Engineering 22 (2023) 82–93

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Intelligent Manufacturing—Article

# A Causal Model-Inspired Automatic Feature-Selection Method for Developing Data-Driven Soft Sensors in Complex Industrial Processes

Yan-Ning Sun [a], Wei Qin [a,b,*], Jin-Hua Hu [a], Hong-Wei Xu [a], Poly Z.H. Sun [a]

[a] School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Institute of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

*Article history:*
Received 9 January 2022
Revised 20 April 2022
Accepted 8 June 2022
Available online 20 August 2022

*Keywords:*
Big data analytics
Machine intelligence
Quality prediction
Soft sensors
Intelligent manufacturing

ABSTRACT

The soft sensing of key performance indicators (KPIs) plays an essential role in the decision-making of complex industrial processes. Many researchers have developed data-driven soft sensors using cutting-edge machine learning (ML) or deep learning (DL) models. Moreover, feature selection is a crucial issue because a raw industrial dataset is usually high-dimensional, and not all features are conducive to the development of soft sensors. A perfect feature-selection method should not rely on hyperparameters and subsequent ML or DL models. Rather, it should be able to automatically select a subset of features for soft sensor modeling, in which each feature has a unique causal effect on industrial KPIs. Therefore, this study proposes a causal model-inspired automatic feature-selection method for the soft sensing of industrial KPIs. First, inspired by the post-nonlinear causal model, we integrate it with information theory to quantify the causal effect between each feature and the KPIs in the raw industrial dataset. After that, a novel feature-selection method is proposed to automatically select the feature with a non-zero causal effect to construct the subset of features. Finally, the constructed subset is used to develop soft sensors for the KPIs by means of an AdaBoost ensemble strategy. Experiments on two practical industrial applications confirm the effectiveness of the proposed method. In the future, this method can also be applied to other industrial processes to help develop more advanced data-driven soft sensors.

© 2022 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

## 1. Introduction

Over the past decade, data-driven intelligence has been rapidly pushing the envelope in the construction of complex industrial processes by taking advantage of the Industrial Internet of Things (IIoT), big data analytics (BDA), and artificial intelligence (AI) technologies [1–3]. In this context, in order to better create products and services from various materials and resources, the performance of complex industrial processes should be further improved in areas such as product quality, production efficiency, energy consumption, and pollutant emissions. However, these key performance indicators (KPIs) usually cannot be measured and analyzed online through existing sensors [4,5]. Offline laboratory analysis introduces high delay, making it challenging to improve industrial production in time [6,7]. Therefore, the soft sensor technique has attracted extensive efforts for the online estimation of industrial KPIs.

The soft sensor technique aims to describe the input–output behavior of the system by constructing mathematical models with easy-to-measure variables as the input and KPIs as the output. It can be roughly classified into two categories: first-principle (white-box) models and data-driven (black-box) models [8,9]. Models in the former category represent the causality of actual systems, which can only work well by *a prior* understanding of the physical or chemistry knowledge [10,11]. As a result, data-driven models, which focus on association relationships without reflecting actual causality, have become the mainstream to develop soft sensors for industrial KPIs [12].

For example, shallow machine learning (ML) models, such as partial least-squares (PLS), support vector regression (SVR), and their extensions, have been employed to learn quality characteristics from the historical data of complex industrial processes. In Ref. [13], an optimized sparse PLS (OSPLS) model is proposed to estimate the product quality of batch process industries. A robust multi-output least-squares SVR (M-LS-SVR) has been proposed for the online estimation and control of molten iron quality indices in blast furnace ironmaking [14]. Furthermore, deep learning (DL)

* Corresponding author.
*E-mail address:* wqin@sjtu.edu.cn (W. Qin).

https://doi.org/10.1016/j.eng.2022.06.019
2095-8099/© 2022 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

models have been widely studied to capture nonlinear features in complex industrial data. Aiming for multisource heterogeneous data, Ren et al. [15] proposed a wide–deep–sequence (WDS) model combining a wide–deep (WD) model and a long short-term memory (LSTM) network to extract quality-related information from both key time-invariant variables and time-domain features. Yuan et al. [16] developed a sampling interval-attention LSTM (SIA-LSTM) to deal with time series with irregularly sampled data in the soft sensing of key quality variables. Ou et al. [17] proposed a stacked autoencoder (SAE) with quality-driven regularization for quality prediction in an industrial hydrocracking process. To reduce the information loss and generalization degradation of DL models, Yuan et al. [18] proposed a layer-wise data augmentation-based SAE (LWDA-SAE) for the soft sensing of boiling points in the hydrocracking process. Considering the issue of input feature selection in DL, Wang et al. [19] proposed a multi-objective evolutionary nonlinear ensemble learning model with evolutionary feature selection (MOENE-EFS) for silicon prediction in blast furnace ironmaking. Furthermore, in Ref. [20], a deep probabilistic transfer learning (DPTL) framework is proposed to tackle the distribution discrepancy and missing data in a multiphase flow process. Although the above methods have achieved acceptable results in some industrial applications, there are still some research gaps.

Feature selection is still a crucial issue, because a raw industrial dataset is usually high-dimensional, and not all features are conducive to the development of soft sensors. Data-driven soft sensor modeling is essential to recognize patterns in industrial data, so as to determine the quantitative relationships between industrial KPIs and their related features (variables). Selecting a compact and informative subset of features can greatly reduce the complexity of models and help us to fully understand the operation mechanisms of complex industrial processes [21–23]. If the selected features are the causal variables of KPIs, data-driven soft sensors will undoubtedly be more interpretable and stable. Otherwise, blindly improving data-driven models will introduce a complex model structure and difficult-to-tune hyperparameters, which are contrary to the principle of Occam's Razor and the reliability requirements of industry [24]. In other words, it is preferable for a feature-selection method to automatically select a subset of features for soft sensor modeling in which each feature has a unique causal effect on industrial KPIs.

Furthermore, actual industrial data is difficult to obtain and expensive, especially for discrete industries, which hinders the industrial applications of data-driven soft sensors. Based on the production behaviors, complex industrial processes can be classified into process and discrete industries [25]. The production behaviors of the process industries are either continuous, such as chemical processes, power generation, and ironmaking, or occur on a batch of indistinguishable materials, such as food processing, paper making, and injection molding. The production behaviors of discrete industries are either physical or mechanical processes for materials, such as engine assembly, semiconductor manufacturing [26], and household appliance manufacturing, in which the materials used are usually the products of other industrial processes [27,28]. Such processes usually have a larger scale, stronger dynamic, and less clear mechanism than those of process industries. The data collection depends almost entirely on the experience of industrial practitioners, so the raw industrial data is more nonlinear, insufficient, and uncertain. In this situation, ensemble ML with fewer parameters, good robustness, and interpretability is more suitable for complex industrial processes with weak mechanisms [29].

Therefore, this study focuses on the following two scientific questions: ① How can the causal effect between each feature and the KPIs in a raw industrial dataset be quantified? ② How can a subset of features for data-driven soft sensor modeling be automatically selected? Causal models such as the Granger causality, the conditional independence test, and structural equations [30–32] have been widely employed for research on finance [33], climate [34], and industry [35–37]. However, there is no research on integrating causal effect and feature selection for data-driven soft sensors. The main challenges are as follows. The conditional independence test will lead to information loss for the soft sensing of KPIs because of the data adequacy assumption. In addition, the Granger causality and structure equation models depend on the hypothesis of the data-generation mechanism. In summary, the main works of this study are listed below.

(1) Inspired by the post-nonlinear causal model, we integrate it with information theory to quantify the causal effect between each feature and the KPIs in a raw industrial dataset. This can avoid the hypothesis of the data-generation mechanism and provide helpful insight for understanding complex industrial processes.

(2) A novel feature-selection method is proposed to automatically select the feature with a non-zero causal effect to construct the subset of features, which can reduce information loss, promote the interpretability of soft sensor models, and help to improve accuracy and robustness.

(3) The constructed subset is used to develop soft sensors for the KPIs by means of an AdaBoost ensemble strategy. We also introduce two actual complex industrial processes: an injection molding process from Foxconn Technology Group in China and a diesel engine assembly process from Guangxi Yuchai Machinery Group Co., Ltd. in China. Experiments on these two industrial applications confirm the effectiveness of the proposed method.

The rest of this paper is structured as follows. Section 2 describes related works. Sections 3 and 4 then provide detailed descriptions of the proposed method. Subsequently, in Section 5, experimental studies are carried out on two actual complex industrial processes. Finally, Section 6 summarizes the conclusions.

## 2. Related works

This section reviews feature-selection and causal discovery methods, which will motivate the problem formulation and basic idea of this study.

### 2.1. Feature-selection method

As circulated in the industry, data and features determine the upper limit of ML, while models and algorithms just approach this upper limit. Feature selection involves selecting a subset of features as the input of ML from a given candidate feature set [38] and is motivated by two reasons. First, even if there is no *a priori* or domain knowledge, feature selection helps us to fully understand the data and provides perceptual insights [39]. Second, it directly realizes feature dimensionality reduction, which effectively reduces the complexity of ML models [40]. In general, two key aspects are involved in feature selection: the subset search strategy and subset evaluation criteria.

#### 2.1.1. Subset search strategy

Take a candidate input feature set $F$ with $M$ input features, where $F = \{X_1, X_2, \ldots, X_M\}$, and $X_i$ ($i$ is the number of $X$ and $i = 1, 2, \ldots, M$) denotes the candidate input feature. There are $2^M$ candidate subsets $S$, where $S \subseteq F$. The objective of the subset search strategy is to select an optimal feature subset $S$ from $F$ [41]. Eq. (1) shows that the forward search strategy first initializes $S$ with an empty set. After that, based on the subset evaluation criterion, one feature is selected from $F$ and added to $S$ in each iteration until the stopping threshold is reached.

$$S = S \cup \{\mathrm{EC}(S \cup X_i, Y) \ > \ \mathrm{ST}, \ X_i \in F \backslash S\} \quad (1)$$

where EC(·) denotes the evaluation criterion; ST denotes the stopping threshold; and $Y$ denotes the output feature.

Another strategy is called backward search, as illustrated in Eq. (2); this first initializes $S = F$. Then, one feature is removed from $S$ in each iteration until the stopping threshold is reached.

$$S = S \backslash \{\mathrm{EC}(S \backslash X_i, \ Y) \ < \ \mathrm{ST}, \ X_i \in S\} \quad (2)$$

These two strategies are greedy because only the local optimality is implemented. Moreover, it is difficult to determine the optimal evaluation criteria and stopping threshold with good interpretability and theoretical basis.

#### 2.1.2. Subset evaluation criteria

Many evaluation criteria have been used to judge whether to retain the candidate feature in each iteration, such as the degree of association and divergence, and the performance of ML models. The variance $\sigma^2$ measures the degree of feature divergence, and does not consider the association between input and output features [42]. The Pearson correlation coefficient (PCC) selects the input features most relevant to the target, and only focuses on the linear association [43]. The maximal information coefficient (MIC) detects the nonlinear association between two variables [44], but more samples are needed, and the total association is easy to underestimate. Feature selection based on the above criteria does not depend on ML models and is also known as a kind of filtering method.

The parameters of ML models, such as the information gain of a decision tree and regression coefficients [45], can also be used as subset evaluation criteria, which measure the importance or weight of features. This kind of method, which is known as embedding, relies on an ML training process with expensive computing costs and is essentially based on associations. Aside from taking the performance maximization as the evaluation criteria, wrapper methods combine ML with optimization algorithms, such as genetic [38], evolutionary [19], and particle swarm algorithms [39], to automatically select optimal feature combinations. Wrapper methods also bring expensive computing costs, and it is easy to cause over-fitting, especially in industrial applications.

### 2.2. Causal discovery method

Discovering causal relations is a fundamental task of scientific research and technological progress; it strictly distinguishes cause and effect variables, revealing the mechanism and guiding decision-making more effectively than an understand of associations can do. Without considering the lagged effect, causal discovery approaches mainly rely on conditional independence tests and structural equation models to learn causal effects from observed data [46].

#### 2.2.1. Conditional independence tests

Given a set of ternary variables {$X$, $Y$, $Z$}, the specific causal structure can be tested by the conditional independence between variables. As illustrated in Fig. 1, if the relation between ternary variables is that $Y$, $X$, and $Z$ are independent, then the causal structure must be Markov equivalent class (Fig. 1(a)). If the relation is that $X$ and $Z$ are independent on their own, but are not independent once $Y$ is introduced, the causal structure must be a V-structure (Fig. 1(b)). On this basis, Peter–Clark (PC) and inductive causation (IC) algorithms, which are suitable for a wide range, learn causal structure through a two-stage process of the causal skeleton and causal direction [47,48].

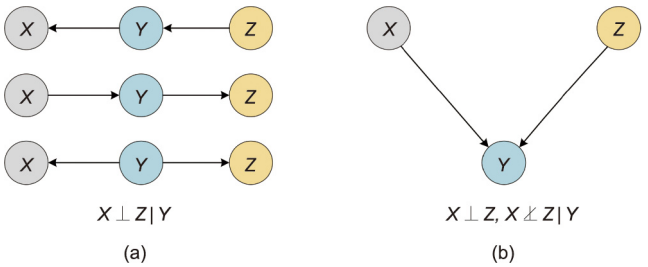


**Fig. 1.** Causality between ternary variables. (a) Markov equivalent class; (b) V-structure.

#### 2.2.2. Structural equation models

The hypothesis of the data-generating mechanism describes how the effect variables are determined by causal variables and causal mechanisms, including the linear non-Gaussian acyclic model (LiNGAM) [49], the additive noise model (ANM) [50], information geometric causal inference (IGCI) [32], and the post-nonlinear model (PNM) [51]. As the most general model, illustrated in Fig. 2, PNM includes the nonlinear influence $f_1$ of the cause $X$, the noise or disturbance $\varepsilon$, and the measurement distortion $f_2$ in the observed effect $Y$. The formula is shown below.

$$Y = f_2(f_1(X) + \varepsilon) \quad (3)$$

where $\varepsilon \perp X$; $f_1$ and $f_2$ are nonlinear functions, and $f_2$ should be invertible.

Due to the limitations of data sufficiency and conditional independence tests, the causal structure obtained from the PC and IC algorithms is not equivalent to the actual physical object. Feature selection based on this causal structure will cause significant information loss, so that the best feature combinations for ML are not attained. In contrast, the PNM in structural equation models can more effectively bridge causal discovery and feature selection.

As for the feature-selection issue, embedding and wrapper methods rely on an ML training process with expensive computing costs. Their performance is directly affected by the selected ML models. Filtering methods, such as variance-based, PCC-based, and MIC-based methods, do not depend on ML models and select a subset of features by manually setting a stopping threshold in advance. A typical stopping threshold includes a specific number of selected features, such as a specific variance value, PCC value, or MIC value. Obviously, it is difficult to determine a stopping threshold with good interpretability and a theoretical basis. Causal discovery brings new light to solve this problem by quantifying the causal effect between each feature and the KPIs in a raw industrial dataset to automatically select a subset of features for data-driven soft sensor modeling. The proposed method is introduced in detail in the next section.

## 3. Causal model-inspired feature selection

### 3.1. PNM with information theory

Given a set of cause variables {$X_1$, $X_2$, …, $X_k$} (where $k$ is the number of variables) and effect variable $Y$, the PNM in Eq. (3) can be extended to Eq. (4).

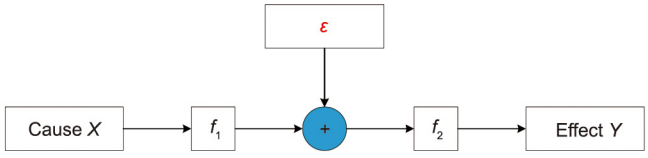


**Fig. 2.** Post-nonlinear causal model. $f_1$, $f_2$: nonlinear functions; $\varepsilon$: the noise or disturbance.

$$Y = f_2(f_1(X_1, X_2, \ldots, X_k) + \varepsilon_k) \tag{4}$$

To discover the causal relations between another variable $X_{k+1}$ and $Y$, Eq. (4) is further extended to Eq. (5).

$$Y = f_2(f_1(X_1, X_2, \ldots, X_k, X_{k+1}) + \varepsilon_{k+1}) \tag{5}$$

If $X_{k+1}$ reduces the noise term, it contains the causal information of $Y$. Thus, the causal effect of $X_{k+1}$ on $Y$ can be quantified by Eq. (6).

$$\mathrm{CE}_{X_{k+1} \to Y} = \frac{1}{2} \log \frac{\sigma^2(\varepsilon_k)}{\sigma^2(\varepsilon_{k+1})} \tag{6}$$

where CE is causal effect.

The problem is that it is necessary to establish and depend on the two regression models, which have high computational complexity and affect accuracy. In addition, the hypothesis of the data-generating mechanism in PNM should be improved.

This study defines the causal effect by means of information theory to solve these problems. In information theory, the Shannon entropy is adopted to measure uncertainty and average information in a discrete random variable $X$ as follows:

$$H(X) = -\sum_x P(x) \log P(x) \tag{7}$$

where $H(\cdot)$ denotes the Shannon entropy; $P(x)$ denotes the probability mass function; and $x$ is the observed value of $X$.

The total uncertainty in the two discrete random variables $X$ and $Y$ can be calculated by joint entropy as follows:

$$H(X, Y) = -\sum_{x,y} P(x, y) \log P(x, y) \tag{8}$$

where $y$ is the observed value of $Y$.

If $X$ is given, the uncertainty in $Y$ can be reduced by considering the information in $X$. Then, the residual uncertainty in $Y$ can be calculated by conditional entropy as follows:

$$H(Y|X) = H(X, Y) - H(X) \tag{9}$$

By substituting Eqs. (7) and (8) into Eq. (9), the conditional entropy can be presented by the probability of $X$ and $Y$. Information theory extends PNM by considering uncertainty instead of variance [30]. In other words, the causal effect can be quantified by measuring the extent to which $X_{k+1}$ reduces the uncertainty of $Y$. As illustrated in Fig. 3, given a set of cause variables $\{X_1, X_2, \ldots, X_k\}$, the residual uncertainty in $Y$ can be calculated by

$$H(Y|X_1, X_2, \ldots, X_k) = H(X_1, X_2, \ldots, X_k, Y) - H(X_1, X_2, \ldots, X_k) \tag{10}$$

When $X_{k+1}$ is further given, the residual uncertainty in $Y$ can be represented as follows:

$$H(Y|X_1, X_2, \ldots, X_k, X_{k+1}) = H(X_1, X_2, \ldots, X_k, X_{k+1}, Y) - H(X_1, X_2, \ldots, X_k, X_{k+1}) \tag{11}$$

Thus, the causal effect of $X_{k+1}$ on $Y$ is obtained as follows:

$$\mathrm{CE}_{X_{k+1} \to Y} = H(Y|X_0, X_1, \ldots, X_k) - H(Y|X_0, X_1, \ldots, X_k, X_{k+1}) \tag{12}$$

Eq. (12) only relies on the information theory to realize regression model-free causal effect quantification. Furthermore, data discretization is a vital data preprocessing technique to calculate the entropy of continuous random variables. In this study, we apply a histogram-based method to discrete data, and the optimal number of bins $n_h$ is estimated by

$$n_h = \max\left(\frac{R}{2 \cdot \mathrm{IQR} \cdot n^{\frac{1}{3}}}, \log_2(n+1)\right) \tag{13}$$

where $R$ is the range of data; IQR is the interquartile range; and $n$ is the number of samples.

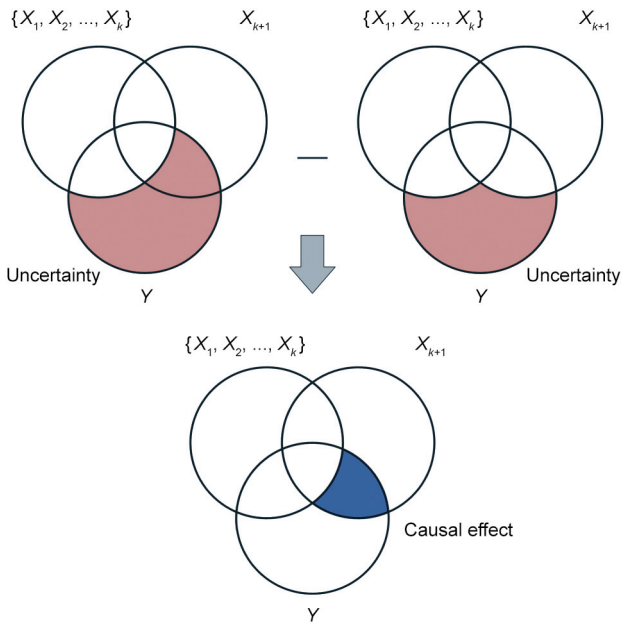


**Fig. 3.** Venn diagram of the improved causal effect, where the first red shadow indicates the residual uncertainty in $Y$ when a set of cause variables $\{X_1, X_2, \ldots, X_k\}$ is given, the second red shadow indicates the residual uncertainty in $Y$ when $X_{k+1}$ is further given, and the blue shadow indicates the causal effect of $X_{k+1}$ on $Y$.

### 3.2. Causal effect-based automatic feature selection

We present a novel feature-selection idea that takes the forward search strategy as the subset search strategy and the causal effect in Eq. (12) as the subset evaluation criteria. Its formal expression is as follows:

$$S = S \cup \{\mathrm{CE}_{X_i \to Y} \neq 0, X_i \in F/S\} \tag{14}$$

Compared with Eqs. (1) and (2), the feature-selection method shown in Eq. (14) only needs to traverse all candidate input features $X_i$ in a specific order, does not need to set a stop threshold, and automatically selects the input feature combination with a non-zero causal effect. In the actual execution process, we determine the traversal order according to the mutual information between each candidate input feature $X_i$ and output feature $Y$. Algorithm 1 gives the pseudo-code of the causal effect-based automatic feature-selection algorithm. The detailed implementation process of this method is also shown in Fig. 4.

**Algorithm 1**. Causal effect-based automatic feature-selection algorithm.

**Input:** Dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ (where $N$ is the number of dataset) with candidate input feature set $F = \{X_1, X_2, \ldots, X_M\}$ and output feature $Y$
**Output:** Causal feature set $S$
1: Initialize: $S = \varnothing$; dataset discretization processing using Eq. (13)
2: **for** $i = 1$ to $M$ **do**
3: Calculate the causal effect $\mathrm{CE}_{X \to Y}$ of $X_i$ on $Y$ using Eq. (12)
4: **if** $\mathrm{CE}_{X_i \to Y} \neq 0$, **then** $S = S \cup \{X_i\}$
5: **end for**
6: **return** $S$

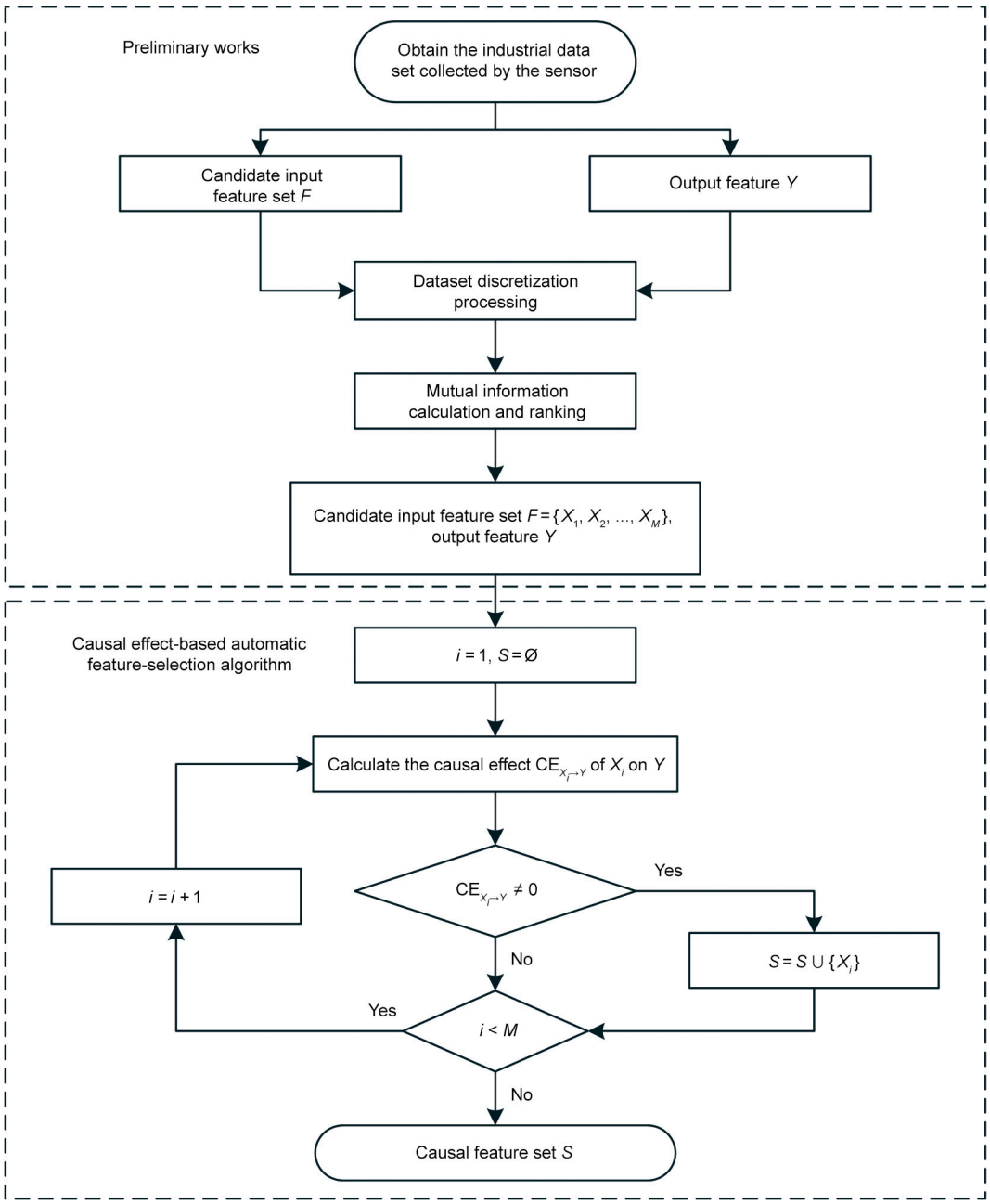


**Fig. 4.** Flow chart of the proposed method.

## 4. AdaBoost decision tree-based soft sensor modeling

In this study, taking the decision tree as the basic learner, an AdaBoost ensemble ML algorithm is employed for the soft sensor modeling of industrial KPIs. It should be pointed out that this model is not designed to outperform all existing models. Instead, we believe that, once the causal information is extracted, data-driven soft sensors can achieve satisfactory accuracy and interpretability. In the future, we will study more advanced ML or DL models for data-driven soft sensors.

### 4.1. Decision tree regressor

A decision tree regressor is mainly a classification and regression tree (CART) algorithm, which can solve classification or regression problems. Take training dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ (where $N$ is the number of samples). When applying a CART to solve the regression problems, based on the idea of bisection recursive segmentation, the optimal segmentation variable $j$ and segmentation point $s$ are selected by using the square error minimization criterion, that is, the following equation is solved:

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{15}$$

where $c_1$ and $c_2$ are output values; $R_1$ and $R_2$ are two regions in the input space.

Then, the input space is divided into two regions $R_1$ and $R_2$ by the variable $j$ and point $s$. Two sub-nodes are generated from this node, containing $N_1$ and $N_2$ samples, respectively.

$$R_1(j,s)=\{x|x^{(j)}\le s\},\ R_2(j,s)=\{x|x^{(j)}>s\} \tag{16}$$

The optimal output values $\hat{c}_1$ and $\hat{c}_2$ in these two regions are further determined as follows:

$$\hat{c}_1 = \frac{1}{N_1}\sum_{x_i \in R_1(j,s)} y_i, \quad \hat{c}_2 = \frac{1}{N_2}\sum_{x_i \in R_2(j,s)} y_i \tag{17}$$

Let the process to recur in turn until the end conditions are met; finally, divide the input space into $W$ regions $R_1$, $R_2$, ..., $R_W$ to generate a decision tree:

$$f(x) = \sum_{w=1}^{W} \hat{c}_w I, \qquad x \in R_w \tag{18}$$

where $I(\cdot)$ is the indicating function and $w$ is the number of regions. If $x \in R_w$, then $I = 1$; otherwise, $I = 0$.

After the regression tree is generated, it is pruned from the bottom to the root node. For each pruning case, a subtree is generated, thus forming a subtree sequence $f_1(x), f_2(x), \ldots, f_n(x)$. Next, use the cross-validation method on the independent verification data set to compare the square error of each subtree with respect to the verification set, and select the optimal decision tree $f_\alpha(x)$ ($\alpha$ is the number of sequence and $\alpha = 1, 2, \ldots, n$).

### 4.2. AdaBoost ensemble learning for soft sensing

As illustrated in Algorithm 2, given $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ as the training set, $W_t = \{w_t(1), w_t(2), \ldots, w_t(N)\}$ ($t$ is the number of iteration and $t = 1, 2, \ldots, T$, where $T$ is the total number of iterations) denotes the weight distribution over $D$ at the $t$th boosting iteration. At later iterations, the weight distribution will be updated by increasing the weight of samples with poor performance and decreasing the weight of those with good performance. The average loss function to measure the performance is given by

$$\bar{L}_t = \sum_{i=1}^{N} L_t(i) w_t(i) \tag{19}$$

where $L_t$ is a loss function with a range of 0–1. Three candidate $L_t$ are presented by Ref. [52]; this study uses the exponential one, as follows:

$$L_t(i) = 1 - \exp\{l_t(i)/\max(l_t(i))\}, i = 1, 2, \cdots, N \tag{20}$$

where $l_t(i) = |f_t(x_i) - y_i|$ is the loss for each training example. The reweighting procedure is formulated as follows:

$$w_{t+1}(i) = w_t(i)\alpha_t^{1-L_t(i)}/Z_t \tag{21}$$

where $\alpha_t = \bar{L}_t/\left(1 - \bar{L}_t\right)$ is the weight updating parameter; $Z_t$ is the normalization factor that makes $W_{t+1}$ a probability distribution. The final AdaBoost regression result can be obtained by

$$f(x) = -\sum_{t=1}^{T} f_t(x) \log \alpha_t \tag{22}$$

**Algorithm 2**. AdaBoost ensemble ML for soft sensor modeling.

**Input:** Training set with $N$ datasets $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, the basic learning algorithm, and the total number of iterations $T$

**Output:** The soft sensor model $f(x)$

1: Initialize: $W_1 = \{w_1(1), w_1(2), \ldots, w_1(N)\}$, $w_1(i) = 1/N$, $i = 1, 2, \ldots, N$

2: **for** $t = 1$ to $T$ **do**

3: Take a sample set $R_t$ from $D$ using distribution $W_t = \{w_t(1), w_t(2), \ldots, w_t(N)\}$

4: Calculate the loss function $L_t(i)$ for each training sample

5: Calculate the average loss $\bar{L}_t$

6: Set $\alpha_t$ to update distribution $W_t$ as $W_{t+1} = \{w_{t+1}(1), w_{t+1}(2), \ldots, w_{t+1}(N)\}$

7 **end for**

8 **return** $f(x)$

## 5. Experimental studies

In this section, the proposed method is validated by experiments on two actual complex industrial processes.

### 5.1. Experimental setup

According to the theoretical derivation, it can be seen that the proposed feature-selection method is a kind of filtering method. In this method, the causal effect is used as the subset evaluation criteria, and the forward search strategy is used to automatically select the subset of features for training the soft sensor model. The proposed method does not need to set a stop threshold. Each feature $X_i$ in this subset has a unique causal effect on $Y$. It can be concluded without verification that other filtering methods lack this advantage.

The performance evaluation of feature selection usually considers two aspects: the number of selected features and the performance of the soft sensors. We hope to use the least number of input features to achieve the best performance of the soft sensors. It is well-known that variance-, PCC-, and MIC-based methods are the simplest and most effective filtering feature-selection methods with good generalization. Therefore, these three methods are taken as benchmarks for comparison purposes. We first use the proposed method to determine the number of selected features (marked as $K$). Then, the stop threshold of three benchmarks is also set to $K$. Finally, the feature subsets obtained by the above four methods are used to train the AdaBoost decision tree-based soft sensor model, and the performance of the soft sensors is compared. During this process, the experimental data of two complex industrial processes are randomly divided into two groups according to a 60:40 proportion; that is, 60% is taken as the training set and 40% is taken as the testing set. The root-mean-square error (RMSE) and the coefficient of determination $R^2$, which are two widely used performance evaluation metrics, are defined by Eqs. (23) and (24), respectively, and are adopted in this study. Eventually, if the RMSE and $R^2$ of our method are better than those of the three benchmarks, the effectiveness of the proposed method can be verified.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N_T} (y_i - \hat{y}_i)^2 / N_T} \tag{23}$$

$$R^2 = 1 - \sum_{i=1}^{N_T} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{N_T} \left(y_i - \bar{y}_i\right)^2 \tag{24}$$

where $N_T$ is the number of samples in the testing set; $y_i$ is the real value of the $i$th sample; $\hat{y}_i$ is the estimated value of the soft sensor model; and $\bar{y}_i$ is the mean of all estimated values.

All the codes of this study are written in Python 3.7. The four most important hyperparameters of the AdaBoost decision tree-based soft sensor model are the maximum depth and minimum samples split of each decision tree regressor, as well as the number of estimators and learning rate of AdaBoost ensemble learning. In two experiments, by fine-tuning up and down near the default value, they are set to 10.0, 5.0, 20.0, and 1.3, respectively. All other hyperparameters use default values. The hardware environment is Intel (R) Core (TM) i7-8700 central processing unit (CPU) @3.20 GHz 32.00G random access memory (RAM).

### 5.2. Experimental study on the injection molding process

The first complex industrial process is the injection molding process from Foxconn Technology Group in China. This process uses an injection molding machine (Fig. 5) to melt the plastic raw materials at a high temperature. It then injects the plastic melt

into the mold at high speed and high pressure; the melt undergoes complex physicochemical changes at a constant pressure to yield plastic products. Through the repeated operation of this process, a large number of the same products can be produced. During this process, the final product quality is measured with a high delay, which seriously affects timely decision-making for ensuring quality stability. Therefore, the injection molding process is used to verify and apply the proposed method. The data of 16 600 production batches were collected, including 86 candidate input features, and the product sizes were used as the KPIs [53].

Based on Section 3, we quantify the causal effects of 86 candidate input features on the product size (mm) of the injection molding process. As shown in Fig. 6, it is found that only nine candidate input features contain causal information about product size; given these nine features, the remaining features have no causal effect on it. Thus, these nine features are utilized as the input features of the soft sensor model to estimate the value of the product size. The nine features are: instantaneous flow ($m^3 \cdot s^{-1}$), cycle time (s), jacking time (s), post-cooling time (s), mold temperature (°C), clamping time (s), ejection time (s), clamping pressure (Pa), and opening time (s).

Table 1 shows the RMSE and $R^2$ of the soft sensor model under different feature-selection methods. We can see that the causal effect-based feature-selection method provides the lowest RMSE and the largest $R^2$. It outperforms the three benchmarks, because the causal information of product size is accurately extracted; in addition, redundant non-causal information is effectively removed. Moreover, compared with the benchmarks, the proposed method does not need to set a stopping threshold and can naturally avoid information loss.

Fig. 7 shows the soft sensor results for product size under different feature-selection methods. It can be seen that the causal effect-based method can more effectively estimate the slight fluctuation of quality than the methods based on three benchmarks. Fig. 8 shows the scatter diagrams and probability density curves of the soft sensor results under different feature-selection methods. It can be seen that the estimated values of the causal effect-based method are closer to the real value. Furthermore, the probability density curve from the causal effect-based method is "thinner" and "taller" than those from the benchmarks, which also proves that the proposed method has better accuracy.

### 5.3. Experimental study on the diesel engine assembly process

The second complex industrial process is the diesel engine assembly process from Guangxi Yuchai Machinery Group Co., Ltd. (China). As shown in Fig. 9, mechanical parts are assembled into diesel engine products through eight sub-assembly lines, including the main assembly line, five sub-assembly lines, the performance test line, and the package line. The consistency of the rated power under the same work conditions is one of the most important KPIs, but its inspection requires time-consuming and high-cost bench testing. We implemented the test on 1763 samples; for each sample, the data of 39 process variables were collected along the assembly process [36,37] and were utilized as the candidate input features to verify and apply the proposed method.

Further verification and application of the proposed method is performed on the diesel engine assembly process. Similarly, the causal effects of 39 candidate input features on the rated power (kW) of the diesel engine products are quantified. As shown in Fig. 10, it is found that only six candidate input features contain causal information about the rated power, while, given these six features, the remaining ones have no causal effect on it. Thus, these six features are utilized as the input features of the soft sensor model to estimate the value of the rated power. This six features are: fuel consumption per 100 kilometers (L), running time (min), fuel consumption rate (%), intercooler inlet pressure (Pa), intercooler inlet temperature (°C), and axial clearance (mm).

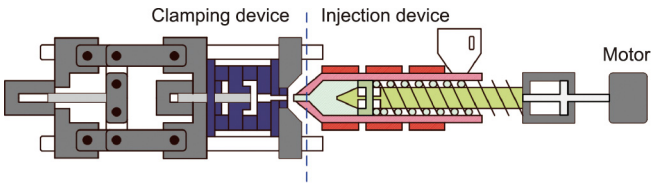


**Fig. 5.** Diagram of the injection molding machine.

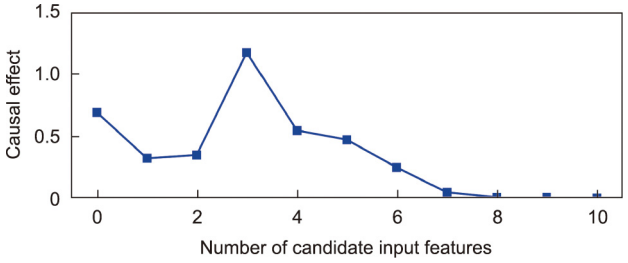


**Fig. 6.** The causal effect of different candidate input features on product size in the injection molding process.

**Table 1**
RMSE and $R^2$ of the soft sensing model under different feature-selection methods in the injection molding process.

| Methods | RMSE (mm) | $R^2$ (%) |
|---|---|---|
| Variance-based | 0.031 | 65.1 |
| PCC-based | 0.031 | 65.2 |
| MIC-based | 0.027 | 73.2 |
| Cause effect-based | 0.023 | 80.4 |

Table 2 shows the RMSE and $R^2$ of the soft sensor model under different feature-selection methods. Again, we can see that the causal effect-based feature-selection method provides the lowest RMSE and the largest $R^2$. It is worth noting that the three benchmarks have very low $R^2$, indicating that it is difficult to explain the output features using the selected variables. Fig. 11 shows the soft sensor results for the rated power under different feature-selection methods. It can be seen that the causal effect-based method can more accurately estimate the value of the rated power than the three benchmarks. Fig. 12 shows the scatter diagrams and probability density curves of the soft sensor results under different feature-selection methods. The estimated values of the causal effect-based method are closer to the real rated power than those of the other methods. In addition, the probability density curve from the causal effect-based method is "thinner" and "taller," which once again proves that the proposed model has better accuracy than the three benchmarks.

According to the results of these two experiments, the following insights can be obtained. The proposed method is effective and universal, and can help us to understand complex industrial processes. In practical industrial applications, this method can select a subset of features from the raw industrial dataset that is compact and informative. For example, among the 86 candidate input features in the injection molding process, only nine candidate input features contain causal information about the product size. Compared with these nine features, other features are irrelevant or redundant for soft sensor modeling. The performance of the soft sensors can be further improved in two ways. One is to develop

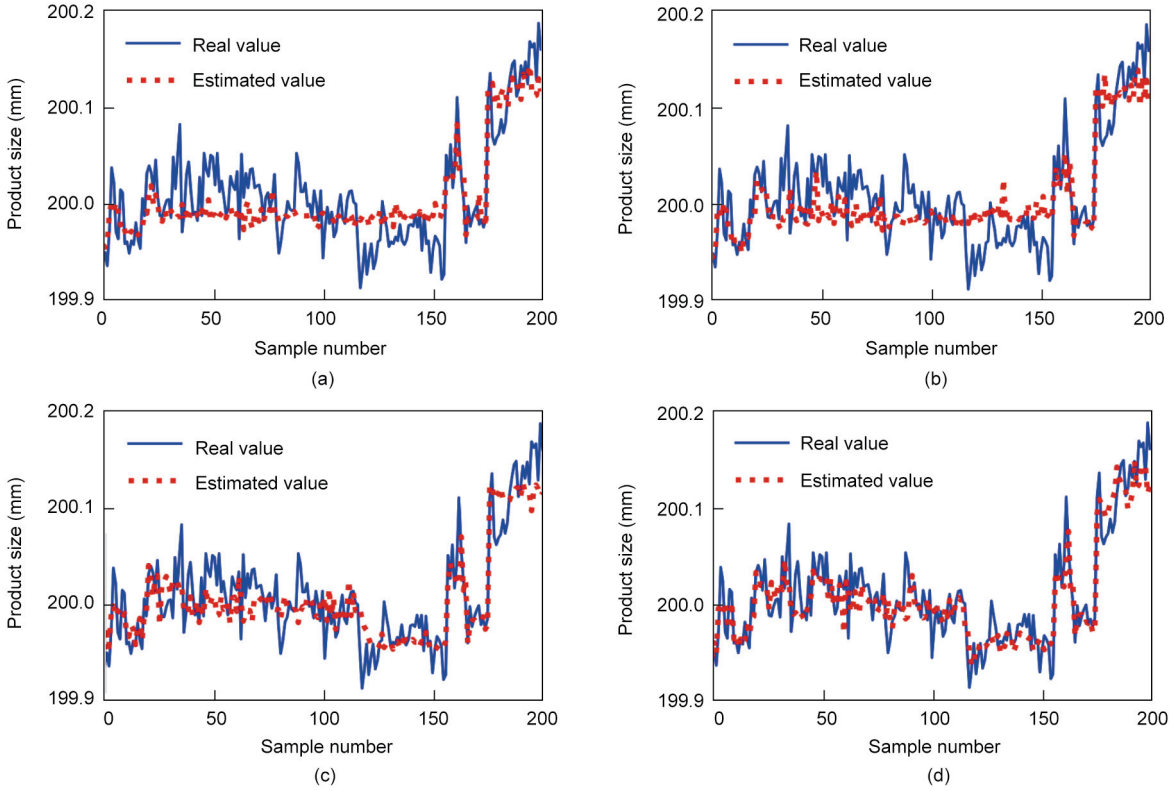


**Fig. 7.** Soft sensor results of product size under different feature-selection methods in the injection molding process. (a) Variance; (b) PCC; (c) MIC; (d) causal effect.

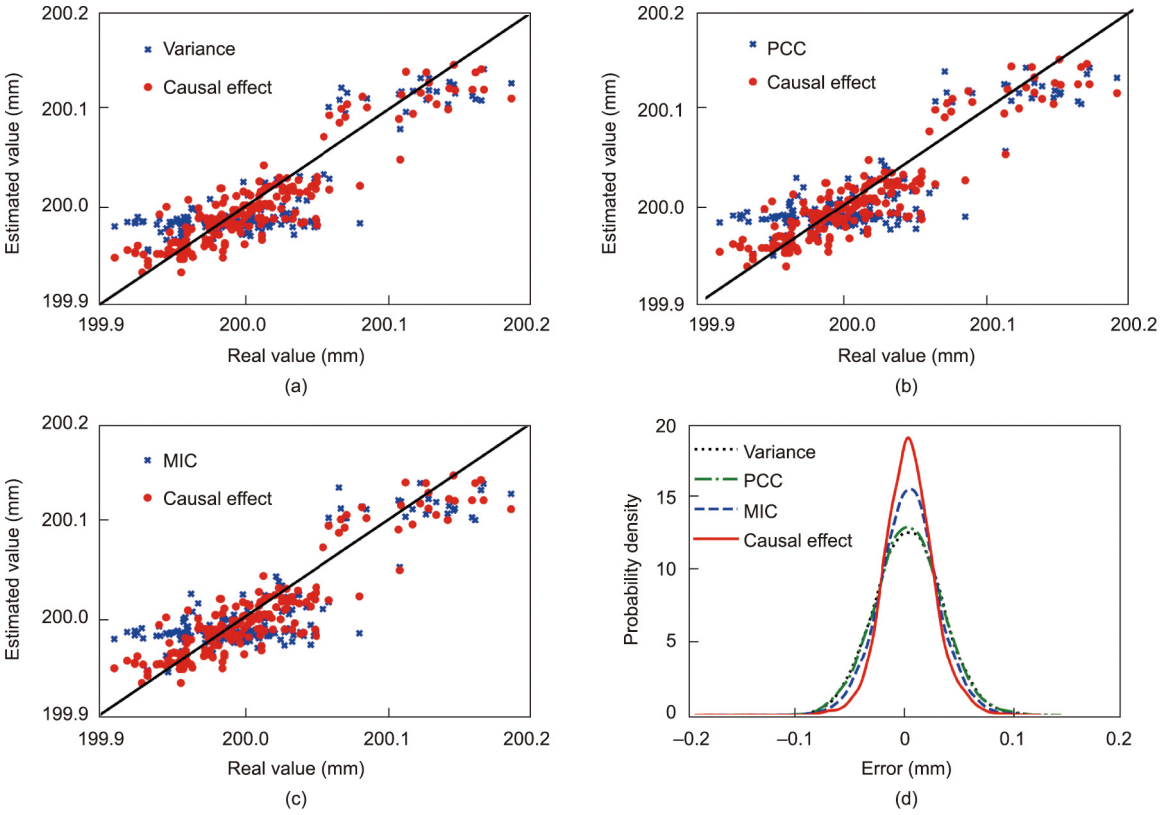


**Fig. 8.** Scatter diagrams and probability density curves of the soft sensor results under different feature-selection methods in the injection molding process.

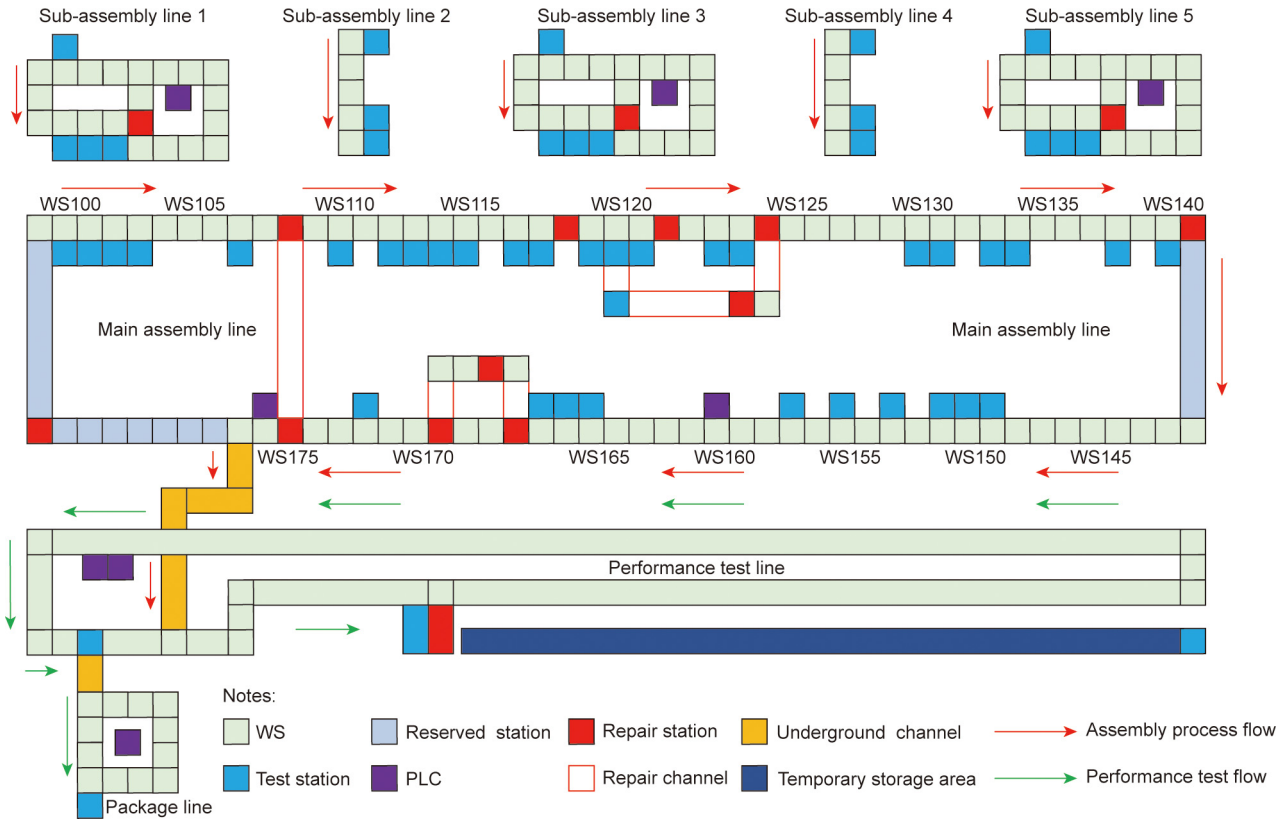


**Fig. 9.** Diagram of the diesel engine assembly process. WS: work station; PLC: programmable logic controller.

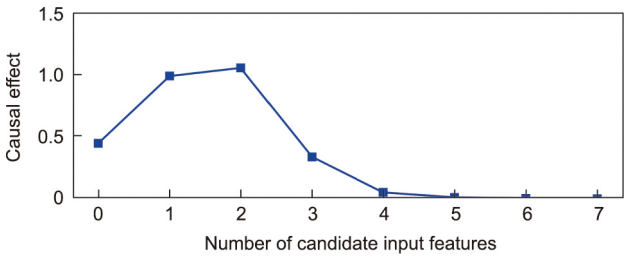


**Fig. 10.** The causal effect of different features on the rated power in the diesel engine assembly process.

**Table 2**
RMSE and $R^2$ of the soft sensor model under different feature-selection methods in the diesel engine assembly process.

| Methods | RMSE (kW) | $R^2$ (%) |
|---|---|---|
| Variance-based | 3.207 | 18.5 |
| PCC-based | 3.078 | 24.9 |
| MIC-based | 3.066 | 25.5 |
| Cause effect-based | 2.215 | 61.1 |

a more advanced data-driven model, which can fit the data distribution of the selected features more fully. Based on our experience, the performance of the existing data-driven models is similar when selecting the same input features. Thus, this paper only introduces an AdaBoost ensemble learning model for soft sensor modeling, while a comparison of different models lies beyond our scope. The other way is obtaining a deeper understanding of the industrial processes by means of first principles, so as to obtain more comprehensive and sufficient data to help train better data-driven models. In other words, although we are developing data-driven methods, research on the first principles of complex industrial processes should not be ignored.

## 6. Conclusions

This study proposes a causal effect-based feature-selection method for developing soft sensors for KPIs in complex industrial processes. Integrating the PNM with information theory, a causal effect quantification method is presented to extract the causal information of KPIs. The proposed method can provide helpful insights into the soft sensing of KPIs and helps to improve the accuracy and interpretability of ML. In addition, decision tree regression with the AdaBoost ensemble is employed for soft sensor modeling and requires almost no fine-tuning of parameters to achieve excellent performance. Our experimental studies on actual industrial processes confirm the effectiveness and promising applications of this method.

However, the PNM is a non-temporal causal model, so this paper does not consider the lagged effect of causality. If the proposed method is applied to time series data, it is necessary to first estimate the causal delay. This is another topic that may be addressed in our future work. In addition, this study focuses on the causal effect-based feature-selection method, while the research on downstream ML models is weak. As we mentioned earlier, under the same input features containing causal information, the improvement of soft sensor results by cutting-edge ML models is limited. Therefore, especially for complex industrial scenarios, our future work will focus on the interval estimation and risk

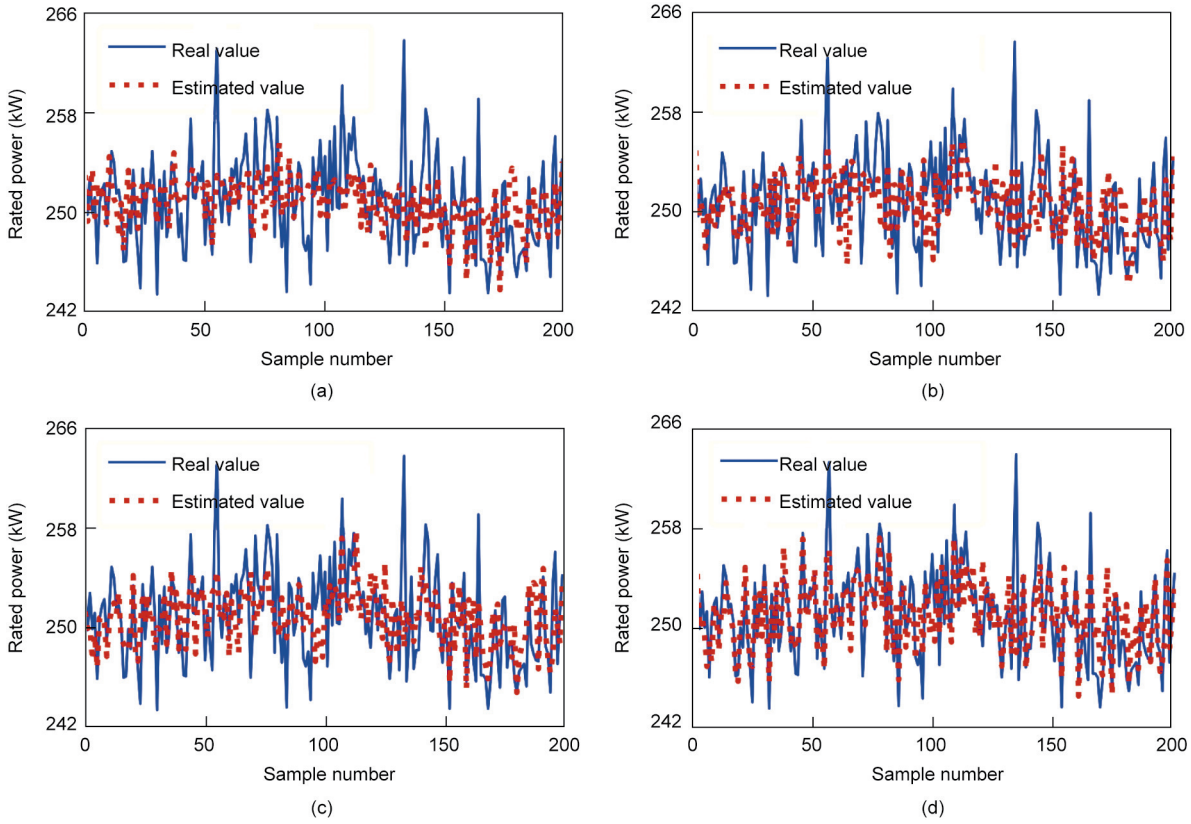


**Fig. 11.** Soft sensor results for product size under different feature-selection methods in the diesel engine assembly process. (a) Variance; (b) PCC; (c) MIC; (d) causal effect.

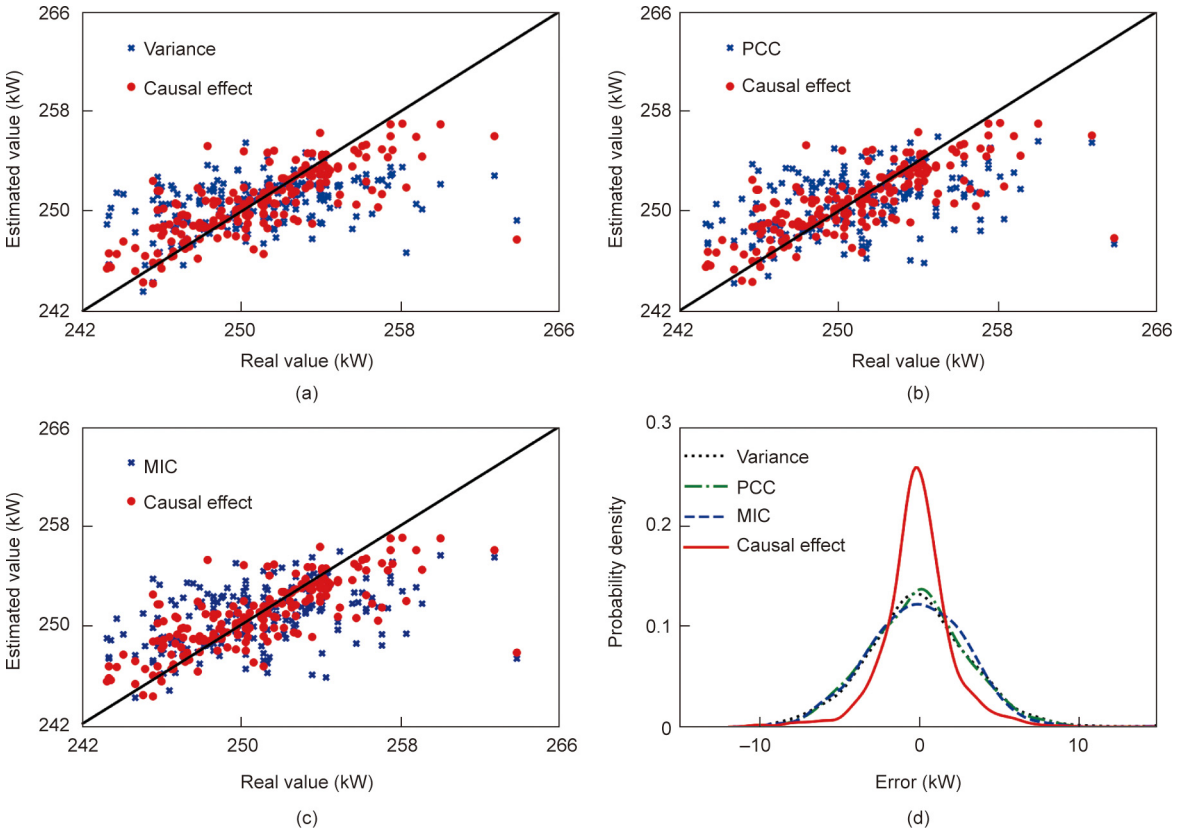


**Fig. 12.** Scatter diagrams and probability density curves of the soft sensor results under different feature-selection methods in the diesel engine assembly process.

assessment of KPI models based on the theory of uncertainty quantification.

## Acknowledgments

We would like to thank the Ministry of Education–China Mobile Research Foundation Project of China (MCM20180703), and the National Key Research and Development Program of China (2020YFB1711100) for financial support. We would also like to thank Foxconn Technology Group for providing us with a historical dataset of the injection molding process, as well as the editors and reviewers for their valuable comments.

## Compliance with ethics guidelines

Yan-Ning Sun, Wei Qin, Jin-Hua Hu, Hong-Wei Xu, and Poly Z.H. Sun declare that they have no conflict of interest or financial conflicts to disclose.

## References

[1] Gao L, Shen W, Li X. New trends in intelligent manufacturing. Engineering 2019;5(4):619–20.
[2] Wang J, Zheng P, Lv Y, Bao J, Zhang J. Fog-IBDIS: industrial big data integration and sharing with fog computing for manufacturing systems. Engineering 2019;5(4):662–70.
[3] Wang J, Ma Y, Zhang L, Gao RX, Wu D. Deep learning for smart manufacturing: methods and applications. J Manuf Syst 2018;48:144–56.
[4] Yuan X, Gu Y, Wang Y, Yang C, Gui W. A deep supervised learning framework for data-driven soft sensor modeling of industrial processes. IEEE Trans Neural Netw Learn Syst 2020;31(11):4737–46.
[5] Liu C, Wang K, Wang Y, Yuan X. Learning deep multimanifold structure feature representation for quality prediction with an industrial application. IEEE Trans Ind Inform 2022;18(9):5849–58.
[6] Ren L, Meng Z, Wang X, Zhang L, Yang LT. A data-driven approach of product quality prediction for complex production systems. IEEE Trans Ind Inform 2021;17(9):6457–65.
[7] Geng Z, Dong J, Chen J, Han Y. A new self-organizing extreme learning machine soft sensor model and its applications in complicated chemical processes. Eng Appl Artif Intell 2017;62:38–50.
[8] Shi J, Zhou S. Quality control and improvement for multistage systems: a survey. IIE Trans 2009;41(9):744–53.
[9] Schrangl P, Tkachenko P, del Re L. Iterative model identification of nonlinear systems of unknown structure: systematic data-based modeling utilizing design of experiments. IEEE Control Syst Mag 2020;40(3):26–48.
[10] Mao J, Chen D, Zhang L. Mechanical assembly quality prediction method based on state space model. Int J Adv Manuf Technol 2016;86:107–16.
[11] Zhou X, Zhang Y, Mao T, Zhou H. Monitoring and dynamic control of quality stability for injection molding process. J Mater Process Technol 2017;249:358–66.
[12] Sun Q, Ge Z. Deep learning for industrial KPI prediction: when ensemble learning meets semi-supervised data. IEEE Trans Ind Inform 2021;17(1):260–9.
[13] Jiang Q, Yan X, Yi H, Gao F. Data-driven batch-end quality modeling and monitoring based on optimized sparse partial least squares. IEEE Trans Ind Electron 2020;67(5):4098–107.
[14] Zhou P, Guo D, Wang H, Chai T. Data-driven robust M-LS-SVR-based NARX modeling for estimation and control of molten iron quality indices in blast furnace ironmaking. IEEE Trans Neural Netw Learn Syst 2018;29(9):4007–21.
[15] Ren L, Meng Z, Wang X, Lu R, Yang LT. A wide–deep-sequence model-based quality prediction method in industrial process analysis. IEEE Trans Neural Netw Learn Syst 2020;31(9):3721–31.
[16] Yuan X, Jia Z, Li L, Wang K, Ye L, Wang Y, et al. A SIA-LSTM based virtual metrology for quality variables in irregular sampled time sequence of industrial processes. Chem Eng Sci 2022;249:117299.
[17] Ou C, Zhu H, Shardt YAW, Ye L, Yuan X, Wang Y, et al. Quality-driven regularization for deep learning networks and its application to industrial soft sensors. IEEE Trans Neural Netw Learn Syst. In press.
[18] Yuan X, Ou C, Wang Y, Yang C, Gui W. A layer-wise data augmentation strategy for deep learning networks and its soft sensor application in an industrial hydrocracking process. IEEE Trans Neural Netw Learn Syst 2021;32(8):3296–305.
[19] Wang X, Hu T, Tang L. A multiobjective evolutionary nonlinear ensemble learning with evolutionary feature selection for silicon prediction in blast furnace. IEEE Trans Neural Netw Learn Syst 2022;33(5):2080–93.
[20] Chai Z, Zhao C, Huang B, Chen H. A deep probabilistic transfer learning framework for soft sensor modeling with missing data. IEEE Trans Neural Netw Learn Syst 2022;33(12):7598–609.
[21] Peng H, Fan Y. Feature selection by optimizing a lower bound of conditional mutual information. Inf Sci 2017;418–419:652–67.
[22] Wang J, Xu C, Zhang J, Zhong R. Big data analytics for intelligent manufacturing systems: a review. J Manuf Syst 2022;62:738–52.
[23] Lee DH, Yang JK, Lee CH, Kim KJ. A data-driven approach to selection of critical process steps in the semiconductor manufacturing process considering missing and imbalanced data. J Manuf Syst 2019;52:146–56.
[24] Sun S, Hu X, Liu Y. An imbalanced data learning method for tool breakage detection based on generative adversarial networks. J Intell Manuf 2021;2021:1–15.
[25] Perši N, Dušak V. Conceptual modelling of continuous discrete production systems. In: Proceedings of the 6th EUROSIM Conference on Modelling and Simulation; 2007 Sep 9–13; Ljubljana, Slovenia. EUROSIM; 2007. p. 1–7.
[26] Xu HW, Qin W, Lv YL, Zhang J. Data-driven adaptive virtual metrology for yield prediction in multi-batch wafers. IEEE Trans Ind Inform 2022;18(12):9008–16.
[27] Diaz CJL, Ocampo-Martinez C. Energy efficiency in discrete-manufacturing systems: insights, trends, and control strategies. J Manuf Syst 2019;52:131–45.
[28] Thiede S, Turetskyy A, Kwade A, Kara S, Herrmann C. Data mining in battery production chains towards multi-criterial quality prediction. CIRP Ann 2019;68(1):463–6.
[29] Finkeldey F, Volke J, Zarges JC, Heim HP, Wiederkehr P. Learning quality characteristics for plastic injection molding processes using a combination of simulated and measured data. J Manuf Process 2020;60:134–43.
[30] Keskin Z, Aste T. Information-theoretic measures for nonlinear causality detection: application to social media sentiment and cryptocurrency prices. R Soc Open Sci 2020;7(9):200863.
[31] Spirtes P, Zhang K. Causal discovery and inference: concepts and recent methodological advances. Appl Inform 2016;3:3.
[32] Janzing D, Mooij J, Zhang K, Lemeire J, Zscheischler J, Daniušis P, et al. Information-geometric approach to inferring causal directions. Artif Intell 2012;182–183:1–31.
[33] Xu L. Machine learning and causal analyses for modeling financial and economic data. Appl Inform 2018;5:11.
[34] Nowack P, Runge J, Eyring V, Haigh JD. Causal networks for climate model evaluation and constrained projections. Nat Commun 2020;11:1415.
[35] Sun Y, Qin W, Zhuang Z, Xu H. An adaptive fault detection and root-cause analysis scheme for complex industrial processes using moving window KPCA and information geometric causal inference. J Intell Manuf 2021;32(7):2007–21.
[36] Sun Y, Qin W, Zhuang Z. Nonparametric-copula-entropy and network deconvolution method for causal discovery in complex manufacturing systems. J Intell Manuf 2022;33(6):1699–713.
[37] Sun Y, Qin W, Zhuang Z. Quality consistency analysis for complex assembly process based on Bayesian networks. Procedia Manuf 2020;51:577–83.
[38] Xu H, Zhang J, Lv Y, Zheng P. Hybrid feature selection for wafer acceptance test parameters in semiconductor manufacturing. IEEE Access 2020;8:17320–30.
[39] Qin W, Zhuang Z, Guo L, Sun Y. A hybrid multi-class imbalanced learning method for predicting the quality level of diesel engines. J Manuf Syst 2022;62:846–56.
[40] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: a new perspective. Neurocomputing 2018;300:70–9.
[41] Han M, Ren W. Global mutual information-based feature selection approach using single-objective and multi-objective optimization. Neurocomputing 2015;168:47–54.
[42] Han Y, Yu L. A variance reduction framework for stable feature selection. Stat Anal Data Min: ASA Data Sci J 2012;5(5):428–45.
[43] Sun YN, Zhuang ZL, Xu HW, Qin W, Feng MJ. Data-driven modeling and analysis based on complex network for multimode recognition of industrial processes. J Manuf Syst 2022;62:915–24.
[44] Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science 2011;334(6062):1518–24.
[45] Mokhtia M, Eftekhari M, Saberi-Movahed F. Feature selection based on regularization of sparsity based regression models by hesitant fuzzy correlation. Appl Soft Comput 2020;91:106255.
[46] Cai RC, Chen W, Zhang K, Hao ZF. A survey on non-temporal series observational data based causal discovery. Chin J Comput 2017;40(6):1470–90. Chinese.
[47] Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet 2019;10:524.
[48] You D, Li R, Liang S, Sun M, Ou X, Yuan F, et al. Online causal feature selection for streaming features. IEEE Trans Neural Netw Learn Syst. In press.
[49] Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-Gaussian acyclic model for causal discovery. J Mach Learn Res 2006;7:2003–30.
[50] Janzing D, Peters J, Mooij J, Schölkopf B. Identifying confounders using additive noise models. 2012. arXiv:1205.2640.
[51] Zhang K, Hyvärinen A. Nonlinear functional causal models for distinguishing cause from effect. In: Wiedermann W, von Eye A, editors. Statistics and causality: methods for applied empirical research. Wiley; 2016. p. 185–201.

[52] Drucker H. Improving regressors using boosting techniques. In: Proceedings of the 14th International Conference on Machine Learning (ICML); 1997 Jul 8–12; Nashville, TN, USA. San Francisco: Morgan Kaufmann Publishers Inc.; 1997. p. 107–15.

[53] Sun YN, Chen Y, Wang WY, Xu HW, Qin W. Modelling and prediction of injection molding process using copula entropy and multi-output SVR. In: Proceedings of 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE); 2021 Aug 23–27; Lyon, France. IEEE; 2021. p. 1677–82.