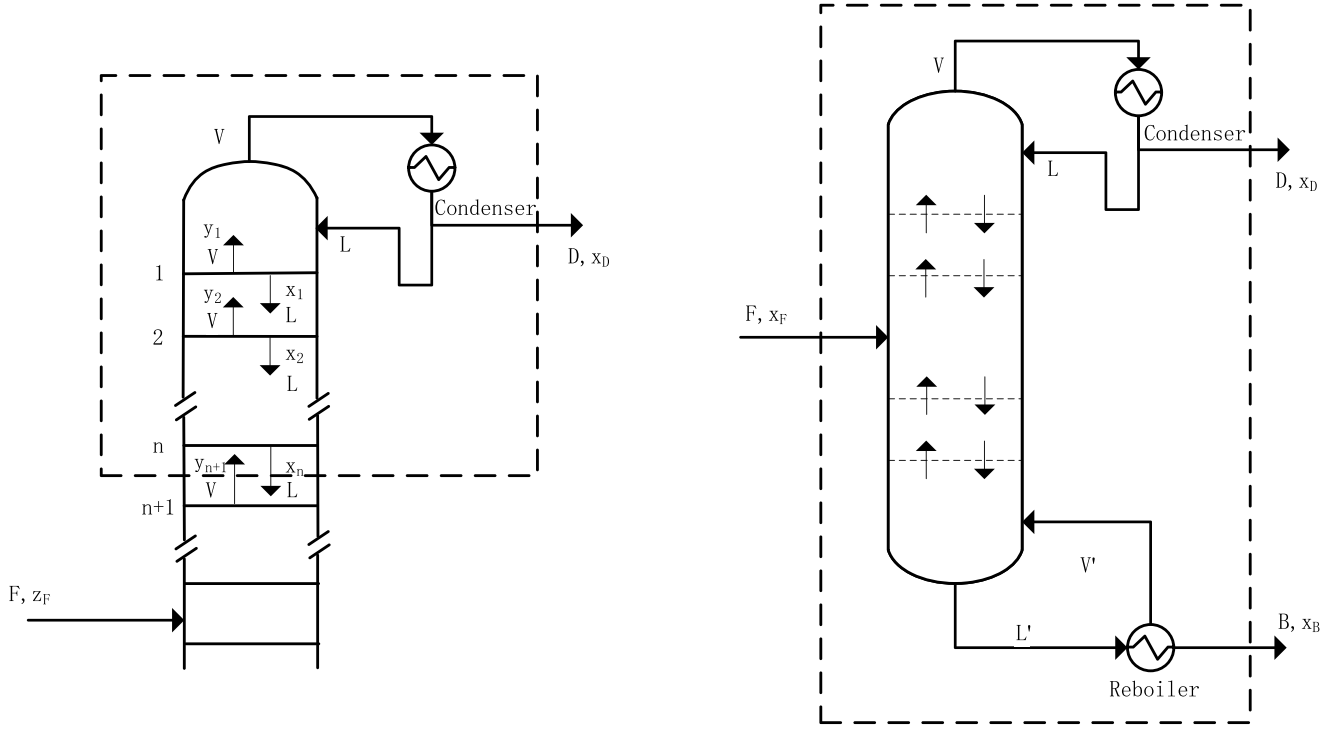


Figure 3. System boundary to derive operating line equation for distillation column.

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

column where impurity concentration in the distillate is of our interest, F/D for the secondary and feed columns are included in the list of secondary process variables. As illustrated in Table 3, we construct two secondary variables, k8 and k9, to monitor the mass balance behavior for the secondary and primary columns, respectively.

Secondary Variables Constructed from Energy Balance. Apart from mass balance, energy balance also plays an important role in distillation column design and operation. Inside a distillation column, thermodynamic equilibrium is achieved between the liquid and vapor phases on each stage. And the standard model for distillation involves simultaneously solving Mass, Equilibrium, Summation, and Heat (MESH) equations for every tray. Heat transfer directly affects the temperature and composition distributions within the distillation column. Therefore, a secondary variable based on heat transfer is adopted as an indicator for the impurity concentration. Such equilibrium involves mass transfer (and thus composition change) between the two phases, which is enabled by mixing and heat transfer. Nevertheless, directly deriving and incorporating detailed energy balance equations for each and every stage in the column is not only computationally expensive but also practically infeasible, as heat cannot be directly measured or determined from primary process variables. Meanwhile, under the reasonable assumption of adiabatic column and constant molar overflow (which translates to similar latent heat of vaporization for all components),[27,28] it can be shown that the amount of heat transferred across a section of the column is proportional to the temperature difference between the two ends of the column section and the flow rate in the section. With this, we construct another secondary variable Q as eq 3

$$Q = L_r(T_{top} - T_{bot}) \tag{3}$$

where L_r is the liquid reflux flow rate and T_{top} and T_{bot} denote the temperature at the top and bottom of the column section of interest, respectively. As shown in Table 3, we construct two secondary variables, k11 and k12, to monitor energy balance within the primary column.

Secondary Variables Constructed from Ideal Gas Law. At low and medium pressures, most gases behave ideally or close to ideal gas. In this case, the density of vapor traffic inside the column is proportional to P/T according to the ideal gas law. Here, P and T denote the pressure and absolute temperature, respectively. As a result, we propose a secondary process variable P/T for the secondary column, namely k10 in Table 3.

Secondary Variables Constructed from Temperature and Pressure Distributions. Finally, monitoring temperature and pressure distributions indicates spatiotemporal dynamics of distillation operation, such as changes in composition and relative volatility. Therefore, instead of following the deep learning approach, such as using spatiotemporal autoencoders,[23] we choose to characterize these spatiotemporal dynamics using secondary variables based on temperature and pressure differences across different locations in the primary and secondary columns. Specifically, secondary variables k1 through k4 listed in Table 3 are constructed to monitor the temperature distribution, whereas k5 and k6 are to monitor the pressure distribution within the column. The idea is that, since temperature and pressure sensor locations are typically fixed, the temperature and pressure gradients within the column, which reflect disturbances and changes in composition distribution inside the column, can be effectively monitored by the temperature and pressure differences taken at different locations.The maximum mutual information coefficient (MIC)[29] was applied to quantify the correlation between the secondary variables and the product impurities, with a value between 0 and 1. Larger values indicate stronger correlation, see Results and Discussion section for more details.

Adaptive Variable Selection for Redundancy Reduction. As mentioned earlier, variable screening and selection helps reduce model complexity and redundancy, enhance computational efficiency, and shed light on process dynamics and fault propagation. Among various variable selection criteria, mutual information (MI) is attractive as it can quantify nonlinear dependencies among variables.[30] Consider any two continuous random variables X and Y, their MI, $\mathrm{MI}(X, Y)$, is given as eq 4

$$\mathrm{MI}(X, Y) = \int_{x\in X}\int_{x\in Y} p(x, y)\log\frac{p(x, y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y \tag{4}$$

where $p(x, y)$ denotes the joint probability distribution of x and y, and $p(x)$ and $p(y)$ correspond to the marginal distributions of $x \in X$ and $y \in Y$, respectively. If there is no statistical correlation between X and Y, then $\mathrm{MI}(X, Y) = 0$. Furthermore, MI is related to the information entropy H by $\mathrm{MI}(X, Y) = H(X) + H(Y) - H(X, Y)$, where $H(X, Y)$ is the joint entropy of X and Y. Generalizing this to high dimensions (for multisensor scenarios in industrial process monitoring) yields the following mutual information expression as eq 5

$$\begin{aligned}&\mathrm{MI}((X_1, X_2, \cdots, X_n), Y)\\ &\quad = H(X_1, X_2, \cdots, X_n) + H(Y) - H(X_1, X_2, \cdots, X_n, Y)\end{aligned} \tag{5}$$

where $W = \{X_1,\cdots,X_n\}$ is a set containing all n primary and secondary process variables and Y is the target variable (i.e., impurity concentration in primary column distillate product).

To adaptively determine the number of (primary and secondary) process variables to be used for soft sensing, we introduce a novel feature selection method based on the normalization of maximum relevance and minimum common redundancy.[31] For a given W, our goal is to identify a subset $S \subseteq W$ that minimizes the redundancy within S. To do this, note that the relative redundancy between for any two process variables $X_i, X_j \in S$ can be defined as eq 6

$$\mathrm{RI}(X_i, X_j) = \frac{\mathrm{MI}(X_i, X_j)}{\max\{\mathrm{MI}(X_i, X_j), \mathrm{MI}(X_i, Y), \mathrm{MI}(X_j, Y)\}} \tag{6}$$

Examining the common redundancy among X_i, X_j, and Y, a new common mutual information CI can be defined as eq 7

$$\mathrm{CI}(X_i, X_j, Y) = \mathrm{RI}(X_i, X_j)\cdot\min\{\mathrm{MI}(X_i, Y), \mathrm{MI}(X_j, Y)\} \tag{7}$$

which, when generalizing to the set level, becomes eq 8

$$\begin{aligned}\mathrm{CI}(X_i, S, Y) = &\frac{\mathrm{MI}(X_i, S)}{\max\{\mathrm{MI}(X_i, S), \mathrm{MI}(X_i, Y), \mathrm{MI}(S, Y)\}}\\ &\min\{\mathrm{MI}(X_i, Y), \mathrm{MI}(S, Y)\}\end{aligned} \tag{8}$$

With this, we can define a redundancy function f for a specific process variable X_i as eq 9

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

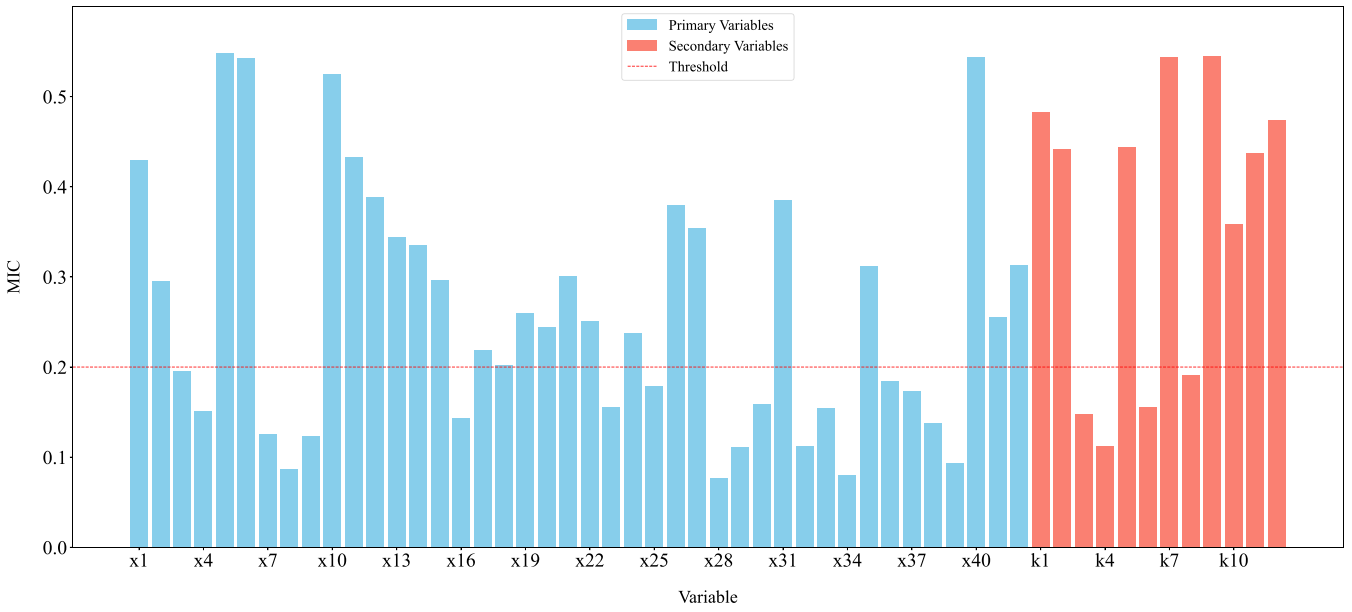


Figure 4. MIC value for each primary (blue bar) and secondary (red bar) process variable. The average MIC value for all primary variables is 0.2580, whereas that for all secondary variables is 0.3611 (40.0% higher than the average MIC value for primary variables). For illustration, we set a threshold MIC value of 0.2.

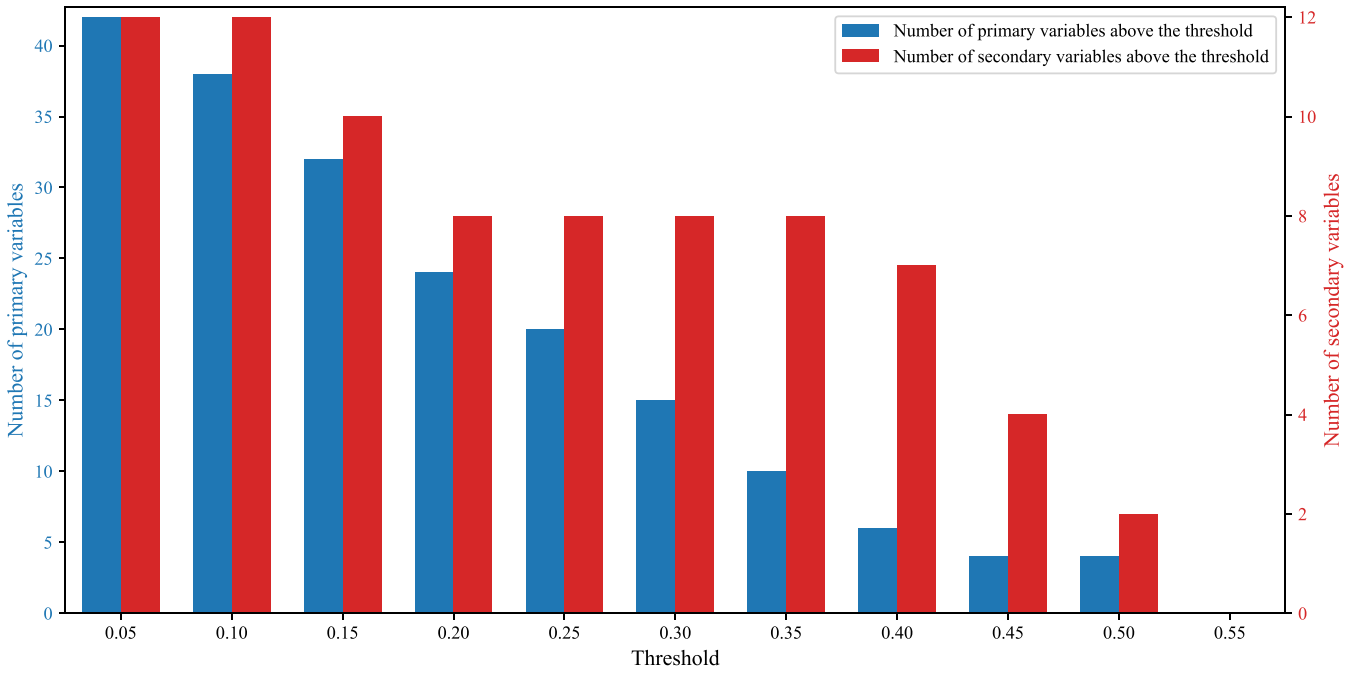


Figure 5. Relationship between MIC threshold value and the number of primary (in blue) and secondary (in red) variables selected for soft sensing.

$$f(X_i) = \frac{\mathrm{MI}(X_i, Y) - \mathrm{CI}(X_i, S, Y)}{\mathrm{MI}(Y, Y)} \tag{9}$$

which favors high relevance with the target variable Y and penalizes common redundancy with other variables in S. Thus, f is a desired measure for ranking process variables. Based on this, we propose an Adaptive Variable Selection method considering Redundancy (AVSR). Starting from an empty set S, AVSR first uses mutual information $\mathrm{MI}(X_i, Y)$ to rank each primary and secondary process variables $X_i \in W$, and the highest ranked process variable is added to S to make it nonempty (hence being able to calculate $\mathrm{CI}(X_i, S, Y)$ and $f(X_i)$). Then, following a greedy search mechanism, AVSR calculates $f(X_i)$ for all $X_i \in W/S$ and the process variable with the highest $f(X_i)$ is selected and added to the subset S. This process is repeated and more process variables are augmented into S until $\mathrm{MI}(S, Y)$, as calculated from eq 5, reaches maximum for the first time. Beyond this point, adding more variables into S will not increase the mutual information any further, suggesting a diminishing return. Here, it is worth noting that the redundancy function $f(X_i)$ is used for ranking and variable selection, whereas the mutual information $\mathrm{MI}(S, Y)$ is used as the stopping criterion. Through the Dow data

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

challenge problem, the effectiveness of AVSR is discussed in Results and Discussion section.

RESULTS AND DISCUSSION

In this section, we systematically evaluate our proposed algorithm using the Dow data challenge data set and compare the results with state of the art approaches. We divide this section into two main parts. First, we validate the effectiveness of our proposed secondary variable construction approach in harnessing process knowledge and improving soft sensing performance. Second, we study the effectiveness of our proposed variable selection method considering redundancy. Specifically, we perform ablation studies to individually evaluate the effectiveness of each component or improvement in our proposed framework to understand how these components holistically contribute to the overall success of the soft sensing framework.

Effectiveness of Secondary Variable Construction Method. First, we validate the effectiveness of secondary process variables constructed in extracting nonredundant features and process dynamics. Here, we use the maximal information coefficient (MIC)[29] as an indicator to assess the correlation among the extracted features.[32] MIC uses binning so as to apply mutual information on continuous random variables as eq 10

$$\mathrm{MIC}(X, Y) = \max_{ab<n^{0.6}} \frac{\mathrm{MI}(X, Y)}{\log \min(a, b)} \tag{10}$$

where n is the sample size of the data set and a and b denote the number of grid partitions for X and Y, respectively. Similar to MI, MIC value of 0 between two variables indicates their mutual independence, whereas as MIC value that approaches 1 suggests the presence of a strong correlation between the two variables.

We calculate the MIC values between any primary or secondary process variable and the target variable (impurity concentration) and summarize the results in Figure 4 and Table 3. The average MIC value for all primary process variables is 0.2580, whereas the average MIC value for the 12 secondary process variables is 0.3611, indicating that the constructed secondary variables perform better in extracting nonredundant underlying features. Furthermore, Figure 5 clearly illustrates the usefulness of introducing these secondary process variables, as given any MIC threshold value, the proportion of secondary variables meeting or exceeding this threshold is always greater than the proportion of primary variables. This shows that secondary variables consistently demonstrate stronger correlations with the target variable compared to primary variables. Overall, these results show that most of these secondary process variables can harness process knowledge that has not been explored or incorporated by the existing primary variables, making these secondary variables necessary and crucial to the accurate estimation of impurity concentration.

To further validate the effectiveness of introducing the constructed secondary variables in soft sensing, we implement three classic machine/deep learning algorithms, namely support vector regression (SVR),[33] artificial neural network (ANN),[34] and long short-term memory network (LSTM),[35] for impurity concentration estimation task. Specifically, SVR excels in handling high-dimensional data, ANN exhibits strong nonlinear approximation capabilities, and LSTM demonstrates superior performance in processing time-series data due to its unique memory structure. Given their unique characteristics, these three models are selected as representatives for evaluating our proposed method. Table 4 quantitatively illustrates how the introduction of secondary process variables improves the impurity concentration estimation accuracy using trained SVR, ANN, and LSTM on the test set. In addition, the estimated and actual impurity concentration under the three methods are shown in Figure 6. We observe that, when secondary variables are incorporated, the estimated impurity concentration profiles match more closely with the actual ones compared to only using primary variables for monitoring. Note that here, we only include primary and secondary process variables whose MIC values are greater than or equal to 0.2 (see Figure 4). This results in 24 out of 42 primary variables (57%) and 8 out of 12 (67%) secondary variables to be selected. It turns out that, in all three soft sensing methods, incorporating secondary process variables improves R^2 and reduces estimation error (in terms of MSE, RMSE, and MAE). Furthermore, we observe that, despite only implementing classic machine/deep learning methods for soft sensing, thanks to our proposed secondary variable construction method, our proposed approach actually achieves great performance that is already better or comparable to the state-of-the-art algorithms in the literature (see Table 2). This result is particularly encouraging and promising, considering that existing state of the arts typically use complex neural network architectures, whereas our proposed approach is quite simple and lightweight to train and deploy. In this case, for models using only primary variables, the ANN has 24, 32, 16, and 1 neurons per layer, while the LSTM has 24, 16, and 1. For models using both primary and secondary variables, the corresponding ANN has 32, 32, 16, and 1 neurons per layer, and the LSTM has 32, 32, and 1 neurons per layer. For the ANN model, we adopted the following hyper parameters: a learning rate of 0.001, batch size of 64, MSE loss function, and the Adam optimizer. The LSTM network was configured with learning rate (0.001) and

Table 4. Incorporating Our Knowledge-Based Secondary Process Variables into Soft Sensing Framework Improves R^2, Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) Metrics Compared to Only Using Existing Primary Variables[a]

soft sensing method	variables included	R^2	MSE	RMSE	MAE
SVR	primary variables only	0.78456	0.16976	0.41202	0.33670
	primary and secondary variables	**0.82404**	**0.13864**	**0.37235**	**0.30045**
ANN	primary variables only	0.74997	0.19701	0.44385	0.34434
	primary and secondary variables	**0.77929**	**0.17390**	**0.41702**	**0.32298**
LSTM	primary variables only	0.81329	0.14712	0.38356	0.30050
	primary and secondary variables	**0.84035**	**0.12580**	**0.35468**	**0.27706**

[a]Note that variables with MIC values greater or equal to 0.2 are selected. For R^2, higher value is better (bold font), whereas for other metrics, lower value is better (bold font).

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

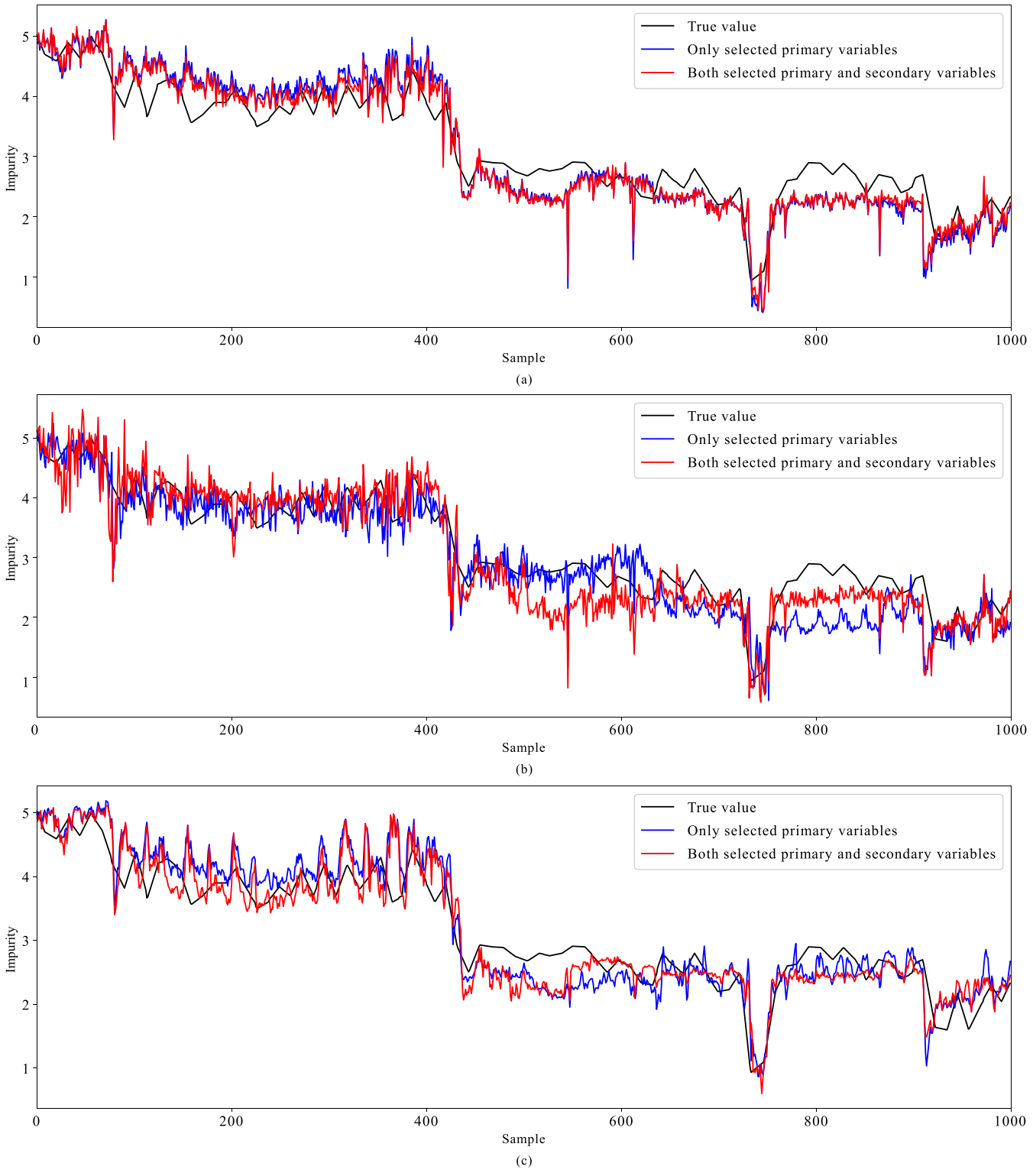


Figure 6. Impurity concentration estimation using (a) SVR, (b) ANN, (c) LSTM on the test set. The blue (respectively red) curve corresponds to the case where only selected primary variables (respectively both selected primary and secondary variables) are considered for soft sensing.

optimizer (Adam), while utilizing a larger batch size of 128 and MSE as the objective function.

Effectiveness of Adaptive Variable Selection Method Considering Redundancy (AVSR). Redundancy among primary and secondary process variables can introduce biases and obscure the underlying process dynamics, thereby deteriorating impurity concentration prediction accuracy. Therefore, we propose an adaptive variable selection method based on mutual information to identify variables with the highest mutual dependence with the target variable (impurity concentration) and lowest common redundancy with other selected variables, as illustrated in eq 9. As discussed before, the AVSR procedure follows a greedy search. To initiate the process, we start from an empty set S, and the process variable

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

with the highest mutual information with the target variable, that is, arg max$_i$ MI(X_i, Y), is selected and added to S. In this case, as illustrated in Table 5 and Figure 4, variable x6, which represents PC make flow, has the highest mutual information value among all 42 primary and 12 secondary process variables and thus is first selected and added to S. Next, we identify the process variable in W/S with the highest $f(X_i)$ and add it to S. This process repeats itself until MI(S, Y) reaches maximum for the first time. In this case, MI(S, Y) reaches maximum when 20 variables are selected (see Table 5). Hence, these 20 variables are used for subsequent soft sensing. Among them, there are 4 knowledge-based secondary variables (20%) and 16 primary variables (80%). Notably, the maximum MI(S, Y) value of 5.413 also equals the entropy of the target variable. According to eq 5, this means $H(S) = H(S, Y)$, meaning that the optimal variable set S contains all the information needed to estimate the impurity concentration.

Table 5. Variables Selected Following the AVSR Approach and Their Order of When the Variable is Added to S[a]

order of when X_i is added to S	variable	MI(S, Y)
1	x6: PC make flow	1.002
2	x43: PC reflux/feed ratio	2.172
3	**k3:** x30 − x31	**3.562**
4	x4: input to PC bed 2 flow	4.836
5	x17: PC tails temperature 1	5.349
6	x7: PC base level	5.405
7	x24: SC tails flow	5.410
8	x35: SC tails concentration	5.412
9	**k4:** x32 − x29	**5.412**
10	x25: SC tray DP	5.412
11	x38: FC calculated DP	5.412
12	x30: SC bed 1 temperature	5.412
13	x44: PC make/reflux ratio	5.412
14	x28: SC base temperature	5.412
15	x16: PC tails temperature	5.412
16	x34: SC tails temperature	5.412
17	x18: PC bed 4 temperature	5.412
18	**k8:** $\frac{x23}{x23 - x24}$	**5.412**
19	x2: PC tails flow	5.412
20	**k10:** $\frac{x27}{x28 + 273.15}$	**5.413**
21−54	...	5.413

[a]Among all 20 variables selected, 4 of them are knowledge-based secondary variables (highlighted in bold).

Furthermore, to validate the effectiveness of our proposed variable selection method, especially the stopping criterion, we conduct an experiment where we monitor the soft sensing performance (in terms of R^2 and MSE) using SVR for different number of variables selected based on maximum $f(X_i)$ principle. From the results shown in Figure 7, we see that the impurity concentration estimation performance first increases, reaches its peak when 20 variables are selected (as given by our AVSR approach), and then deteriorates as more variables are included. This is expected as when the number of variables reaches a certain threshold, introducing more variables no longer provides new information but actually results in redundancy among the variables, thereby leading to a decline in predictive accuracy. This demonstrates the effectiveness of the proposed variable selection method.

Table 6 summarizes the improvement in impurity concentration estimation accuracy on the test set thanks to AVSR under different soft sensing methods. In this case, for models based on MIC variable screening, the ANN has 32, 32, 16, and 1 neurons per layer, while the LSTM has 32, 32, and 1. For models based on AVSR variable screening, the corresponding ANN has 20, 16, 8, and 1 neurons per layer, and the LSTM has 20, 16, 8, and 1 neurons per layer. For the ANN model, we adopted the following hyper parameters: a

Table 6. SVR Results with Knowledge Variables of Different Variable Selection Method

soft sensing method	variable selection	R^2	MSE	RMSE	MAE
SVR	MIC ≥ 0.2	0.82404	0.13864	0.37235	0.30045
	AVSR	**0.85393**	**0.11509**	**0.33926**	**0.26303**
ANN	MIC ≥ 0.2	0.77929	0.1739	0.41702	0.32298
	AVSR	**0.79171**	**0.16412**	**0.40512**	**0.31694**
LSTM	MIC ≥ 0.2	0.84035	0.1258	0.35468	0.27706
	AVSR	**0.89414**	**0.08341**	**0.28881**	**0.21857**

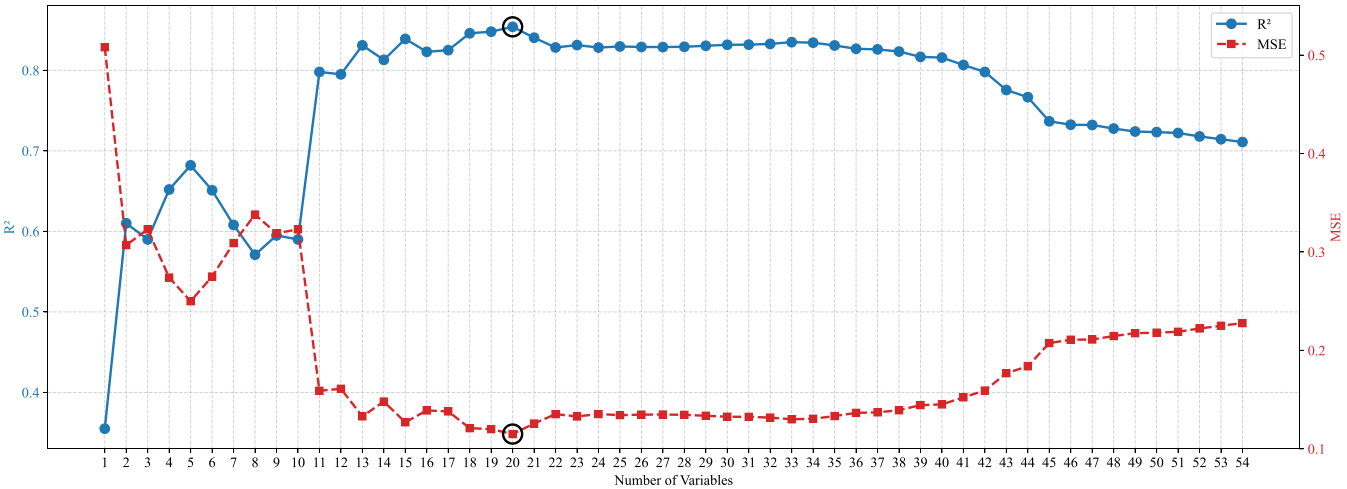


Figure 7. R^2 and MSE results using SVR for different numbers of variables selected for soft sensing. Predictive accuracy is the best when 20 variables are selected for monitoring, which matches with our AVSR result.

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

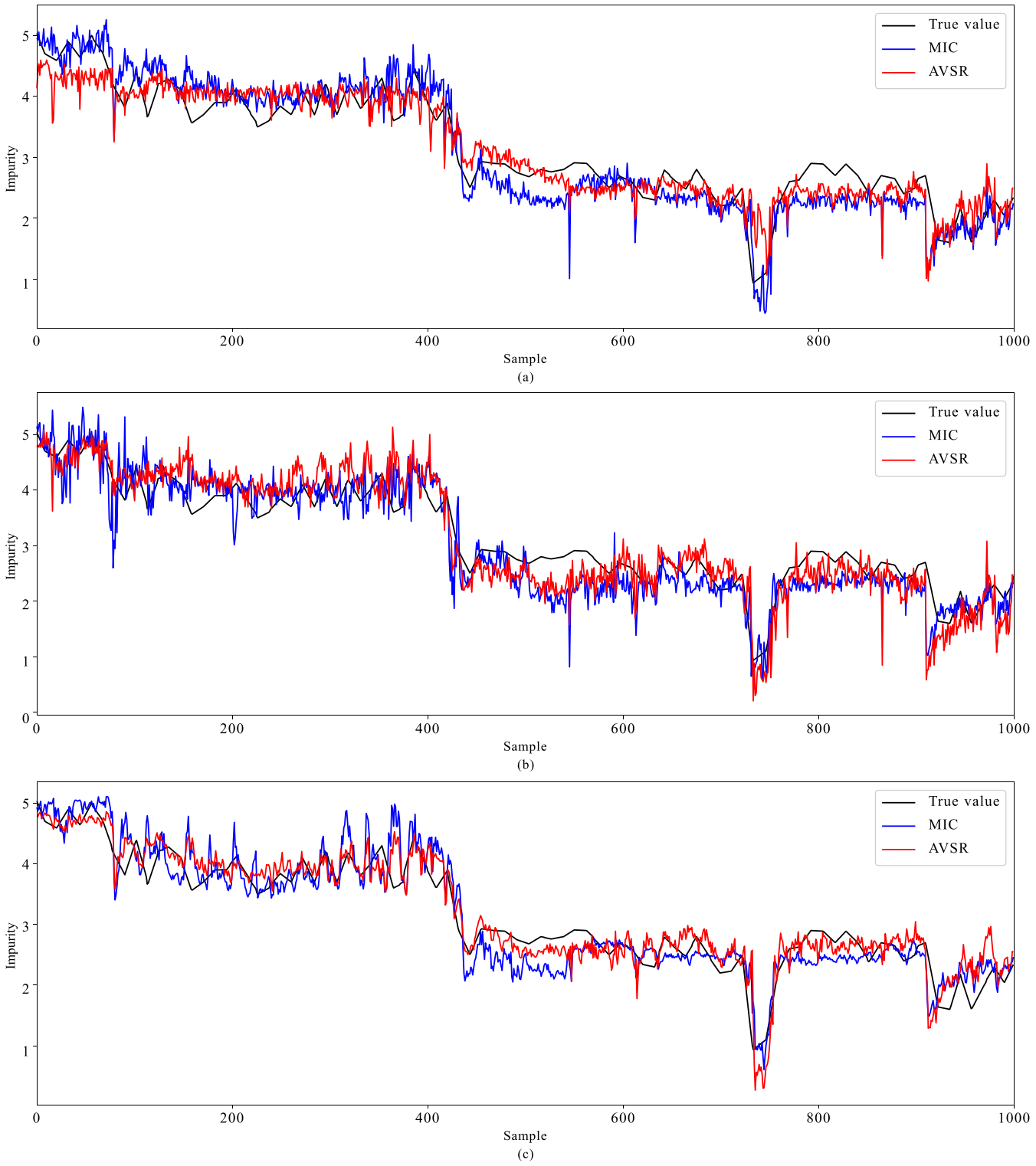


Figure 8. Impurity concentration estimation using (a) SVR, (b) ANN, (c) LSTM on the test set. The red curve adopts AVSR, whereas the blue curve selects variables whose MIC ≥ 0.2 (see Figure 4).

learning rate of 0.001, batch size of 64, MSE loss function, and the Adam optimizer. The LSTM network was configured with learning rate (0.001) and optimizer (Adam), while utilizing a larger batch size of 128 and MSE as the objective function. In addition, the estimated and actual impurity concentration under the three methods are shown in Figure 8. Compared to only using a static threshold (MIC ≥ 0.2) for variable selection, the estimated impurity concentration profiles obtained using our proposed AVSR approach match more closely with the actual ones under all three soft sensing methods. Clearly, our proposed knowledge-based secondary variable construction and adaptive variable selection method considering redundancy synergistically work together to improve soft sensing performance. This enables classic machine/deep learning soft sensing methods to have comparable or even higher performance than state-of-the-art

https://doi.org/10.1021/acs.iecr.5c01644
Ind. Eng. Chem. Res. 2025, 64, 15363−15376

algorithms in the literature (see Table 2). In particular, we show that, when using standard LSTM as the soft sensing algorithm, we successfully achieve the highest R^2 result reported in the literature. This result is insightful because, for the first time, we show that, by smartly designing new data preprocessing steps to incorporate useful process knowledge and select useful process variables, superior soft sensing performance can be achieved without having to use complex, stacked deep learning architectures. Since data preprocessing steps can be done offline and do not involve extensive computations, the overall soft sensing framework is simple and lightweight to implement. More importantly, we remark that one can seamlessly integrate our proposed methods with any soft sensing algorithm as plug-ins, thereby demonstrating its wide adaptability and broad applicability to a range of process monitoring applications.

CONCLUSIONS

In this article, we present a simple, lightweight soft sensing framework to tackle the Dow data challenge problem as well as a slew of process monitoring applications. This framework consists of a simple yet powerful knowledge-based secondary variable construction module and an adaptive variable selection module that minimizes variable redundancy. As part of the data preprocessing steps, both modules work synergistically to construct and select the most essential and useful process variables based on process knowledge and underlying data structure for soft sensor design and process monitoring, thereby significantly reducing computational burden. Using the Dow data challenge problem as an illustrative case study, we show that our proposed secondary variable construction method successfully brings in new process knowledge (in the form of simple design equations) that is previously underexplored by existing physical (primary) sensor measurements. We also show that our proposed AVSR approach can systematically identify and eliminate variable redundancies by leveraging the underlying information and relations embedded in primary and secondary process variables. Together, these innovations have led to the best result reported in the literature in the Dow data challenge even when using a standard LSTM network for soft sensing. Overall, we believe that the methods and results presented in this work suggest a new direction of soft sensing research in that developing more effective data preprocessing steps that can better incorporate process knowledge, increase information gain, and reduce variable redundancy can be as important as developing more advanced data-driven soft sensing architectures.

ASSOCIATED CONTENT

Data Availability Statement

The source code will be made available upon reasonable requests.

AUTHOR INFORMATION

Corresponding Authors

Jingde Wang − *College of Chemical Engineering, Beijing University of Chemical Technology, Beijing 100029, China*; Email: jingdewang@mail.buct.edu.cn

Wei Sun − *College of Chemical Engineering, Beijing University of Chemical Technology, Beijing 100029, China*; orcid.org/0000-0003-4027-3751; Email: sunwei@mail.buct.edu.cn

Authors

Wantong Fang − *College of Chemical Engineering, Beijing University of Chemical Technology, Beijing 100029, China*

Cheng Ji − *College of Chemical Engineering, Beijing University of Chemical Technology, Beijing 100029, China*

Fangyuan Ma − *College of Chemical Engineering, Beijing University of Chemical Technology, Beijing 100029, China; School of Chemical Engineering, Oklahoma State University, Stillwater, Oklahoma 74078, United States*

Zheyu Jiang − *School of Chemical Engineering, Oklahoma State University, Stillwater, Oklahoma 74078, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.iecr.5c01644

Author Contributions

[§]W.F. and C.J. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

W.F., C.J., F.M., J.W., and W.S. acknowledge funding support from the National Natural Science Foundation of China under grant number 22278018. F.M. and Z.J. acknowledge funding support from the U.S. National Science Foundation (NSF) under award number 2331080. F.M. and Z.J. acknowledge funding support from Oklahoma Center for Advancement of Science and Technology (OCAST), Oklahoma Applied Research Support (OARS) program under grant number AR24-069.

REFERENCES

(1) Jiang, Z.Online Monitoring and Robust, Reliable Fault Detection of Chemical Process Systems. In *Computer Aided Chemical Engineering*; Elsevier, 2023; Vol. *52*, pp 1623−1628.

(2) Jiang, Y.; Yin, S.; Dong, J.; Kaynak, O. A review on soft sensors for monitoring, control, and optimization of industrial processes. *IEEE Sens. J.* **2021**, *21*, 12868−12881.

(3) Sun, Q.; Ge, Z. A survey on deep learning for data-driven soft sensors. *IEEE Trans. Ind. Inf.* **2021**, *17*, 5853−5866.

(4) Sharmin, R.; Sundararaj, U.; Shah, S.; Griend, L. V.; Sun, Y.-J. Inferential sensors for estimation of polymer quality parameters: Industrial application of a PLS-based soft sensor for a LDPE plant. *Chem. Eng. Sci.* **2006**, *61*, 6372−6384.

(5) Chitralekha, S. B.; Shah, S. L. Application of support vector regression for developing soft sensors for nonlinear processes. *Can. J. Chem. Eng.* **2010**, *88*, 696−709.

(6) Yan, W.; Tang, D.; Lin, Y. A data-driven soft sensor modeling method based on deep learning and its application. *IEEE Trans. Ind. Electron.* **2017**, *64*, 4237−4245.

(7) Zhu, W.; Ma, Y.; Zhou, Y.; Benton, M.; Romagnoli, J. *Computer Aided Chemical Engineering*; Elsevier, 2018; Vol. *44*, pp 2245−2250.

(8) Wu, X.; Chen, J.; Xie, L.; Chan, L. L. T.; Chen, C.-I. Development of convolutional neural network based Gaussian process regression to construct a novel probabilistic virtual metrology in multi-stage semiconductor processes. *Control Eng. Pract.* **2020**, *96*, No. 104262.

(9) Hong, S.; An, N.; Cho, H.; Lim, J.; Han, I.-S.; Moon, I.; Kim, J. A Dynamic Soft Sensor Based on Hybrid Neural Networks to Improve Early Off-spec Detection. *Eng. Comput.* **2023**, *39*, 3011−3021.

(10) Bangi, M. S. F.; Kao, K.; Kwon, J. S.-I. Physics-informed neural networks for hybrid modeling of lab-scale batch fermentation for β-carotene production using Saccharomyces cerevisiae. *Chem. Eng. Res. Des.* **2022**, *179*, 415−423.

(11) Liu, Z.; Cai, W.; Xu, Z.-Q. J. sMulti-scale deep neural network (MscaleDNN) for solving Poisson-Boltzmann equation in complex

domains, arXiv:2007.11207. arXiv.org e-Print archive, 2020. https://arxiv.org/abs/2007.11207.
(12) Wu, Z.; Wang, H.; He, C.; Zhang, B.; Xu, T.; Chen, Q. The application of physics-informed machine learning in multiphysics modeling in chemical engineering. *Ind. Eng. Chem. Res.* **2023**, *62*, 18178−18204.
(13) Karniadakis, G. E.; Kevrekidis, I. G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422−440.
(14) Ma, F.; Ji, C.; Xu, M.; Wang, J.; Sun, W. Spatial correlation extraction for chemical process fault detection using image enhancement technique aided convolutional autoencoder. *Chem. Eng. Sci.* **2023**, *278*, No. 118900.
(15) Chen, Z.; Ge, Z. Knowledge automation through graph mining, convolution, and explanation framework: A soft sensor practice. *IEEE Trans. Ind. Inf.* **2022**, *18*, 6068−6078.
(16) Ma, F.; Ji, C.; Wang, J.; Sun, W.; Tang, X.; Jiang, Z. MOLA: Enhancing Industrial Process Monitoring Using a Multi-Block Orthogonal Long Short-Term Memory Autoencoder. *Processes* **2024**, *12*, No. 2824, DOI: 10.3390/pr12122824.
(17) Braun, B.; Castillo, I.; Joswiak, M.; Peng, Y.; Rendall, R.; Schmidt, A.; Wang, Z.; Chiang, L.; Colegrove, B. Data science challenges in chemical manufacturing *IFAC Prepr.* 2020.
(18) Qin, S. J.; Guo, S.; Li, Z.; Chiang, L. H.; Castillo, I.; Braun, B.; Wang, Z. Integration of process knowledge and statistical learning for the Dow data challenge problem. *Comput. Chem. Eng.* **2021**, *153*, No. 107451.
(19) Qin, S. J.; Liu, Y. A stable Lasso algorithm for inferential sensor structure learning and parameter estimation. *J. Process Control* **2021**, *107*, 70−82.
(20) Liu, Y.; Qin, S. J. A novel two-step sparse learning approach for variable selection and optimal predictive modeling. *IFAC-PapersOnLine* **2022**, *55*, 57−64.
(21) Qin, S. J.; Liu, Y.; Tang, S. Partial least squares, steepest descent, and conjugate gradient for regularized predictive modeling. *AIChE J.* **2023**, *69*, No. e17992.
(22) Barton, M.; Lennox, B. Model stacking to improve prediction and variable importance robustness for soft sensor development. *Digital Chem. Eng.* **2022**, *3*, No. 100034.
(23) Zhu, X.; Damarla, S. K.; Huang, B. In *Spatiotemporal Stacked Autoencoder Based Soft Sensor Modeling for the Dow Data Challenge Problem*, 2023 IEEE Smart World Congress (SWC); IEEE, 2023; pp 1−6.
(24) Xu, J.; Lei, C.; Xu, D.; Zhu, X. In *Time Series Slow Feature Extraction for Dow Data Problem*, 2024 36th Chinese Control and Decision Conference (CCDC); IEEE, 2024; pp 184−188.
(25) Meng, X.; Liu, Q.; Yang, C.; Zhou, L.; Cheung, Y.-M. A novel deep learning-based robust dual-rate dynamic data modeling for quality prediction. *IEEE Trans. Ind. Inf.* **2024**, *20*, 1324−1334.
(26) Liu, Y.; Qin, S. J. Knowledge-informed sparse learning for relevant feature selection and optimal quality prediction. *IEEE Trans. Ind. Inf.* **2023**, *19*, 11499−11507.
(27) McCabe, W.; Smith, J.; Harriott, P. *Unit Operations of Chemical Engineering*, 7th ed.; McGraw-Hill: New York, 2005.
(28) Mathew, T. J.; Tawarmalani, M.; Agrawal, R. Relaxing the constant molar overflow assumption in distillation optimization. *AIChE J.* **2023**, *69*, No. e18125.
(29) Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M.; Sabeti, P. C. Detecting novel associations in large data sets. *Science* **2011**, *334*, 1518−1524.
(30) Ji, C.; Ma, F.; Wan, J.; Sun, W. A Conditional Entropy Based Feature Selection for Soft Sensor Development in Chemical Processes. *Chem. Eng. Trans.* **2023**, *103*, 61−66.
(31) Che, J.; Yang, Y.; Li, L.; Bai, X.; Zhang, S.; Deng, C. Maximum relevance minimum common redundancy feature selection for nonlinear data. *Inf. Sci.* **2017**, *409-410*, 68−86.
(32) Kinney, J. B.; Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 3354−3359.
(33) Smola, A. J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199−222.
(34) López, O. A. M.; López, A. M.; Crossa, J. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*; Springer, 2022; pp 379−425.
(35) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735−1780.